A Book Chapter Example

Your Name

Contents

Chapter 1

An Introduction to Nonparametric Methods: Schistosomiasis

Using statistics is no substitute for thinking about the problem -Douglas Montgomery[1]

Randomization tests, permutation tests, and bootstrap methods are quickly gaining in popularity as methods for conduct statistical inference. Why? These nonparametric methods require fewer assumptions and provide results that are often more accurate than those from traditional techniques using well-known distributions (such as the normal, t, or F distribution). These methods are based on computer simulations instead of distributional assumptions and thus are particularly useful when the sample data are skewed or if the sample size is small. In addition, nonparametric methods can be extended to other parameters of interest, such as the median or standard deviation, while the well known parametric methods described in introductory statistics courses are often restricted to just inference for the population mean.

We begin this chapter by comparing two treatments for a potentially deadly disease called Schistosomiasis (shis-tuh-soh-mahy-uh-sis). We illustrate the basic concepts behind nonparametric methods by using randomization tests to:

- Provide an intuitive description of statistical inference.
- Conduct a randomization test by hand
- Use software to conduct a randomization test
- Compare one-sided and two-sided hypothesis tests

[1]Douglas Montgomery, Design and Analysis of Experiments, Fifth edition, Wiley, 2003, page 21.

- Making connections between randomization tests and conventional terminology

After working through the schistosomiasis investigation, you will have the opportunity to analyze several other data sets using randomization tests, permutation tests, bootstrap methods, and rank-based nonparametric tests.

1.1 Investigation: Can a New Drug Reduce the Spread of Schistosomiasis?

Schistosomiasis is a disease occurring in humans caused by parasitic flatworms called schistosomes (skis'-tuhsohms). Schistosomiasis affects about 200 million people worldwide and is a serious problem in sub-Saharan Africa, South America, China, and Southeast Asia. The disease can cause death, but more commonly results in chronic and debilitating symptoms, arising primarily from the body's immune reaction to parasite eggs lodged in the liver, spleen, and intestines.

Currently there is one drug, praziquantel (prā'zĭ-kwän'tĕl'), in common use for treatment of schistosomiasis; it is cheap and effective. However many organizations are worried about relying on a single drug to treat a serious disease which affects so many people worldwide. Drug resistance may have prompted a 1990s outbreak in Senegal, where cure rates were low. In 2007, several researchers published work involving a promising drug called K11777 that, in theory, might also treat schistosomiasis.

In this chapter, we will analyze data from this study where the researchers wanted to find out whether K11777 helps to stop schistosome worms from growing. In one phase of the study, 10 female laboratory mice and 10 male laboratory mice were deliberately infected with the schistosome parasite. Seven days after being infected with schistosomiasis, each mouse was given injections every day for 28 days. Within each sex, 5 mice were randomly assigned to a treatment of K11777 whereas the other 5 mice formed a control group injected with an equal volume of plain water. At day 49, the researchers euthanized the mice and measured both the number of eggs and the numbers of worms in the mice livers. Both numbers were expected to be lower if the drug was effective.

Table 1.1 gives the worm count for each mouse. An individual value plot of the data is shown in Figure 1.1. Notice that the treatment group has fewer worms than the control group for both females and males.

Table 1.1: Worm count data for the schistosomiasis study. Treatment is a regimen of K11777 injections from day 7 to day 35. Control is the same regimen, but with a water solution only.

Female		Male	
Treatment	Control	Treatment	Control
1	16	3.0	31
2	10	5.0	26
2	10	9.0	28
10	7	10.0	13
7	17	6.0	47
Mean 4.4	**12**	**6.6**	**29**

NOTE There is a difference between individual value plots and dotplots. In dotplots (such as Figures 1.3 and 1.4 shown later in this chapter), each observation is represented by a dot along a number line (x-axis). When values are close or the same, the dots are stacked. Dotplots can be used in place of histograms when the sample size is small. Individual value plots, as shown in Figure 1.1, are used to simultaneously display each observation for multiple groups. They can be used instead of boxplots to identify outliers and distribution shape, especially when there are relatively few observations.

Activity: *Describing the Data*

1. Use Figure 1.1 to visually compare the number of worms for the treatment and control groups for both the male and the female mice. Does each of the four groups appear to have a similar center and a similar spread? Are there any outliers (extreme observations that don't seem to fit with the rest of the data)?
2. Calculate appropriate summary statistics (e.g., the median, mean, standard deviation, and range) for each of the four groups. For the female mice, calculate the difference between the treatment and control group means. Do the same for the male mice.

The descriptive analysis in Questions 1 and 2 points to a positive treatment effect: K11777 appears to have

reduced the number of parasitic worms in this sample. But descriptive analysis is usually only the first step in ascertaining whether an effect is real; we often conduct a significance test or create a confidence interval to determine if chance alone could explain the effect.

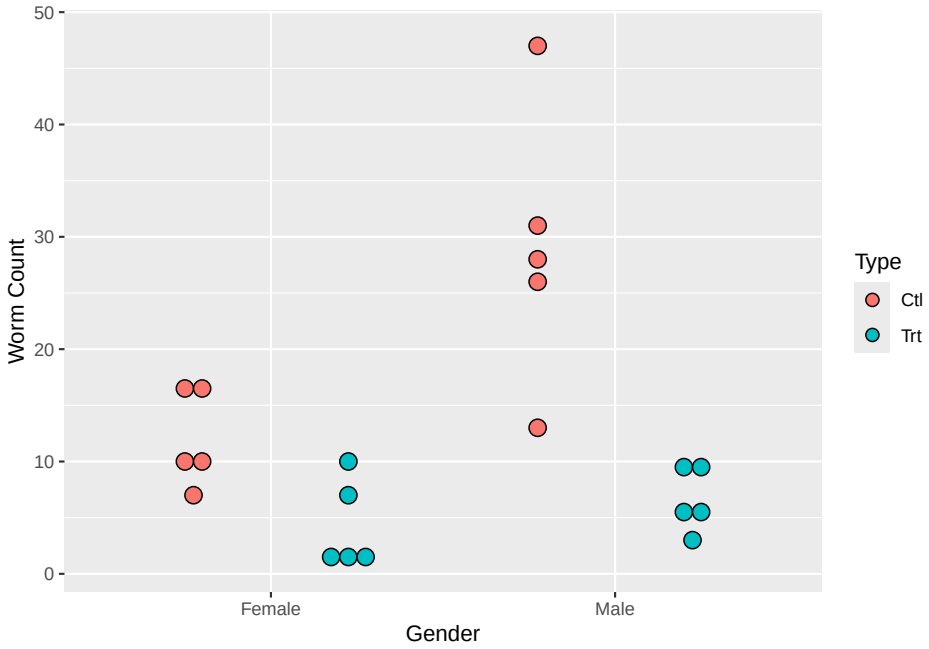

Figure 1.1: Individual value plot of the worm count data

Most introductory statistics courses focus on hypothesis tests that involve using a normal, t-, chi-square or F-distribution to calculate the p-value. These tests are often based on the central limit theorem. In the

schistosomiasis study, there are only five observations in each group. This is a much smaller sample size than is recommended for the central limit theorem, especially given that Figure 1.1 indicates that the data may not be normally distributed. Since we cannot be confident that the sample averages are normally distributed, we will use a distribution-free test, also called a nonparametric test. Such tests do not require the distribution of our sample statistic to have any specific form and are often useful in studies with very small sample sizes.

MATHEMATICAL NOTE: For any population with mean m and finite standard deviation s, the central limit theorem states that the sample mean x from an independent and identically distributed sample tends to follow the normal distribution if the sample size is large enough. The mean of x is the same as the population mean, m, while the standard deviation of x is s/1n, where n is the sample size.

We will use a form of nonparametric statistical inference known as a randomization hypothesis test to analyze the data from the schistosomiasis study. **Ran-**

domization hypothesis tests are significance tests that simulate the random allocation of units to treatments many times in order to determine the likelihood of observing an outcome at least as extreme as the one found in the actual study.

Key Concept: Parametric tests (such as z-tests, t-tests, or F-tests) assume that data come from a population that follows a probability distribution or use the central limit theorem to make inferences about a population. **Nonparametric tests** (such as randomization tests) do not require assumptions about the distribution of the population or large sample sizes in order to make inferences about a population.

We will introduce the basic concepts of randomization tests in a setting where units (mice in this example) are randomly allocated to a treatment or control group. Using a significance test, we will decide if an observed treatment effect (the observed difference between the mean responses in the treatment and control) is "real" or if "random chance alone" could plausibly explain the observed effect. The null hypothesis states that "random chance alone" is the reason for the observed effect. In this initial discussion, the alternative hypothesis will be onesided because we want to show that the true treatment mean (μtreatment) is less than the true control mean (μcontrol). Later, we will expand the discussion to consider modifications needed to deal with two-sided alternatives.

1.2 Statistical Inference Through a Randomization Test

Whether they take the form of significance tests or confidence intervals, inferential procedures rest on the fundamental question for inference: "What would happen if we did this many times?" Let's unpack this question in the context of the female mice in the schistosomiasis study. We observed a difference in means of 7.6 = 12.00 - 4.40 worms between control and treatment groups. While we expect that this large difference reflects the effectiveness of the drug, it is possible that chance alone could explain this difference. This "chance alone" position is usually called the null hypothesis and includes the following assumptions:

- The number of parasitic worms found in the liver naturally varies from mouse to mouse.
- Whether or not the drug is effective, there clearly is variability in the responses of mice to the infestation of schistosomes.

- Each group exhibits this variability, and even if the drug is not effective, some mice do better than others.
- The only explanation for the observed difference of 7.6 worms in the means is that the random allocation randomly placed mice with larger numbers of worms in the control group and mice with smaller numbers of worms in the treatment group.

In this study, the null hypothesis is that the treatment has no effect on the average worm count, and it

is denoted as | H_0: μcontrol = μtreatment Another way to write this null hypothesis is H_0: the treatment has no effect on average worm count

The research hypothesis (the treatment causes a reduction in the average worm count) is called the alternative hypothesis and is denoted H_a (or H_1). For example, H_a: mcontrol 7 mtreatment Another way to write this alternative hypothesis is Ha: the treatment reduces the average worm count Alternative hypotheses can be "one-sided, greater than" (as in this investigation), "one-sided, less-than" (the treatment causes an increase in worm count), or "two-sided" (the treatment mean is different, in one direction or the other, from the control mean). We chose to test a one-sided hypothesis because there is a clear research interest in one direction. In other words, we will take action (start using the drug) only if we can show that K11777 reduces the worm count.

Key Concept: The fundamental question for inference: Every statistical inference procedure (parametric or nonparametric) is based on the question "How does what we observed in our data compare to what would happen if the null hypothesis were actually true and we repeated the process many times?" For a randomization test comparing responses for two groups, this question becomes "How does the observed difference between groups compare to what would happen if the treatments actually had no effect on the individual responses and we repeated the random allocation of individuals to groups many times?"

Activity: *Conducting a Randomization Test by Hand*

3. To get a feel for the concept of a p-value, write each of the female worm counts on an index card. Shuffle the 10 index cards, and then draw five cards at random (without replacement). Call these five cards the treatment group and the five remaining cards the control group. Under the null hypothesis (i.e. the treatment has no effect on worm counts), this allocation mimics precisely what actually happened in our experiment, since the only cause of group differences is the random allocation. | Calculate the mean of the five cards representing the treatment group and the mean of the five cards representing the control group. Then find the difference between the control and treatment group means that you obtained in your allocation. To be consistent, take the control group mean minus the treatment group mean. Your work should look similar to the following simulation:

[[[Fig_CT]]]

4. If you were to do another random allocation, would you get the same difference in means? Explain.

5. Now, perform nine more random allocations, each time computing and writing down the difference in mean worm count between the control group and the treatment group. Make a dotplot of the 10 differences. What proportion of these differences are 7.6 or larger?

6. If you performed the simulation many times, would you expect a large percentage of the simulations to result in a mean difference greater than 7.6? Explain.

