Chapter 6:

Comparing Two Groups

This chapter continues testing simple hypotheses concerning the population mean. In Chapter 5, we introduced tests and confidence intervals covering means of a *single* population. In this chapter, we do the same, but for the *difference in means* between two populations. Along with Normality, the usual assumption is that the measurements in each of the two groups are independent of each other.

❧❧❧

The mayor of İstanbul decided that the average response times for fire stations in the city were too long. A Representative suggested a new GPS system. Unfortunately, the GPS system cost several million Lira. Thus, the mayor needs to know that it will work well enough to pay for it out of tax income.

To fully test the effect of the GPS system, the mayor randomly divided the 39 districts into two groups. He rented the GPS system and had the fire stations in the first group of 15 districts use it. The fire stations in the remaining 24 districts continued using the old method.

6.1: Two independent samples; equal variance

Let us assume that you have *two* categories of individuals. For each individual, you measure a specific characteristic and the group membership. This can be as simple as measuring the height of several people to determine if men or women are taller, or it can be as complex as measuring a latent variable based on a variety of different measures on two different populations of individuals. The key is that you have a single measurement, a group membership (male or female), and you want to compare the averages of the two populations.

To solve this, we can make a few assumptions: The heights are independent. The heights are distributed Normally *in each category*. The height variance is the same in both populations. That is, we are assuming the two populations differ *only in their means*.[1] In symbols, we assume:

two populations

$$X_i \sim \mathcal{N}\left(\mu_x, \sigma^2\right) \qquad \text{and} \qquad Y_i \sim \mathcal{N}\left(\mu_y, \sigma^2\right)$$

Creating the test statistic follows the natural course. Since we want to test for a difference in population means, we should use a test statistic based on the difference in sample means. That leaves the question of the distribution. From our assumption, we know

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma^2}{n_x}\right) \qquad \text{and} \qquad \overline{Y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma^2}{n_y}\right)$$

Thus, the distribution of the difference in means is

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\right)$$

Now, we have a test statistic and its distribution. If we know σ^2, then the test statistic is

population variance

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}}}$$

which, as expected, has a standard Normal distribution.

[1] Again, these assumptions must be tested.

unrealistic

However, as before, knowing the population variance without knowing the population means is unrealistic. It is most likely that we will need to estimate the population variance with the sample variance, as we did in Section 5.3. This estimation produces the test statistic

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}},$$

which is commonly written as

$$t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \tag{6.1}$$

pooled variance

Here, s_p^2 is the estimated variance of the entire population, called the *pooled* variance because you are pooling both samples together.

Notice that this test statistic also has the basic form of all of the parametric test statistics in the previous chapter: a difference divided by its standard error. Here, since we are assuming that the populations have a common variance, we are using a weighted average of the two sample variances, called the pooled variance:

$$s_p^2 := \frac{(n_x - 1)s_x^2 + (n_Y - 1)s_y^2}{n_x + n_y - 2} \tag{6.2}$$

To test the statistic, you need to know its degrees of freedom. In the previous section, it was $n - 1$ for one group. Here, since there are two populations, it is $n - 2$, which can be written as $(n_x - 1) + (n_y - 1)$.

research hypothesis

Example 6.1: You decide to test the hypothesis that men and women are the same height (on average). To do this, you measure the heights of 10 men and 15 women. The measured heights for the men were 68, 71, 69, 70, 73, 72, 70, 67, 72, and 68 inches; for women, 63, 65, 65, 62, 68, 62, 63, 68, 65, 64, 65, 65, 70, 65, and 65 inches.

equal variances

In this sample, the men had an average height of 70in, with a variance of 4 in^2. The women had an average height of 65in, with a variance of 5 in^2. Assuming the population height variances are equal, does the data support the hypothesis at the $\alpha = 0.05$ level?

Solution: First, let us clearly state the null and alternative hypotheses:

$$\begin{aligned} H_0: \quad & M \sim \mathcal{N}(\mu, \sigma^2), \text{ and} \\ & F \sim \mathcal{N}(\mu, \sigma^2) \\ H_A: \quad & M \sim \mathcal{N}(\mu_M, \sigma^2), \text{ and} \\ & F \sim \mathcal{N}(\mu_F, \sigma^2), \text{ with } \mu_M \neq \mu_F. \end{aligned}$$

Next, we are given the necessary information. Let us substitute it into our formulas (Eqns 6.1 and 6.2). First, the pooled variance:

$$\begin{aligned} s_p^2 &:= \frac{(n_M - 1)s_M^2 + (n_F - 1)s_F^2}{n_M + n_F - 2} \\ &= \frac{(9)4 + (14)5}{23} \\ &\approx 4.6087, \text{ and} \\ s_p &= 2.14679 \end{aligned}$$

Second, the test statistic

$$\begin{aligned} t &= \frac{\bar{m}_1 - \bar{f}_2}{s_p \sqrt{\frac{1}{n_M} + \frac{1}{n_F}}} \\ &= \frac{70 - 65}{2.14679 \sqrt{\frac{1}{10} + \frac{1}{15}}} \\ &= \frac{5}{0.87642} \\ &\approx 5.70501 \end{aligned}$$

The number of degrees of freedom are $\nu = 10 + 15 - 2 = 23$. Thus, from the tables, the critical value is 2.0687. As our test statistic $t = 5.70501 > 2.0687 = \text{cv}$, we can reject the null hypothesis that men and women are the same height, on average. Figure 6.1 indicates the same conclusion. ◇

test statistic

Since we have a computer, we can actually go a bit farther. We can calculate the p-value. Recall that the definition of the p-value is the probability

p-value

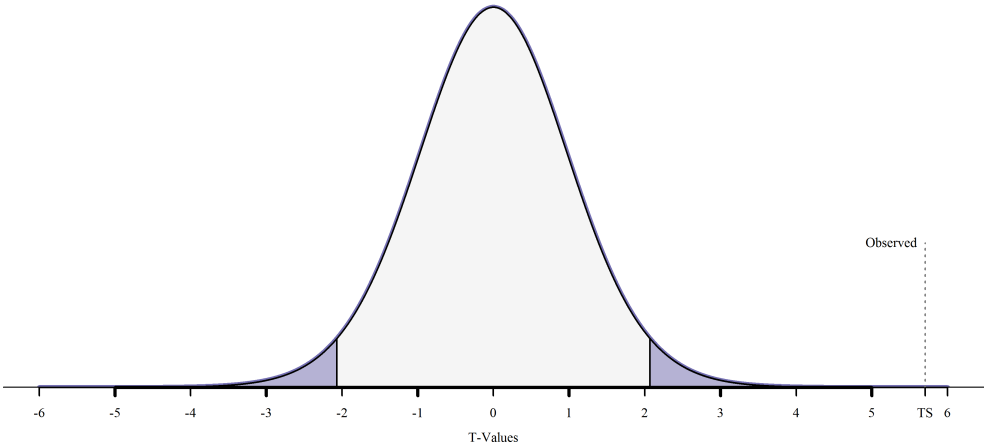

Figure 6.1: *Plot of the null distribution in Example 6.1. Note that the test statistic (TS) is located in the rejection region. As such, we reject the null hypothesis as being sufficiently unlikely and conclude that the genders do not have the same heights.*

of observing data this extreme or more so, **given the null hypothesis is true**. In R, the code to calculate the p-value in this example is

```
2*pt(5.705, df=23, lower.tail=FALSE)
```
(6.3)

