❼ Randomization Methods for Hypothesis Testing

Covered in this chapter:

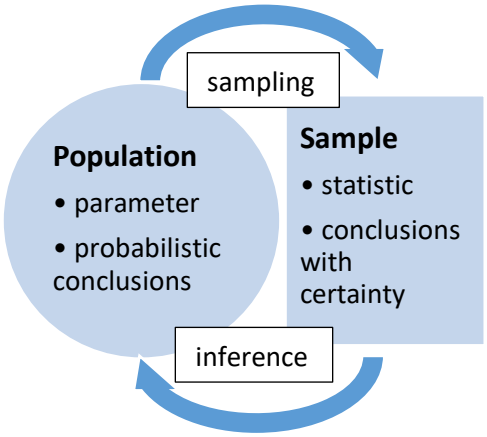
- Dimensions of statistical inference
- Hypothesis testing with a randomization test
- Monte Carlo and permutation tests
- p-values and statistical conclusion validity

7.1 Introduction to statistical inference

In the first half of this book, we focused on using statistics to describe and explore data. We now shift our focus to using statistics to make inferences from data. These different roles of statistics—to describe and to infer—are bound up with the relationship between sample and population. Let's work through the dynamics of that relationship first (see Figure 7.1). Descriptive analysis is done in sample data, and any research question that can be addressed by the data can be answered *with certainty*—for the sample. For example, in our sample of 200 Kindergarten we can *know* the mean reading score for girls, the median math score for boys, or the number of students with a disability.

But sample data comes from a population, and often the sample is only a very small piece of the population. We can't *know* the mean reading score for the population of girls from sample data. Furthermore, samples differ. From the population of all Kindergarten students in the U.S., countless samples of 200 can be drawn, and each will be a unique subset of the population. Each of those samples will produce a different answer to the research question, and we have no idea how close our sample's "answer" is to the correct answer (the value of the parameter) is. As a result, conclusions made about the population from sample data are uncertain. Inferential statistics help us quantify that uncertainty with probabilities, allowing us to make probabilistic conclusions about population values or statements (such as the mean reading score for *all* Kindergarten girls in the U.S.). These probabilistic conclusions are called **estimates**.

Figure 7.1 The population – sample dynamic in statistical inference.



In this dynamic relationship between population and sample, the ultimate goal of research is to *know* population parameters and characteristics. Given that we only have samples and sample data to work with, the process of developing accurate conclusions about population parameters is a long-term project. Nothing is ever "proven" by the evidence in a single study. Rather, statistical inference is a disciplined process of accumulating and evaluating statistical evidence from sample data to make increasingly confident assertions about the state of affairs in the population.

7.2 Dimensions of statistical inference

There are four types of knowledge about population parameters and relationships between variables that we establish by inference. These are presented in Table 7.1, along with the method(s) used to establish them, and the kind of evidence that we evaluate in making probabilistic conclusions about each dimension. Let's work though these. The first two dimensions were introduced in Chapter 3, so we already know a bit about them. **External validity** is a fancy term for generalizability—the confidence we have that sample findings are likely to be true in the population. External validity refers to confidence that: a) the sample finding applies to potential participants who were not represented by (or included in) the sample but are part of the population, and b) the sample finding is not an artifact of the particular measures and methods used in the study. **Causal conclusion validity** is confidence that an effect or relationship observed in sample data reflects a true underlying causal connection between the variables in the population, and that all plausible alternative explanations have been ruled out. As we learned in Chapter 3, these inferential dimensions are not statistical per se; they have much more to do with the methods used in the study.

Table 7.1 Dimensions of statistical inference

Dimension	Method	Evidence
1 External validity Confidence that study findings based on sample data generalize to population of interest.	Representative sampling	random sampling large N concern with representativeness of subgroups
2 Causal conclusion validity Confidence that observed difference is caused by the predictor variable alone.	Experimentation	manipulated predictor variable (i.e., experimentally-arranged groups) control of all common-causal variables or alternative explanations
3 Statistical conclusion validity Confidence that the difference or relationship observed in sample data is a true (non-random, non-artifactual) phenomenon.	Null hypothesis testing Significance testing	low probability (p value) under the null hypothesis of an effect or relationship as large or larger than the one observed in the study
4 Effect size validity Confidence regarding the magnitude and direction of the true difference or relationship.	Parameter estimation Confidence intervals	precision in parameter estimation effect size statistic appropriate to effect

The next two dimensions require statistical methods, and will be the focus of the next two chapters. **Statistical conclusion validity** refers to confidence that an effect or relationship observed in sample data is not simply a random event, but reflects a true relationship in the population. As we observed earlier, millions of samples can be taken from a population and their estimates of some relationship will differ. We therefore need a method for evaluating the relationship in our sample data against the range of outcomes than occur randomly. If the sample evidence suggests that a true (i.e., non-random) relationship exists in the population, we want to attach a probability to that conclusion. The inference that a relationship *exists* in the population based on sample data does not conclude anything about that relationship's direction or size. **Effect size validity** is confidence that a (non-random) relationship observed in sample data has a particular direction and size in the population. For this inferential task we need to be able to estimate, with quantifiable precision, the direction and size of the relationship in the population that we observed in our sample data. Finally, let's reiterate that developing accurate knowledge (i.e., highly probable inferences) about population relationships requires evidence from more than one sample or study. We now turn to the focus of this chapter—how to use randomization methods to assess statistical conclusion validity.

7.3 Randomization test

How do we use statistics to develop statistical conclusion validity? In this section we learn the basic elements of a randomization test, through a simple example, and then extend those basic elements to understand Monte Carlo and permutation tests, which are variations on the randomization test. All are examples of **resampling** methods, which is a group of statistical methods that repeatedly sample from study data to develop distributions of outcomes that enable statistical inference. Let's see how resampling is used in the randomization test.

Imagine we have a brood of 6 newborn chicks and we want to test the effect of a special diet on weight. Because a randomization test depends on random assignment, let's assume the chicks are randomly assigned to diet condition (normal or special) in the study and after 6 weeks their weight is measured in grams (g). The study results show that the chicks on the normal diet weighed 136g, 141g, and 179g (mean weight = 152g), and the chicks on the special diet weighed 160g, 181g, and 229g (mean weight = 190g).

Is there evidence that the diet had an effect on chick weight? First let's acknowledge that if you had a large population of 6-week old chicks, the mean weights of random samples of 3 chicks would differ: some groups of chicks would be lighter or heavier than others—just by chance—because weight varies. So mere sampling variation could explain any outcome, including our study's outcome. What we need to know is *how likely* it is for two randomly-arranged samples of 3 chicks to have mean weights as different as (or more different than) 152g and 190g.

The randomization test was developed by Sir Ronald Fisher in the 1920s to evaluate the null hypothesis associated with some treatment or intervention. A null hypothesis (H_o) is a statement that the treatment has no effect. In our example, H_o would be "Diet has no effect on chick weight." The alternative hypothesis (H_A) would include all other outcomes. In our example, H_A would be "Diet has an effect on chick weight." The logic of a randomization test starts with assuming that the null hypothesis is true. If H_o is true (and there is really no effect of special diet compared to normal diet on weight), then the chicks' weights are **exchangeable**—that is, any weight value observed in the normal diet group could just as easily have shown up in the special diet group and vice versa. Under the null hypothesis, we should be able to simply randomize our 6 chick weights into 2 groups of 3 scores each because all 6 scores are exchangeable. A randomization study produces all possible random

arrangements of N scores into groups of k scores (in our study, N=6 and k=3). For you math nerds, the number of permutations of N scores into samples of size k and N-k can be calculated as follows:

Formula 7.1 $$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

For each permutation* or random arrangement of scores the statistic of interest is calculated. That procedure generates a **reference distribution** of statistics (in this case we're interested in the mean difference) under H_o. Table 7.2 shows the 20 permutations of 6 scores into 2 groups of 3 from our example, as well as the group means and mean difference for each. Notice that our study data (highlighted below) is one of those random arrangements.

*In math, permutations imply different orderings of objects, but in statistics we use the term synonymous with random arrangement. We don't consider the same scores but in a different order to be a unique random arrangement, mainly because they produce the same statistic.

Table 7.2 Permutations of sample data where N=6 and k=3

Permutation	Group 1	M_1	Group 2	M_2	$M_1 - M_2$
1	136 141 160	145.7	179 181 229	196.3	-50.7
2	136 141 179	152	160 181 229	190	-38
3	136 141 181	152.7	160 179 229	190	-36.7
4	136 141 229	168.7	160 179 181	173.3	-4.7
5	136 160 179	158.3	141 181 229	183.7	-25.3
6	136 160 181	159	141 179 229	183	-24
7	136 160 229	175	141 179 181	167	8
8	136 179 181	165.3	141 160 229	176.7	-11.3
9	136 179 229	181.3	141 160 181	160.7	20.7
10	136 181 229	182	141 160 179	160	22
11	141 160 179	160	136 181 229	182	-22
12	141 160 181	160.7	136 179 229	181.3	-20.7
13	141 160 229	176.7	136 179 181	165.3	11.3
14	141 179 181	167	136 160 229	175	-8
15	141 179 229	183	136 160 181	159	24
16	141 181 229	183.7	136 160 179	158.3	25.3
17	160 179 181	173.3	136 141 229	168.7	4.7
18	160 179 229	189.3	136 141 181	152.7	36.7
19	160 181 229	190	136 141 179	152	38
20	179 181 229	196.3	136 141 160	145.7	50.7

Next we convert our observed statistic, which is often referred to as a **test statistic**, to a probability. We use the reference distribution from the randomization procedure, and find the probability of an outcome that is at least as large as our observed outcome (test statistic = -38g). In Table 7.2 we see 4 scores ≥ 38g (in the positive or negative direction), so the probability of observing a mean difference ≥ 38g is 4/20, or .20. Why do we disregard the sign of the mean difference when computing the probability? Because a randomization test is a test of H_o, and the alternative to "no effect" is "any effect." This probability (written formally as $p(\text{observed} \mid H_o) = 4/20 = .2$) is called a **p-value.** The p-value is a *conditional* probability, with the vertical line in the statement above indicating "conditioned upon" or "given." So, conditioned upon H_o being true, the probability of observing a mean difference as large or larger (in either direction) than the difference we observed in our study is .2.

7.4 Interpreting the results of a randomization test

Interpreting the results of a randomization test hinge on three principles: proper understanding of a p-value, whether the assumption of exchangeability is met in the original study data, and what statistical alternatives are included in the alternative hypothesis. Let's take each of these principles, think them through, and see how they affect conclusions from a randomization test.

p-value. A p-value is a broadly misunderstood statistic. To understand the meaning of, and to properly interpret, a p-value we use the guidelines set out by Sir Ronald Fisher.

- **A p-value is a continuous measure of evidence against H_o**. The lower the p-value, the stronger the evidence against H_o.
- **A p-value is just one piece of evidence**. A p-value must be combined with other pieces of evidence, ideally across multiple studies, to help draw conclusions from data. A p-value from a single study doesn't "prove" anything: a low p-value is evidence against the validity of H_o but does not establish H_o as false.
- **A p-value is not evidence of H_A**. A low p-value does not speak to the validity of H_A or establish it as true. Remember, if H_o is false, there are many alternative hypotheses to be considered as potentially responsible for the observed outcome. And, again, evidence from multiple studies is needed to validate a particular H_A.

- **A p-value of .05 is a reasonable cutoff**. According to Fisher, $p < .05$ is a reasonable cutoff for evidence that calls the validity of H_o into question and allows us to further consider the alternative hypotheses.

You may have heard the terms "significant" or "statistically significant," and likewise "nonsignificant." These terms are ubiquitous in scientific articles, and media coverage of science, and are used to characterize study findings where the p-value was below some threshold value set to declare a finding significant. We avoid the use of that term for two reasons. First, a p-value is a continuous measure of evidence. When it is used to confer a binary status ("significant" or "nonsignificant") we abandon the true continuous nature of the underlying evidence. Declarations of "significance" based on a p-value also convey much more certainty about the outcome than is warranted, or that Fisher ever intended in setting up the principles of statistical inference. Second, the term "significant" in common usage implies "meaningful" or "important." A p-value is merely a probability of some observed outcome (or greater) occurring under the null hypothesis; it says nothing about the importance of the outcome. But the common and statistical meanings of "significance" are often conflated, resulting in misleading interpretations of inferential tests and p-values.

Our randomization test of the effect of special diet compared to normal diet on weight in 6 chicks produced a p-value of .2. This indicates that there was no evidence against the null hypothesis, *in that particular study*. Remember, a p-value from single study does not confirm H_o (i.e., there is absolutely no effect of special diet on chick weight), so another study (with different chicks, different bags of feed, different environmental condition, etc.) might find something different.

Exchangeability. A randomization test assumes exchangeability in the original scores, meaning that all random arrangements of scores are equally likely. This is an important assumption because the randomization test gives equal weight to each permutation of scores in the calculation of the p-value. If all permutations of scores are *not* equally likely, then the p-value is not accurate and mischaracterizes the amount of evidence against H_o. When the data come from an experiment, the assumption of exchangeability is justified. Indeed, a randomization test mirrors the random assignment of participants (i.e., chicks) to condition (i.e., special vs normal diet) in a study. In that case, as we learned in Chapter 2, causal conclusions can be made about randomization test results. Although our chick study did not result in much evidence against H_o, if it had, we could have interpreted that

difference in weights as caused by the diet variable. In the next section we look at how a randomization study changes when the assumption of exchangeability is not justified.

Statistical alternatives under H_A. As we stated earlier, the alternative hypothesis (H_A) in a randomization test includes all "non-null" outcomes. This is why our H_A was stated "Diet has an effect on chick weight." It is important to recognize that a treatment or intervention can be reflected in location statistics (e.g., the mean difference), variability statistics (e.g., variance ratio), and even symmetry statistics. *A randomization test can be conducted using any outcome statistic, but interpretations of the test are limited to that statistic*. If, for example, we expect that feeding chicks a special diet will affect all chicks about the same, then a location statistic like the mean difference would capture that effect. If we suspected however that diet would make some chicks heavier than others, then a variability statistic might better capture that effect. Our interpretations of randomization test results, then, should acknowledge the particular statistical lens used to view the outcome. With that said, we learn more about an intervention's effect on an outcome if we view it through different statistical lenses. We will demonstrate this below, as well as in Chapter 9, when we use randomization tests in data analysis.