The reasoning in the previous activity leads us to the randomization test and an interpretation of the

fundamental question for inference. The fundamental question for this context is as follows: "If the null hypothesis were actually true and we randomly allocated our 10 mice to treatment and control groups many times, what proportion of the time would the observed difference in means be as big as or bigger than 7.6?" This long-run proportion is a probability that statisticians call the **p-value** of the randomization test. The p-values for most randomization tests are found through simulations. Despite the fact that simulations do not give exact p-values, they are usually preferred over the tedious and time-consuming process of listing all possible outcomes. Researchers usually pick a round number such as 10,000 repetitions of the simulation and approximate the p-value accordingly. Since this p-value is an approximation, it is often referred to as the **empirical p-value**.

Key Concept: Assuming that nothing except the ran-

dom allocation process is creating group differences, the p-value of a randomization test is the probability of obtaining a group difference as large as or larger than the group difference actually observed in the experiment.

Key Concept:

The calculation of an empirical p-value requires these steps:

- Repeat the random allocation process a number of times (N times).
- Record, each time, whether or not the group difference exceeds or is the same as the one observed in the actual experiment (let X be the number of times the group difference exceeds or is the same as the one observed).
- Compute X/N to get the p-value, the proportion of times the difference exceeds or is the same as the observed difference.

NOTE: Many researchers include the observed value as one of the possible outcomes. In this case, N = 9999 iterations are typically used and the p-value is calculated as (X + 1)/(9999 + 1). The results are very similar whether X/10,000 or (X + 1)/(9999 + 1) is used. Including the observed value as one of the possible allocations is a more conservative approach and protects against getting a p-value of 0. Our observation from the actual experiment provides evidence that the true p-value is greater than zero.

1.3 Performing a Randomization Test Using a Computer Simulation

While physical simulations (such as the index cards activity) help us understand the process of computing an empirical p-value, using computer software is a much more efficient way of producing an empirical p-value based on a large

number of iterations. If you are simulating 10 random allocations, it is just as easy to use index cards as a computer. However, the advantage of a computer simulation is that 10,000 random allocations can be conducted in almost the same amount of time it takes to simulate 10 allocations. In the following steps, you will develop a program to calculate an empirical p-value.

Activity: *Using Computer Simulations to Conduct a Hypothesis Test*

7. Use the technology instructions provided on the CD to insert the schistosomiasis data into a statistical software package and randomly allocate each of the 10 female worm counts to either the treatment or the control group.
8. Take the control group average minus the K11777 treatment group average.
9. Use the instructions to write a program, function, or macro to repeat the process 10,000 times. Count the number of simulations where the difference between the group averages (control minus K11777) is greater than or equal to 7.6, divide that count by 10,000, and report the resulting empirical p-value.
10. Create a histogram of the 10,000 simulated differences between group means and comment on the shape of the histogram. This histogram, created from simulations of a randomization test, is called an empirical randomization distribution. This distribution describes the frequency of each observed difference (between the control and treatment means) when the null hypothesis is true.
11. Based on your results in Questions 9 and 10 and assuming the null hypothesis is true, about how frequently do you think you would obtain a mean difference as large as or larger than 7.6 by random allocation alone?
12. Does your answer to Question 11 lead you to believe the "chance alone" position (i.e., the null hypothesis that the mean worm count is the same for both the treatment and the control), or does it lead you to believe that K11777 has a positive inhibitory effect on the schistosome worm in female mice? Explain.

Figure 1.2 shows a histogram resulting from the previous activity. A computer simulation of Question 9

resulted in a p-value of 281/10,000 = 0.0281. This result shows that random allocation alone would produce a mean group difference as large as or larger than 7.6 only about 3% of the time, suggesting that something other than chance is needed to explain the difference in group means. Since the only other distinction

between the groups is the presence or absence of treatment, we can conclude that the treatment causes a reduction in worm counts.

We conducted four more simulations, each with 10,000 iterations, which resulted in p-values of 0.0272,

0.0282, 0.0268, and 0.0285. When the number of iterations is large, the empirical randomization distribution (such as the histogram created in Question 10) provides a precise estimate of the likelihood of all possible values of the difference between the control and treatment means. Thus, when the number of iterations is large, well-designed simulation studies result in empirical p-values that are fairly accurate. The larger the number of iterations (i.e., randomizations) within a simulation study, the more precise the p-value is.

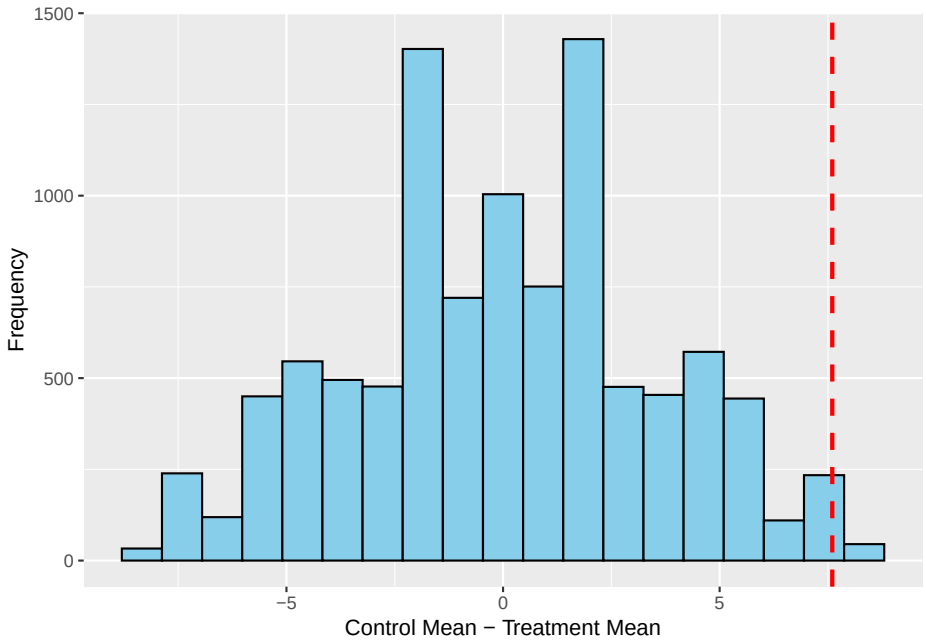

Figure 1.2: Histogram showing the results of a schistosomiasis simulation study. In this simulation, 281 out of 10,000 resulted in a difference greater than or equal to 7.6.

Because the sample sizes in the schistosomiasis study are small, it is possible to apply mathematical

methods to obtain an **exact p-value** for this randomization test. An exact p-value can be calculated by writing down the set of all possibilities (assuming each possible outcome is equally likely under the null hypothesis) and then calculating the proportion of the set for which the difference is at least as large as the observed difference. In the schistosomiasis study, this requires listing every possible combination in which five of the 10 female mice can be allocated

to the treatment (and the other five assigned to the control). There are 252 possible combinations. For each of these combinations, the difference between the treatment and control means is then calculated. The exact p-value is the proportion of times in which the difference in the means is at least as large as the observed difference of 7.6 worms. Of these 252 combinations, six have a mean difference of 7.6 and one has a mean difference greater than 7.6 (namely 8.8). Since all 252 of these random allocations are equally likely, the exact p-value in this example is $7/252 = 0.0278$. However, most real studies are too large to list all possible samples. Randomization tests are almost always adequate, providing approximate p-values that are close enough to the true p-value.

CAUTION: Conducting a two-sample t-test on the female mice provides a p-value of 0.011. This p-value of 0.011 is accurate only if the observed test statistic (i.e., the difference between means) follows appropriate assumptions about the distribution. Figure 1.2 demonstrates that the distributional assumptions are violated. While the randomization test provides an approximate p-value "close to 0.0278," it provides a much better estimate of the exact p-value than does the two-sample t-test. Note that each of the five simulations listed gave a p-value closer to the exact p-value than the one given by the two-sample t-test. *Be careful not to trust a p-value provided by statistical software unless you are certain the appropriate assumptions are met.*

Key Concept: The larger the number of randomizations within a simulation study, the more precise the p-value is. When sample sizes are small or sample data clearly are not normal, a p-value derived from a randomization test with 10,000 randomizations is typically more accurate than a p-value calculated from a parametric test (such as the t -test).

Sometimes we have some threshold p-value at or below which we will reject the null hypothesis and

conclude in favor of the alternative. This threshold value is called a significance level and is usually denoted by the Greek letter alpha (α). Common values are $\alpha = 0.05$ and $\alpha = 0.01$, but the value will depend heavily on context and on the researcher's assessment of the acceptable risk of stating an incorrect conclusion. When the study's p-value is less than or equal to this significance level, we state that the results are statistically significant at level A. If you see the phrase "statistically significant" without a specification of α the writer is most likely

assuming $\alpha = 0.05$, for reasons of history and convention alone. However, it is best to show the p-value instead of simply stating a result is significant at a particular α-level.

1.4 Two-Sided Tests

The direction of the alternative hypothesis is derived from the research hypothesis. In this K11777 study, we enter the study expecting a reduction in worm counts and hoping the data will bear out this expectation. It is our expectation, hope, or interest that drives the alternative hypothesis and the randomization calculation. Occasionally, we enter a study without a firm direction in mind for the alternative, in which case we use a two-sided alternative. Furthermore, even if we hope that the new treatment will be better than the old treatment or better than a control, we might be wrong—it may be that the new treatment is actually worse than the old treatment or even harmful (worse than the control). Some statisticians argue that a conservative objective approach is to always consider the two-sided alternative. For a **two-sided test**, the p-value must take into account extreme values of the test statistic in either direction (no matter which direction we actually observe in our sample data)

Key Concept: The direction of the alternative hypothesis does not depend on the sample data, but instead is determined by the research hypothesis before the data are collected.

We will now make our definition of the p-value more general to allow for a wider variety of significance

testing situations. The **p-value** is the probability of observing a group difference as extreme as or more extreme than the group difference actually observed in the sample data, assuming that there is nothing creating group differences except the random allocation process.

Activity: *A Two-Sided Hypothesis Test*

13. Run the simulation study again to find the empirical p-value for a two-sided hypothesis test to determine if there is a difference between the treatment and control group means for female mice.
14. Is the number of simulations resulting in a difference greater than or equal to 7.6 identical to the number of simulations resulting in a difference less than or equal to -7.6? Explain why these two values are likely to be close but not identical.
15. Explain why you expect the p-value for the two-sided alternative to be about double that for the onesided alternative. Hint: You may want to

look at Figure 1.2

16. Using the two-sided alternative hypothesis, the two-sample t-test provides a p-value of 0.022.[2] This p-value would provide strong evidence for rejecting the assumption that there is no difference between the treatment and the control (null hypothesis). However, this p-value should not be used to draw conclusions about this study. Explain why.

For the above study, a simulation involving 100,000 iterations provided an empirical p-value of 0.0554.

Again, because this particular data set is small, all 252 possible random allocations can be listed to find that the exact two-sided p-value is $14/252 = 0.0556$.

1.5 What Can We Conclude from the Schistosomiasis Study?

The key question in this study is whether K11777 will reduce the spread of a common and potentially deadly disease. The result that you calculated from the one-sided randomization hypothesis test should have been close to the exact p-value of 0.0278. This small p-value allows you to reject the null hypothesis and conclude that the worm counts are lower in the female treatment group than in the female control group. In every study, it is important to consider how random allocation and random sampling impact the conclusions.

Random allocation: The schistosomiasis study was an **experiment** because the units (female mice)

were randomly allocated to treatment or control groups. To the best of our knowledge this experiment controlled for any outside influences and allows us to state that there is a cause and effect relationship between the treatment and response. Therefore, we can conclude that K11777 did cause a reduction in the average number of schistosome parasites in these female mice.

Random sampling: Mice for this type of study are typically ordered from a facility that breeds and raises lab

mice. It is possible that the mice in this study were biologically related or were exposed to something that caused their response to be different from that of other mice. Similarly, there are risks in simply assuming that male mice have the same response as females, so the end-of-chapter exercises provide an opportunity to conduct a separate test on the male mice. Since our sample of 10 female mice was not selected at random from the population of all mice, we should question whether the results from this study hold for all mice.

