6
Testing Hypotheses

The salt-and-pepper of inferential statistics is estimation and testing hypotheses. In the last chapter, we talked about estimation and making certain inferences about the world. In this chapter, we will be talking about how to test the hypotheses on how the world works and evaluate the hypotheses using only sample data.

In the last chapter, I promised that this would be a very practical chapter, and I'm a man of my word; this chapter goes over a broad range of the most popular methods in modern data analysis at a relatively high level. Even so, this chapter might have a little more detail than the lazy and impatient would want. At the same time, it will have way too little detail than what the extremely curious and mathematically inclined want. In fact, some statisticians would have a heart attack at the degree to which I skip over the math involved with these subjects—but I won't tell if you don't!

Nevertheless, certain complicated concepts and math are beyond the scope of this book. The good news is that once you, dear reader, have the general concepts down, it is easy to deepen your knowledge of these techniques and their intricacies—and I advocate that you do before making any major decisions based on the tests introduced in these chapters.

Null Hypothesis Significance Testing

For better or worse, **Null Hypothesis Significance Testing** (**NHST**) is the most popular hypothesis testing framework in modern use. So, even though there are competing approaches that—at least in some cases—are better, you need to know this stuff up and down!

Okay—Null Hypothesis Significance Testing—those are a bunch of big words. What do they mean?

NHST is a lot like being a prosecutor in the United States' or Great Britain's justice system. In these two countries—and a few others—the person being charged is presumed innocent, and the burden of *proving* the defendant's guilt is placed on the prosecutor. The prosecutor then has to argue that the evidence is inconsistent with the defendant being innocent. Only after it is shown that the extant evidence is unlikely if the person is innocent, does the court rule a guilty verdict. If the extant evidence is weak, or is likely to be observed even if the dependent is innocent, then the court rules not guilty. That doesn't mean the defendant is innocent (the defendant may very well be guilty!)—it means that either the defendant was guilty, or there was not sufficient evidence to prove guilt.

With simple NHST, we are testing two competing hypotheses: the null and the alternative hypotheses. The *default* hypothesis is called the null hypothesis—it is the hypothesis that our observation occurred from chance alone. In the justice system analogy, this is the hypothesis that the defendant is innocent. The alternative hypothesis is the opposite (or complementary) hypothesis; this would be like the prosecutor's hypothesis.

The *null hypothesis* terminology was introduced by a statistician named R. A. Fischer in regard to the curious case of Muriel Bristol: a woman who claimed that she could discern, just by tasting it, whether milk was added before tea in a teacup or whether the tea was poured before the milk. She is more commonly known as the *lady tasting tea*.

Her claim was put to the test! The lady tasting tea was given eight cups; four had milk added first, and four had tea added first. Her task was to correctly identify the four cups that had tea added first. The null hypothesis was that she couldn't tell the difference and would choose a random four teacups. The alternative hypothesis is, of course, that she had the ability to discern wither the tea or milk was poured first.

It turned out that she correctly identified the four cups. The chances of randomly choosing the correct four cups is 70 to 1, or about 1.4%. In other words, the chances of that happening under the null hypothesis is 1.4%. Given that it is so very unlikely to have occurred under the null hypothesis, we may choose to *reject* the null hypothesis. If the null and alternative hypotheses are mutually exclusive and collectively exhaustive, then a rejection of the null hypothesis is tantamount to an acceptance of the alternative hypothesis.

We can't say anything for certain, but we can work with probabilities. In this example, we wanted to prove or disprove the lady tasting tea's claims. We did not try to evaluate the probability that the lady could tell the difference; we assumed that she could not and tried to show that it was unlikely that she couldn't, given her stellar performance on the assessment.

So, here's the basic idea behind NHST as we know it so far:

1. Assume the opposite of what you are testing.
2. (Try to) show that the results you receive are unlikely given that assumption.
3. Reject the assumption.

We have heretofore been rather *hand-wavy* about what constitutes sufficient *unlikelihood* to reject the null hypothesis and how we determine the probability in the first place. We'll discuss this now.

In order to quantify how likely or unlikely the results we receive are, we need to define a *test statistic*—some measure of the sample. The sampling distribution of the test statistic will tell us which test statistics are most likely to occur by chance (under the null hypothesis) with repeated trials of the experiment. Once we know what the sampling distribution of the test statistic looks like, we can tell what the probability of getting a result as extreme as we got is. This is called a *p-value*. If it is equal to or below some pre-specified boundary, called an *alpha level* (α level), we decide that the null hypothesis is a bad hypothesis and embrace the alternative hypothesis. Largely, as a matter of tradition, an alpha level of .05 is used most often, though other levels are occasionally used as well. So, if the observed result would only occur 5% or less of the time (p-value < .05), we consider it a sufficiently unlikely event and reject the null hypothesis. If the .05 cut-off sounds rather arbitrary, it's because it is.

So, here's our updated and expanded *basic idea* behind NHST:

1. Formulate a set of two hypotheses: a null hypothesis (often denoted as H0) and an alternative hypothesis (often denoted H1)
   - H0: there is no effect
   - H1: there is an effect
2. Compute the test statistic.
3. Given the sampling distribution of the test statistic under the null hypothesis, you can calculate the probability of obtaining a test statistic equal to or more extreme than the one you calculated. This is the *p-value*. Find it.
4. If the probability of obtaining a test statistic being equal to or more extreme than the one you calculated is sufficiently unlikely (equal to or less than your alpha level), then you may reject the null hypothesis.
5. If the null and alternative hypotheses are collectively exhaustive, you may embrace the alternative hypothesis.