7.5 Subtypes of the randomization test

The randomization test is an umbrella term that includes two variants, the **Monte Carlo test** and the **permutation test**. Although they are often referred to as "randomization tests," it is important to know how they differ because a Monte Carlo test involves different computations, and a permutation test different interpretations, than the randomization test covered above. We demonstrate these two types of randomization studies in data analysis in Chapter 9, but for now let's distinguish them conceptually.

Monte Carlo test. A true randomization test uses all possible random arrangements of scores to construct the reference distribution, which in turn produces an *exact* p-value. In most studies, "all possible random arrangements" is a very large number and often too large to be computationally feasible. For example, in a study that randomly assigned 50 participants to one of two groups (25 per group)—which is really not a large study as studies go—there are 1.26×10^{14}, or 126 trillion, random arrangements of scores. It would take most desktop computers several hours to work through the calculations of such a large randomization test. Monte Carlo methods operate on the same principle as

random sampling: a random sample of sufficient size generates an accurate picture of the population, allowing sample-based findings to be confidently generalized to the population. In this case we're interested in accurately capturing the true reference distribution, so a Monte Carlo test randomly samples from the population of all random arrangements to generate an approximation of the true reference distribution. Monte Carlo samples of around 5000 generate estimated reference distributions that are very accurate, and provide estimated p-values that are very close to exact p-values. Because very few studies are as small as our chick weight study (N=6), most randomization studies actually use Monte Carlo methods, generating estimated rather than exact reference distributions and p-values.

Permutation test. A true randomization test assumes exchangeability in the original study data. This is assured by random assignment of participants to conditions (i.e., experimental research), in which scores from one condition could have just as easily been in the other condition. In studies where the predictor variable is selected rather than experimentally-arranged (i.e., non-experimental research), scores from one condition are *not* exchangeable with scores from the other condition. In Chapter 2 we compared girls' and boys' scores on a mathematics standardized test. Samples of female and male Kindergartners are already different when they enter the study, so their scores are not exchangeable. The lack of random assignment in the study does not preclude the use of randomization methods for hypothesis testing. A permutation test, then, is simply a randomization study of non-experimental data. The use of "permutation" here recognizes the lack of exchangeability. A study that doesn't randomize participants to groups cannot establish exchangeability. A permutation study still produces a reference distribution (of, say, the mean difference between girls and boys on the math test) which in turn generates a p-value (exact if true randomization, estimated if Monte Carlo) that speaks to the null hypothesis. Causal conclusions about the "effect" of gender on math ability are, however, precluded.

7.6 Randomization tests in R

In this section we learn how to set up and run a randomization test in R. The focus will be understanding the logic of the coding needed to carry out the resampling procedure, and generating and using the reference distribution to do null hypothesis testing. The goal is that once you understand

the basic architecture of the randomization test as it is expressed in R code, you can adapt it to address any hypothesis testing problem with any statistic.

We will use the Melanoma data in the `MASS` package for our example, which is data on 205 patients with malignant melanoma. First we call up the dataframe information and documentation so we can see how variables are coded and their measurement units.

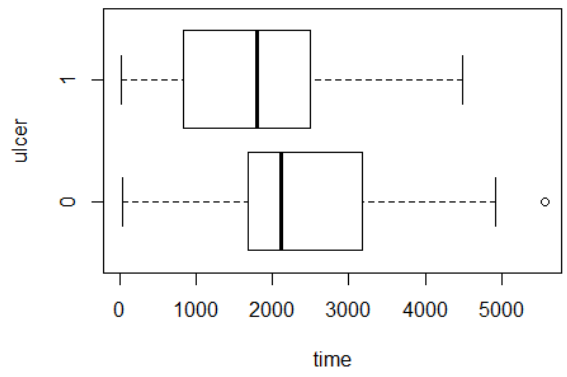
```
library(MASS)
str(Melanoma)
```

```
## 'data.frame':    205 obs. of  7 variables:
##  $ time     : int  10 30 35 99 185 204 210 232 232 279 ...
##  $ status   : int  3 3 2 3 1 1 1 3 1 1 ...
##  $ sex      : int  1 1 1 0 1 1 1 0 1 0 ...
##  $ age      : int  76 56 41 71 52 28 77 60 49 68 ...
##  $ year     : int  1972 1968 1977 1968 1965 1971 1972 1974 1968 1971 ...
##  $ thickness: num  6.76 0.65 1.34 2.9 12.08 ...
##  $ ulcer    : int  1 0 0 0 1 1 1 1 1 1 ...
```

```
?Melanoma
```

Let's concern ourselves with the question of the relationship between **ulcer** status (1=presence, 0=absence) and survival **time** (measured in days). Because we have a categorical predictor and a numeric outcome, refer back to Chapter 3 for a refresher on how to do descriptive data analysis. To prepare for the randomization study we'll do a couple key bits of that here, generating boxplots and mean survival times by ulcer group. We will also need to know the ulcer group sample sizes (or group ns), which we can get with the `table()` function.

```
boxplot(time~ulcer,data=Melanoma, horizontal = T)
```




```
tapply(Melanoma$time,Melanoma$ulcer,mean)
```

```
##        0        1
## 2414.965 1817.811
```

```
table(Melanoma$ulcer)
```

```
##
##   0   1
## 115  90
```

This quick descriptive analysis shows that melanoma patients with an ulcer, compared to those with no ulcer, have shorter survival times on average. The randomization test tests the null hypothesis, so let's set up our hypotheses for this test.

H_o: There is no relationship between ulcer status and melanoma survival time.

H_A: There is a relationship between ulcer status and melanoma survival time.

Notice that I used the phrasing "relationship between" rather than "effect on" in the hypothesis statements. This is to recognize that ulcer groups were not experimentally arranged in the original study. Rather, melanoma patients developed an ulcer or not (or they already had an ulcer prior to the study or not), and those with an ulcer were compared to those with no ulcer. We can imagine those groups of patients differed in many other ways than just the presence or absence of an ulcer. Lacking experimental manipulation of the predictor variable in the original study (which in this case would involve randomly assigning people to be given an ulcer), we cannot infer any causal effect of ulcer on survival time. These observations about the study design also confirm the lack of exchangeability in the

data, requiring a permutation study. We will be using Monte Carlo methods because the number of unique random arrangements (or permutations) of 205 scores into groups of 115 and 90 is on the order of 10^{58} and not computationally feasible. Let's walk through the randomization test in steps:

Step 1. Set up test statistic. The first step in the randomization (technically, here, a Monte Carlo permutation) test is to decide on the outcome statistic we will use to test H_o. The decision depends somewhat on how you think the predictor will affect the outcome, and what's the best statistical lens to "see" that effect. Let's use the mean difference as our test statistic. Below we find the mean difference in survival time between ulcer conditions, and save it in an object (**stat**) for later use. We also arrange the calculation so that the mean difference is a positive number; why we do that will be clear in Step 4.

```
stat=(mean(Melanoma$time[Melanoma$ulcer == 0]))-(mean(Melanoma$time[Melanoma$ulcer
== 1]))
stat
```

```
## [1] 597.1541
```

The mean difference in survival time between patient with and without an ulcer is about 597 days. We want to know how probable it is to get a difference that large or larger if H_o is true, or in other words if having an ulcer is unrelated to survival time.

Step 2. Set up resampling loop. The next step in the randomization study is to set up the loop by which we will sample from the sample data, randomize the sample values into groups, calculate the statistic of interest, store that value, and repeat the process. We also want to make sure that our loop conducts this process under the condition that H_o is true. Below is a for-loop set up to conduct that resampling task. Since this introduces some new R functions and concepts, let's walk through the code line by line.

```
set.seed(1008)
N=5000
meandiff=numeric(N)
for (i in 1:N) {
  data <- sample(Melanoma$time,205,replace=FALSE)
  grp1 <- data[1:115]
  grp2 <- data[116:205]
  meandiff[i] <- mean(grp1)-mean(grp2)
}
```

- The `set.seed()` function is only there to ensure that when you run the code you get the same results as everyone else. Anytime we use the `sample()` function, it will produce

independent and unique sets of objects for each person. So for this example only I wanted everyone to end up with the same results, hence the set.seed.

- N defines the number of times we run the operations in the loop. For a Monte Carlo study, the number of samples needed to create an accurate reference distribution is recommended to be 5000. N is also used for sample size (file this under: there are only so many Greek and Roman letters for statisticians to use, so they get used for different things). Nevertheless, this could be confusing, so we need to be careful in how we report the analysis.

- The `numeric()` function creates a numeric vector, called **meandiff** in this example, which will eventually contain N values. Think of it as a container in which you're storing each statistic generated by one pass through the resampling procedure.

- The `for()` function is the engine of a for-loop, which is a repeating series of operations that result in a piece of output. Each pass through the loop is an i (or iteration), and we pass through the loop N times. For each i, we:

    - Take a random sample of our time data. Because we are sampling *without* replacement (replace=FALSE) the `sample()` function will select a random time value, store it in the object, then select another random value from the remaining 204 cases, etc. until all 205 cases have been selected. In other words, sampling without replacement essentially creates a random *ordering* of the 205 values in the time variable. We store that random arrangement of values in an object called **data**.

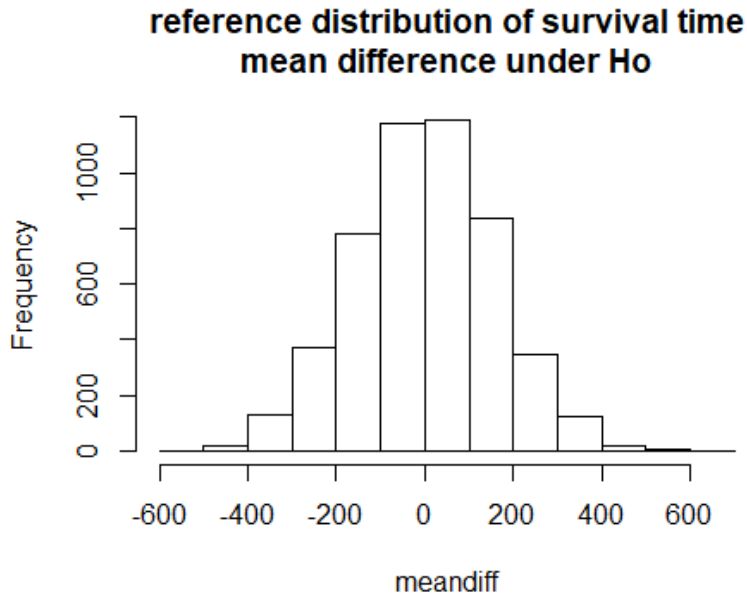
    - Divide the 205 values into two groups of the size of the original ulcer groups. The first 115 cases are called **grp1** and the remaining 90 cases are called **grp2**. It is important to maintain the original group sample sizes in a randomization study. This process is no different than shuffling a 205-card deck and dealing the cards into piles of 115 and 90. The randomizing of time values into groups is what we expect to happen under H_o: group 1 and group 2 will differ, but only randomly.

    - Calculate the mean difference from our randomized groups and deposit that value in the numeric **meandiff** container created above.

This process repeats N times, each time with a new random sample of weight values.

Step 3. Generate reference distribution. The reference distribution is the distribution of mean difference statistics for a random sample of 5000 permutations of 206 scores into groups of 115 and 90. We don't need to look at a histogram of the reference distribution to use it, but the histogram does help us to see and appreciate how the outcome statistic behaves under H_o. We can see that most sample mean differences cluster around zero, which is logical if we remind ourselves that under H_o ulcer status is unrelated to survival time. Larger mean differences in both directions are progressively less probable occurrences under H_o.

```
hist(meandiff, main="reference distribution of survival time
mean difference under Ho")
```



Step 4. Calculate p value. The p-value for our example is the probability of observing a mean difference at least as large as our test statistic, under H_o. Since *all* the outcomes in the **meandiff** container occurred under H_o, we simply need to count those outcomes that exceeded 597.2 and divide by N to get the probability. Remember that we want to count values that are larger than the test statistic in both the positive and negative direction, because H_A includes the alternative hypotheses that ulcer, compared to no ulcer, will be associated with both less and more survival time. In R we can do this by taking the absolute value (with the `abs()` function) of the resampled mean differences and

using the `length()` function to count how many are greater than or equal to the test statistic, then divide that total by N.

```
pvalue=length(which(abs(meandiff)>=stat))/N
pvalue
```

```
## [1] 2e-04
```

Having converted our test statistic to a probability under H_o, we find p=.0002. This constitutes strong evidence against H_o, encouraging us to explore alternative hypotheses. The data from the original study suggest that the presence of an ulcer in melanoma patients is associated with shorter (rather than longer) survival times. This is of course one of many plausible alternative hypotheses, mainly because the original study was not experimental and thus could not eliminate those alternative explanations with random assignment of patients to ulcer condition.

Let's look at one more example of a randomization test by using the median difference as our outcome statistic. Referring back to the boxplots generated above, we see two things: one, the presence of an outlier in the high end of the no ulcer group, which has an effect on the mean of that group; and two, the group medians appear to be less different than the group means were. So, as a way to enhance **statistical conclusion validity** with regard to the conclusion that *typical* survival times are higher in patients with no ulcer compared to patients with an ulcer, let's test that hypothesis with a different location statistic—the median difference. H_o and H_A remain the same in this test.