More importantly, the results have not shown that this new drug will have the same impact on humans as it does on mice. In addition, even though we found

[2]When we do not assume equal variances Minitab uses 7 degrees of freedom providing a p-value of 0.022 while R uses 7.929 degrees of freedom resulting in a p-value of 0.0194.

that K11777 does cause a reduction in worm counts, we did not specifically show that it will reduce the spread of the disease. Is the disease less deadly if only two worms are in the body instead of 10? Statistical consultants aren't typically expected to know the answers to these theoretical, biological, or medical types of questions, but they should ask questions to ensure that the study conclusions match the hypothesis that was tested. In most cases, drug tests require multiple levels of studies to ensure that the drug is safe and to show that the results are consistent across the entire population of interest. While this study is very promising, much more work is needed before we can conclude that K11777 can reduce the spread of schistosomiasis in humans.

A Closer Look: Nonparametric Methods

1.6 Permutation Tests versus Randomization Tests

The random allocation of experimental units (e.g., mice) to groups provides the basis for statistical inference in a randomized comparative experiment. In the schistosomiasis K11777 treatment study, we used a significance test to ascertain whether cause and effect was at work. In the context of the random allocation study design, we called our significance test a randomization test. | In **observational studies**, subjects are not randomly allocated to groups. In this context, we apply the same inferential procedures as in the previous experiment, but we commonly call the significance test a **permutation test** rather than a randomization test.[3] More importantly, in observational studies, the results of the test cannot typically be used to claim cause and effect; a researcher should exhibit more caution in the interpretation of results.

NOTE: The permutation test does not require that the data (or the sampling distribution) follow a normal distribution. However, the null hypothesis in a permutation test assumes that samples are taken from two populations that are similar. So, for example, if the two population variances are very different, the p-value of a permutation test may not be reliable. However, the two-sample t-test (taught in most introductory courses) allows us to assume unequal variances.

Key Concept: Whereas in experiments units are randomly allocated to treatment groups, observational studies do not impose a treatment on a unit. Because the ran-

[3]This text defines a randomization test as a permutation test that is based on random allocation. Some statisticians do not distinguish between permutation tests and randomization tests. They call simulation studies permutation tests, whether they are based on observational studies or experiments.

dom allocation process protects against potential biases caused by extraneous variables, experiments are often used to show causation.

Age Discrimination Study

Westvaco is a company that produces paper products. In 1991, Robert Martin was working in the engineering department of the company's envelope division when he was laid off in Round 2 of several rounds of layoffs by the company.3 He sued the company, claiming to be the victim of age discrimination. The ages of the 10 workers involved in Round 2 were: 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The ages of the three people laid off were 55, 55, and 64.

Figure 1.3 shows a comparative dotplot for age by layoff category. This dotplot gives the impression that

Robert Martin may have a case: It appears as if older workers were more likely to be laid off. But we know enough about variability to be cautious.

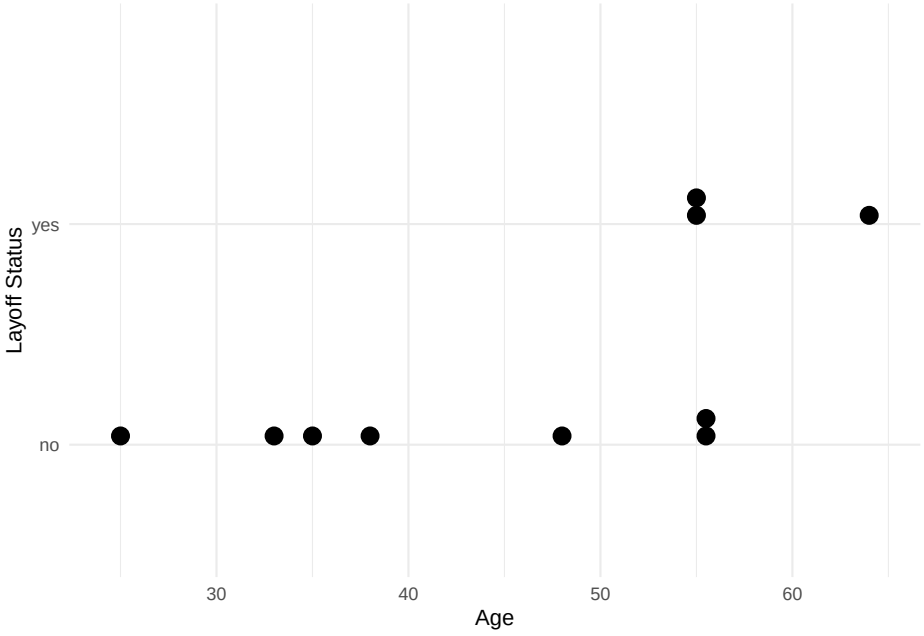

Figure 1.3: Dotplot of age in years of worker versus layoff (whether he or she was laid off)

Extended Activity: *Is There Evidence of Age Discrimination?*

Data set: `Age` 17. Conduct a permutation test to determine whether the observed difference between means is likely to occur just by chance. Use `Age` as the response variable and `Layoff` as the explanatory variable. Here we are interested in only a one-sided hypothesis test to determine if the mean age of people who were laid off is higher than the mean age of people who were not laid off.

18. Modify the program/macro you created in Question 17 to conduct a one-sided hypothesis test to determine if the median age of people who were laid off is higher than the median age of people who were not laid off. Report the p-value and compare your results to those in Question 17.

Since there was no random allocation (i.e., people were not randomly assigned to a layoff group),

statistical significance does not give us the right to assert that greater age is *causing* a difference in being laid off. The null hypothesis in this context becomes "The observed difference could be explained as if by random allocation alone." That is, we proceed as any practicing social scientist must when working with observational data. We "imagine" an experiment in which workers are randomly allocated to a layoff group and then determine if the observed average difference between the ages of laid-off workers and those not laid off is significantly larger than would be expected to occur by chance in a randomized comparative experiment. | While age could be the cause for the difference—hence proving an allegation of age discrimination— there are many other possibilities (i.e., extraneous variables), such as the educational levels of the workers, their competence to do the job, and ratings on past performance evaluations. Rejecting the "as if by random allocation" hypothesis in the nonrandomized context can be a useful step toward establishing causality; however, it cannot establish causality unless the extraneous variables have been properly accounted for. | In the actual court case, data from all three rounds of layoffs were statistically analyzed. The analysis showed some evidence that older people were more likely to be laid off; however, Robert Martin ended up settling out of court.

1.7 Permutation and Randomization Tests for Matched Pairs Designs

The ideas developed in this chapter can be extended to other study designs, such as a basic two-variable design called a matched pairs design. In a matched pairs design, each experimental unit provides both measurements in a study with two treatments (one of which could be a control). Conversely, in the completely randomized situation of the schistosomiasis K11777 treatment study, half the units were assigned to control and half to treatment; no mouse received both treatments.

Music and Relaxation

Grinnell College students Anne Tillema and Anna Tekippe conducted an experiment to study the effect of music on a person's level of relaxation. They hypothesized that fast songs would increase pulse rate more than slow songs. The file called Music contains the data from their experiment. They decided to use a person's pulse rate as an operational definition of the person's level of relaxation and to compare pulse rates for two selections of music: a fast song and a slow song. For the fast song they chose "Beyond" by Nine Inch Nails, and for the slow song they chose Rachmaninoff's "Vocalise." They recruited 28 student subjects for the experiment.

Anne and Anna came up with the following experimental design. Their fundamental question

involved two treatments: (1) listening to the fast song and (2) listening to the slow song. They could have randomly allocated 14 subjects to hear the fast song and 14 subjects to hear the slow song, but their more efficient approach was to have each subject provide both measurements. That is, each subject listened to both songs, giving rise to two data values for each subject, called a matched pairs. Randomization came into play when it was decided by a coin flip whether each subject would listen first to the fast song or the slow song.

NOTE: There are several uses of randomness mentioned in this chapter. The emphasis of this chapter is on the use of **randomization tests** for statistical inference. Most introductory statistics courses discuss random **sampling** from a population, which allows the results of a specific study to be generalized to a larger population. In experiments, units are **randomly allocated to groups** which allows researchers to make statements about causation. In this example, Anne and Anna **randomize the order** to prescribe two conditions on a single subject.

Specifically, as determined by coin flips, half the subjects experienced the following procedure:

[one minute of rest; measure pulse (prepulse)] > [listen to fast song for 2 minutes; measure pulse for second minute (fast song pulse)] > [rest for one minute] > [listen to slow song for 2 minutes; measure pulse for second minute (slow song pulse)].

The other half experienced the procedure the same way except that they heard the slow song first and

the fast song second. | Each subject gives us two measurements of interest for analysis: (1) fast song pulse minus prepulse and (2) slow song pulse minus prepulse. In the data file, these two measurements are called `Fastdiff` and `Slowdiff`, respectively.

Figure 1.4 shows a dotplot of the 28 `Fastdiff`-minus-`Slowdiff` values. Notice that positive numbers

predominate and the mean difference is 1.857 beats per minute, both suggesting that the fast song does indeed heighten response (pulse rate) more than the slow song. We need to confirm this suspicion with a randomization test.

To perform a randomization test, we mimic the randomization procedure of the study design. Here,

the randomization determined the order in which the subject heard the songs, so randomization is applied to the two measurements of interest for each subject. To compute a p-value, we determine how frequently we would obtain an observed difference as large as or larger than 1.857.

```
#> Bin width defaults to 1/30 of the range of the data. Pick
#> better value with `binwidth`.
```

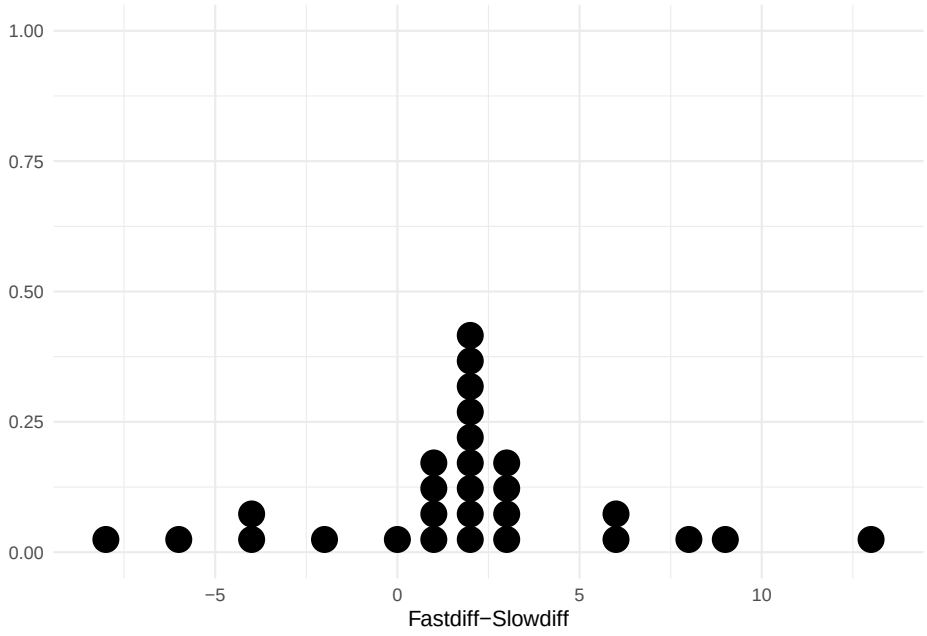

Figure 1.4: Dotplot of the difference in pulse rates for each of the 28 subjects.

Extended Activity: *Testing the Effect of Music on Relaxation*

Data set: `Music`

19. Before they looked at the data, Anne and Anna decided to use a one-sided

test to see whether fast music increased pulse rate more than slow music. Why is it important to determine the direction of the test before looking at the data?

20. Create a simulation to test the Music data. Use the technology instructions provided to randomly multiply a 1 or a -1 by each observed difference. This randomly assigns an order ('Fastdiff - Slowdiff' or 'Slowdiff - Fastdiff'). Then, for each iteration, calculate the mean difference. The p-value is the proportion of times your simulation found a mean difference greater than or equal to 1.857.

    (a) Create a histogram of the mean differences. Mark the area on the histogram that represents your p-value.

    (b) Use the p-value to state your conclusions in the context of the problem. Address random allocation and random sampling (or lack of either) when stating your conclusions.

CAUTION: The type of randomization in Question 20 does not account for extraneous variables such as a great love for Nine Inch Nails on the part of some students or complete boredom with this band on the part of others (i.e., "musical taste" is a possible confounder that randomizing the order of listening cannot randomize away). There will always be a caveat in this type of study, since we are rather crudely letting one Nine Inch Nails song "represent" fast songs.