This gives a p-value of approximately $8.263 \times 10^{-6} = 0.000008263$, which is tiny compared to our usual $\alpha = 0.05$. In fact, even if we had chosen $\alpha = 0.0001$, we could still safely reject the null hypothesis that men and women have the same average height. And, we did this based on just 25 data points.

two-tailed test

CDF

In Code Snippet 6.3, we multiply the p-value by 2 because this is a two-tailed test; that is, the null hypothesis only concerns equality. Also, the R function, `pt()` returns the cumulative probability function of the t distribution; that is, it returns $\mathbb{P}[T \leq t]$. In other words,

$$\texttt{pt(x)} = \mathbb{P}[T \leq x]$$

This function is easier to remember if you remember that the 'p' stands for cumulative '*p*robability'.

In lieu of dealing with bare probability functions, we can use the full power of our statistical environment to do the calculations for us. In general, there are three steps to analysis: import the data, test the assumptions, and test the hypothesis.

Here are the three steps in R for this example's data

```
male   = c(68,71,69,70,73, 72,70,67,72,68)
female = c(63,65,65,62,68, 62,63,68,65,64, 65,65,70,65,65)

shapiro.test(male)
shapiro.test(female)

t.test(male,female, var.equal=TRUE)
```

The first two lines import the data into R, storing the heights of these males into the `male` variable and the heights of these females into the `female` variable. The second two lines test the assumption that the measurements are Normally distributed in each group. The final line performs the t-test. Note that the results of these lines are the same as for the discussion above:

```
        Two Sample t-test

data:  male and female
t = 5.705, df = 23, p-value = 8.263e-06
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 3.186982 6.813018
sample estimates:
mean of x mean of y
       70        65
```

In addition to performing the test, we also have a 95% confidence interval for the difference in the population averages: males are from 3.19 to 6.81 inches taller than women, on average. How does the output tell us this? Since `male` precedes `female` in the data line, we know the confidence interval is for `male` – `female`.

confidence interval

6.2: Two independent samples; unequal variance

unequal variances

Now, you may be asking, what happens if I am not sure that the variances in the two populations are equal? As the previous formula was based on the assumption that the variances were equal, relaxing that requirement changes the formula. Actually, the formula is much simpler and interpretable:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \tag{6.4}$$

Notice that this formula also has the standard form of a t-test. The difference is in the denominator.

Before we heave a sigh of relief and ask why we don't always use Formula 6.4, referred to as Welch's t-test (1947), we have to concern ourselves with the degrees of freedom for the t-statistic. Here is where the complexity arises. The current best approximate solution, the Welch-Satterthwaite equation (Satterwaithe 1946; Welch 1947), is to approximate the degrees of freedom with

Welch-Satterthwaite

$$\nu = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(s_x^2/n_x\right)^2}{n_x - 1} + \frac{\left(s_y^2/n_y\right)^2}{n_y - 1}} \tag{6.5}$$

Thankfully, you only have to tell the computer to use this form; you do not have to calculate it yourself. However, you have to *know* to tell the computer.

p-value

Many statistical packages (including R, SPSS™and SAS™) provide the p-values under the assumption that the variances are equal *and* without that assumption. They will also give you p-values on the null hypothesis that the variances are equal. However, it is actually much simpler than that.

If you use the t-test enough times, you will notice that when the population variances are equal, the t-test that assumes equal variances gives the same answer as the t-test that does not, on average. When the population variances are not equal, the unequal-variance t-test is the appropriate test. In other words, you should always use the unequal-variance t-test (Formula 6.5) unless you are *absolutely* sure the variances are equal.

certainty

Warning: *Well, I should place a warning here. Remember, there are* assumptions *underlying the t-test. The most important is that the measurements in the two populations are Normally distributed. If that is not true (or close to being true), then you should not use either t-test unless the sample size is 'large enough.' This is a rather important assumption, as the next section demonstrates.*

Furthermore, there is a very *slight gain in power when using the equal variance test as opposed to the unequal variance test. But, the gain is slight and disappears if the population variances are not equal.*

power

Example 6.2: Let us return to Example 6.1. In that example, we assumed the population variances were the same. Let us perform the same test, without making this assumption.

Solution: The only change is in the fifth line. Here are the three steps in R:

```
male   = c(68,71,69,70,73, 72,70,67,72,68)
female = c(63,65,65,62,68, 62,63,68,65,64, 65,65,70,65,65)

shapiro.test(male)
shapiro.test(female)

t.test(male,female)
```

default

The default for the t-test function is to not make the assumption of equal variances. The substantive results are the same as in Example 6.1. There is significant evidence that heights differ between the genders ($t = 5.84; \nu = 20.91; p \ll 0.0001$). In fact, we are 95% confident that men are between 3.22 and 6.78 inches taller than women, on average. Figure 6.2 illustrates this. Here is the t-test output from R:

confidence interval

```
        Welch Two Sample t-test

data:  male and female
t = 5.8387, df = 20.914, p-value = 8.667e-06
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 3.218677 6.781323
sample estimates:
mean of x mean of y
       70        65
```

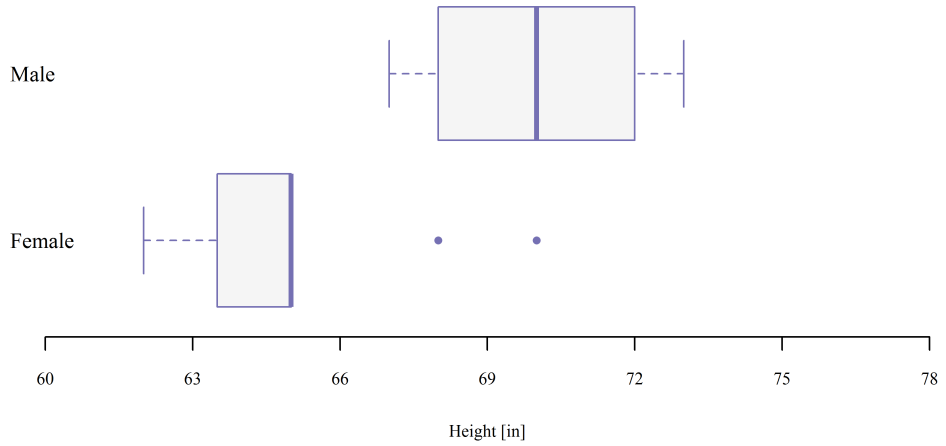

Figure 6.2: *A box-and-whiskers plot of the two groups of heights. Note that the male heights appear to be significantly higher than female heights, which is what the t-test indicated.*