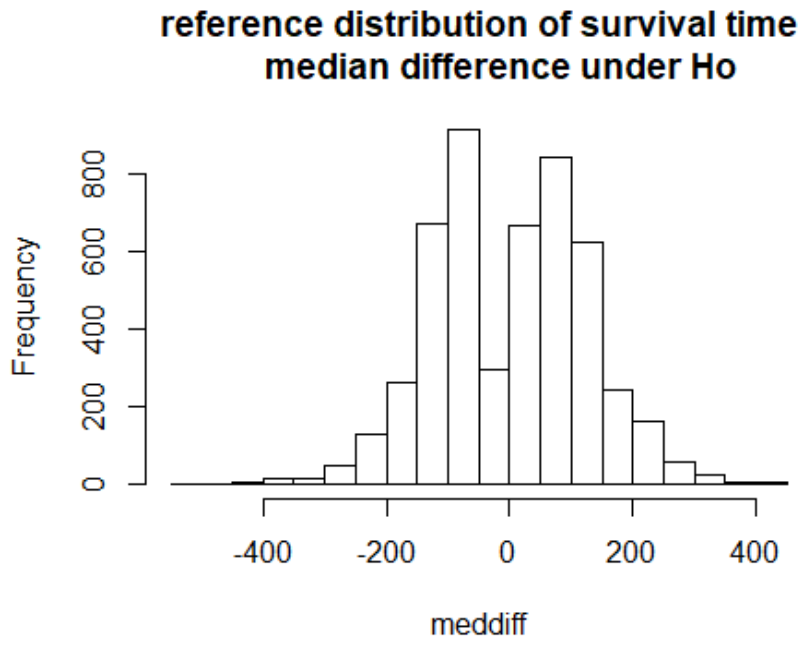
We combined all the steps of the randomization study for this example, but the changes to the code are noted below with #annotations.

```
#new test statistic
stat=(median(Melanoma$time[Melanoma$ulcer == 0]))-
  (median(Melanoma$time[Melanoma$ulcer == 1]))
stat
```

```
## [1] 303.5
```

```
N=5000
meddiff=numeric(N)     #newly named numeric vector object
set.seed(1008)
for (i in 1:N) {
  data <- sample(Melanoma$time,205, replace=FALSE)
  grp1 <- data[1:115]
  grp2 <- data[116:205]
  meddiff[i] <- median(grp1)-median(grp2)      #median difference
}
```

```
hist(meddiff, main="reference distribution of survival time
     median difference under Ho")       #call meddiff object
```




```
pvalue=length(which(abs(meddiff)>=stat))/N       #call meddiff object
pvalue

## [1] 0.012
```

Notice that the median difference (303.5 days) is roughly half of the size of the mean difference. Nevertheless, that difference is still a highly improbable event under H_o (p=.012), constituting moderate to strong evidence against H_o. Together, these two randomization studies—each with a different location statistic—provide consistent evidence that the observed differences in typical survival times of ulcer and no-ulcer patients are not random. They also illustrate that statistical conclusion validity (see Table 6.1) is best established with an accumulation of evidence.

7.7 Writing up the results of a randomization test

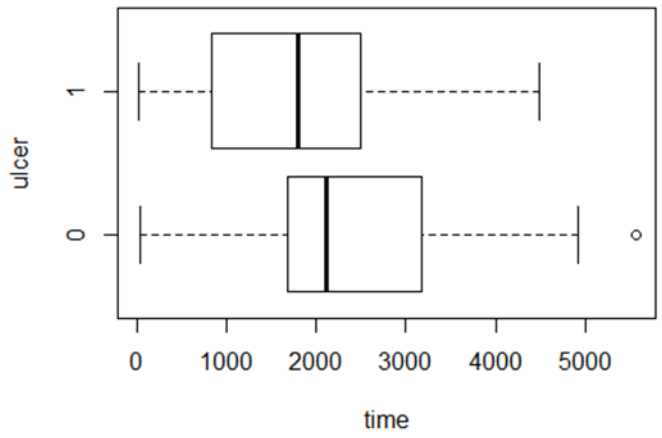
We will do much more with randomization tests in data analysis in Chapter 9, testing hypotheses in studies with various predictor and outcome variable types, with and without exchangeability. To end this chapter, here is a written summary of the analysis we did on the ulcer and survival time data. Just a reminder: we don't know how these patients were sampled, so we can't assume random sampling and thus these findings lack external validity. We established earlier that the lack of random

assignment, and exchangeability, in the original study weakens the causal inference we might have made about the *effect* of having an ulcer on survival time.

Results

The relationship between the presence or absence of an ulcer on survival time (in days) was examined in a sample of patients with malignant melanoma (N=205) with a randomization test of the median difference. The median survival times for patients with, and without, an ulcer were 1799.5 days and 2103.0 days, respectively (see Figure 1). The randomization test created a reference distribution of 5000 median differences under the null hypothesis of no relationship between ulcer status and survival time. The observed median difference (303.5 days) had an associated p-value of .012. This was strong evidence against H_o, and suggests that the presence of an ulcer is associated with shorter survival times in melanoma patients.

Figure 1. Survival time (days) by ulcer group (0=absence, 1=presence)



7.8 Problems

7.8.1. What dimension of statistical inference is related to:

a. sampling in the original study?
b. whether the original study is an experiment?
c. distinguishing between random and non-random study outcomes?
d. estimating the true magnitude of an effect or relationship?
e. showing that study outcomes are reliable?

7.8.2. What is the assumption of exchangeability and how is it consequential in randomization studies?

7.8.3. Describe the difference between a randomization study and a Monte Carlo study?

7.8.4. What is a reference distribution and what role does it serve in hypothesis testing with a randomization test?

7.8.5. Can exchangeability be assumed in data from non-experimental studies? Explain why not.

7.8.6. What is the difference between the null and alternative hypotheses in a randomization test?

❽ Bootstrapping Methods for Parameter Estimation

Covered in this chapter:

- The logic of bootstrapping
- Parameter estimation and confidence intervals
- Bootstrapped confidence intervals
- Bootstrapping for estimation in R

8.1 Introduction to estimation

Statistical inference involves making probabilistic conclusions, based on sample data, about population parameters and relationships between variables. In chapter 6 we learned about hypothesis testing and **statistical conclusion validity**, which involves observing sample-based evidence against H_o and therefore evidence of some non-random effect or relationship. Statistical conclusion validity does not, however, speak to the magnitude of that effect or relationship. This is an important issue because statistical conclusion validity can occur even when the effect size is trivially small, if certain conditions are present (e.g., large N, small sample variances, or a combination of both). So "statistical significance" does not imply that the observed effect or relationship is large or meaningful in practical terms. We also should be aware that treatment effects or relationship sizes will vary, even across hypothetical identical studies, because samples vary. **Effect size validity** is the statistical inference dimension that involves making probabilistic statements about the direction and magnitude of an effect or relationship based on sample data. It is also worth reminding ourselves that although statistical conclusion validity indicates evidence for *some* non-random effect or relationship, we cannot assume that the predictor variable is solely responsible for that effect (i.e., causal conclusion validity). That inference, of course, depends on the control over alternative explanations that were designed into the study that produced the data.

The *true* size and direction of an effect or relationship—meaning the size and direction of the effect in the population, could it be known—is a **parameter**. Establishing effect size validity then is a process of estimating that parameter based on sample data. In this chapter we will learn the statistical tools and methods for parameter estimation, including a resampling method called **bootstrapping**.

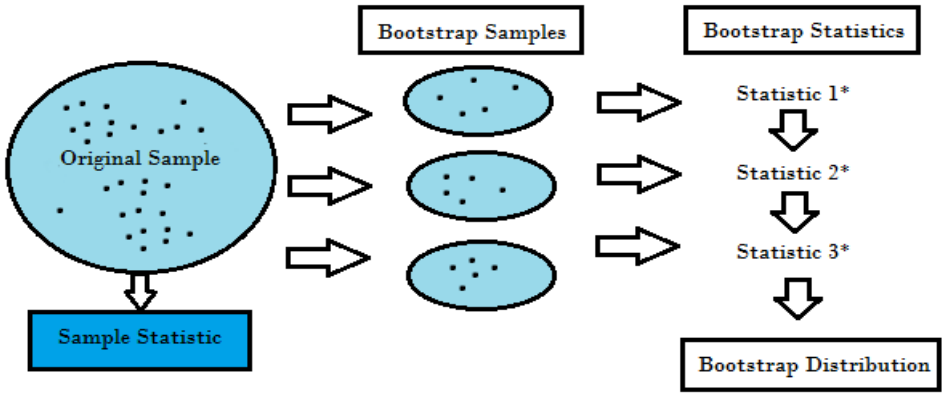
Bootstrapping is to parameter estimation as the randomization test is to null hypothesis testing. Through bootstrapping we will create a reference distribution of outcomes—called a bootstrap distribution—that will allow us to construct a probabilistic estimate of the parameter.

8.2 The logic of bootstrapping

The logic of bootstrapping operates on the law of large numbers, which states that for any statistical experiment (e.g., a coin toss, a dice roll), the mean of a large number of repeated experiments is very close to the expected value of the outcome. For example we know that the true probability (or expected value) of tossing heads on a coin is .5. If you toss a coin 10,000 times, the number of heads will be very close to .5. Now, if we make the statistical experiment a random sample and the outcome a statistic (e.g., mean, median) the law of large numbers says that if we repeatedly sample from a population, each time calculating a statistic, the mean of those sample statistics will be very close to the statistic's expected value. The problem is that we rarely have access to a population from which to sample, and that's where bootstrapping comes in.

In **bootstrapping** we treat our sample data as a surrogate, or stand-in, population. Indeed, *any* sample is a surrogate of the population from which it came, although as we learned in Chapter 2 random samples—particularly large random samples—are more representative of their parent populations than convenience samples. The primary assumption underlying statistical inference by bootstrapping is that our sample is reasonably representative of the parent population. Bootstrapping involves **sampling with replacement** from our surrogate population (the sample data), calculating and saving a statistic from that sample, and repeating the process a large number of times. Sampling with replacement means that a value is chosen from the data, recorded, and then "thrown back" into the sample. As a result, bootstrapped samples can select the same value multiple times, and that's OK. In fact, sampling with replacement is what enables bootstrapping to create samples that reflect the surrogate population. Bootstrapping uses the **plug-in principle**, which states that each bootstrapped sample statistic "plugs in" for the parameter—the unknown population value of interest. This resampling process generates a distribution of bootstrapped statistics, called a bootstrap distribution, which will enable us to estimate the parameter. These principles are illustrated in Figure 1.

Figure 8.1. Bootstrapping



Source: https://www.statisticshowto.com/bootstrap-sample/

It is important to recognize that a bootstrapped distribution is always centered around the sample statistic, not the population parameter. Bootstrapping isn't a method for magically "seeing" the population parameter from our sample data, nor will it help us improve the accuracy of a parameter estimate. It bears repeating: the best parameter estimates come from large, random samples from the population. The accuracy of any estimate depends on the how closely the sample reflects the population—if your sample is biased, its estimates will be biased. What bootstrapping *does* do is tell us how accurate, or precise, our sample-based estimate is. A bootstrap distribution approximates the true sampling distribution of a statistic; this is good because we rarely can sample repeatedly from the population and get a picture of that "true" sampling distribution. So bootstrapping is very useful for getting estimates of standard errors of statistics and creating confidence intervals. More about these concepts next.

8.3 Parameter estimation

Let's revisit our Chapter 3 example for a minute. In that study we compared samples of boy and girl Kindergarten students on a math achievement test. The point of doing the study was to be able to say something—based on that particular sample of 200 students—about the *true* difference between girls' and boys' math achievement (meaning in the population of *all* Kindergarten students). Research is almost always interested in parameters, and parameters are almost always unknown, hence the need to estimate them. We are aware however that another sample of 200 Kindergarten students would

yield a different estimate, a third sample yet another estimate, and so on. Parameter estimation must address this problem—that sample estimates of a parameter vary—and it does so with a confidence interval.

A **confidence interval** is a parameter estimation method that identifies an interval, defined by lower and upper values, that we have some confidence includes the parameter. In interval estimation, "some confidence" is a precise value that is set by the analyst or scientist. That confidence level is often set at .95, by convention, which results in an interval of values that we are 95% confident includes the parameter. We will come back to the issue of interpreting a confidence interval in a bit. For now, we need to understand how a confidence interval is constructed, and where the quantities that contribute to the calculation of a confidence interval come from.

There are three contributors to a confidence interval, as shown in the conceptual formula below: a statistic, the standard error of that statistic, and a multiplier that establishes the confidence level for our interval estimate of the parameter.

statistic ± (standard error of the statistic * confidence level multiplier)

The product of the standard error and the confidence level multiplier produces the **margin of error**, a statistic is reported a lot in fields that use survey and polling research. The margin of error is then subtracted from, and added to, the statistic to generate the lower and upper values of the interval. Next we look more closely at each of the terms in the confidence interval.

Statistic. Parameter estimation begins with a statistic, obviously. Our Chapter 3 example was interested in the mean difference between populations of girls and boys on a math achievement test. So we could simply use a single value—the mean difference from our sample data—to estimate the parameter; this is called **point estimation**. The value of a point estimate from a large random sample is hard to overstate. Point estimation does not account for the error inherent in any sample statistic, however. Sample statistics vary depending on the particular sample you happen to get in your study. Point estimates don't tell us how much they vary, or could be expected to vary, across hypothetical samples—you know, all the samples you *didn't* get in your study. Interval (unlike point) estimates do incorporate an estimate of the statistic's error, which is what we turn to next.

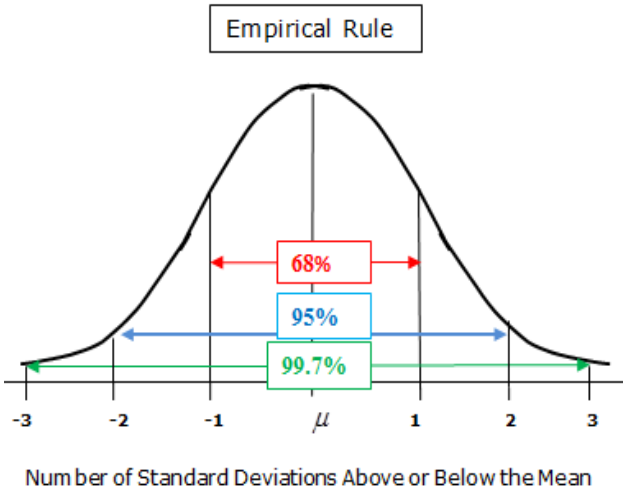
Standard error. The standard error of a statistic is the average amount of variability in the statistic when observed over many random samples of the same size from the population of interest. Because that variability is due to the random differences in samples, it is called error. But the standard error is conceptually equivalent to the standard deviation, which you recall from Chapter 2 quantifies the average deviation from the mean in a sample of scores. The standard error, then, is simply the standard deviation of a distribution of statistics. Every statistic has a standard error. The standard error of a statistic tells us how much variability, on average, we could expect if we repeatedly drew samples from the population and calculated the statistic in each. So the standard error is essential to interval estimation because it quantifies the uncertainty inherent in using a particular sample statistic to estimate the parameter, knowing that another sample would produce a different value of the statistic.