1.8 The Bootstrap Distribution

Bootstrapping is another simulation technique that is commonly used to develop confidence intervals and hypothesis tests. Bootstrap techniques are useful because they generalize to situations where traditional methods based on the normal distribution cannot be applied. For example, they can be used to create confidence intervals and hypothesis tests for any parameter of interest, such as a median, ratio, or standard deviation. Bootstrap methods differ from previously discussed techniques in that they sample **with replacement** (randomly draw an observation from the original sample and put the observation back before drawing the next observation). | Permutation tests, randomization tests, and bootstrapping are often called **resampling techniques** because, instead of collecting many different samples from a population, we take repeated samples (called resamples) from just one random sample.

Extended Activity: *Creating a Sampling Distribution and a Bootstrap Distribution*

Data set: `ChiSq`

21. The file ChiSq contains data from a highly skewed population (with mean 0.9744 and standard deviation 1.3153). a. Take 1000 simple random samples of size 40 and calculate each mean (x). Plot the histogram of the 1000 sample means. The distribution of sample means is called the sampling distribution. b. What does the central limit theorem tell us about the shape, center, and spread of the sampling distribution in this example? c. Calculate the mean and standard deviation of the sampling distribution in Part A. Does the sampling distribution match what you would expect from the central limit theorem? Explain.

22. Take one simple random sample of size 40 from the ChiSq data. a. Take 1000 resamples (1000 samples of 40 observations with replacement from the one simple random sample). b. Calculate the mean of each resample (x*) and plot the histogram of the 1000 resample means. This distribution of resample means is called the bootstrap distribution. c. Compare the shape, center, and spread of the simulated histograms from Part B and Question 21 Part A. Are they similar?

23. Instead of using the sample mean, create a sampling distribution and bootstrap distribution of the standard deviation of the ChiSq data using a sample size of 40. Compare the shape, center, and spread of the simulated histograms and compare the mean and standard deviation of the distributions.

Key Concept: The bootstrap method takes one simple random sample of size n from a population. Then many resamples (with replacement) are taken from the original simple random sample. Each resample is the same size as the original random sample. The statistic of interest is calculated from each resample and used to create a bootstrap distribution.

In many real-world situations, the process used in Question 21 is not practical because collecting more

than one simple random sample is too expensive or time consuming. While the approach in Question 22 is computer intensive, it is simple and convenient since it uses only one simple random sample. The key idea behind bootstrap methods is the assumption that the original sample represents the population, so resamples from the one simple random sample can be used to represent samples

from the population, as is done in Question 22. Thus, the bootstrap distribution provides an approximation of the sampling distribution.

Most traditional methods of statistical inference involve collecting one sample and calculating the sample

mean. Then, based on the central limit theorem, assumptions are made about the shape and spread of the sampling distribution. In Question 22 we used one sample to calculate the sample mean and then used the bootstrap distribution to estimate the shape and spread of the sampling distribution.

The central limit theorem tells us about the shape and spread of the sample mean. A key advantage of

the bootstrap distribution is that it works for any parameter of interest. Thus, the bootstrap distribution can be used to estimate the shape and spread for any sampling distribution of interest.

CAUTION: When sample sizes are small, one simple random sample may not represent the population very well. However, with larger sample sizes, the bootstrap distribution does represent the sampling distribution.

Figure 1.5 shows the sampling distribution and the bootstrap distribution when a sample size of 10 is used to estimate the mean of the `ChiSq` data. Notice that the spreads for both histograms are

roughly equivalent. The central limit theorem tells us that the standard deviation of the sampling distribution (the distribution of $\bar{x}$) should be $\sigma/\sqrt{n} = 1.3153/\sqrt{10} = 0.4159$. The standard deviation of the bootstrap distribution is 0.4541, which is a reasonable estimate of the standard deviation of the sampling distribution. In addition, both graphs have similar, right-skewed shapes. The strength of the bootstrap method is that it provides accurate estimates of the shape and spread of the sampling distribution. In general, histograms from the bootstrap distribution will have a similar shape and spread as histograms from the sampling distribution.

```
#> [1] "1.465279886"
```

[[[Fig1,5]]]

The bootstrap method does not improve our estimate of the population mean. The mean of the sampling distribution in Question 21 will typically be very close to the population mean. But the mean of the bootstrap distribution in Question 22 typically will not be as accurate, because it is based on only one simple random sample. Ideally, we would like to know how close the statistic from our original sample is to the population parameter. A statistic is biased if it is not centered at the value of the population parameter. We can use the bootstrap distribution to estimate the bias of a statistic. The difference between

the original sample mean and the bootstrap mean is called the **bootstrap estimate of bias**.

Key Concept: The estimate of the mean (or any parameter of interest) provided by the bootstrap distribution is not any better than the estimate provided by the observed statistic from the original simple random sample. However, the shape and spread of the bootstrap distribution will be similar to the shape and spread of the sampling distribution. The bootstrap technique can be used to estimate sampling distribution shapes and standard deviations that cannot be calculated theoretically.

1.9 Using Bootstrap Methods to Create Confidence Intervals

A **confidence interval** gives a range of plausible values for some parameter. This is a range of values surrounding an observed estimate of the parameter—an estimate based on the data. To this range of values we attach a level of confidence that the true parameter lies in the range. An alpha-level, α, is often used to specify the level of confidence. For example, when $\alpha = 0.05$, we have a 100(1 - α), = 95% confidence level. Thus, a 100(1 - α), confidence interval gives an estimate of where we think the parameter is and how precisely we have it pinned down.

1.10 *Bootstrap t Confidence Intervals{-}

If the bootstrap distribution appears to be approximately normal, it is typically safe to assume that a t-distribution can be used to calculate a 100(1 - α), confidence interval for μ, often called a bootstrap t confidence interval:

$$\bar{x} \pm t^* \left(S^*\right) \tag{1.1}$$

where S^* is the standard deviation of the bootstrap distribution and t^* is the critical value of the t-distribution with n - 1 degrees of freedom.

The one simple random sample of size n = 10 used to create the bootstrap distribution in Figure 1.5b has a mean of $\bar{x} = 1.238$ and a standard deviation of s = 1.490. The bootstrap distribution in Figure 1.5b has a mean of $\bar{x}^* = 1.249$ and a standard deviation of $S^* = 0.4541$. Notice that Formula (1.1) uses the

mean from the original sample but uses the bootstrap distribution to estimate the spread. If we *incorrectly assume* that the sampling distribution in Figure 1.5 is normal, a 95% bootstrap t confidence interval for μ is given by

$$\bar{x} \pm t^* (S^*) = 1.238 \pm 2.262(0.4541) \tag{1.1}$$

where $t^* = 2.262$ is the critical value corresponding to the 97.5th percentile of the t-distribution with n - 1 = 9 degrees of freedom. Thus, the 95% confidence interval for μ is (0.211, 2.265).

MATHEMATICAL NOTE: The bootstrap t confidence interval is similar to the traditional one-sample t confidence interval. The key difference is that the bootstrap distribution estimates the standard error of the statistic with S* instead of $s/\sqrt{n}$. When the data are not skewed and have no clear outliers, parametric tests are very effective with relatively small sample sizes (10–30 observations may be enough to use the t-distribution). The following formula uses the t-distribution to calculate a 100(1 - α), confidence interval for the mean of a normal population:

$$\bar{x} + t^* \left(\frac{s}{\sqrt{n}} \right) \tag{1.2}$$

where $s/\sqrt{n}$ is the standard error of x and t* is the critical value of the t-distribution with n - 1 degrees of freedom. Using the original sample of size 10 with mean 1.238 and standard deviation 1.490, we find that a 95sample means only when the sampling distribution is approximately normal. If the data are skewed, even sample sizes greater than 30 may not be large enough to make the sampling distribution appear normal.

With skewed data or small sample sizes (if the original data are not normally distributed), parametric methods (which are based on the central limit theorem) are not appropriate. In Figure 1.5 we see that the sampling distribution is skewed to the right. *Thus, with a sample size of 10, neither the traditional onesample t confidence interval nor the bootstrap t confidence interval is reliable in this example.* However, with a sample size of 40, the histograms in Questions 21 and 22 should tend to look somewhat normally distributed.

Bootstrap Percentile Confidence Intervals

Bootstrap percentile confidence intervals are found by calculating the appropriate percentiles of the bootstrap distribution. To find a 100(1 - α) confidence

interval, take the $\alpha/2$ * 100 percentile of each tail of the bootstrap distribution. For example, to find a 95% confidence interval for μ, sort all the observations from the bootstrap distribution and find the values that represents the 2.5th and 97.5th percentiles of the bootstrap distribution. The 2.5th percentile of the bootstrap distribution in Figure 1.5b is 0.546, and the 97.5th percentile is 2.282. Thus, a 95% confidence interval for μ is (0.546, 2.282). Notice that the percentile confidence interval is not centered at the sample mean. Since the bootstrap distribution is right skewed, the right side of the confidence interval (2.282 - 1.238 = 1.044) is wider than the left side of the confidence interval (1.238 - 0.546 = 0.692). This lack of symmetry can influence the accuracy of the confidence interval.

Key Concept: A bootstrap percentile confidence interval contains the middle 100(1 - a)If the bootstrap distribution is symmetric and is centered on the observed statistic (i.e., not biased), percentile confidence intervals work well.

When to Use Bootstrap Confidence Intervals

Bootstrap methods are extremely useful when we cannot use theory, such as the central limit theorem, to approximate the sampling distribution. Thus, bootstrap methods can be used to create confidence intervals for essentially any parameter of interest, while the central limit theorem is limited to only a few parameters (such as the population mean).[4] However, bootstrap methods are not always reliable.

Small sample sizes still produce problems for bootstrap methods. When the sample size is small, (1) the sample statistic may not accurately estimate the population parameter, (2) the distribution of sample means is less likely to be symmetric, and (3) the shape and spread of the bootstrap distribution may not accurately represent those of the true sampling distribution.

In addition, bootstrap methods do not work equally well for all parameters. For example, the end-ofchapter

exercises show that bootstrapping often provides unreliable bootstrap distributions for median values because the median of a resample is likely to have only a few possible values. Thus, confidence intervals for medians should be used only with large (n $\geq$ 100) sample sizes.

[4]Theoretical methods allow distributional tests for more than just the population mean. However, for purposes of this text it is sufficient to understand that distributional methods tend to be more complicated and are limited to testing only a few parameters that could be of interest.

It is not easy to determine whether bootstrap methods provide appropriate confidence intervals. The bootstrap t and bootstrap percentile confidence intervals are often compared to each other. While the percentile confidence interval tends to be more accurate, neither of the two should be used if the intervals are not relatively

close. If the bootstrap distribution is skewed or biased, other methods should be used to find confidence intervals. More advanced bootstrap methods (such as BCa and tilting confidence intervals) are available that are generally accurate when bias or skewness exists in the bootstrap distribution.[5]

Extended Activity:*Estimating Salaries of Medical Faculty*

Data set: `MedSalaries`. The file `MedSalaries` is a random sample of n = 100 salaries of medical doctors who were teaching at United States universities in 2009.

24. Create a bootstrap distribution of the mean by taking 1000 resamples (with replacement). Create a bootstrap t confidence interval and a bootstrap percentile distribution to estimate the mean salaries.
25. Create a bootstrap distribution of the standard deviation by taking 1000 resamples (with replacement). Create a bootstrap t confidence interval and a bootstrap percentile distribution to estimate the population standard deviation.
26. Use Formula (1.2) to create a 95
27. Explain why Formula (1.2) cannot be used to create a 95

1.11 Relationship Between the Randomization Test and the Two-Sample t-Test

R.A. Fisher, perhaps the preeminent statistician of the 20th century, introduced the randomization test in the context of a two-group randomly allocated experiment in his famous 1935 book, *Design of Experiments*.[6] At that time he acknowledged that the randomization test was not practical because of the computational intensity of the calculation. Clearly, 1935 predates modern computing. Indeed, Efron and Tibshirani describe the permutation test as "a computer-intensive statistical technique that predates computers."[7] Fisher went on to assert that the classical two-sample t-test (for independent samples) approximates the randomization test very well. Ernst cites references to several approximations to the randomization tests using classical and computationally tractable methods that have been published over time.[8]

If you have seen two-sample tests previously, it is likely to have been in the context of what Ernst calls the population model, which he distinguishes from the randomization model. In a **population model**, units are selected at random from one or more populations. Most observational studies are population models. One simple case of a population model involves comparing two separate population means. In this case, we can take two independent simple random samples and use the classic two-sample t-test to make the comparison.