Note that the degrees of freedom is no longer an integer. This is due to the Welch-Satterthwaite approximation of Formula 6.5. ◇

6.3: Testing the Assumption

two-sample

As with the one-sample t-test, the assumption of the two-sample t-test has to do with Normality. The assumption is that the measurements *in each group* came from a Normal distribution. Testing the assumption requires partitioning the measurements into the two groups and performing a Normality test on each subset. This we did in Example 6.2 (although the measurements were already partitioned).

Example 6.3: Using the `football` datafile, let us determine if the Big 12 scores more points on average than the SEC. This data comes from all regular season games in 2009.

Solution: The dependent variable is the number of points scored in each game (`score`). The grouping variable (independent variable) is the team's

grouping variable

conference.

As we wish to draw inferences on two population averages, we would like to use the two-sample t-test. We would like this as it has higher power than any non-parametric test — assuming its assumptions are met.

parametric test

The assumption of the two-sample t-test is that the measurements in each group came from a Normal distribution. Before we can test this, we need to partition the scores by conference. In R, I run

Normality

```
b12 = score[conference=="Big 12"]
sec = score[conference=="SEC"]
```

Now, I have two variables of scores. The first variable, `b12`, contains all game scores of the Big 12 teams. The second variable, `sec`, contains all game scores of the SEC teams.

partition

With this, we merely perform a Shipiro-Wilk test on each of the two variables: `shapiro.test(b12)` and `shapiro.test(sec)`. These tests indicate that this data violates the assumption of the t-test ($p_b = 0.01846$; $p_s = 0.03925$). As such, we will need to use some other test (Section 6.4). ◇

Note: The code to partition the scores by conference is good to know as it is used frequently. Note the double equals inside the bracket. These indicate you are *testing* the conference variable to determine which have the value "Big 12". This is where the partitioning is happening.

testing equality

This is covered again in Example 6.5, below.

6.4: Non-Parametric Means Tests I

The tests of means (thus far) have all assumed that the underlying populations were distributed Normally. This assumption is never true, and the Central Limit Theorem does not save us unless the sample sizes are quite large or the distributions are approximately Normal. So, what do we do if the sample sizes are small and the samples are not sufficiently Normal? In those cases, we can use non-parametric methods.

Normality

Non-parametric tests *do* make assumptions about the underlying distribution, but those assumptions do not require a *specific* distribution. When comparing two samples, the Mann-Whitney test (also known as the two-

non-parametric tests

sample Wilcoxon test) requires that the two samples differ *only* in their measure of center.

6.4.1 Two-sample Mann-Whitney test

Just as there is a two-sample t-test used to compare means of two populations whose measures are Normally distributed, there is a non-paramteric alternative when the measurements are not Normally distributed. It is called the two-sample Wilcoxon test (or the Mann-Whitney test). The assumption of the Mann-Whitney test is that the two samples are identically distributed, except for the center.

Example 6.4: Do democratic States have a higher external debt than autocratic States? My friend asserts that autocratic States have a lower external debt than democratic States. Thus, his stated hypothesis is

research hypothesis

$$H_A : \mu_a < \mu_d$$

Notice that this is the *alternative* hypothesis. The null hypothesis *always* includes the "no effect" position. As my friend stated autocracies have a *lower* external debt, he is stating the alternative, $\mu_a < \mu_d$. Had he said "Autocracies do not have a greater external debt," his statement would be $\mu_a \leq \mu_d$, which includes the null position ($\mu_a = \mu_d$), and would be the null.

To test his hypothesis (actually, to test the null hypothesis), he selected a random sample world States and measured their external debts (Table 6.1). Assuming his sample is representative, does reality support his assertion?

Solution: The steps for the Mann-Whitney test are quite similar to those of the one-sample Wilcoxon test (see Section 5.5.1). Before we begin, as always, determine the null hypothesis, the hypothesis that includes the "no effect" position.[2]

null hypothesis

The first step of the Wilcoxon test is to rank all of the values, from either largest to smallest or smallest to largest. Once they are ranked, you add up the ranks of either group. For this example, let us add the ranks of

[2] This comment is actually *very* important. Except when we are making power calculations, we only test the null hypothesis. This is because the null contains all of the information about the distribution we are using in our test. This is why I have often written the null hypothesis in distributional form.

Do not ever forget that these tests are based on the distribution assumed, not necessarily on the stated null hypothesis. The key is to match your statement with the distributional hypothesis.

State	External Debt	Rank	Type
Australia	920	11	D
Brazil	216	7	D
China	347	9	A
Iraq	50	1	A
Kazakhstan	93	4	A
Norway	548	10	D
Pakistan	52	2	A
Saudi Arabia	72	3	A
South Korea	334	8	D
Ukraine	104	5	A
United Arab Emirates	129	6	A

Table 6.1: *External debt (x_i), in billions for selected States. Data from the CIA(2009). This table accompanies Example 6.4. In this data, m, the number of democratic states under consideration, is 4 and n, the number of autocratic states under consideration, is 7.*

the autocratic States (it is easier to add smaller integers). With that, our test statistic is $W = 30$ (using this method), and our sample sizes are $m = 4$ and $n = 11$. Looking at the Mann-Whitney table, we find our p-value is $p = 0.01$. As this is a one-sided test, we do not need to double that p-value. As $p = 0.01 < \alpha = 0.05$, we can reject the null hypothesis and conclude that the data support my friend's contention that democratic States have a higher average external debt than autocratic States. ◇

Again, in R, determining the p-value is very straight-forward. The applicable function is the same as for the one-sample test. You just pass it two samples instead of one. Thus, for this example, you would use: **p-value**

```
a = c(347,50,93,52, 72,104,129)
d = c(920,216,54,334)
wilcox.test(a,d, alternative="less")
```

The output gives the test statistic ($W = 2$) and the p-value ($p = 0.01212$).

Note: There is a little disagreement in the literature (and in the statistical software) as to what should be the test statistic. Some assert that the larger of the sum of the ranks should be the test statistic. Others assert that it should be the smaller of the sum. Others do the calculations on the

difference in the sums. In the end, it really does not matter. The different programs use test distributions adjusted for their specific test statistics. Thus, the p-values will be the same across software.