If we have the statistic, how do we find the standard error of the statistic? Well, there are two ways. First, some standard errors can be calculated from a formula. For example, the standard error of the mean ($SE_{\bar{x}}$) can be calculated from the sample standard deviation and sample size as follows: $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}$. This formula produces only an *estimated* standard error of the mean, and the accuracy of that estimate depends on assumptions (i.e., that the data are randomly sampled and normally distributed in the population) that are often not met in real data. An inaccurate standard error estimate will then pass its inaccuracy on to the confidence interval. The second method is to bootstrap the standard error of the statistic whose parameter we are interested in estimating, which we cover in the next section. We *must* use a bootstrapped standard error if we want to estimate with a statistic that doesn't have a derived standard error formula. We *can* use bootstrap standard errors for estimating with means and proportions (statistics that have standard error formulas), and we might actually want to if the assumptions underlying those formulas are not met.

Confidence level multiplier. Let's go back to a concept you learned in high school called the **empirical rule** (see Figure 8.1). The empirical rule applies the properties of the standard normal distribution (i.e., a normal distribution with μ = 0 and σ = 1) to some random variable, which ideally is also normally distributed. The empirical rule tells us the percentage of observations in that random variable in reference to common standard deviation (or z score) cutoffs. For example, about 68% of the observations of a normally-distributed random variable are expected to be between z ± 1. If you turn that around to create a tool for estimating μ, you can see that the interval formed by z ± 1 includes the

most likely 68% of the values of μ. In other words, the multiplier 1*SD produces a 68% confidence interval for estimating the parameter. If we want a more confident estimate, say 95% confidence, we would need an interval that included the most likely 95% of the values of the parameter, or 2*SD. So the multiplier is just the value of the standard (in this case, z) score that sets the confidence level of the interval estimate by including that particular percentage of the most likely values of the parameter (in this case, μ).

Figure 8.1 The empirical rule



Source: https://www.softschools.com/math/probability_and_statistics/the_normal_distribution_empirical_rule/

Normal-theory interval estimation. Let's apply these ideas to the problem of estimating a population mean (μ) from sample data using the classic normal-theory method. Below is the conceptual formula introduced earlier for building a confidence interval; under that is the normal-theory formula for building a 95% confidence interval to estimate the population mean.

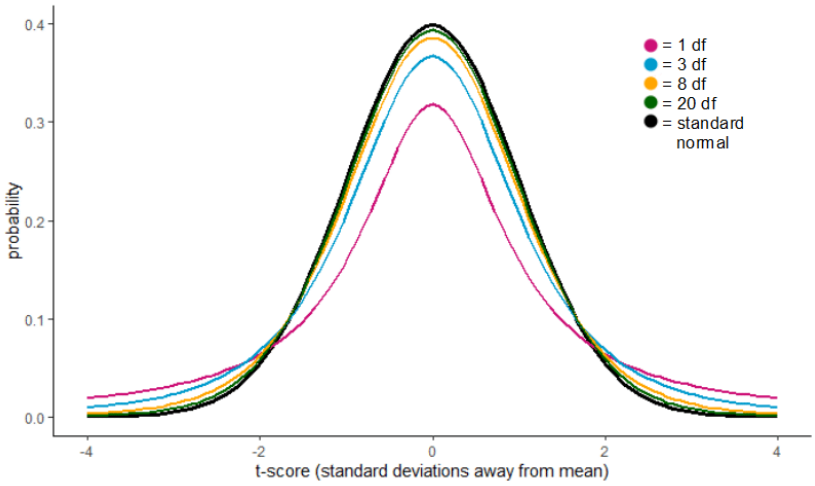
statistic ± (standard error of the statistic * confidence level multiplier)

$\bar{X} \pm (SE_{\bar{X}} * t_{.95})$, where $SE_{\bar{X}}$ is estimated by $\frac{s_x}{\sqrt{n}}$

Notice that the standard error of the mean is estimated from sample values. Why do we estimate it?—because normal theory specifies that the actual standard error of the mean is given by $\frac{\sigma_x}{\sqrt{n}}$, but we rarely know what σ_x is, so we estimate it from sample data. This helps us understand why our multiplier

is a t score rather than a z score. The t distribution—which is really a family of normal distributions (see Figure 8.2)—is used to estimate the population mean in situations when we have small-sample data and don't know σ. The various t distributions are indexed by their degrees of freedom (df), which is simply n-1 in this example. So, when n is very large, the t and standard normal (or z) distributions are nearly identical, and offer similar multipliers. As sample size gets smaller, however, the t distributions increasingly reflect more variability (because small-n estimates vary more), which means that the t multiplier will be larger than the z multiplier for a given confidence level. Table 8.3 presents some examples of multipliers from the z and t distributions for forming a confidence interval. The larger multipliers for confidence intervals based on small samples produce less precise estimates, as we will see shortly.

Figure 8.2 The t distribution



Source: Bevans, B. (November, 2020). https://www.scribbr.com/statistics/t-distribution/

Table 8.3 Multipliers from the z and t distributions for several sample sizes

	multiplier for:	
	90% CI	95% CI
z (large n assumed)	± 1.68	± 1.96
t (n=30, df=29)	± 1.70	± 2.05
t (n=20, df=19)	± 1.73	± 2.09
t (n=10, df=9)	± 1.83	± 2.26

The normal-theory confidence interval using a t multiplier is widely used for parameter estimation. And for good reason: its estimation accuracy (i.e., does it contain the parameter at the expected confidence level?) is very good, if the sample is large (say, n > 100) and the sample data are normally distributed and come from a random sample of the population. Those conditions aside, the normal-theory estimation method has one big limitation: it only works if you have a formula for calculating the standard error of the statistic from sample data. Many researchers use means in their research, and therefore want to estimate the population mean. Conveniently, there is a formula for calculating the estimated standard error of the mean (see above). But that is not the case for many other statistics that we might want to use in research, such as the median or risk difference. In the next section we learn how bootstrapping extends the usefulness of normal-theory estimation methods and opens up alternative methods that are not bound to normal theory.

8.4 Bootstrapped confidence intervals

In this section we introduce two confidence interval methods that use bootstrapping: the **t interval with bootstrap standard error** and the **bootstrap percentile interval**. These two methods are the hammer and screwdriver of the confidence interval toolbox; they're intuitive, easily calculated, and accurate in most circumstances. Let's learn the R code for bootstrapping and how we use that to construct confidence intervals.

8.4.1 The **t interval with bootstrap standard error** follows the formula covered in the previous section (reproduced below) with one exception.

$$\bar{X} \pm (SE_b * t_{.95})$$

We substitute the bootstrapped standard error of the mean (SE_b) for the normal-theory standard error of the mean ($SE_{\bar{x}}$). This substitution allows the t interval to be used to estimate parameters whose standard error is not possible to calculate with a formula (e.g., the median). For this first example, we'll stick with the sample mean, and construct a bootstrapped interval to estimate the population mean (μ). Let's revisit the Chapter 2 example of math achievement in Kindergarten students, and imagine we're interested in the mean math achievement score in the population of all Kindergarten girls. Since that's an unknown parameter it will have to be estimated; recall that the sample data is in a variable called **c1rmscal**.

```
ecls=read.table(file="ecls200.txt",header=TRUE)
str(ecls)
```

```
## 'data.frame':    200 obs. of  77 variables:
##  $ id       : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender   : int  1 1 1 1 2 1 2 1 1 2 ...
##  $ race     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ c1rrscal : int  25 22 22 28 29 14 23 23 38 33 ...
##  $ c1rrscala: int  25 22 22 28 29 14 23 23 38 33 ...
##  $ c1rrtsco : int  54 50 51 58 59 38 52 52 67 63 ...
##  $ c1rmscal : int  22 19 23 26 19 9 24 21 31 29 ...
##  $ c1rmtsco : int  55 52 57 60 52 31 58 54 65 62 ...
```

Let's set up the code to find SE_b. A bootstrapped standard error is simply the standard deviation of a distribution of bootstrapped statistics. Since we are interested in the bootstrapped standard error of the mean, we create a distribution of means from bootstrapped samples. Here's a run-through of the resampling operation needed to do that, explaining each step of the R code below.

First, we define three objects.

- **N** is the number of bootstrap samples we want to draw. How do we set N? The overriding concern is that our bootstrapped estimate of the standard error is reliable—meaning that we want to be confident that independent bootstrapping procedures (if they were to be done) using the same data produce very close estimates, so that in turn we have confidence in our particular estimate of SE_b. The experts recommend that N=1000 is fine for estimating standard errors. There's no cost to setting N higher (e.g., 3000, 5000), given the power of standard desktop computers, but Ns > 1000 won't return too much more accuracy.

- **n** is the bootstrap sample size, or how many cases we will draw for each bootstrap sample. The plug-in principle requires that we use the original sample n in our bootstrapping, so that the bootstrapped distribution accurately reflects the population sampling distribution (for that n), and SE_b reflects $SE_{\bar{x}}$.

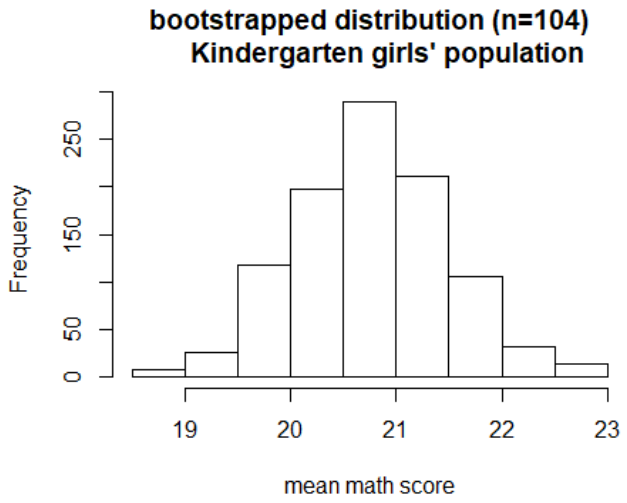
- We set up a container to hold the N means this bootstrapping operation will produce, and give it a name. In the example below that container is called **boot**, but it's just a name and could be called anything.

Second, we set up a for-loop to organize the resampling operation: i stands for a particular bootstrap sample, and the operation inside the loop (defined by these brackets: { }) is done for each of our N samples. The for-loop below has two operations.

- First we randomly sample n values (with replacement) from the math achievement variable (**c1rmscal**), specifying only the girls' data (gender = 1). Those n values are stored in an object called **bootsamp**. You can think of this as one bootstrap sample.
- Next we find the mean of the sample created in the first line, and pass that mean to the numeric container set up earlier.

The loop then repeats that two-step process for all N bootstrap samples. Once we have run the loop and created N means from bootstrapped samples, we can generate a histogram of the bootstrap distribution. Although not necessary for interval estimation, the histogram helps us visualize what the true population sampling distribution (of the mean in this case) looks like. Finally, it is essential that we save the standard deviation of the bootstrapped distribution, which is the bootstrapped standard error (SE_b), in an object for later use.

```
#bootstrap distribution of the sample mean
N=1000
n=104
boot=numeric(N)
for (i in 1:N) {
  bootsamp <- sample(ecls$c1rmscal[ecls$gender == 1],n,replace=T)
  boot[i] <- mean(bootsamp)
}
hist(boot, main="bootstrapped distribution (n=104)
     Kindergarten girls' population",xlab="mean math score")

SEb=sd(boot)
```



Now let's use R to find the t interval with bootstrap standard error confidence interval to estimate the mean math achievement score in the population (μ) of Kindergarten girls. First we find the sample mean (stored in an object called **xbar**). Having found SE_b above, we find the appropriate t-score multiplier with the `qt()` function, and use those two quantities to calculate the margin of error (MOE, stored in an object called **moe**). Then we subtract the MOE from, and add it to, the sample mean for the lower and upper limits of the CI.

```
xbar=mean(ecls$c1rmscal[ecls$gender == 1])
moe=qt(0.975,n-1)*SEb
xbar-moe

## [1] 19.3155

xbar+moe

## [1] 22.16527

#or do in one operation
xbar+c(-1,1)*moe

## [1] 19.31550 22.16527
```

The t with bootstrap standard error estimation method produces a symmetrical interval: the range of plausible values of the parameter is the same below and above the statistic. This reflects the shape of a true population sampling distribution that is assumed to be normal. Accordingly, the t with bootstrap standard error interval works best when the sample data (and, presumably, the population) are approximately normally distributed. We will learn what "works best" means later.

8.4.2 The **bootstrap percentile interval** follows from an intuitive idea. For a 95% confidence interval the most plausible values of the parameter are the middle 95% of the values of the bootstrapped distribution. Conversely, the least likely values of the parameter are the 5% of the values in the lower and upper tails of the bootstrapped distribution. By that logic, we can construct an interval consisting of percentile values from the bootstrapped distribution. For a 95% confidence interval we simply identify the 2.5th and 97.5th percentile values. Regardless of the shape of the distribution, these values define the middle 95% of the bootstrapped statistics.

All we need to generate this interval is a bootstrapped distribution of the statistic, which we already have. From that, we simply find the appropriate percentile values using `quantile()` function.

```
#95% percentile CI
quantile(boot,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 19.37500 22.10601
```

These two bootstrapped interval estimation methods produced very similar intervals. That won't always be the case, but when it is either interval is acceptable to report as your estimate. We'll get to interpreting a confidence interval after one more example.