In a ***randomization model**, a fixed number of experimental units are randomly allocated to treatments. Most experiments are randomization models. In randomization models such as the schistosomiasis example,

the two samples are formed from a collection of available experimental units that are randomly divided into two groups. Since there are a fixed number of units, the groups are not completely independent. For example, if one of the 10 male mice had a natural resistance to schistosomiasis and was randomly placed in the treatment group, we would expect the control group to have a slightly higher worm count. Since the two groups are not completely independent, the assumptions of the classic two-sample t-test are violated. Even if the sample sizes in the schistosomiasis study were much larger, the randomization test would be a more appropriate test than the two-sample t-test. However, empirical evidence has shown that the two-sample t-test is a very good approximation to the randomization test when sample sizes are large enough. We are fortunate that, in this age of modern computing, we no longer have to routinely compromise by using the t-test to approximate the randomization test.

Key Concept: Historically, the two-sample t -test was used to approximate the p-value in randomization models because randomization tests were too difficult to compute. However, now that computers can easily simulate random assignment to groups, randomization tests should be used to calculate p-values for randomization models, especially if sample sizes are fairly small.

1.12 Wilcoxon Rank Sum Tests for Two Independent Samples

The **Wilcoxon rank sum test**, also called the two-sample **Mann-Whitney test**, makes inferences about the difference between two populations based on data from two independent random samples. This test ranks observations from two samples by arranging them in order from smallest to largest.

Focusing on ranks instead of the actual observed values allows us to remove

Table 1.2: Randomly selected pitchers and first baseman from 2005 National League baseball teams.

Team	Position	Name	Salary(\$)
Milwaukee Brewers	Pitcher	Obermueller, Wes	342000
Houston Astros	Pitcher	Backe, Brandon	350000
Atlanta Braves	Pitcher	Sosa, Jorge	650000
Atlanta Braves	Pitcher	Thomson, John	4250000
Cincinnati Reds	First Baseman	Casey, Sean	7800000
Arizona Diamondbacks	First Baseman	Green, Shawn	7833333
San Diego Padres	First Baseman	Nevin, Phil	9625000
New York Mets	Pitcher	Glavine, Tom	10765608
Colorado Rockies	First Baseman	Helton, Todd	12600000
Philadelphia Phillies	First Baseman	Thome, Jim	13166667

Table 1.3: Ranking the 10 randomly selected 2005 National League baseball players.

Position	Pr	Pr	Pr	Pr	FB	FB	FB	Pr	FB	FB
Salary	342	350	650	4250	7800	7833	9625	10766	12600	13167
Rank	1	2	3	4	5	6	7	8	9	10

assumptions about the normal distribution. Rank-based tests have been used for many years. However, rank-based methods (discussed

in this section and the next section) are much less accurate than methods based on simulations. In general, randomization tests, permutation tests, or bootstrap methods should be used whenever possible.

The following example examines whether pitchers and first basemen who play for National League baseball teams have the same salary distribution. The null and alternative hypotheses are written as,

H_0: the distribution of the salaries is the same for pitchers and first basemen
H_a: the distribution of the salaries is different for pitchers and first basemen

Table 1.2 shows the salaries of five pitchers and five first basemen who were randomly selected from all National League baseball players. Table 1.3 ranks each of the players based on 2005 salaries.

Note that if two players had exactly the same salary, standard practice would be to average the ranks of the tied values.

For the Wilcoxon rank sum test, we define the following terms:

- n_1 is the sample size for the first group (5 for the pitcher group in this

example)
- n_2 is the sample size for the second group (5 for the first baseman group in this example)
- N= $n_1 + n_2$
- W, the Wilcoxon rank sum statistic, is the sum of the ranks in the first group (1 + 2 + 3 + 4 + 8 = 18)

If the two groups are from the same continuous distribution, then W has a mean,

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{5(11)}{2} = 27.5 \tag{1.3}$$

and standard deviation[9]

$$\sigma_W = \sqrt{\frac{n_1 n_2(N+1)}{12}} = \sqrt{\frac{(5)(5)(11)}{12}} = 4.787 \tag{1.4}$$

If W is far from μ_W, then the Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions—that is, rejects H_0 (no difference in distribution of salaries) in favor of H_a (salary distributions are different based on position). The p-value is the probability of observing a sample statistic, W, at least as extreme as the one in our sample. Since 18 is less than the hypothesized mean, 27.5, the p-value for the two-sided test in this example is found by calculating 2 * P(W $\leq$ 18).

MATHEMATICAL NOTE: Computer software such as R, S-plus, or SAS tends to use the exact distribution of W, though Minitab uses a normal approximation for this test. If the data contain ties, the exact distribution for the Wilcoxon rank sum statistic changes and the standard deviation of W should be adjusted. Statistical software will typically detect the ties and use the normal distribution (using the adjusted standard deviation) instead of an exact distribution.[10]

Extended Activity: *Wilcoxon Rank Sum Tests*

Data set: `NLBB Salaries`

25. Using a software package, conduct the Wilcoxon rank sum test to determine if the distribution of salaries is different for pitchers than for first basemen.
26. Find 2 X P(W $\leq$ 18) assuming W N(27.5, 4.787). How does your answer compare to that fromQuestion 25?

Table 1.4: Randomly selected catchers from 2005 National League baseball teams.

Team	Position	Name	Salary(\$)
Pittsburgh Pirates	Catcher	Ross, David	338500
Los Angeles Dodgers	Catcher	Phillips, Jason	339000
Atlanta Braves	Catcher	Perez, Eddie	625000
Washington Nationals	Catcher	Bennett, Gary	750000
Pittsburgh Pirates	Catcher	Santiago, Benito	2150000

27. Use a two-sided two-sample t-test (assume unequal variances) to analyze the data. Are your conclusions the same as in Question 25? Create an individual value plot of the data. Are any distributional assumptions violated? Which test is more appropriate to use for this data set?

At first it may seem somewhat surprising that first basemen tend to make more than pitchers. However, in 2005 there were 19 first basemen and 215 pitchers in the National League. Many pitchers did not play much and got paid a low salary, whereas all 19 first basemen were considered quite valuable to their teams.

1.13 Kruskal-Wallis Test for Two or More Independent Samples

The ***Kruskal-Wallis* test** is another popular nonparametric test that is often used to compare two or more independent samples. Like ANOVA, a more common parametric test that will be discussed in later chapters, the Kruskal-Wallis test requires independent random samples from each population. When the data clearly deviate from the normal distribution, the Kruskal-Wallis test will be more likely than a one-way ANOVA to identify true differences in the population. The null and alternative hypotheses for the Kruskal-Wallis test are:

H_0: the distribution of the response variable is the same for all groups H_a: some responses are systematically higher in some groups than in others

The Kruskal-Wallis test is also based on ranks. The ranks are summed for each group, and when these group sums are far apart, we have evidence that the groups are different. While the calculations for the Kruskal-Wallis test statistic are provided here, we suggest using statistical software to conduct this significance test. Continuing the baseball salaries example, Table 1.4 displays salaries of five randomly selected catchers from 2005 National League baseball teams.

For the Kruskal-Wallis test, we define the following terms:

- n_1 is the sample size for the first group (5 for the pitcher group)
- n_2 is the sample size for the second group (5 for the first baseman group)
- n_3 is the sample size for the third group (5 for the catcher group)
- $\mathrm{N} = n_1 + n_2 + n_3$
- R_i is the sum of the ranks for the ith group ($R_1 = 35$, $R_2 = 62$, and $R_3 = 23$)

The Kruskal-Wallis test statistic is calculated as,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{(15)(16)}\left(\frac{35^2}{5} + \frac{62^2}{5} + \frac{23^2}{5}\right) - 3(16) = 7.98 \tag{1.5}$$

The exact distribution of H under the null hypothesis depends on each ni, so it is complex and time consuming to calculate. Even most statistical software packages use the chi-square approximation with I - 1 degrees of freedom to obtain p-values (where I is the number of groups).

NOTE: When the chi-square approximation is used, each group should have at least five observations.

Extended Activity: *Kruskal-Wallis Test*

Data set: `NLBB Salaries`

28. Using a software package, run the Kruskal-Wallis test (use all three groups with samples of size 5 per group) to determine if the distribution of salaries differs by position. Create an individual value plot of the data. Do the data look normally distributed in each group?

Mathematical Note: If the spread of each group appears to increase as the center (mean or median) increases, transforming the data—such as by taking the log of each response variable—will make the data appear much more normally distributed. Then parametric techniques can often be used on the transformed data. In the baseball salary example, the data are highly right skewed in at least two groups. While a log transformation on salaries is helpful, there is still not enough evidence that the transformed salaries are normally distributed. Thus, nonparametric methods are likely the most appropriate approach to testing whether there is a difference in the distribution of salaries based on position.

Nonparametric tests based on rank are usually less powerful (less likely to reject the null hypothesis) than the corresponding parametric tests. Thus, you are

less likely to identify differences between groups when they really exist. If you are reasonably certain that the assumptions for the parametric procedure are satisfied, a parametric procedure should be used instead of a rank-based nonparametric procedure. Many introductory texts suggest that, in order to conduct a parametric test, you should have a sample size of 15 in each group and no skewed data or outliers.

1.14 Multiple Comparisons

In introductory texts, statistical inference is often described in terms of drawing one random sample, performing one significance test, and then stating appropriate conclusions—analysis done, case closed. However, there are many situations where inference is not that simple. Performing multiple statistical tests on the same data set can create several problems.

Using a significance level of $\alpha = 0.05$ (i.e., rejecting H_0 in favor of the alternative when the p-value is less than or equal to 0.05) helps to ensure that we won't make a wrong decision. In other words, one time out of 20 we expect to incorrectly reject the null hypothesis. But what if we want to do 20 or more tests on the same data set? Does this mean that we're sure to be wrong at least once? And if so, how can we tell which findings are incorrect? The following activities explore how researchers can protect themselves from drawing conclusions from statistical findings that could be the result of random chance.

Extended Activity: *Comparing Car Prices*

29. Open the 'Car1' data set and conduct three two-sided hypothesis tests to determine if there is a difference in price. Compare the means: Pontiac versus Buick (test 1), Cadillac versus Pontiac (test 2), and Cadillac versus Buick (test 3). Provide the p-value for each of these three tests. Which tests have a p-value less than 0.05?

30. Assuming the null hypotheses are true, each of the three tests in Question 29 has a 5% chance of inappropriately rejecting the null hypothesis. However, the probability that at least one of the three tests will inappropriately reject the null hypothesis is 14.26%. Assuming that the null hypothesis is true and that each test is independent, complete the following steps to convince yourself that this probability is correct.

    (a) Each test will either reject (R) or fail to reject (F). List all eight possible outcomes in the table below.

Case	Test_1	Test_2	Test_3	Probability
1	F	F	F	
2	F	F	R	
3	F	R	F	
4				
5				
6				
7				
8				

(b) The probability that each test rejects is $P(R) = 0.05$, and the probability that each test fails to reject is $P(F) = 0.95$. For example, the probability that all three tests fail to reject is $0.95^3 = 0.8574$. The probability that the first two fail to reject and the third does reject is $0.95 \times 0.95 \times 0.05 = 0.0451$. Complete the table. Verify the probabilities sum to 1. The probability that at least one test rejects is $1 - 0.8574 = 0.1426$.

31. Repeat Question 30 using ($\alpha = 0.10$). What is the probability that at least one of the three tests will inappropriately reject the null hypothesis?

32. To compare all four car makes, six hypothesis tests will be needed. List all six null hypotheses. Assuming independence and that the null hypothesis is true, what is the probability that at least one of the six tests will inappropriately reject the null hypothesis at $\alpha = 0.05$?