Example 6.5: The `football` dataset contains the points scored in all of the 2009 games played by Big 12 and SEC teams. My friend is a big fan of the SEC. Of course, I am a OSU Cowboys fan. My friend stated that the average number of points scored by the SEC is greater than that of the Big 12 (in 2009). Is my friend correct?

Go Pokes!

Note that the data are from 2009 and my friend would like to generalize this sample to all SEC and Big 12 football games (the population).

Solution: First, we must import the `football` dataset into `R`, giving it a useful name:

```
fb <- read.csv("football.csv")
```

This command imports the file named `football.csv` located in the *current working directory* and stores it in the variable `fb`. Once the data are downloaded, we attach the dataset with `attach(fb)`.[3] Now, we can access the variables in this dataset more easily.

There are two groups being discussed: Big 12 football schools and SEC football schools. We need to compare the game scores of these two groups. We would prefer to use the two-sample t-test as it is more powerful than the Mann-Whitney test.

The assumption of the two-sample t-test is that the measurements in each sample (group) are Normally distributed. To test this assumption, we first need to separate the scores into the two groups. In this example, the measurement variable is `score` and the grouping variable is `conference`. Thus, we can create the vector of Big 12 scores by

```
b12 = score[conference=="Big 12"]
```

Similarly, we create the vector of SEC scores by

```
sec = score[conference=="SEC"]
```

With this, we now have the two samples (variables) to compare: `b12` and `sec`.

[3] Please read the note at the end of the chapter regarding the `attach()` function.

Now, testing each group of scores for Normality is straight-forward; we use the Shapiro-Wilk test on each group.

```
shapiro.test(b12)
shapiro.test(sec)
```

Note that *at least one* group violates the assumption of Normality. Thus, we should not use the t-test. We use the Mann-Whitney test instead.

The question asks about whether SEC teams scored more points, on average, than Big 12 teams. While we would have liked to use the two-sample t-test, at least one of the two groups was not Normally distributed. Thus, we will use the Mann-Whitney test. The command to perform the non-parametric test is

```
wilcox.test(sec,b12, alternative="greater")
```

R tells us the value of the W statistic as well as the p-value, which is what we need. The results of the Mann-Whitney test indicate that there is *not* sufficient evidence against the null hypothesis ($W = 10106$; $p = 0.6449$). As such, we are unable to reject the null hypothesis and conclude that the two conferences do not score significantly different numbers of points in their games, on average. ◇

> ***Note***: Some will (and *should*) point out that my friend is relying on the unstated assumption that 2009 is a "typical" year for the SEC and the Big 12. In serious research, this needs to be explored and shown to be a reasonable assumption. If it is not, then we cannot generalize the results from 2009 to other years. **generalize**

Because there were *only two* groups in the dataset, we could have used a shortcut `wilcox.test( score ~ conference )`.

The code `score ~ conference` is a "formula" in R: The dependent variable is to the left of the tilde; the independent variable(s), to the right. If there are more than two groups represented, you cannot use this shortcut. **formula**

Area ID	MFRI	Biome Type
1	1	Xeric Shrubland
2	2	Xeric Shrubland
3	3	Xeric Shrubland
4	4	Xeric Shrubland
5	15	Xeric Shrubland
6	26	Temperate Broadleaf Forest
7	33	Temperate Broadleaf Forest
8	80	Temperate Broadleaf Forest
9	100	Temperate Broadleaf Forest
10	125	Temperate Broadleaf Forest
11	150	Temperate Broadleaf Forest
12	240	Temperate Broadleaf Forest

Table 6.2: *Mean fire return intervals for 12 areas in the United States. The biome type is also provided. This data is for Example 6.6.*

Example 6.6: A biome is an ecological community with a similar climactic condition. There are many ways of grouping biomes, they all focus on different aspects of the ecosystem. I prefer the World Wild Fund for Nature (WWF) classification system, which enumerates 14 different terrestrial biomes.

mfri

The mean fire return interval is the average time between major fires in the ecological community.

research hypothesis

I hypothesize that the mean fire return interval for xeric shrublands is shorter than that for temperate broadleaf forests. To test this hypothesis, I randomly sampled five xeric shrubland areas and seven temperate broadleaf forest areas. I measured the mean fire return interval for each of the 12 areas (Table 6.2).

Solution: The research (stated) hypothesis is

$$H_R : \mu_x < \mu_t$$

Because it does not contain an equality (=, ≤, or ≥), it is our alternative hypothesis. Thus, our null and alternative hypotheses are

$$H_0 : \mu_x \geq \mu_t$$
$$H_A : \mu_x < \mu_t$$

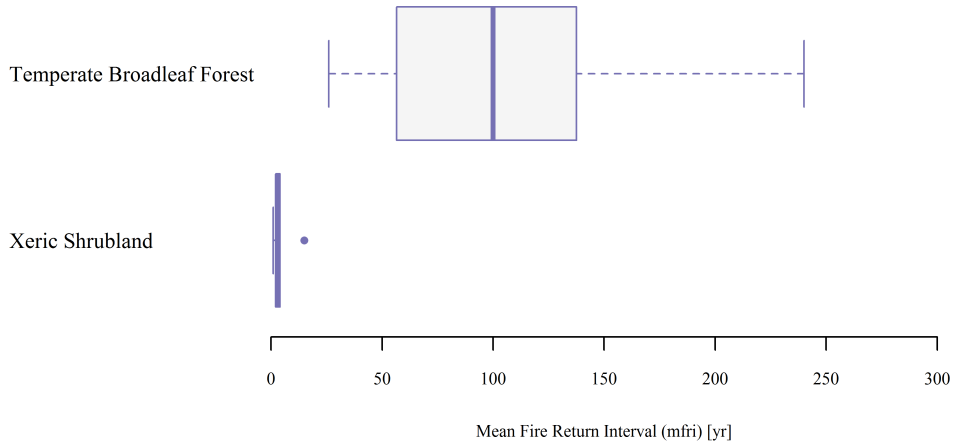

Figure 6.3: *A box-and-whiskers plot of the mean fire return interval for two biomes, xeric shrubland and temperate broadleaf forest. Note the significant difference in the measures of center.*

parametric test

Normality

non-parametric test

With those hypotheses stated, we can continue to selecting the best test. Because it is more powerful, I would prefer to use the two-sample t-test. It assumes that the measurements in each group are Normally distributed. The Shapiro-Wilk test, however, indicates that the mean fire return intervals for the xeric shrubland are not Normally distributed ($p = 0.022 < 0.05 = \alpha$). Thus, we will use the Mann-Whitney test.