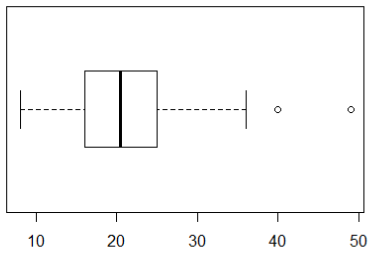
8.5 Example of bootstrapped interval estimation

Confidence intervals for estimating population means (as in the example above) are very common in research. That is in part due to the fact that the estimated standard error of the mean is easy to calculate from sample statistics ($\frac{s}{\sqrt{n}}$). However, research questions can involve many other statistics than the mean, and most of those statistics don't have a formula for calculating their standard error. The bootstrapped confidence intervals we learned above can be used to estimate parameters associated with other statistics too. Let's use bootstrapping to develop confidence intervals for estimating the population median.

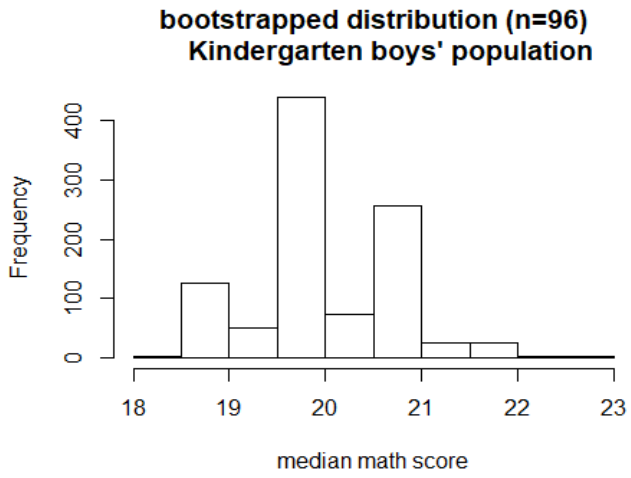
This example uses the math achievement data from the Kindergarten boys. We can see from the boxplot below that there are two outliers in the boys' scores. Because the mean is a non-robust location statistic (see Chapter 2), and vulnerable to the influence of extreme values, it might make

sense to use the median rather than the mean to describe typical or “average” math achievement in boys. If we want to estimate the median value in the population of Kindergarten boys however, we will need to bootstrap the standard error of the median.

```
boxplot(ecls$c1rmscal[ecls$gender == 1],horizontal=T)
```




```
table(ecls$gender)

##   1   2
## 104  96

#bootstrap distribution of the sample median
N=1000
n=96
boot=numeric(N)
for (i in 1:N) {
  bootsamp <- sample(ecls$c1rmscal[ecls$gender == 2],n,replace=T)
  boot[i] <- median(bootsamp)
}
hist(boot, main="bootstrapped distribution (n=96)
     Kindergarten boys' population", xlab="median math score")

SEb=sd(boot)
```



We simply adapt the code used in the previous example, and highlighted in yellow above are the elements that were changed for this estimation problem. To start we find the boys' sample size (n=96). Then we change the for-loop to select the boys' data based on their value in the gender variable, and request the median for each bootstrap sample. The histogram of the bootstrapped distribution of medians is shaped much differently than the distribution of the mean in our first example. What does that mean for the accuracy of a normal theory method like the t interval? We will consider those issues soon. Before we do, let's generate both intervals for estimating the median math score in the population of Kindergarten boys. In the example with the girls' data, the two intervals were identical (at least to one decimal place); here we see intervals that are not as similar. This illustrates the effect of a bootstrap distribution that is not normally distributed.

```
#confidence intervals
med=median(ecls$c1rmscal[ecls$gender == 2])
moe=qt(0.975,n-1)*SEb
med+c(-1,1)*moe               #t interval
```

```
## [1] 18.56857 21.43143
```

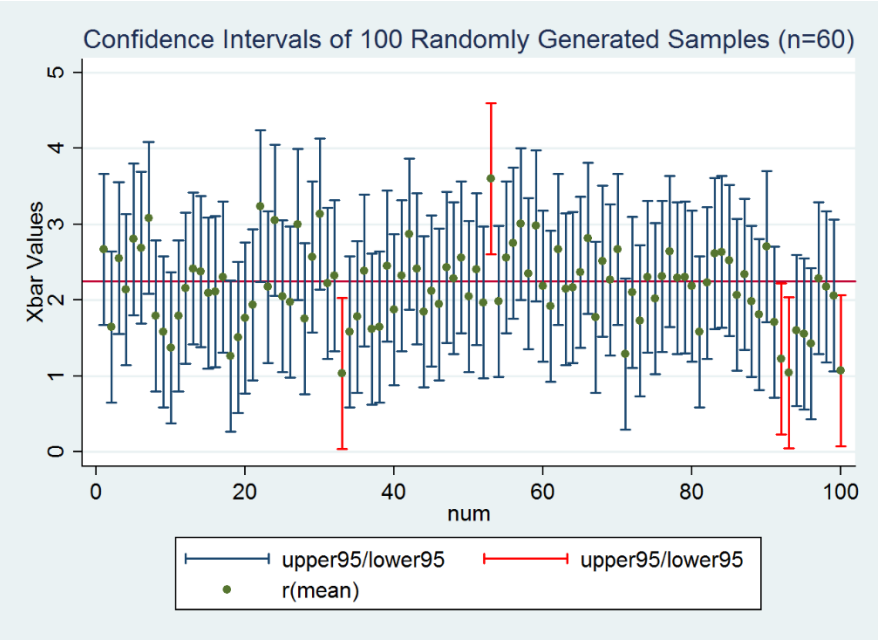
```
quantile(boot,c(0.025,0.975))  #percentile interval
```

```
##  2.5% 97.5%
##  19.0  21.5
```

8.6 How to interpret a confidence interval

A confidence interval consists of a range of plausible values of the (unknown) parameter—a range that is defined by lower and upper values—to which we attach a level of confidence. Confidence intervals are often interpreted with phrasing like "we are 95% confident that the interval 'contains' or 'includes' the parameter." However, a proper understanding of a confidence interval requires that you recognize that your particular interval is just one of many that would have arisen from hypothetical bootstrapping procedures from the same data, and yours could have been any one of those other intervals. Imagine an experiment in which random samples (of size n) are drawn from a population with a known mean, and a confidence interval for the mean generated from each sample. The results from such an experiment are displayed in Figure 8.3.

Figure 8.3 Confidence intervals of the mean from random samples from a population with μ = 2.25



Source: https://sites.nicholas.duke.edu/statsreview/671-2/

What do we learn about confidence intervals from this display? First, sample means (green dots), and in turn confidence intervals, vary randomly. Second, a few means vary so much from μ that their intervals don't even include the parameter (those shown in red). Indeed, for a 95% confidence interval we expect that over a long series of random sample-generated intervals, 95% of them will *cover* the parameter (include it within the interval range) and 5% will not. That's what is shown above; out of 100 random samples, confidence intervals from 5 do not cover the parameter. Third, most sample means cluster closely around μ (2.25) and so across many hypothetical samples there will emerge some consensus on most plausible values of μ. In other words, not all values within a particular confidence interval are equally likely values of the parameter.

With those principles in mind, let's take another shot at confidence interval interpretation. Confidence in parameter estimation is indeed a long term proposition. It really doesn't make any sense to say we have 95% confidence in a *particular* interval. When we say that we "have 95% confidence in" an interval of values containing the parameter, we mean that 95/100 intervals generated from hypothetical samples like ours will include the parameter. That interpretation, however, is not easy to render into a simple sentence, which is one reason why simpler (but somewhat inaccurate) interpretations gain currency.

Earlier we learned two bootstrap methods and used them to estimate a parameter: a t interval and a percentile interval. Having introduced the concept of confidence interval *coverage* above, let's now address the question of how we evaluate these confidence interval methods. An accurate interval is one whose coverage equals its confidence and whose "misses" are the same at both ends of the interval. For example, a 95% confidence interval should cover the parameter in exactly 95% of the samples from a simulation study (like the one illustrated in Figure 8.3) and have 2.5% error rates on each side (estimating too high and too low at the same rate). Researchers have concluded that both the t interval and percentile intervals perform well in large samples. In small samples the t interval performs better, but both estimation methods have a narrowness bias, meaning that the true coverage is less than 95%. If the bootstrap distribution is not normally distributed, the percentile interval performs better.

Revisiting the example above in which we estimated the median math score in the Kindergarten boys, the t and percentile intervals provide slightly different values. We have large N data, but given that the the bootstrap distribution of medians is not normally distributed, the percentile interval is the more accurate estimator.

8.7 Factors that affect confidence intervals

In this section we explore two factors that affect bootstrapped confidence intervals: **sample size (n)** and **confidence level**. Remember, anything that affects the margin of error (MOE) changes the precision of an interval estimate. Factors that reduce the MOE lead to more precise estimates, whereas factors that increase the MOE lead to less precise estimates.

8.7.1 To see how **sample size** affects a confidence interval let's do a simulation in which we find the bootstrapped standard error (SE_b) while systematically varying sample size. The code for the first simulation (n=100) is shown. A normal population was set up to sample from, with mean = 5 and standard deviation = 2. Then we simply applied the loop used in previous examples to generate a bootstrapped distribution of means for a particular sample size and find the standard deviation of that distribution. You can run it, as well as the other three simulations, but remember that your results will differ randomly from mine because you are doing independent sampling from the population data. The results of all four simulations are presented in Table 8.4—what do they show? First, notice the inverse relationship between n and SE_b. Larger samples generate means (or whatever statistic you're using to

estimate) that vary less, which in turn yields smaller values of SE_b and smaller MOEs. Sample size is also related to the t multiplier used in the t interval, which has an additional effect on the MOE: larger n is related to smaller t values, which yield smaller MOEs. Larger sample sizes also produce more precise percentile confidence intervals by the same logic. Means that vary less will result in bootstrapped distributions in which the 2.5th and 97.5th percentile values are closer together.

```
dat=rnorm(10000, 5, 2)   #set up a normal population with mu=5, sigma=2
N=1000
n=100
boot=numeric(N)
for (i in 1:N) {
  bootsamp <- sample(dat,n,replace=T)
  boot[i] <- mean(bootsamp)
}
#bootstrapped standard error of mean
SEb=sd(boot)
SEb

## [1] 0.1982948
```

Table 8.4 Bootstrapped standard errors of the mean and margin of error for several sample sizes from a normal population.

n	SE_b	t multiplier for 95% confidence (df = n-1)	MOE
100	0.20	1.98	0.40
50	0.28	2.01	0.56
25	0.40	2.06	0.83
10	0.63	2.26	1.43

8.7.2 Thus far in this chapter we have discussed 95% confidence intervals, but in fact the **confidence level** attached to an interval estimate is the analyst's or researcher's decision. A 95% confidence level is a common convention, but other values may be preferred depending on how the interval is used. For example, imagine that we have a treatment whose effectiveness in the population is estimated by an interval. A value of 0 would indicate that the treatment is not effective, and if 0 is a plausible value (is included in the interval estimate) then the treatment would not be offered to the public. However, if the consequences of offering the treatment are high (say the treatment has some nasty side effects)

then we would want to be very confident that the treatment is indeed effective—so that's a situation for setting confidence high, perhaps 99%. On the other hand, if the consequences of *not* offering the treatment are high (say some people could die without it) then we might not need to be so confident about the treatment's effectiveness because a treatment that might be effective is better than nothing—so that's a scenario for setting confidence lower, perhaps 85% or 90%. Those are considerations for setting confidence level when intervals are used to make decisions with practical consequences. When intervals are used simply to estimate a parameter, 95% is an acceptable all-purpose confidence level.

You can easily see how the confidence level affects interval precision. In a t interval with bootstrap standard error, the confidence level determines the t multiplier. Higher confidence levels produce larger t values, which in turn translate to less precise intervals. For example, referring to the table above (and you can explore this in R with the `qt()` function), if we wanted to generate a t interval with 99% confidence, the t multiplier (for n=100/df=99) would be 2.63. With a bootstrap percentile interval, the same result occurs: higher confidence levels produce quantiles that are more extreme, leading to less precise intervals.

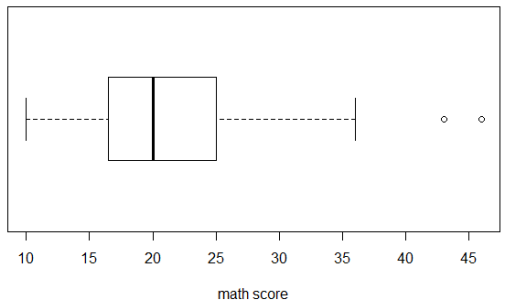
8.8 Writing up a parameter estimation study

We will do much more with interval estimation in data analysis in Chapter 9, producing bootstrapped confidence intervals to estimate effects or group differences in studies with various predictor and outcome variable types, and thus estimating with various statistics. To end this chapter, here is a written summary of the study done to estimate math achievement in the population of Kindergarten boys.

Results

Math achievement in the population of Kindergarten boys was estimated from the math achievement test scores of a random sample of boys (n=96). Exploration of the sample data revealed slight positive skew and the presence of two outliers (see Figure 1). As a result, the median (med=20) was used to estimate describe average math achievement. Two estimation methods were used to estimate the population median with 95% confidence. The t interval with bootstrap standard error (95% CI: 18.6, 21.4) and the bootstrap percentile interval (95% CI: 19.0, 21.5) produced similar estimates. We have a high level of confidence that the true median math achievement in the population of Kindergarten boys is between 19 and 21.5.

Figure 1. Boxplot of math achievement scores



8.9 Problems

Use the `Melanoma` data (205 patients with malignant melanoma) in the `MASS` package for problems 8.9.1-8.9.3.

8.9.1. Generate a bootstrapped percentile confidence interval (with 95% confidence) for estimating the mean survival time in the population of patients. Interpret.

8.9.2. Generate separate bootstrapped percentile confidence intervals (with 95% confidence) for estimating the mean survival time in the populations of female (sex=0) and male (sex=1) patients respectively. Interpret each. What do these confidence intervals say about the survival times of women and men?