Extended Activity: *The Least-Significant Differences Method and the Bonferroni Method*

Data set: `Car1`

When the significance level is controlled for each individual test, as was done in Question 29, the process is often called the **least-significant differences method (LSD)**. Notice that using a = 0.05 for all tests has some undesirable properties, especially when a large number of tests being conducted. If 100 independent tests were conducted to compare multiple groups (and there really were no differences), the probability of incorrectly rejecting at least one test would be 1 - 0.95^{100} = 0.994. Thus, using α= 0.05 as a critical value for 100 comparisons will almost always lead us to incorrectly conclude that some results are significantly different.

One technique that is commonly used to address the problem with multiple comparisons is called the ***Bonferroni method**. This technique protects against the probability of false rejection by using a cutoff value of α/K, where K is the number of comparisons. In Question 29, there are three comparisons (i.e.,three hypothesis tests). Thus, a cutoff value of 0.05/3 = 0.01667 should be used. In

other words, when there are three comparisons as in Question 29, the Bonferroni method rejects the null hypothesis when the p-value is less than or equal to 0.01667. Using the least-significant differences method ($\alpha = 0.05$), as was done in Question 29, we would conclude that the prices of Buicks and Chevrolets are significantly different, but using the Bonferroni method we would fail to reject in all three tests.

33. Repeat Question 30 using the Bonferroni cutoff value of $0.05/3 = 0.016667$ instead of $\alpha = 0.05$. Find the probability that at least one of the tests rejects.

34. Using all four groups of cars and $\alpha = 0.05$ (cutoff of $0.05/6$), do any of the six tests reject the null hypothesis with the Bonferroni method?

35. If there were seven groups, 21 hypothesis tests would be needed to compare all possible pairs. Using $\alpha = 0.05$ and the Bonferroni's method (reject H_0 if the p-value is less than $0.05/21 = 0.00238$) what is the probability that at least one of the tests would reject?

MATHEMATICAL NOTE: Other terms that are commonly discussed with multiple comparisons are **familywise type I error** and ***comparisonwise type I error**. Bonferroni's method is an example of a technique that maintains the familywise type I error. With the familywise type I error 0.05, assuming that there really is no difference between any of the K pairs, there is only a 5differences method is used to maintain a comparisonwise type I error rate: Assuming that a particular null hypothesis test is true, there is a 5

Choosing a Critical Value

The α-level represents the probability of a **type I error**. A type I error can be considered a false alarm: Our hypothesis test has led us to conclude that we have found a significant difference when one does not exist. However, it is important to recognize that it is also possible to make a **type II error**, which means our hypothesis test failed to detect a significant difference when one exists. In essence, a type II error can be thought of as an alarm that failed to go off.

Notice that if the Bonferroni method is used with all six tests, the critical value for each individual test is $0.05/6 = 0.00833$. Thus, this method often fails to detect real differences between groups, leaving us open to a high rate of type II error while protecting us against type I errors.

Neither the least-significant differences nor the Bonferroni method is ideal. Caution should be used with both techniques, and neither technique should be used with numerous comparisons. The key is to recognize the benefits and limitations of each technique and to properly interpret what the results of each technique

tell us. Some researchers suggest limiting the number of tests, using both techniques, and letting the reader decide

which conclusions to draw. Both techniques are commonly used when there are fewer than 10 comparisons. However, a researcher should always decide which comparisons to test before looking at the data.

Chapter Summary

This chapter described the basic concepts behind randomization tests, permutation tests, bootstrap methods, and rank-based nonparametric tests. **Parametric tests** (such as z-tests, t-tests or F-tests) assume that data follow a known a probability distribution or use the central limit theorem to make inferences about a population. ***Nonparametric tests** do not require assumptions about the distribution of the population or the central limit theorem in order to make inferences about a population.

The ***null hypothesis**, denoted H_0, states that in a study nothing is creating group differences except the random allocation process. The research hypothesis is called the **alternative hypothesis** and is denoted H_a (or H_1). The p-value is the likelihood of observing a statistic at least as extreme as the one observed

from the sample data when the null hypothesis is true. A threshold value, called a **significance level**, is denoted by the Greek letter alpha (α). When a study's p-value is less than or equal to this significance level, we state that the results are **statistically significant at level** α. Exact p-values are often difficult to calculate, but ***empirical p-values** can often be simulated through a randomization or permutation test. The empirical p-value will become more precise as the number of randomizations within a simulation study increases.

The steps in a **randomization test** are as follows:

- An experiment is conducted in which units are assigned to a treatment and an observed sample statistic is calculated (such as the difference between group means).
- Software is used to simulate the random allocation process a number of times (N iterations).
- For each iteration, the statistic of interest (difference between group means) is recorded, with X being the number of times the statistic in the iteration exceeds or is the same as the observed statistic in the actual experiment.
- X/N is computed to find the p-value, the proportion of times the statistic exceeds or is the same as the observed difference.

A ***permutation test** is a more general form of the randomization test. The steps in both tests are identical, except that permutation tests do not require

random allocation. Randomization tests and permutation tests can provide very accurate results. These tests are preferred over parametric methods when the sample size is small or when there are outliers in a data set. Since real data sets tend not to come from exactly normal populations, it is important to recognize that even p-values from parametric tests are approximate (but typically accurate as long as the sample sizes are large enough, the data are not skewed, there are no outliers, and the data are reasonably normal). A graph such as a boxplot or individual value plot should always be created to determine if parametric methods are appropriate. Randomization tests are gaining popularity because they require fewer assumptions and are just as powerful as parametric tests.

Bootstrap methods take many (at least 1000) resamples with replacement of the original sample to create

a bootstrap distribution. If the bootstrap distribution is symmetric and unbiased, bootstrap t or bootstrap percentile confidence intervals can be used to approximate 100(1 - α)%, confidence intervals.

The steps in creating **bootstrap confidence intervals** are as follows:

- One sample of size n is taken from a population and the statistic of interest is calculated.
- Software is used to take resamples (with replacement) of size n from the original sample a number of times (N iterations). For each iteration, the statistic of interest is calculated from the resample.
- The **bootstrap distribution**, which is the distribution of all N resample statistics, is used to estimate the shape and spread of the sampling distribution.
- A **bootstrap t confidence interval** is found by calculating $\bar{x} \pm t^*(S^*)$ where S^* is the standard deviation of the bootstrap distribution and t^* is the critical value of the t(n - 1) distribution with 100(1 - α)%, of the area between - t* and t*.
- A 100(1 - α), bootstrap percentile confidence interval is found by taking the α / 2 * 100 percentile of each tail of the bootstrap distribution.

Bootstrap confidence intervals based on small samples can be unreliable. The bootstrap t or percentile confidence interval may be used if,

- the bootstrap distribution does not appear to be biased,
- the bootstrap distribution appears to be normal, and
- the bootstrap t and percentile confidence intervals are similar.

Simulation studies can easily be extended to testing other terms, such as the median or variance, whereas most parametric tests described in introductory statistics classes (such as the z-test and t-test) are restricted to testing for the mean. Simulation studies are an extremely useful tool that can fairly easily be used to calculate accurate p-values for research hypotheses when other tests are not appropriate.

Before computationally intensive techniques were easily available, rank-based nonparametric tests, such

as the **Wilcoxon rank sum** test and the **Kruskal-Wallis test**, were commonly used. These tests do not require assumptions about distributions, but they tend to be less informative because ranks are used instead of the actual data. Both the Mann-Whitney test and the Kruskal-Wallis test assume that sample data are from independent random samples whose distributions have the same shape and scale. Each sample in the Kruskal-Wallis test should consist of at least five measurements. Rank-based nonparametric tests tend to be less powerful (less likely to identify differences between groups) than parametric tests (when assumptions do hold) and resampling methods. When the sample sizes are small and there are reasons to doubt the normality assumption, rank-based nonparametric tests are recommended over parametric tests. Randomization tests and permutation tests are typically preferred over parametric and rank-based tests. Their p-values are often more reliable, and they are more flexible in the choice of parameter tested.

One final note of caution: Even though it is possible to analyze the same data with a variety of parametric

and nonparametric techniques, statisticians should never search around for a technique that provides the results they are looking for. Conducting multiple tests on the same data and choosing the test that provides the smallest p-value will cause the results to be unreliable. If possible, determine the type of analysis that will be conducted before the data are collected.

Exercises

E1. Is it important in the schistosomiasis study for all 20 mice to come from the same population of mice? Why or why not?

E2. Assume the researchers in this study haphazardly pulled the female mice from a cage and assigned the first five to the treatment and the last five to the control. Would you trust the results of the study as much as if five mice were randomly assigned to each group?

E3. A recent study in the northwest United States found that children who watched more television were more likely to be obese than children who watched less television. Can causation be inferred from this study?

E4. What is the difference between a random sample and a randomized experiment?

E5. Explain the difference between a population model and a randomization model.

E6. Explain how the independence assumption of the two-sample t-test is violated in a randomization model.

E7. If the sample size is large, will the histogram of the sample data have a shape similar to that of the normal distribution? Explain.

E8. If the sample size is large, will the sample mean be normally distributed? Explain.

E9. Why should boxplots or other graphical techniques be used to visualize data before a parametric test is conducted?

E10. Suppose that in our study of schistosomiasis in female mice the p-value was 0.85. Would you be able to conclude that there was no difference between the treatment and control means?

E11. **Using Other Test Statistics**
Data set: 'Mice'. One major advantage of randomization/permutation tests over classical methods is that they easily allow the use of test statistics other than the mean.

1. Modify the program/macro you created in Question 9 to measure a difference in group medians instead of a difference in means for the female mice. Report the p-value and compare your results to those for Question 9.
2. You might also wonder if there is a difference in the variability in the groups. Modify the macro you created in Question 9 to test whether the variances of the female groups are equal. Report the p-value and state your conclusions

E12. **Testing Male Mice**
Data set: 'Mice'.

1. Using the data for the male mice, run a simulation to decide whether K11777 inhibits schistosome viability (i.e., reduces worm count) in male mice. Describe the results, including a histogram of the simulation results, the p-value, and a summary statement indicating your conclusion about the research question of schistosome viability.
2. Modify the program/macro you created in Part A to measure a difference in group medians instead of a difference in means for the male mice. Report the p-value and compare your results to those for Part A.
3. You might also wonder if there is a difference in the variability in the groups. Modify the macro you created in Part A to test if the variances of each male group are equal. Report the p-value and state your conclusions.

E13. **Bird Nest Study**

Data set: 'Birdnest'. This data set was collected in the spring of 1999 for a class project by Amy Moore, a Grinnell College student. Each record in the data set represents data for a species of North American passerine bird. Passerines are "perching birds" and include many families of familiar small birds (e.g., sparrows and warblers) as well as some larger species like crows and ravens, but do not include hawks, owls, water fowl, wading birds, and woodpeckers. Moore took all North American passerines for which complete evolutionary data were available, which comprised 99 of the 470 species of passerines in North America (part of her study used this evolutionary information). One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species, nest type was categorized as either closed or open. Although nests come in a variety of types (see the 'Nesttype' variable), in this data set "closed" refers to nests with only a small opening to the outside, such as the tree-cavity nest of many nuthatches or the pendant-style nest of an oriole. "Open" nests include the cup-shaped nest of the American robin.

1. Moore suspected that closed nests tend to be built by larger birds, but here we will treat the alternative as two-sided, since her suspicion was based on scanty evidence. Use comparative dotplots or boxplots and summary statistics to describe the relationship between average body length and nest type (the 'Closed' variable). (Note: 'Closed' = 1 for closed nests; 'Closed' = 0 for open nests.) Does it appear that Moore's initial suspicion is borne out by the data? of the simulation results, the p-value, and a summary statement indicating your conclusion about the research question of schistosome viability.

2. Run a permutation test using a two-sided alternative to determine if type of nest varies by body length and interpret your results. Be sure to state your conclusions in the context of the problem and address how random allocation and random sampling (or lack of either) impact your conclusions.

E14. **Twins Brain Study**

Data set: 'Twins'. In a 1990 study by Suddath et al., reported in Ramsey and Schafer,[12] researchers used magnetic resonance imaging to measure the volume of various regions of the brain for a sample of 15 monozygotic twins, where one twin was affected with schizophrenia and the other was unaffected. The twins were from North America and comprised eight male pairs, and seven female pairs ranging in age from 25 to 44 at the time of the study. The sizes in volume (cm^3) of the hippocampus are in the file called 'Twins'.