According to this test, since $p = 0.0013 < 0.05 = \alpha$, we reject the null hypothesis and conclude that the average mean fire return intervals for xeric shrubland are shorter than those for temperate broadleaf forests ($W = 0; p = 0.0013$). In fact, we are 95% confident that the average mean fire return interval for xeric shrublands is from 25 ti 149 years *shorter* than that of the temperate broadleaf forests. Figure 6.3 illustrates this. ◇

The R code for this analysis is

```
xer = c( 1,2,3,4,15 )
tbf = c( 26,33,80,100,125,150,240 )

# Test Normality
shapiro.test(xer)
shapiro.test(tbf)
```

```
# Test null hypothesis
wilcox.test(xer,tbf, alternative="less")

# Obtain the symmetric 95% confidence interval
wilcox.test(xer,tbf, conf.int=TRUE)
```

assumption

Non-parametric tests do not assume the *specific* distribution of the measures. They do, however, make other assumptions. In order to use the Wilcoxon test, you must assume that the the underlying distribution is symmetric. In order to use the Mann-Whitney test, you must assume the two samples are identically distributed (except for the center). I will leave it as an exercise for you to investigate the effect of different distributions on the applicability of the Wilcoxon-type tests.

6.5: Non-Parametric Means Tests II*

The two-sample t-test requires that the sample means of each group have a Normal distribution. This means that either the measurements come from a Normally-distributed population or that the sample size is large enough for the Central Limit Theorem to be useful (Appendix C). The Mann-Whitney test, a non-parametric test, requires that the two populations differ only in the middle; the distributions are otherwise the same. When the mean and the variance are independent, this assumption may be easily met. When the mean and the variance are dependent, this assumption cannot be met (see, for instance, Sections A.5, B.4, and B.5, among others). In such cases, the mean and the variance are functions of each other. This is a direct violation of the assumptions of the Mann-Whitney test.

permutation test

In such cases, we still have an option, a decidedly less powerful option—the permutation test.

6.5.1 The Permutation Test

The key to the permutation test is to see that, under the null hypothesis, the two samples come from the *same* population. The grouping is artificial in that any other grouping will produce similar results. Permutation tests permute the sample into all possible groupings, measure the test statistic for each of these other groupings, and compare the observed test statistic to this distribution of possible test statistics.

The number of permutations increases exponentially with the sample size. Thus, large samples require too much time to perform full permuta-

tion tests. A slightly weaker alternative is the **randomization** test. This test randomly permutes the combined sample multiple times to obtain the distribution of possible test statistics.

randomization test

Here is the process:

1. Given: two samples, x_1 and x_2, of size n_1 and n_2
2. Calculate $\overline{x}_1 - \overline{x}_2$, which is the test statistic
3. Repeat the following a sufficiently large number of times. This is the loop.
    a) Randomly permute the combined sample into two new samples of size n_1 and n_2
    b) Measure the test statistic of this partitioning
4. Compare the observed test statistic with this distribution of test statistics

Here, I provide the raw code to illustrate the above steps. However, there are packages in `R` that perform permutation (randomization) tests. If you wish, you can skip to that section.

6.5.2 The Code*

Let us see how a randomization test can be created in `R`. Let us start with two samples, `x1` and `x2`:

```
x1 = c(10.76, 15.05, 17.01, 5.07)
x2 = c(19.50, 8.16, 10.38, 6.75, 12.72)
```

Note that the sample sizes are different. This is entirely permissible.

```
n1 = length(x1)
n2 = length(x2)
```

Now, we calculate the observed test statistic.

```
obs = mean(x1) - mean(x2)
```

Now, we do the loop. For ease, let us do the loop 10,000 times. Recall that the purpose of the loop is to approximate the distribution of the test statistic. Thus, we will need to store these *un*observed test statistics in a variable. For want of a better name, let this variable be `TS`.

With this, the opening of the loop will be the lines

```
TS = numeric()
for( i in 1:10000) {
```

Inside the loop, we need to randomly assign n_1 of the values to group 1 and n_2 of the values to group 2. These two lines randomly permute the values into group 1.

```
pmt = sample(n1+n2,n1)
grp1 = c(x1,x2)[pmt]
```

This line puts the rest into group 2

```
grp2 = c(x1,x2)[-pmt]
```

Those three lines constitute Step 3a. Step 3b is to calculate the test statistic for this particular partitioning.

```
TS[i] = mean(grp1) - mean(grp2)
```

This ends the loop, so the next line is

```
}
```

After this loop runs 10,000 times, we can plot the distribution of test statistics

```
hist(TS)
```

We can also calculate a p-value

```
2*min(mean(TS>=obs),mean(TS<=obs))
```

The entire code is given here, for convenience.

```
x1 = c(10.76, 15.05, 17.01, 5.07)
x2 = c(19.50, 8.16, 10.38, 6.75, 12.72)
n1 = length(x1)
n2 = length(x2)

obs = mean(x1) - mean(x2)

TS = numeric()
for( i in 1:10000) {
  pmt = sample(n1+n2,n1)
  grp1 = c(x1,x2)[pmt]
  grp2 = c(x1,x2)[-pmt]
  TS[i] = mean(grp1) - mean(grp2)
}
```

```
hist(TS)

2*min(mean(TS>=obs),mean(TS<=obs))
```

Running this code gives a p-value of 0.8946. As this value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis; we were unable to detect a difference in the means of these two populations.

> ***Note***: As with all non-parametric tests, the permutation test is of low power. In fact, as with all tests based solely on the data, this test has very low power. But, sometimes, this test is all one can use.

6.5.3 The perm Package

The above code is very general and portable. The only lines that need to be changed are the first two, the lines specifying the samples. Everything else can be left alone.

However, there are some advances with this test that are beyond the scope of this text. As such, you should use on of the packages devoted to the permutation (and randomization) test. In R, there are several and include coin, permtest and perm. Here, I show you how to use the perm package.

permTS

For comparing two populations, the function is permTS. The perm represents "permutation test," the TS, "two samples." There are only two required slots, corresponding to the two samples. A third slot allows you to perform one-sided hypothesis tests. A fourth slot allows you to use different varieties of the permutation test.

Thus, `permTS(x1,x2)` will perform a two-sided variety of the above permutation test, while `permTS(x1,x2, alternative="less")` will perform a one-sided variety.

There are four varieties available, one approximation and three exact forms. The approximation uses the permutation Central Limit Theorem. The three exact methods use the network algorithm, Monte Carlo, and a complete enumeration. Thus, the following will perform a permutation tests using each of the above methods:

```
permTS(x1,x2, method="pclt")
permTS(x1,x2, method="exact.network")
permTS(x1,x2, method="exact.mc")
permTS(x1,x2, method="exact.ce")
```

Running these show that the p-values are not all the same. This is due to different methods for estimating the p-value. None of the four is guaranteed to be always better than the others. There will *rarely* be a substantive difference, so it rarely matters.