8.9.3. Generate separate t with bootstrapped standard error confidence intervals (with 95% confidence) for estimating the mean survival time in the respective populations of patients with (ulcer=1) and without (ulcer=0) an ulcer. Interpret each. What do these confidence intervals say about the relationship between having an ulcer and melanoma survival time?

8.9.4. In the `cats` (MASS) data, estimate the mean body weight in the population of cats with 95% confidence intervals, using both t and percentile methods. Interpret the intervals. Explain the difference between the methods.

8.9.5. Generate separate percentile confidence intervals (with 95% confidence) for estimating the body weight in the respective populations of male (Sex=M) and female (Sex=F) cats. Interpret each. What do these confidence intervals say about the relationship between sex and cat body weight?

8.9.6. Using the `birthwt` (MASS) data estimate the median birthweight in the population from which the sample of mothers came using a t with bootstrapped standard error confidence intervals (with 95% confidence). Then, generate separate t confidence intervals (with 95% confidence) for estimating the median birthweight in the respective populations of mothers who smoked (smoke=1) and didn't smoke (smoke=0) during pregnancy. Interpret each. What do these confidence intervals say about the relationship between mothers' smoking status during pregnancy and the infant birthweight?

❾ Using Resampling Methods for Statistical Inference

Covered in this chapter:

- The general linear model as a guide for statistical modeling
- The statistical model of a study and its implications for data analysis
- Examples of randomization and bootstrapping in various statistical models

9.1 Introduction

This chapter pulls together statistics for bivariate description and summary (Chapters 3-6) with the resampling methods used for inference around statistical conclusion validity (Chapter 7) and effect size validity (Chapter 8) and applies them in various data analytic examples. This chapter has three objectives:

1. Learn how to diagnose the statistical model of a study, and translate that model to data analysis.
2. Combine descriptive and inferential methods in a data analytic context and show, through examples, how various descriptive statistics like the mean difference or the risk ratio are used in hypothesis testing and parameter estimation with resampling tools.
3. Present a process for the systematic analysis that follows the inferential framework presented in Chapter 7, and serves as a guide for your data analysis on assignments or projects.

As in previous chapters, this chapter will also address the written communication of results from studies with a continuous outcome variable in a way that reflects the systematic data analytic process above.

9.2 Statistical inference framework

In this chapter we can now apply the inferential concepts introduced in Chapter 6 as a framework for using statistics and statistical methods in data analysis and in reporting the results of data analysis. That framework (see Table 9.1) is presented below.

Table 9.1. Inferential framework for data analysis

Dimension	Inferential task	Evidence
1 External validity	Do findings based on sample data generalize to population of interest?	• random sampling • large N
2 Causal conclusion validity	Is the observed relationship between predictor and outcome caused by the predictor variable, and only the predictor variable?	• experimental design • control of all alternative explanations
3 Statistical conclusion validity	Is the observed relationship between predictor and outcome a non-random observation?	• low probability (p value) under H_o of a relationship as large or larger than the observed relationship
4 Effect size validity	What is the direction and size (and range of likely sizes) of the relationship between predictor and outcome in the population?	• precision in parameter estimation • effect size statistic appropriate to hypothesized effect

Recall from Chapter 7 that there are four types of knowledge about population parameters and relationships between variables that we establish by inference—that is, using a probabilistic guess or estimate that is based on sample data. Here's a general rule that underpins each of the inferential tasks above: *knowledge about some quality of the population is only as good as the subset of the population that you have at your disposal*. Poor sampling erodes confidence in all subsequent inferences, which is why **external validity**—the inferential task that depends on sampling methods—comes first in the framework above. Without adequate evidence of external validity, we can't very well apply insights from subsequent steps in the inferential process to any particular population. Because statistical analysts (like you!) often receive the data after the study that produced the data is done, we have little control over external validity. We should nevertheless incorporate the evidence for external validity into our subsequent analysis and inferences. In a similar way, analysts have little control over **causal conclusion validity** because it is inherent to the *design* of the study that produced the data, and generally that study has been completed once we get the data. Evidence for causal conclusion validity directly informs exchangeability assumptions, which in turn affects the way we interpret our hypothesis tests and parameter estimates, so it is important to find out about the design of the study that produced the data we are analyzing.

Statistical conclusion validity and **effect size validity** are inferential tasks that require data and, in turn, statistical methods; those methods have been the focus of this book. Even though the analyst has more responsibility for these inferential tasks, there are still study design issues that can affect these inferences. Here are two such issues:

1. **The outcome variable measure can affect the variability of the outcome data**. Imagine a researcher using a rubber ruler to measure the stride length of a sample of participants. Each measurement would include some random bit of ruler stretch, which would put random variability into the observations, which in turn affects hypothesis testing and parameter estimation. Many outcome measurement instruments used in the social sciences (of the self-report, paper and pencil variety) are like rubber rulers. Measurement methods can surely introduce bias (or systematic variability) into outcome measurements, and as a result affect statistical conclusion validity. But even absent bias, random sources of variability (i.e., measurement error) can disturb our statistical methods. So the data you get to analyze can include various amounts of measurement error, which is hard to compensate for after the study is done.
2. **The study may not be able to detect the relationship or effect it is trying to detect**. A study can have sampling and study design whose quality is beyond reproach, but if there is no relationship or effect in the population, the study probably won't (and, in fact, *shouldn't*) detect one. A study—like a microscope lens—has a particular power to be able to "see" or detect relationships. If you put a small cellular organism on a slide and try to view it with a magnifying glass, you won't see anything. If you put the slide under the powerful lens of a microscope, you'll see the organism. It's not the organism's fault that you didn't see it with the magnifying glass; seeing the "something" that's there requires a lens of appropriate power. The same quantitative logic applies to studies: if a relationship of some size between predictor and outcome exists in the population (and, presumably, in the sample too if it's a good sample), a study will only be able to detect that relationship if that study has enough power. Study power affects statistical conclusion validity, and we learn more about statistical power in Chapter 9.

Establishing statistical conclusion validity and effect size validity requires the correct use of hypothesis testing and parameter estimation methods that we have learned so far in this book, and the careful interpretation of the results of those methods.

9.3 Statistical model

Before we get to the data analytic examples in this chapter, we need to introduce the concept of the statistical model. We are at a point in this book where we have a workbench of statistics for summarizing variables of all measurement types, knowledge of those statistics' properties and applications, as well as a set of methods for null hypothesis testing (randomization, permutation, Monte Carlo tests) and parameter estimation (bootstrapped confidence intervals). Statisticians and analysts are commonly presented with data and a question, and expected to "answer" that question with the data. But how do we know what statistics and statistical methods—from all those available on our workbench—to use in data analysis? The answer is in the **statistical model**, which is a statement of the variables on the x and y "sides" of an equation as well as the measurement scales of those variables. A statistical model takes a particular research question and its constituent variables, then, and re-expresses it in a form called the **general linear model**. The general linear model takes the form

$$Y = \beta_0 + \beta_1 X$$

where Y is the outcome variable, β_0 is the intercept or the value of Y when X=0, X is the predictor variable, and β_1 is the estimated relationship between X and Y.* The intercept is not consequential in our identifying the statistical model inherent in a research question and its variables inasmuch as β_0 is a constant and is only useful when we want to use the linear equation for prediction.

*If this looks familiar to the linear regression model from Chapter 5, it should—because a regression model is one specific expression of the general linear model.

Let's take our example from Chapter 3 and re-express it in general linear model form. The research question was "Do Kindergarten girls and boys differ on a test of math achievement?"

$$mathach = \beta_0 + \beta_1 gender$$

The model above identifies gender as the predictor and math achievement as the outcome. To specify the particular statistical model we also need to know the type of data on each side of the equation; in other words, how are math score and gender measured? Math achievement is a continuous numeric measure and gender is a categorical (with 2 categories) variable. *The combination of x and y variables and their measurement scale yield a statistical model.* In Chapters 3-6 we learned the statistics used in four specific, and very common, forms of the general linear model. Those model forms are organized in Table 9.2, with some alternate names sometimes used for them. The form of the statistical model (i.e.,

what variable and data type is on each side of the model equation) is much more important than its name, because the model form guides the statistics and statistical methods used to estimate β_1 in each.

Table 9.2 Common forms of the general linear model

Model form	**Traditional alternate names**	**Chapter**
$numeric\ Y = \beta_0 + \beta_1 categorical\ X$	ANOVA* model	3
$categorical\ Y = \beta_0 + \beta_1 categorical\ X$	Proportions model	4
$numeric\ Y = \beta_0 + \beta_1 numeric\ X$	Regression model	5
$categorical\ Y = \beta_0 + \beta_1 numeric\ X$	Logistic model	6

*ANOVA is short for analysis of variance, which refers to a set of statistical methods have traditionally been used in models of that type.

In our example above, β_1 would be the sample-based estimate of the relationship between gender and math achievement score. Recall from Chapter 3 that there are lots of possible statistics with which to explore this relationship, including statistics that capture location, variance, and symmetry. Each statistic is a sort of lens through which to view that x-y relationship, and is a data analytic decision that is informed by both the study purpose and what emerges in the data itself. If we used the mean difference to examine the relationship between gender and math achievement, then β_1 would be the estimated size of the mean difference between girls' and boys' math scores in the population of Kindergarten students.

Knowledge of the statistical model also helps in diagnosing whether research questions can or can't be answered in the data. For example, what if the research question was "Do Kindergarten girls and boys differ in their rates of passing a math achievement test? That is a very similar sounding question, but in fact calls for a different statistical model—one with a categorical (e.g., pass/not pass) outcome variable. If the data do not contain that variable, or such a variable cannot be computed from existing variables in the dataset, then that question cannot be addressed.

The general linear model is a powerful tool for organizing statistical models by their constituent variables and data types, and for providing guardrails for our data analytic work. It can be extended to accommodate much more complex models than those we are learning in this course, but these four basic models are the foundation for all those other models.

9.4 Resampling for Inference in an ANOVA model

With those preliminaries covered, let's work through a data analytic example with data from a study with the first statistical model in Table 9.2: **numeric outcome and categorical predictor**. In this example, as in those that follow, the goal is to integrate the descriptive and inferential statistical tools we have acquired in previous chapters into a more complete data analytic exercise, pulling together the available evidence (from the sampling method, study design, and data analysis) for validity dimensions 1-4 in Table 8.1.

For this exercise imagine we have been given the ECLS Kindergarten data and asked to find out if "disabled and non-disabled students differ in their average reading ability," using the variable (**c1rrscal**) highlighted below as the outcome variable and the predictor (**p1disabl**, and coded 1=disabled and 2=nondisabled).

Before we get into the data analysis, let's assess the evidence for external validity and causal conclusion validity. Given that we were not involved in the study design and data collection and don't have first-hand knowledge of the sampling method, we try to get as authoritative information about the study's sampling and method as we can. In this case the data came from a large, nationally representative probability sample (see https://nces.ed.gov/ecls/kindergarten.asp). Although the example below uses a small random subset of the original study data because it is easier to visualize and learn from, in principle we have strong evidence for the generalizability of the findings. As regards causal conclusion validity, we must likewise learn about the study design from as authoritative sources (e.g., study website, research article) as possible. In this case we know the data came from a survey study; the study did not experimentally arrange participants into disability status groups (indeed, such a thing would be highly unethical!). In other words, the study design does not control any alternative explanations for the effect of disability status on reading achievement. As a result, we have very little evidence for attributing any observed difference in reading ability to the *causal* effect of disability status. We now turn to the data analysis and the assessment of statistical conclusion validity (Is the reading ability difference between disabled and non-disabled Kindergartner students nonrandom?) and effect size validity (What is the size of the reading ability difference disabled and non-disabled Kindergartner students?).

We begin with descriptive analysis of the predictor variable in the form of frequency and proportions tables, and the outcome variable in the form of boxplots by group. What do we learn from these? About 82% of the students are non-disabled. The boxplots show that both group's scores have positive skew, with the non-disabled group's scores being more skewed. Reading scores are more variable among non-disabled than disabled students. Additionally, there are four outliers (at least as diagnosed by the boxplot rule) in the non-disabled group data, three of which are very high scores. Remember from Chapter 2 that outliers need to be checked for accuracy and for evidence that those observations are part of the population interest. Let's assume for this exercise that those three students' scores are correct, and they are ordinary Kindergarten students. In light of the greater positive skew and the presence of several very high scores in the non-disabled group, we should prepare to do our null hypothesis testing with a robust location statistic in addition to the mean.

```
ecls=read.table(file="ecls200.txt",header=TRUE)
str(ecls)
```

```
## 'data.frame':    200 obs. of  77 variables:
##  $ id       : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender   : int  1 1 1 1 2 1 2 1 1 2 ...
##  $ race     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ c1rrscal : int  25 22 22 28 29 14 23 23 38 33 ...
##  $ c1rrscala: int  25 22 22 28 29 14 23 23 38 33 ...
```

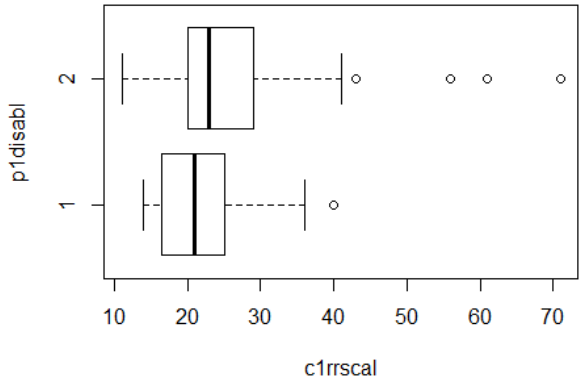
```
table(ecls$p1disabl)
```

```
##   1   2
##  36 164
```

```
prop.table(table(ecls$p1disabl))
```

```
##    1    2
## 0.18 0.82
```

```
boxplot(c1rrscal~p1disabl,data=ecls, horizontal = T)
```



Hypothesis testing. Turning to the hypothesis testing, let's remind ourselves what the hypotheses are for our randomization study. Although we don't need to include these statements in our write-up of the analysis, it is a good reminder that the lack of control in the original study over alternative explanations (to a causal effect of disability status on reading) leaves us with numerous plausible alternative hypotheses.