1. Should the data be analyzed as match pairs or be treated as if there were two independent samples?

2. Use appropriate graphics and summary statistics to describe the difference in brain volume for affected and unaffected twins.

3. Use the appropriate permutation test to ascertain if the difference in brain volume described in Part B is the result of schizophrenia or if it could be explained as a chance difference. Report your p-value and summarize your conclusion.

E15. **Comparing Parametric and Nonparametric Tests**
Data set: 'Birdnest'and 'Music'.

1. Using a t-test, compute the two-sided p-value for the bird nest study in Exercise E.13. and compare the results to what you found with the randomization test.

2. Using a t-test, compute the one-sided p-value for the music study in Question 20 and compare the results to what you found with the randomization test.

E16. **Means versus Medians in Rank-Based Tests**
Data set: 'SameMean'. Rank-based nonparametric tests do not answer the same question as the corresponding parametric procedure. Many people assume that these nonparametric tests are testing for group medians. This is not always true. Rank-based tests can be interpreted as testing for the median only if the shapes and scales of the populations are the same. The following exercise illustrates this point by providing an example where the medians and the means are identical but nonparametric tests will reject the null hypothesis. | Use the 'SameMean' data to conduct the Kruskal-Wallis test. Calculate the mean and median for each group. What conclusions can you draw from the data?

E17. **Rank Based Bird Nest Tests**
Data set: 'Birdnest'.

1. Use the Wilcoxon rank sum test to conduct a significance test for the bird nest study discussed in Exercise E.13.

2. Use the Kruskal-Wallis test to conduct a significance test for the bird nest study. Determine whether the distribution of bird size (response is Length) is the same for each nest type. Note that when the chi-square approximation is used, each group should have at least five observations. You may need to create an "other" group to combine all nest types with sample sizes less than five.

E18. **Bootstrap Confidence Intervals**
Data set: 'ChiSq'. Take a simple random sample of size 40 from the 'ChiSq' data file.

1. Create a bootstrap distribution of the mean (or use the distribution you created in Question 22). Calculate a 95

2. Create a bootstrap distribution of the mean (or use the distribution you created in Question 22). Calculate a 95percentile confidence intervals for the mean reliable?

3. Create a bootstrap distribution of the standard deviation (or use the distribution you created in Question 23). Calculate a 95

4. Create a bootstrap distribution of the standard deviation (or use the distribution you created in Question 23). Calculate a 95deviation. Are the bootstrap t and percentile confidence intervals for the standard deviation reliable?

E19. **Medians and Trimmed Means in Bootstrap Confidence Intervals** Data set: 'ChiSq'.

1. Take a simple random sample of size n = 40 from the ChiSq data. Create a bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the bootstrap distribution and explain why bootstrap confidence intervals are unlikely to be reliable.

2. Take a second simple random sample of size n = 40 from the ChiSq data. Create a second bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the second bootstrap distribution. With a sample size of 40, why are bootstrap distributions of medians unlikely to be normal?

3. Bootstrap distributions for medians are unlikely to be normally distributed, and means tend to be influenced by outliers. The trimmed mean is a common measure of center that tends to better represent the average value with bootstrap methods. Trimmed means are calculated by first trimming the upper and lower values of the sample. For example, the 25of the middle 50

   | Take a simple random sample of size n = 40 from the ChiSq data. Create a bootstrap distribution of the 25for each resample calculate the mean of the middle 20 observations (remove the smallest 10 and largest 10 values in each resample). Create a histogram of the 1000 trimmed means and describe the shape of this bootstrap distribution. Create a bootstrap t confidence interval and a bootstrap percentile confidence interval to estimate the 25

E20. **Medians and Trimmed Means in Bootstrap Confidence Intervals** Data set: 'MedSalaries'.

1. The file 'MedSalaries' is a random sample of salaries of medical doctors who were teaching at United States universities in 2009. Create a

bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap t confidence interval or a bootstrap percentile confidence interval for the median?

2. Create a bootstrap distribution of the 25(with replacement). In other words, calculate the mean of the middle 50 observations from each resample. Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap t confidence interval or a bootstrap percentile confidence interval for the 25

3. Create a bootstrap distribution of the 5replacement). In other words, calculate the mean of the middle 90 observations from each resample. Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap t confidence interval or a bootstrap percentile confidence interval for the 5

4. Calculate a bootstrap t confidence interval and bootstrap percentile confidence interval for each of the preceding parts of this exercise if the bootstrap distribution indicates that it is appropriate.

E21. **Multiple Comparisons**
Data set: 'NLBB Salaries'.

1. Conduct a permutation test to determine if there is a difference in mean salaries between pitchers and first basemen. Report the p-value and your conclusions based on an individual α-level of 0.05.

2. Conduct a permutation test to determine if there is a difference in mean salaries between pitchers and catchers. Report the p-value and your conclusions based on an individual α-level of 0.05.

3. Conduct a permutation test to determine if there is a difference in mean salaries between first basemen and catchers. Report the p-value and your conclusions based on an individual a-level of 0.05.

4. If each of the previous three tests uses an a-level of 0.05, what is the true probability that at least one of the tests will inappropriately reject the null hypothesis?

5. What is the individual critical value if you use the Bonferroni method with an overall (familywise) α-level of 0.05? Do any of your previous conclusions in the preceding parts of this exercise change if you test for an overall (familywise) comparison? Explain.

Chapter 2

Making Connections: The Two-Sample t-Test, Regression, and ANOVA

In theory, there's no difference between theory and practice. In practice, there is.
-Yogi Berra[1]

Statistics courses often teach the two-sample t-test, linear regression, and analysis of variance (ANOVA) as very distinct approaches to analyzing different types of data. However, this chapter makes connections among these three techniques by focusing on the statistical models. Statistical software has made it easy to calculate statistics and p-values. But without understanding the underlying model assumptions, it is easy to draw incorrect conclusions from the sample data. As studies become more complex, models become fundamental to drawing appropriate conclusions. In this chapter, a simple student experiment involving games and several additional studies are used to do the following:

- Compare the underlying statistical models for the two-sample t-test, linear regression, and ANOVA
- Discuss the model assumptions for each of these three tests
- Create and interpret normal probability plots
- Transform data in order to better fit the model assumptions
- Discuss the mathematical details of each hypothesis test and corresponding confidence interval

[1]Yogi Berra was an American League Baseball player and manager. This quote has also been attributed to computer scientist Jan L. A. van de Snepscheut.

2.1 Investigation: Do Distracting Colors Influence the Time to Complete a Game?

In 1935, John Stroop published a paper presenting his research on the reaction time of undergraduate students identifying ink colors.2 He found that students took a longer time identifying ink colors when the ink was used to spell a different color. For example, if the word "yellow" was printed in blue ink, students took longer to identify the blue ink because they automatically read the word "yellow." Even though students were told only to identify the ink color, the automatized behavior of reading interfered with the task and slowed their reaction time.[2] *Automatized behaviors* are behaviors that can be done automatically without carefully thinking through each step in the process. Stroop's work, demonstrating that automatized behaviors can act as a distracter for other desired behaviors, is so well known that the effect is often called the *Stroop effect*.

Several students in an introductory statistics class wanted to develop a final project that would test the impact of distracters. They decided to conduct a study to determine if students at their college would perform differently when a distracting color was incorporated into a computerized game. This game challenges people to place an assortment of shaped pegs into the appropriate spaces as quickly as possible. Before any data were collected, these students developed a clear set of procedures.

- 40 students would be randomly selected from the college.[3]
- 20 students would be assigned to the standard game and 20 would be assigned to a game with a color distracter. The student researchers would flip a coin to randomly assign subjects to a treatment. Once 20 subjects had been assigned to either group, the rest would automatically be assigned to play the other game.
- Subjects would see a picture of the game and have the rules clearly explained to them before they played the game. An example of both games is shown in Figure 2.1.
- Subjects would play the game in the same area with similar background noise to control for other possible distractions.
- The response variable would be the time in seconds from when the participant pressed the "start game" button to when he or she won the game.

  NOTE It is important to recognize that each subject in this study was assigned to exactly one treatment, either the standard game or the color distracter game. Some researchers may point out that a

[2]Note that many psychologists would call this procedural knowledge instead of automatized behavior. Both are processes that can be done without conscious thought, but automatized behaviors are processes that cannot be slowed down, do not decline with age, and show no gender differences.

[3]Since it was not possible to force college students to be involved in this study, these researchers randomly selected students from an online college directory until they had 40 students who were willing to play the game.

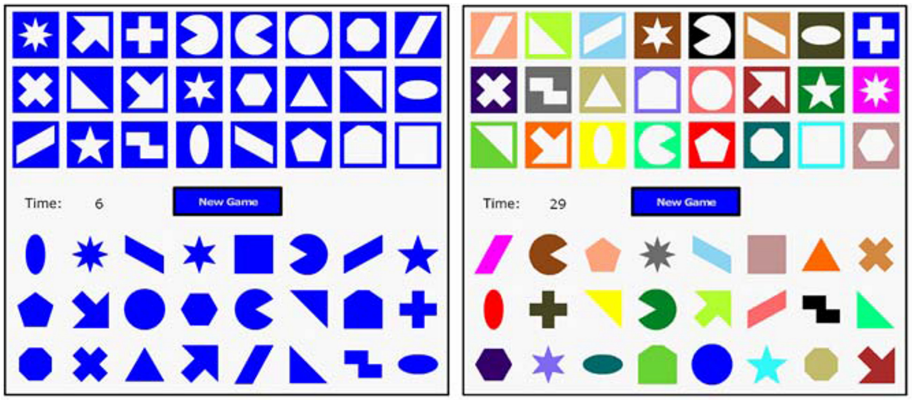

Figure 2.1: An image of the electronic Shapesplosion game with and without color distracters. The instructions for the game were to click and drag each peg to the space with the matching shape.
(#fig:fig2.1)

paired design (where each subject was assigned to both treatments) might have been more efficient. However, for the purposes of this chapter, this study will be treated as the students originally designed it: a study comparing two independent samples.

2.1.1 Understanding the Study Design

1. For this study, identify the units, the population for which conclusions can be drawn, the explanatory variable, and the response variable.
2. Is this study an experiment or an observational study? Explain.
3. The researchers hoped to determine if distracting colors influenced college students' response times when playing a computerized game. Write out in words and symbols appropriate null and alternative hypotheses. Let μ_1 represent the true mean response time of the color group and μ_2 the true mean response time of the standard group. Use a two-sided alternative hypothesis for this question.
4. Create an individual value plot or a boxplot of the Games1 data from this study. Describe the graph. For example, does it look as if the groups have equal means or equal standard deviations? Are there any unusual observations in the data set? Calculate the mean and standard deviation of the color distracter responses, $\bar{y}_1$ and s_1 , as well as the mean and standard deviation of the standard game responses, $\bar{y}_2$ and s_2.

observed		mean		error			
value		response		term			
(random)		(not random)		(random)			
↓		↓		↓			
$y_{1,j}$	=	μ_1	+	$\epsilon_{1,j}$			for $j = 1, 2, ... ,n_1$

2.2 The Two-Sample t-Test to Compare Population Means

2.2.1 The Statistical Model

Generally, **statistical models** have the following form:

observed value = mean response + random error

The statistical model describes each observed value in a data set as the sum of a mean response for some subgroup of interest (often called a group mean) and a random error term. The mean response is fixed for each group, while the random error term is used to model the uncertainty of each individual outcome. The random error term for each individual outcome cannot be predicted, but in the long run there is a regular pattern that can be modeled with a distribution (such as the normal distribution).

The key question in this study is whether or not the two types of games have different average completion times. The two-sample t-test starts with the assumption that the two group means are equal. This is often written as the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ or, equivalently, $H_0 : \mu_1 = \mu_2$.

The underlying model used in the two-sample t-test is designed to account for these two group means (μ_1 and μ_2) and random error. The statistical model for the first population, the color distracter group, is:

where j is used to represent each observation in the sample from the first population. For example, $y_{1,9}$ represents the 9th observation in the first group (the color distracter group). In this data set, there were 20 observations taken from the first population; thus, $n_1 = 20$.