6.6: Further Examples

To further illustrate some of these processes, this section provides several additional examples.

Example 6.7: The `HeartOfTheValleyTriathalon` dataset contains a sample of the intermediate and the finishing times for participants in the May 26, 2014, Heart of the Valley Triathlon held in Corvallis, Oregon. One of the racers hypothesized that males finished faster, on average, than females. Test her hypothesis.

Solution: Her (research) hypothesis is $\mu_m < \mu_f$. As this does not contain the "equals" position, it is also the alternative hypothesis.

As we are testing the average racing times for two populations, we would like to use the two-sample t-test as it is the most powerful of those available to us. However, it requires that the measurements in each group come from a Normally-distributed population. To test this, let us use the Shapiro-Wilk test. According to this test, both samples pass ($p_m = 0.2414; p_f = 0.2971$). Thus, we can use the two-sample t-test.

According to this test, there is a significant difference in times between the two genders ($p < 0.0001$), with the male times being significantly faster than female times, on average. The following is the code used.

```
tri = read.csv("http://rfs.kvasaheim.com/data/
   HeartOfTheValleyTriathalon.csv")
attach(tri)

maletime   = TOTALTIME[GENDER=="M"]
femaletime = TOTALTIME[GENDER=="F"]

shapiro.test(maletime)
shapiro.test(femaletime)

t.test(maletime,femaletime, alternative="less")
```

The first two lines import the data from the Internet and attach it. The second two lines partition the finishing times into male finishing times and female finishing times. The third two lines perform the Shapiro-Wilk test on the two samples. Finally, the last line performs the two-sample t-test, which tests the null hypothesis against the alternative hypothesis that male times are lower, on average, than female times.

Because the p-value was less than $\alpha = 0.05$, we rejected the null hypothesis in favor of the alternative. We were able to detect a difference in average times between men and women in this triathlon. ◇

Example 6.8: The `crime` dataset contains a sample of a lot of variables, including population of the state in 2000 and the state's census region. An associate of mine hypothesized that the average state population in the South is greater than that in the Midwest. Test this hypothesis.

Solution: The research hypothesis is $\mu_s > \mu_m$. Since this does not contain the equals position, it is also the alternative hypothesis.

As we are comparing an average for two population, I would like to use the two-sample t-test as it is the most powerful of those available to us. However, it requires that the measurements in each group come from a Normally-distributed population. To test this, let us use the Shapiro-Wilk test. According to this test, the Midwest sample passes ($p_m = 0.2113$), but the South sample does not ($p_s = 0.00145$). Thus, we cannot use the two-sample t-test.

We can use the Mann-Whitney test, but that assumes the two populations differ *only in their centers*. Histograms of these two samples suggest that the two may differ in more than just their means (Figure 6.4). However, the graphical evidence is equivocal. Thus, let us perform both the Mann-Whitney test and a randomization test to better understand the differences of the two populations.

According to the Mann-Whitney test, we are unable to detect a difference in the two populations ($p = 0.5087$). The two-sample randomization test (using Monte Carlo) concurs ($p = 0.409$). Thus, we fail to reject the null hypothesis and are unable to detect a difference in the average state populations in these two regions of the United States.

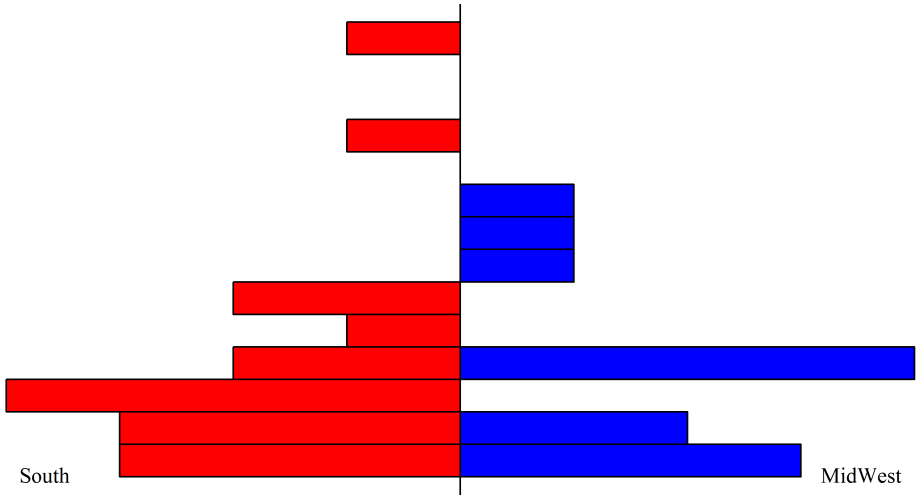

Figure 6.4: *A back-to-back histogram for the populations in each of the two census regions. This is done to gauge whether the two samples come from populations that differ only in their location parameter (center).*

The following is the code used for this analysis. It also includes the code for the back-to-back histogram (Figure 6.4).

```
cr = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(cr)

pops = pop00[census4=="South"]
popm = pop00[census4=="Midwest"]

shapiro.test(pops)
shapiro.test(popm)

wilcox.test(pops, popm, alternative="greater")
permTS(pops, popm, alternative="greater", method="exact.mc"
    )
```

As with the previous example, the first two lines import the data from the Internet and attach it. The second two lines partition the state populations into southern populations and midwestern populations. The third pair of lines perform the Shapiro-Wilk test on the two samples. Finally, the last pair of lines perform the Mann-Whitney test and the randomization test.

The following gives the code for the back-to-back histogram. Note that it requires you "source" a function found online on the book's website.

```
source("http://rfs.kvasaheim.com/Rfctns/histb2b.R")
histb2b( pops,  popm, yaxt="n",bty="n", direction=2, names=
   c("South","MidWest"), breaks=15)
```

◇

Note: Back-to-back histograms are very useful in determining if two variables have the same distribution (except for the center). When examining them, make sure you pay attention to the variance and the skew. The Mann-Whitney test is actually robust to violations of its assumption. Thus, you do not have to be *too* strict in making sure the two distributions are the same.

Example 6.9: The `crime` dataset contains a sample of a lot of variables, including the gross state product (GSP) per capita (average income in the state) in 2000 and the census region of the state. An associate of mine hypothesized that the average GSP per capita in the South is less than that in the Midwest. Test this hypothesis.

Solution: The research hypothesis is $\mu_s < \mu_m$. Since this does not contain the equals position, it is also the alternative hypothesis.

As we are comparing an average for two population, I would like to use the two-sample t-test as it is the most powerful of those available to us. However, it requires that the measurements in each group come from a Normally-distributed population. To test this, let us use the Shapiro-Wilk test. According to this test, neither sample comes from a Normally-distributed population ($p_m = 0.0159; p_s = 0.0001049$). Thus, we cannot use the two-sample t-test.