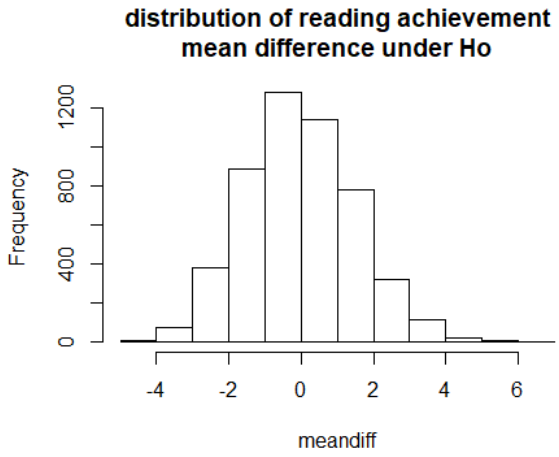
H_o: There is no relationship between disability status and reading achievement.
H_a: There is a relationship between disability status and reading achievement.

Let's begin by conducting a Monte Carlo permutation (see assumption of exchangeability below) test of the mean difference. Below we compute the statistic we are using to test H_o, arranging the statistic to be positive strictly for interpretive convenience. The mean reading score in the non-disabled students is 2.85 points greater than the mean reading score in the disabled students. Then, we set up the loop to conduct the randomization test: we sample without replacement, effectively creating a random ordering of the 200 values in each iteration. Then we simply divide those observations into two groups of the size of the original samples of disabled and non-disabled students. Remember that although we are randomizing observations to group, the assumption of exchangeability is not met in the original study, meaning that we can't attribute observed differences in reading ability to disability status alone. Following that we produce histogram of the distribution of mean differences under H_o, and calculate the p value associated with any (positive or negative) difference greater than the observed mean difference.

```
stat=(mean(ecls$c1rrscal[ecls$p1disabl == 2]))-(mean(ecls$c1rrscal[ecls$p1disabl ==
1]))
stat
```

```
## [1] 2.855014
```

```
N=5000
meandiff=numeric(N)
for (i in 1:N) {
  data <- sample(ecls$c1rrscal,200,replace=FALSE)
  grp1 <- data[1:36]
  grp2 <- data[37:200]
  meandiff[i] <- mean(grp1)-mean(grp2)
}
hist(meandiff, main="distribution of reading achievement
mean difference under Ho")
```




```
pvalue=length(which(abs(meandiff)>=stat))/N
pvalue
```

```
## [1] 0.0644
```

With p = .064, there is very weak evidence of disabled and non-disabled Kindergarten students differing on a reading achievement test. What evidence there is suggests that non-disabled students have higher reading scores than disabled students, but remember that the mean difference is overly influenced by the several high scores in the non-disabled student group.

Parameter estimation. Although there is very little evidence of statistical conclusion validity in the data—meaning we don't have enough evidence to conclude that the small reading difference observed between the groups is significant (i.e., non-random)—we nevertheless complete the analysis by investigating the *size* of the group difference in reading achievement with bootstrapped confidence

interval estimates of the mean difference and median difference in reading achievement in the population of Kindergarten disabled and non-disabled students.

The bootstrapping and confidence interval operations below make a few changes from the basic method learned in Chapter 7 to accommodate the estimation problem in this example: let's work through those changes.

- We sample with replacement *within each group* and store those bootstrapped samples of 36 disabled and 164 nondisabled students, respectively, in different objects.
- We then find the difference between the means (or medians) of those bootstrapped samples, creating one bootstrapped statistic per iteration. We maintain the order of subtraction (non-disabled minus disabled) simply to be consistent with our previous work.
- For the t interval statistic we use the line of code from our earlier permutation tests for computing the mean (or median) difference but give the objects different names.

The remainder of the operations apply the methods we learned, and as learned, in Chapter 6. The t interval with bootstrapped standard error and bootstrapped percentile intervals (both with 95% confidence) are highlighted in yellow below for each statistic.

```
#mean difference
N=2000
boot=numeric(N)
for (i in 1:N) {
  grp1 <- sample(ecls$c1rrscal[ecls$p1disabl == 1],36,replace=T)
  grp2 <- sample(ecls$c1rrscal[ecls$p1disabl == 2],164,replace=T)
  boot[i] <- mean(grp2)-mean(grp1)
}

SEb=sd(boot)

#t interval
mdiff=(mean(ecls$c1rrscal[ecls$p1disabl == 2]))-(mean(ecls$c1rrscal[ecls$p1disabl =
= 1]))
n=200
moe=qt(0.975,n-1)*SEb
mdiff+c(-1,1)*moe

## [1] 0.4079506 5.3020765

#percentile interval
quantile(boot,c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.4860942 5.2696985
```

Although the intervals produce similar estimates, we will interpret them separately. According to the t interval the most plausible values of the "true" (meaning, in the population of all Kindergarten students) mean difference between non-disabled and disabled students on a reading achievement test range between .41 points and 5.3 points. All of those plausible values have non-disabled students scoring higher than disabled students. The percentile interval establishes the plausible (with 95% confidence, that is) values of the parameter between .49 points and 5.3 points.

We can see that although there is some evidence against H_o in the data, leading to a tentative conclusion of statistical significance, the likely difference between these two populations of students is small, and therefore not "significant" in the sense of being important or large. This illustrates the independence of inference for statistical conclusion validity and for estimating the true effect size. We should take care not to conflate these inferential tasks, because they speak to much different questions.

9.5 Resampling for Inference in a Proportions Model

The second statistical model in Table 9.2 is one with a **categorical outcome and a categorical predictor**. In this book we address models with 2-level categorical variables only. That may seem like a limitation, but it is quite a useful model. Many variables—both outcome and predictor—that may be measured with multiple categories are nevertheless reducible because: a) their essence can captured with two categories, or b) there are too few responses in some categories to analyze. For example, father's education was measured in the original ECLS study as a 9-category variable, with categories ranging from "8th grade or below" to "doctorate or professional degree." However, the large majority of respondents were in only three of those categories (high school diploma, some college, or college degree). So creating a new 2-level variable ("high school diploma", "anything beyond high school") captures the essence of the responses. This issue arises with outcome variables too, such as might have occurred if school type had been measured with a "public" vs "private-religious or parochial" vs "private-nonreligious" categories. If for example only a handful of study respondents attended private nonreligious Kindergartens, that 3-category variable could easily be reduced to a 2-category variable while preserving it in concept. The ability to recode a variable, therefore, is a skill that is commonly

needed in data analysis; we will demonstrate that in the example below. Finally, let's recognize that even a 9-level categorical variable is not a quantitative predictor; true quantitative predictors require a different statistical model be analyzed (see Table 9.2).

Let's revisit the problem from Chapter 4 where we described the relationship between father's education level (x) and the type of school the child attended (y). As explained above, the father's education variable is recoded to create a new low (high school diploma or less) vs. high (some college or higher) variable. The code below shows that we do two other things in that recoding process: apply labels to the categories and make the new variable a factor. These changes help make the contingency table easier to read. Before we get to the statistical inference tasks, let's address the question of causal conclusion validity. The "treatment" variable here is having a father with high, compared to low, education. But children were not randomly assigned to this treatment, and thus any difference we might observe in their rates of attending a private school cannot be attributed to the father's education alone. Many alternative explanations can be offered (e.g., high education dads have higher income) for the outcome which were not controlled in the ECLS study. So we have very little evidence of causal conclusion validity.

Hypothesis testing. In this example we are using the risk ratio (RR) to analyze the relationship between father's education and school type. Other statistics that are common to this statistical model (i.e., risk difference, odds ratio) provide different lenses for summarizing this relationship, and those could also be used in the inferential procedures below. First we produce the contingency table (remembering to keep x in columns and y in rows) to generate the conditional proportions of students going to private school, which are then used to find the RR. The direction of the relationship is such that children whose fathers have high, compared to low, education levels are about 2.5 time more likely to attend a private school.

```
library(car)

ecls$daded=as.factor(recode(ecls$wkdaded, "1:3='low'; 6:9='high'; else=NA"))
ecls$schtype=as.factor(recode(ecls$s2kpupri,"2='0';'else'"))
ecls$schtype=factor(ecls$schtype,level=c(0,1),labels=c("private","public"))
tab=table(ecls$schtype,ecls$daded)
addmargins(tab)

##
##           high low Sum
##   private   17  10  27
```

```
##   public    38  72 110
##   Sum       55  82 137
```

```
prop.table(tab,margin=2)
```

```
##
##                high       low
##   private 0.3090909 0.1219512
##   public  0.6909091 0.8780488
```

```
ptab=prop.table(tab,margin=2)
RR=ptab[1,1]/ptab[1,2]
RR   #risk ratio
```

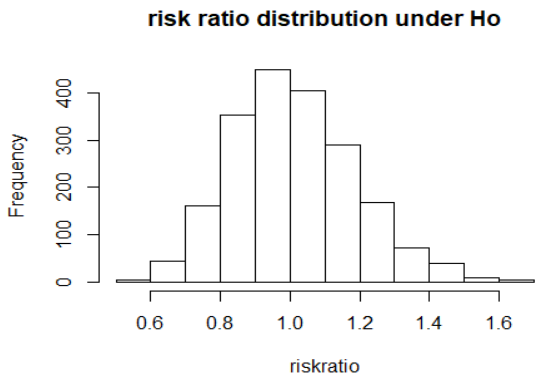
```
## [1] 2.534545
```

Next we set up the permutation test for assessing evidence against the null hypothesis; let's work through the code below.

- First we define three objects: our observed risk ratio (**obs**), the number of permutations to be run (**N**), and a container (**riskratio**) to hold the N risk ratios generated with the loop.
- Next we sample 137 cases without replacement from a vector that contains only 0 and 1. Note that we are not resampling from our data. But wait—our data is just a vector of 137 0s and 1s (where 0=public school and 1=private school) so we're accomplishing the same thing. Remember that under H_o there is no relationship between x and y, so the cell frequencies and conditional proportions are not important. In effect this step simply creates one random ordering of 137 0s and 1s, and we store that random variable in an object (**data**).
- Next we divide this random vector of 0s and 1s into two groups of the sizes of our original high (n=55) and low education (n=82) groups. This creates two groups of 0s and 1s which under the null hypothesis should only be randomly different with regard to the proportion of 1s (or private schools) in each.
- Finally, using the logic explained in Chapter 4 that a mean of a vector of 0s and 1s is the proportion of the 1 category, we calculate the RR and store it in the container.

This loop then gets repeated N times to create a distribution of RRs (for groups of size 55 and 82, respectively) under the null hypothesis. That reference distribution is plotted below. Naturally, the mean of that distribution is very close to 1.0, which is what we would expect under H_o.

```
obs=RR
N=2000
riskratio=numeric(N)
for (i in 1:N) {
  data <- sample(0:1,137, replace=T)
  grp1 <- data[1:55]
  grp2 <- data[56:137]
  riskratio[i] <- mean(grp1)/mean(grp2)
}

hist(riskratio,main="risk ratio distribution under Ho")
```




```
pvalue=length(which(riskratio>=obs))/N
pvalue

## [1] 0
```

You can see from the histogram that an RR of 2.53 is a very unlikely outcome under H_o. In fact, out of 2000 permuted samples in this particular simulation, a risk ratio of 2.53 or greater occurred exactly 0 times, and this is why we get a p value* of zero. Although a real p value can never be zero, we should interpret this estimate as a very small p value (e.g., p<.001), which constitutes strong evidence against H_o.

*This is technically a 1-tailed or 1-sided p value. In previous examples we have computed 2-sided p values because of the ease of the computing involved and to establish the fundamental logic of a p value. For example, taking the absolute value a mean difference in our p value computation allows us to find the probability of a difference that large or larger *in either direction*. Remember, under H_o an outcome of some magnitude is equally likely to be positive or negative. In this example however our p value is associated only with RRs ≥ 2.53. The equivalent "negative" value is RR=.40, which indicates that

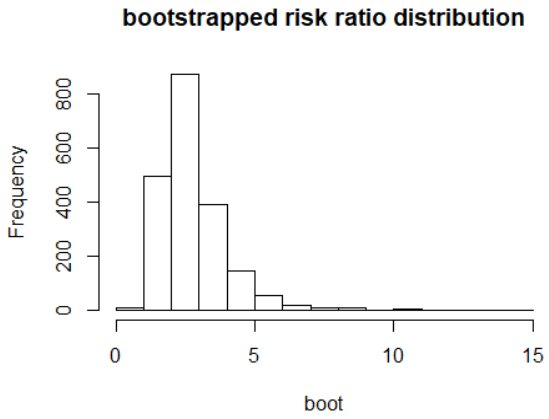
children whose fathers have low, compared to high, education levels are 60% less likely to attend a private school. Under H_o these two RRs are equally probable outcomes. The 2-sided p value in this example would be approximately two times the 1-sided p value.

Parameter estimation. There is strong evidence of statistical conclusion validity in the data. Our sample data suggests (RR=2.53) that the "effect" of having a father with high, compared to low, education on attending a private school is positive and substantial. Nevertheless we need to estimate a range of plausible values of the parameter—the true relationship between father's education and private school attendance—with a confidence interval. Let's see how we do this in R.

- First we create data vectors of our x variable to resample from. Here's the logic: we want to preserve the true relationship that exists between father's education and private school attendance, so now cell frequencies and conditional proportions *do* matter. But we also must acknowledge that our particular sample is but one of many possible estimates of that relationship. So by resampling from our column data, which is anchored by a particular conditional proportion, we are generating a large number of samples of children with high (or low) education fathers, any one of which we could have observed in our study. The two objects **high** and **low** simply reproduce the column data from the frequency table above by replicating 0s and 1s.