This model states that the color distracter game is expected to be centered at the constant value μ_1 . In addition, each observation is expected to have some variability (random error) that is typically modeled by a normal distribution with a mean equal to zero and a fixed variance s2. Similarly, each observation from the second group, the standard game, can be modeled as the sum of μ_2 plus a random error term, $\epsilon_{2,j}$:

$$y_{2,j} = \mu_2 + \epsilon_{2,j} \quad \text{for} \quad j = 1, 2, ..., n_2 \tag{2.1}$$

where $n_2 = 20$, μ_2 is the mean of the standard group, and the $\epsilon_{2,j}$ are random variables (typically from a normal distribution) with a mean equal to zero and variance σ^2 . Often, this statistical model is more succinctly written as:

$$y_{i,j} = \mu_i + \epsilon_{i,j} \quad \text{for} \quad j = 1, 2 \text{ and } j = 1, 2, ..., n_2 \quad \text{where} \quad \epsilon_{i,j} \sim N(0, \sigma^2) \quad (2.1)$$

> **MATHMATICAL NOTE** You may recall from your introductory statistics course that adding a constant to each random variable in a population does not change the shape or spread of the population. Since each mean response (μ_i) is fixed (i.e., a constant value), Equation **??** can be used to show that $y_{i,j} \sim N(\mu_i, \sigma^2)$.

This model has one assumption that you may not have made when previously conducting a two-sample t-test. Equation **??** states that all $\epsilon_{i,j}$ come from a normally distributed population with a mean of zero and variance s2 . This is called the equal variance assumption. Some introductory statistics courses discuss only a two-sample t-test that does not require the equal variance assumption. The equal variance assumption is made here because it makes sense for this experiment, the data support it (s_1 is close to s_2), and it allows a direct comparison to ANOVA and regression models.

In Equation **??**, the mean response of the model is the population mean (μ_1 or μ_2). Just as a sample mean, $\bar{y}_i$, is used to estimate the population means, μ_i, residuals are used to estimate the random error terms. **Residuals** are the difference between the observed response and the estimated mean response. For example, the random error term $\epsilon_{1,12} = +\bar{y}_{1,12} - \mu_1$ is estimated by $\hat{\epsilon}_{1,12} = +\bar{y}_{1,12} - \bar{y}_1$.

> **NOTE** A **statistic** is any mathematical function of the sample data. **Parameters** are actual population values that cannot be known unless the entire population is sampled. The mean response is based on population parameters. If a sample data set is used, we do not know the population parameters. Sample statistics (such as the sample mean, $\bar{y}$, and the sample standard deviation, s) are used to estimate population parameters (μ and σ). Statisticians often use a hat on top of a parameter to represent an estimate of that parameter. For example, an estimate of the population standard deviation is written $s = \hat{\sigma}$, and an estimate for a mean is written $\bar{y}_1 = \hat{\mu}_1$ or $\bar{y}_2 = \hat{\mu}_2$.

2.2.2 Statistical Models for the Two-Sample t-Test

5. Assume that we have two very small populations that can be written as $y_{1,1} = 15$, $y_{1,2} = 17$, $y_{1,3} = 16$, $y_{2,1} = 11$, $y_{2,2} = 9$, $y_{2,3} = 10$. Find μ_1, μ_2, $\epsilon_{1,1}$, $\epsilon_{1,3}$, and $\epsilon_{2,1}$.

Notice the double subscripts on the observed responses: $y_{1,1}$ is read as "y one one." The first subscript tells us that the observation was from the first group, and the second subscript tells us the observation number. For example, $y_{1,j}$ is the jth observation from the first group.

6. Use the game study and the data in the file Games1 to identify n_1, n_2, $y_{1,12}$, $y_{2,12}$, $\epsilon_{1,12}$, and $\epsilon_{2,12}$, where $y_{1,12}$ represents the 12th observation from group 1 (the color distracter group). Note that since this is a sample, not a population, we do not know μ_1 or μ_2 , but we can estimate them with $\bar{y}_1 = \hat{\mu}_1$ and $\bar{y}_2 = \hat{\mu}_2$.

2.2.3 Model Assumptions for the Two-Sample t-Test

Several implicit assumptions are built into the model for the two-sample t-test shown in Equation **??**:

- Constant parameters: The population values in this model (μ_1 , μ_2 , and σ) do not change throughout the study.
- Additive terms: The model described in Equation **??** shows that the observed responses are the sum of our parameters and error terms. For example, we are not considering models such as $y_{i,j} = \mu_i * \epsilon_{i,j}$.
- $\epsilon_{i,j} \sim N(0, \sigma^2)$. This assumption has many key components:
- The error terms are independent and identically distributed (iid).
- The error terms follow a normal probability distribution.
- The error terms have a mean of zero. This implies that the average of several observed values will tend to be close to the true mean. In essence, there is no systematic bias in the error terms.
- The population variance σ^2 is the same for both groups (color distracter and standard games) being tested.

The first assumption tells us about the mean response. The parameter estimate ($\bar{y}_i$) would not be meaningful if the true parameter value (μ_i) were not constant throughout the study. The second assumption simply states the types of models we are building. In later chapters with more complex models, we will discuss how to use residual plots to determine if the model is appropriate. In this chapter, we will focus on the assumptions about the error terms

> **MATHMATICAL NOTE** In later chapters, we will show that a curved pattern in a residual versus fit plot suggests that an additive model may not be appropriate. In this example, there are only two fitted values (i.e., expected values), so we cannot see any curved patterns. When the additive assumption is violated, residual plots may also indicate different standard deviations, a nonnormal distribution, or lack of independence. Transforming the data to a new scale can often make the additivity assumption (and several of the other assumptions) more appropriate.

The statistical model described in Equation **??** assumes that $\epsilon_{i,j}$ are modeled as

independent and identically distributed (iid) random variables. The independent error term assumption states that there is no relationship between one observation and the next. For example, knowing that the 8th subject in a group played the game more quickly than average does not provide any information about whether the 7th or 9th person in the group will be above or below the average.

The identically distributed assumption states that each error is assumed to come from the same population distribution. Thus, each subject from a particular group is from the same population. If any error term based on a particular observation comes from a different population, the two-sample t-test will not be valid. For example, elementary school students may have different expected completion times for the Shapesplosion game than college students. It would be inappropriate to include younger students in a study where the population was assumed to be college students

Model assumptions for the residuals should always be checked with plots of the data. The extended activities will describe normality tests in more detail, but in most situations a simple graph of the residuals will suffice. The two sample t-test actually requires only that the sample means (each $\bar{y}_{i,j}$) be normally distributed. The central limit theorem allows us to assume this is true if group sample sizes are similar and large ($n_1 \geq 15$ and $n_2 \geq 15$) and there does not appear to be any extreme skewness or outliers in the residuals.

Since residuals are defined as the difference between each observed value and the corresponding group mean, they should always sum to zero. Thus, we cannot check residuals to determine whether each of the error terms is centered at zero. The assumption that the error terms are centered at zero is really stating that there are no other sources of variability that may be biasing our results. In essence, the only difference between the two population means is explained by the mean response.

To check the assumption that the two populations have the same variance, an informal test can be used. If the ratio of the sample standard deviations is less than 2, we can proceed with the analysis.[4]

Informal Test for Equal Variances

$$\text{if} \quad \frac{\max(s_1, s_2)}{\min(s_1, 2)} < 2 \quad \text{or, equivalently, if} \quad \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} < 4$$

then we do not have enough evidence to conclude that the population variances are different.

[4]Some texts suggest rejecting the equal variance assumption when the ratio is greater than 3 instead of 2. If the ratio is close to 2 (or 3), many statisticians would suggest conducting a more formal F-test for equal variances.

Several key observations should be made about the individual value plot shown in Figure 2.2:

- The mean completion time is higher for the color distracter group than for the standard group.
- Neither group appears to have clear outliers, skewness, or large gaps.
- The spread (variance) of the two groups appears to be similar.

Key Concept Every statistical hypothesis test has basic underlying conditions that need to be checked before any valid conclusions can be drawn.

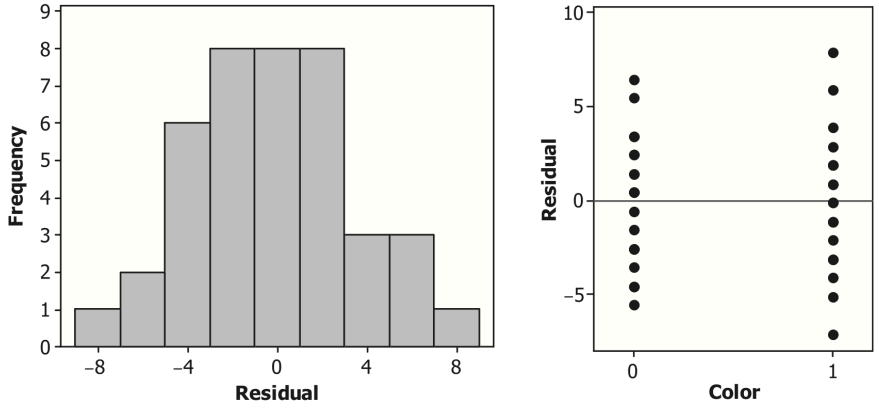

Figure 2.2: Individual value plot of the data from the color distracter and standard games.
(#fig:fig2.2)

Checking Assumptions for the t-Test

7. Calculate the residuals in the Games1 data. Plot a histogram of the residuals (or create a normal probability plot of the residuals). Do the residuals appear to be somewhat normally distributed?
8. Use the informal test to determine if the equal variance assumption is appropriate for this study.
9. The variable StudentID represents the order in which the games were played. Plot the residuals versus the order of the data to determine if any patterns exist that may indicate that the observations are not independent.
10. Use statistical software to conduct a two-sample t-test (assuming equal variances) and find the p-value corresponding to this statistic. In addition, use software to calculate a 95% confidence interval for the difference between the two means (μ_1 - μ_2). Equation **??** and the extended activities provide details on conducting these calculations by hand. If H0 : μ_1

$= \mu_2$ is true, the p**-value** states how likely it is that random chance alone would create a difference between two sample means $(\bar{y}_1 - \bar{y}_2)$ at least as large as the one observed. Based on the p, what can you conclude about these two types of games?

2.3 The Regression Model to Compare Population Means

2.3.1 The Linear Regression Model

The simple linear regression model discussed in introductory statistics courses typically has the following form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, 2, ..., n \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \tag{2.2}$$

A **simple linear regression** model is a straight-line regression model with a single explanatory variable and a single response variable. For this linear regression model, the mean response $(\beta_0 + \beta_1 x_i)$ is a function of two parameters, β_0 and β_1, and an explanatory variable, x. The random error terms, ϵ_i, are assumed to be independent and to follow a normal distribution with mean zero and variance σ^2.

In Equation **??**, we used double subscripts: $i = 1, 2$ was used to show that there were two distinct groups and $j = 1, 2, ..., n_i$ was used to identify each of the $n_1 = n_2 = 20$ items within the two groups. In the regression model, there is only one set of subscripts: $i = 1, 2, ..., n$, where $n = 40 = n_1 + n_2$. Instead of having two distinct means in the model (μ_1 and μ_2), as in the two-sample t-test, we have one regression model where the parameters, β_0 and β_1, are fixed. The categorical explanatory variable, x, indicates game type.

A procedure commonly used to incorporate categorical explanatory variables, such as the game type, into a regression model is to define **indicator variables**, also called **dummy variables**, that will take on the role of the x variable in the model. Creating dummy variables is a process of mapping the column of categorical data into 0 and 1 data. For example, the indicator variable will have the value 1 for every observation from the color distracter game and 0 for every observation from the standard game. Most statistical software packages have a command for automatically creating dummy variables.

> **NOTE** Typically an indicator variable is created for each category. Thus, there would be an indicator variable called Color equal to 1 for the color distracter game and 0 otherwise and another indicator variable called Standard equal to 1 for the standard game and 0 for all other categories. Notice that there is complete redundancy between the two indicator variables: Knowing the value of the Color variable automatically tells us the value of the Standard variable for each subject. Thus, only one of the indicator variables is needed in this model. Although this study has only two categories of games (color and standard), it is common for a categorical explanatory variable to have more than two categories. Chapter 3 provides the opportunity to use indicator variables when there are multiple categories.