We can use the Mann-Whitney test, but that assumes the two populations differ only in their centers. Histograms of these two samples strongly suggest that the two may differ in more than just their means (Figure 6.5)., with the variance of the Midwestern GSPs per capita being much smaller than that of the South. Thus, let us perform a randomization test.

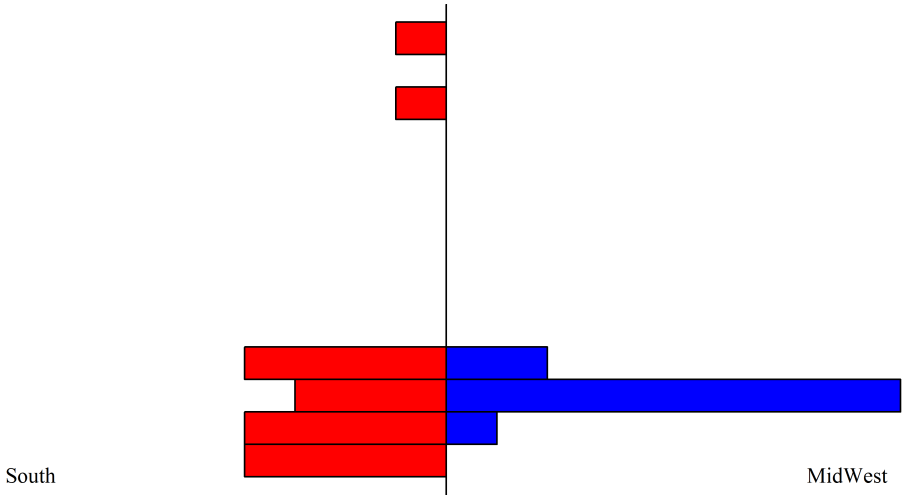

Figure 6.5: *A back-to-back histogram for the GSP per capita in each of the two census regions. This is done to gauge whether the two samples come from populations that differ only in their location parameter (center).*

According to the two-sample randomization test (using the permutation central limit theorem), we cannot detect a difference in the average GSP per capita in the two census regions ($p = 0.7142$).

Again, here is the code for the analysis.

```
cr = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(cr)

gsps = gsp00[census4=="South"]
gspm = gsp00[census4=="Midwest"]

shapiro.test(gsps)
shapiro.test(gspm)

permTS(gsps, gspm, alternative="less", method="pclt")
```

As with the previous example, the first two lines import the data from the Internet and attach it. The second two lines partition the GSPs per capita into southern and midwestern samples. The third pair of lines perform the Shapiro-Wilk test on the two samples. Finally, the last line performs the randomization test. ◇

6.7: Conclusion

In this chapter, you have learned how to perform more tests of means: those for comparing two independent samples. You have also again examined two classes of tests: parametric (assumes a specific distribution for your data) and non-parametric (does not assume a specific distribution for your data).

Non-parametric tests are useful if your data has an obviously non-Normal distribution or if the sample size is small. However, the weakness of all non-parametric tests is that they have lower power than the parametric tests. As such, when the parametric assumptions are not met, one should run the non-parametric test.

Frequently, we wish to compare more than two groups. In such cases, we can repeatedly use the t-test. However, we need to adjust for the fact that we are performing multiple tests on the same data. The disadvantage is that adjustments for multiple comparisons tend to reduce the power of the test. The advantage is that we already know how to perform these tests.

If we do not wish to lose power and perform multiple comparisons, we must use either an analysis of variance test *or* the Kruskal-Wallis non-parametric test. In a future chapter, we learn about these two tests as well as data transformation, which may allow us to use parametric tests when the original data is severely non-Normal, and *post-hoc* tests, which allow us to determine *which* groups are different.

6.8: End of Chapter Materials

6.8.1 R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

Statistics:

length(x) Returns the number of values in the vector x, if x is a vector. Returns the length of the character string x, if x is a character string.

shapiro.test(x) This performs a Shapiro-Wilk test, which determines if the provided sample comes from the Normal distribution.

t.test(·) This function preforms a t-test of the provided data. The four types of t-tests can be specified as

`t.test(x, mu=)`	1-sample t-test
`t.test(x,y)`	2-sample t-test, unequal variances
`t.test(x,y, var.equal=TRUE)`	2-sample t-test, equal variances
`t.test(x,y, paired=TRUE)`	2-sample, paired t-test

wilcox.test(x,y) Performs the Mann-Whitney test, comparing the medians of two samples.

permTS(x,y) Performs a permutation or randomization test comparing the centers of two samples. This requires loading the `perm` package.

Probability:

rexp(n) Returns n random numbers from the specified Exponential distribution: `rexp(100,r=3)` gives 100 random numbers drawn from an $\mathcal{E}xp(\lambda = 3)$ distribution.

dnorm(x) Returns the likelihood (or *d*ensity) for an x-value according to the specified Normal distribution: `dnorm(1,m=3,s=6)` returns the value of the pdf at 1 corresponding to the $\mathcal{N}(\mu = 3, \sigma = 6)$ distribution, 0.062897.

pnorm(x) Returns the cumulative *p*robability for an x-value according to the specified Normal distribution: `pnorm(1.96,m=0,s=1)` returns the value of the CDF at 1.96 corresponding to the $\mathcal{N}(\mu = 0, \sigma = 1)$ distribution, 0.975.

qnorm(p) Returns the value of x corresponding to the p-value provided according to the specified Normal distribution: `qnorm(0.95,m=5,s=1)` returns the x-value such that $\mathbb{P}[X < x] = 0.95$, where X is distributed as $\mathcal{N}(\mu = 5, \sigma = 1)$.

rnorm(n) Returns n *r*andom numbers from the specified Normal distribution: `rnorm(100,m=3,s=6)` gives 100 random numbers drawn from a $\mathcal{N}(\mu = 3, \sigma = 6)$ distribution.

pt(x) Returns the cumulative *p*robability for an x-value according to the specified Student's t distribution: `tnorm(1.96, df=11)` returns the value of the CDF at 1.96 corresponding to the $t(\nu = 11)$ distribution, 0.962. If you would rather calculate the area between your value and ∞ (i.e. $\mathbb{P}[X > x]$), use the parameter `lower.tail=FALSE`. Otherwise, the area is calculated between your value and $-\infty$, as usual.

Graphing:

abline() Draws a line on a currently open plot: `abline(h=3)` draws a horizontal line at $y = 3$; `abline(v=6)` draws a horizontal line at $x = 6$; `abline(a=3,b=1)` draws a line with intercept $a = 3$ and slope $b = 1$.

hist(x) Calculates (and draws) a histogram corresponding to the variable x.

histb2b(x,y) Draws a back-to-back histogram for variables x and y. To use this, the function must first be sourced from the book's website.