- We select a bootstrap sample from each object, careful to use sample sizes that reflect the original study data. Then we assign those samples to objects (**grp1**, **grp2**) and calculate the RR with the data in those samples; this creates one bootstrapped risk ratio.

```
high=c(rep(1,17),rep(0,38))
low=c(rep(1,10),rep(0,72))

N=2000
boot=numeric(N)
for (i in 1:N) {
  grp1 <- sample(high,55,replace=T)
  grp2 <- sample(low,82,replace=T)
  boot[i] <- mean(grp1)/mean(grp2)
}

hist(boot,main="bootstrapped risk ratio distribution")
```




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

```
##     2.5%    97.5%
## 1.277922 5.716039
```

The histogram of the bootstrapped risk ratio distribution is naturally centered around our sample-based point estimate (RR=2.53). However, it is very positively skewed. This is because RRs can't be lower than 0, but can be as large as the random variation in the statistic allows them to be, and a small number of outcomes are (by chance) very large. Because this bootstrapped distribution is not normal, the percentile confidence interval has much better estimating properties than the t interval. The percentile 95% confidence interval indicates that the size of the true relationship can plausibly range from children being 1.3 times and 5.7 times more likely to attend private school if their father has high, compared to low, education.

9.6 Resampling for Inference in a Regression Model

The third statistical model in Table 9.2 is one with a **numeric outcome and a numeric predictor**. Regression models are used a lot in non-experimental research, where the research goal is to predict an outcome rather than to determine what causes the outcome. Data from experimental studies *can* be analyzed with a regression model however, so it is still important to examine the study method for evidence of causal conclusion validity. In Chapter 5 we described the relationship between reading scores in the fall Kindergarten and the spring 1st grade semesters with a least-squares regression coefficient. Below we take up the inferential elements of the analysis of the relationship between Kindergarten and 1st grade reading ability. Before we do, let's remind ourselves that we have strong evidence for external validity in this study, due to the probability sampling in the ECLS study, and weak

evidence for causal conclusion validity, due to the study's lack of control over alternative explanations for 1st grade reading ability other than Kindergarten reading ability.

Hypothesis testing. By way of review, below we estimate the least squares regression model for the variables in our problem, and retrieve the regression coefficient from the `lm()` model object. Note that there are two regression coefficients in a regression model—the intercept and the slope (β_0 and β_1, respectively). The code below retrieves the 2nd of those coefficients (β_1), which is a key function needed for our resampling work.

```
reg=lm(c4rrscal~c1rrscal,data=ecls)
reg$coefficients[2] #retrieve regression coeff

##  c1rrscal
## 0.9521881
```

Below is the code for doing our permutation study to test the null hypothesis testing, followed by the bootstrapping procedure for estimating the parameter. These operations share a logic, so let's go over that now. In the proportions model from earlier, remember that cell frequencies and conditional proportions contained the information about the bivariate relationship, and column frequencies and proportions did not. Similarly, in a regression model we have paired observations (each participant provides a score on x and y) and the information about the relationship is contained in the paired observations. Logically, under H_o we disregard the pairing and essentially randomly assign Kindergarten and 1st grade reading scores to participants. In parameter estimation the pairing is essential to preserve, which is why we bootstrap rows from our data (a row being a pair of x,y observations in which the x-y relationship is preserved).

Let's look at the familiar steps in the permutation test of H_o.

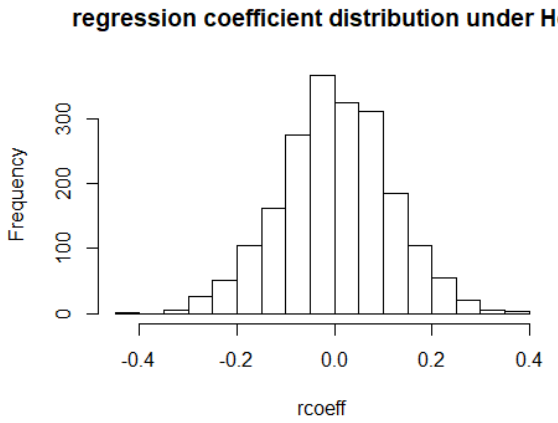
- We set up objects for our observed statistic, number of permutations, and a container to hold the resampled statistics.

- Next we draw random samples of fall Kindergarten and spring 1st grade reading scores (purposefully ignoring the pairings of observations in the original data) and create data vectors with those scores. Sampling without replacement creates random orderings of the 200 scores in each variable.

- We estimate the least squares regression model in that data, and scrape out the regression coefficient.

That process is repeated N times, resulting in a distribution of regression coefficients under the null hypothesis which, naturally, is centered on beta=.00. We calculate the p value (and here it's a 2-sided p value) with our customary method. As in the previous example, none of the resampled statistics in this particular permutation study was greater than .95, hence the p value of 0. Although we know the true p value is not 0, it is nevertheless a very small p value, which provides strong evidence against H_o.

```
beta=reg$coefficients[2]
N=2000
rcoeff=numeric(N)
for (i in 1:N) {
  y=sample(ecls$c4rrscal,200,replace=F)
  x=sample(ecls$c1rrscal,200,replace=F)
  mod<-lm(y~x)
  rcoeff[i] <- mod$coefficients[2]
}

hist(rcoeff,main="regression coefficient distribution under Ho")
```

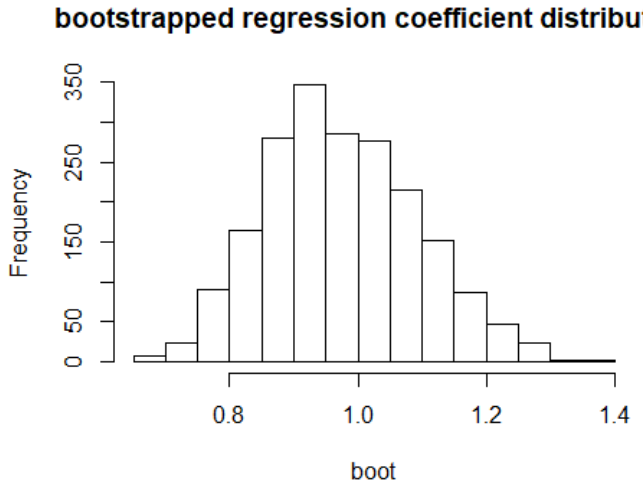



```
pvalue=length(which(abs(rcoeff)>=beta))/N
pvalue

## [1] 0
```

Parameter estimation. With strong evidence of statistical conclusion validity in the data, we now turn to effect size validity—which is the task of estimating the direction and size of the relationship between Kindergarten and 1st grade reading in the population of Kindergarten students. Below is the

bootstrapping procedure, set up to preserve the inherent relationship between x and y by boostrapping *pairs* of x,y observations. Each bootstrap sample is one we could have had observed in our original sample data, and delivers a bootstrapped regression coefficient. This histogram of that distribution is centered on our sample statistic (β_1=.95). Since the bootstrap distribution is reasonably normal, the t and percentile confidence interval are equally good estimators. The t with bootstrapped standard error interval indicates that the size of the true relationship between Kindergarten and 1st grade reading achievement can plausibly (with 95% confidence) range from .72 and 1.18, meaning 1st grade reading scores increase on average between .72 and 1.18 points with each point increase in Kindergarten scores. The bootstrap percentile interval delivers a similar estimate.

```
N=2000
boot=numeric(N)
for (i in 1:N) {
  dat <- ecls[sample(nrow(ecls),200,replace=T),]
  mod <- lm(c4rrscal~c1rrscal,data=dat)
  boot[i] <- mod$coefficients[2]
}

hist(boot,main="bootstrapped regression coefficient distribution")
```




```
#t interval
SEb=sd(boot)
n=200
moe=qt(0.975,n-1)*SEb
beta+c(-1,1)*moe

## [1] 0.7202816 1.1840946
```

```
#percentile interval
quantile(boot,c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.7653076 1.2168231
```

9.7 Resampling for Inference in a Logistic Model

The final statistical model to cover is the logistic model, which is any model with a **categorical outcome and a numeric predictor**. As with the proportions model covered earlier, we are limiting our focus in this chapter to logistic models with 2-category outcome variables. In the Chapter 6 example we described the relationship between family socioeconomic status (x) and school type (y); let's complete that example by doing the statistical inference tasks with resampling methods. By now, you can assess the evidence for external validity and causal conclusion validity on your own.

To set this up, the school type variable was recoded earlier in this chapter (0=private, 1=public), and its frequency table is below. The predictor variable (**wksesl**) is a continuous, integer measure of socioeconomic status (SES) that ranges in this sample from 32 to 66, with a higher number indicating higher SES. As we did in Chapter 5, the logistic model is estimated below and the regression coefficient retrieved from that model output. The regression coefficient is a log-odds so we convert it to a more user-friendly statistic, the odds ratio (OR). The OR indicates that the odds of going to a public school decrease by about 9% for each unit increase on the SES scale.

```
table(ecls$schtype)
```

```
## private  public
##      45     155
```

```
mod=glm(schtype~wksesl,data=ecls,family=binomial)
exp(mod$coefficients[2])
```

```
##    wksesl
## 0.9120166
```

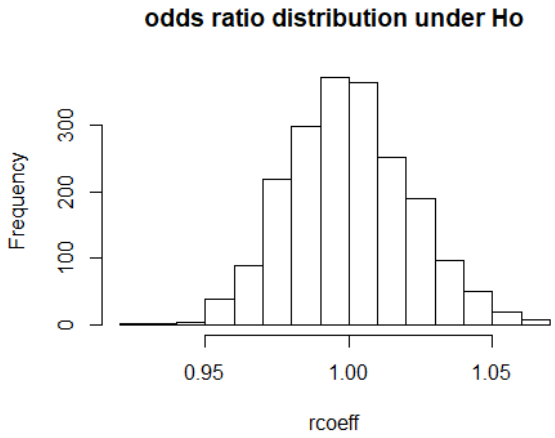
Hypothesis testing. Setting up the randomization test of the null hypothesis and the bootstrapping procedure for parameter estimation in R reflect the same logic as explained above for the regression model because a logistic model also has paired data points (i.e., each student contributes a family SES

score and a school type). The randomization procedure disregards those paired observations, whereas the bootstrapping procedure preserves them. Accordingly, the steps in the permutation test of H_o are very similar to those in the regression model

- We set up objects for our observed statistic, number of permutations, and a container to hold the resampled statistics.
- We then create random vectors of x and y data. The school type data is simply a vector of 200 0s and 1s, inasmuch as school type under H_o is random. Note that we need to sample with replacement from that vector because there are only two elements in it. The random SES data vector is created per our previous randomization tests.
- We estimate the logistic regression model in the randomized data, and retrieve the regression coefficient, converted to an odds ratio in the process.

That process is repeated N times, resulting in a distribution of odds ratios under the null hypothesis which is centered on OR=1.00. We calculate the p value with our customary method, although for convenience we are calculating a 1-sided p value. The particular inequality in the `rcoeff≤OR` statement must reflect which side of 1.0 the observed OR is on. Since our OR=.91, we must find all the outcomes less than or equal to that. As in the previous examples, none of the resampled ORs in this particular permutation study was less than .91, hence the p value of 0. Although we know the true p value is not 0, it is nevertheless a very small p value, which provides strong evidence against H_o.

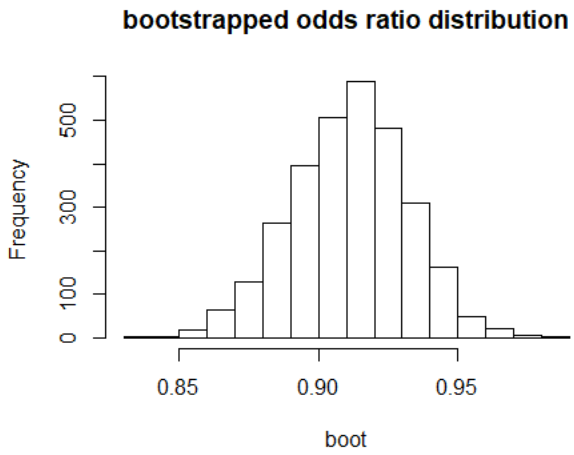
```
OR=exp(mod$coefficients[2])
N=2000
rcoeff=numeric(N)
for (i in 1:N) {
  y=sample(0:1,200, replace=T)
  x=sample(ecls$wksesl,200,replace=F)
  mod<-glm(y~x,family=binomial)
  rcoeff[i] <- exp(mod$coefficients[2])
}

hist(rcoeff,main="odds ratio distribution under Ho")
```




```
pvalue=length(which(rcoeff<=OR))/N
pvalue

## [1] 0
```

Parameter estimation. With strong evidence of statistical conclusion validity in the data, let's estimate the direction and size of the relationship between family SES and school type in the Kindergarten students population. The bootstrapping procedure preserves the relationship between x and y by boostrapping pairs of x,y observations. The t and percentile intervals provide nearly identical estimates, and indicate that with 95% confidence the odds of attending a public school range are reduced from 5% to 13% for each additional point SES scale in the Kindergarten population.

```
N=3000
boot=numeric(N)
for (i in 1:N) {
  dat <- ecls[sample(nrow(dat),200,replace=T),]
  mod <- glm(schtype~wksesl,data=dat,family=binomial)
  boot[i] <- exp(mod$coefficients[2])
}
hist(boot,main="bootstrapped odds ratio distribution")
```




```
#t interval
SEb=sd(boot)
n=200
moe=qt(0.975,n-1)*SEb
OR+c(-1,1)*moe
```

```
## [1] 0.8706223 0.9534110
```

```
#percentile interval
quantile(boot,c(0.05,0.95))
```

```
##        5%       95%
## 0.8761112 0.9446143
```

9.8 Conclusion

We have learned that resampling methods—randomization tests and bootstrapping—are powerful and flexible tools for addressing two important questions of statistical inference:

1. Is there a relationship between and predictor and outcome (beyond what we would expect to happen randomly, that is)?
2. What is the size and direction of that effect in the population?

Both of these tasks use sample data as a surrogate population, resampling from it to generate a distribution of outcomes that allow us to make probabilistic inferences about the questions above. Resampled distributions can approximate the "real" outcome distributions quite well (although they're more hypothetical than real) *if* the original sample is reasonably representative of the population. We have also learned that inferences regarding statistical conclusion validity and effect size validity, which

were the focus in this chapter, exist within a larger inferential framework that includes inferences about external validity, causation, and replicability. Responsible data analysis asks that we assess the evidence, as best we can, bearing on each of these inferential questions.