Mathematics:

abs(x) Returns the magnitude of the argument: `abs(-3)` = 3.

sqrt(x) Returns the positive square root of the argument: `sqrt(9)` = 3.

PROGRAMMING:

attach(d) Connects the dataset *d* to the current working environment so that one does not need to use '$' notation to access its variables and values. This is rather handy if you are only using one dataset in your analysis. If, however, you are using several, then it becomes rather easy to forget that the value you are requesting may not be the one you actually want. As such, use this with care.

for(){} Creates a loop in your script, allowing statements contained within the braces to be performed more than once. This statement is invaluable when performing Monte Carlo analysis.

function(){} Creates a user-defined function, whose parameters (required or options) are contained in the parentheses immediately following `function`, and whose statements are contained in the braces following `function`.

levels(x) Returns the levels of the categorical variable `x`.

names(d) Returns the variables contained in the *d* variable, which can be a dataframe, a list, or a matrix/array.

read.csv(f) Imports a dataset from *f*, the specified file location. If the first row (header) of the dataset contains variable names, you may specify the optional parameter `header=TRUE` in the function call; otherwise, you must specify `header=FALSE`.

6.8.2 Exercises and Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 6 folder. For each of the following problems, please save the associated `R` script in the chapter folder as `ext0x.R`, where `x` is the problem number.

Summary:

1. There *does* exist a two-sample z-test. Why is it not covered in this chapter? What would its drawback be?
2. Why does one want to use the two-sample t-test over the Mann-Whitney test? Why would one use the Mann-Whitney test instead of the two-sample t-test?
3. What is the assumption of the two-sample t-test? How would one test it (name the test)?
4. How does the work-flow for the one-sample t-test differ from that of the two-sample t-test? What is the reason for that difference?
5. When should the Mann-Whitney test be used and when should the Wilcoxon test be used?
6. Why are permutation (randomization) tests sometimes needed? Why should they be used only as a last resort?

Data:

7. In Example 6.5, we first divided the data into two subsets and then performed the Mann-Whitney test. Import the `football1` dataset and use the following line of code in lieu of creating two separate subsets: `wilcox.test(score ~ conference)`. Save this script in your chapter folder as `ext03.R`.
   a) What are the differences in the output?
   b) Is `score` an independent or a dependent variable?
   c) Is `conference` an independent or a dependent variable?
8. Let us examine the `patrickHenry` datafile. A research hypothesizes that the female students at Patrick Henry College score higher on the

SAT Mathematics test, on average, than the male students. Provide 95% confidence intervals for the means of the SAT Mathematics test score for the two genders and for the difference between the two genders. Do the data support the researcher's contention? Provide an appropriate box-and-whiskers plot to illustrate your point.

9. Let us again examine the `patrickHenry` datafile. A research hypothesizes that the female students at Patrick Henry College score higher on the SAT Verbal test, on average, than the male students. Provide 95% confidence intervals for the means of the SAT Verbal test score for the two genders and for the difference between the two genders. Do the data support the researcher's contention? Provide an appropriate box-and-whiskers plot to illustrate your point.

10. Using the `patrickHenry` datafile, can we conclude that the average female has a higher GPA than the average male? Explain fully using statistics and graphics.

11. Using the `studentHeight` datafile, can we conclude that the average male is taller than the average female? Explain fully using statistics and graphics.

Monte Carlo:

12. Create a random dataset (of size 100) from the Normal distribution, with mean 4 and standard deviation 1. Create a second dataset (of size 500) from an Exponential distribution, with mean 4 (rate, $\lambda = 0.25$). Use a seed value of 3. Test the null hypothesis that these two distributions have the same mean. Save this script in your chapter folder as `ext01.R`.

    a) If you wanted to use the parametric test, is the sample size large enough?

    b) Which test should you use?

    c) Does that test reject the null hypothesis?

    d) What is the appropriate conclusion based on the test results?

    e) Knowing what you know about the *actual* variables, are the two population means equal?

13. Create a random dataset (of size 10) from the Normal distribution, with mean 4 and standard deviation 1. Create a second dataset (of size 10) from a Gaussian distribution, with mean 4.1 and standard deviation 1. Use a seed value of 3. Test the null hypothesis that these two distributions have the same mean. Save this script in your chapter folder as `ext02.R`.

    a) If you wanted to use the parametric test, is the sample size large enough?

    b) Which test should you use?

    c) Does that test reject the null hypothesis?

    d) What is the appropriate conclusion based on the test results?

    e) Knowing what you know about the actual variables, are the two population means equal?

6.8.3 Applied Research

This section offers some applied research works that are connected with the topics in this chapter.

- Charles Bérubé and Pierre Mohnen. (2009) "Are Firms That Receive R&D Subsidies More Innovative?" *The Canadian Journal of Economics / Revue canadienne d'Economique*. **42**(1):206–25.
- Matthew S. Bothner, Edward Bishop Smith, and Harrison C. White. (2010) "A Model of Robust Positions in Social Networks." *American Journal of Sociology*. **116**(3): 943–92.
- Kevin Denny and Orla Doyle. (2009) "Does Voting History Matter? Analysing Persistence in Turnout." *American Journal of Political Science*. **53**(1): 17–35.
- Esther Godson and John D. Stednick. (2010) "Modeling Post-Fire Soil Erosion." *Fire Management Today* **70**(3): 32–36.
- Matthijs Kalmijn. (2010) "Consequences of Racial Intermarriage for Children's Social Integration." *Sociological Perspectives*. **53**(2): 271–86.
- John R. Lott, Jr. (2009) "Non-Voted Ballots, the Cost of Voting, and Race." *Public Choice*. **138**(1/2):171–97.
- Tonya L. Putnam. (2009) "Courts without Borders: Domestic Sources of U.S. Extraterritoriality in the Regulatory Sphere" *International Organization*. **63**(3): 459–90.

6.8.4 References and Additional Readings

This section provides a list of statistical works. Those works cited in the chapter are here. Also here are works that complement the chapter's topics.

- Lee J. Bain and Max Engelhardt. (1992) *Introduction to Probability and Mathematical Statistics*, 2nd edn. Brooks/Cole: Belmont, CA.
- William C. Navidi. (2006) *Statistics for Engineering and Scientists*, 2nd edn. McGraw-Hill: New York.
- Franklin E. Satterthwaite. (1946) "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin* **2**(6): 110–114.
- Samuel S. Shapiro and Martin B. Wilk. (1965) "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* **52**(3–4): 591–611.
- Bernard L. Welch. (1947) "The generalization of "Student's" problem when several different population variances are involved," *Biometrika* **34**(1–2): 28–35.