The illustrative example that's going to make sense out of all of this is none other than the gambit of Larry the Untrustworthy Knave that we met in *Chapter 4, Probability*. If you recall, Larry, who can only be trusted some of the time, gave us a coin that he alleges is fair. We flip it 30 times and observe 10 heads. Let's hypothesize that the coin is unfair; let's formalize our hypotheses:

- H0 (null hypothesis): the probability of obtaining heads on this coin is .5
- H1 (alternative hypothesis): the probability of obtaining heads on this coin is not .5

Let's just use the number of heads in our sample as the test statistic. What is the sampling distribution of this test statistic? In other words, if the coin were fair, and you repeated the flipping-30-times experiment many times, what is the relative frequency of observing particular numbers of heads? We've seen it already! It's the *binomial distribution*. A binomial distribution with parameters `n=30` and `p=0.5` describes the number of *heads* we should expect in 30 flips.

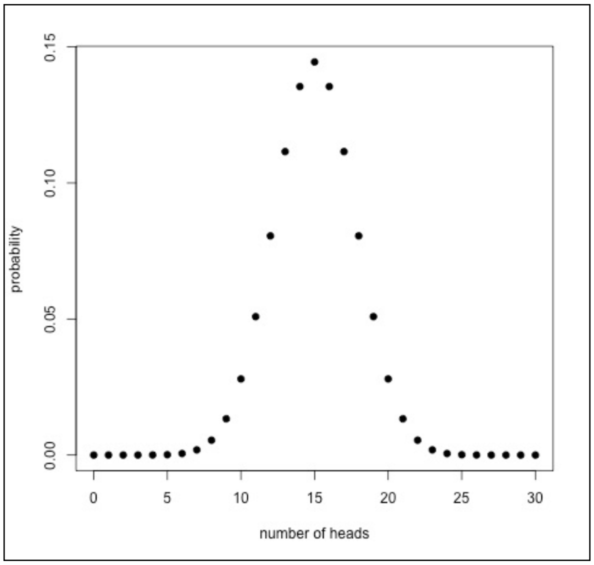

Figure 6.1: The sampling distribution of our coin-flip test statistic (the number of heads)

As you can see, the outcome that is the most likely is getting 15 heads (as you might imagine). Can you see what the probability of getting 10 heads is? Fairly unlikely, right?

So, what's the p-value, and is it less than our pre-specified alpha level? Well, we have already worked out the probability of observing 10 or fewer heads in *Chapter 4, Probability*, as follows:

```
> pbinom(10, size=30, prob=.5)
[1] 0.04936857
```

It's less than .05. We can conclude the coin is unfair, right? Well, yes and no. Mostly no. Allow me to explain.

One and two-tailed tests

You may reject the null hypothesis if the test statistic falls within a region under the curve of the sampling distribution that covers 5% of the area (if the alpha level is .05). This is called the *critical region*. Do you remember, in the last chapter, we constructed 95% confidence intervals that covered 95% percent of the sampling distribution? Well, the 5% critical region is like the opposite of this. Recall that, in order to make a symmetric 95% of the area under the curve, we had to start at the .025 quantile and end at the .975 quantile, leaving 2.5% percent on the left tail and 2.5% of the right tail uncovered.

Similarly, in order for the critical region of a hypothesis test to cover 5% of the most extreme areas under the curve, the area must cover everything from the left of the .025 quantile and everything to the right of the .975 quantile.

So, in order to determine that the 10 heads out of 30 flips is statistically significant, the probability that you would observe 10 or fewer heads has to be less than .025.

There's a function built right into R, called `binom.test`, which will perform the calculations that we have, until now, been doing by hand. In the most basic incantation of `binom.test`, the first argument is the number of *successes* in a Bernoulli trial (the number of heads), and the second argument is the number of trials in the sample (the number of coin flips).

```
   > binom.test(10,30)

           Exact binomial test

   data:  10 and 30
   number of successes = 10, number of trials = 30, p-value = 0.09874
   alternative hypothesis: true probability of success is not equal to
 0.5
```

```
95 percent confidence interval:
 0.1728742 0.5281200
sample estimates:
probability of success
             0.3333333
```

If you study the output, you'll see that the p-value does not cross the significance threshold.

Now, suppose that Larry said that the coin was not biased towards tails. To see if Larry was lying, we only want to test the alternative hypothesis that the probability of heads is less than .5. In that case, we would set up our hypotheses like this:

- H0: The probability of heads is greater than or equal to .5
- H1: The probability of heads is less than .5

This is called a *directional hypothesis,* because we have a hypothesis that asserts that the deviation from chance goes in a particular direction. In this hypothesis suite, we are only testing whether the observed probability of heads falls into a critical region on only one side of the sampling distribution of the test statistic. The statistical test that we would perform in this case is, therefore, called a *one-tailed test*—the critical region only lies on one tail. Since the area of the critical region no longer has to be divided between the two tails (like in the two-tailed test we performed earlier), the critical region only contains the area to the *left* of the .05 quantile.

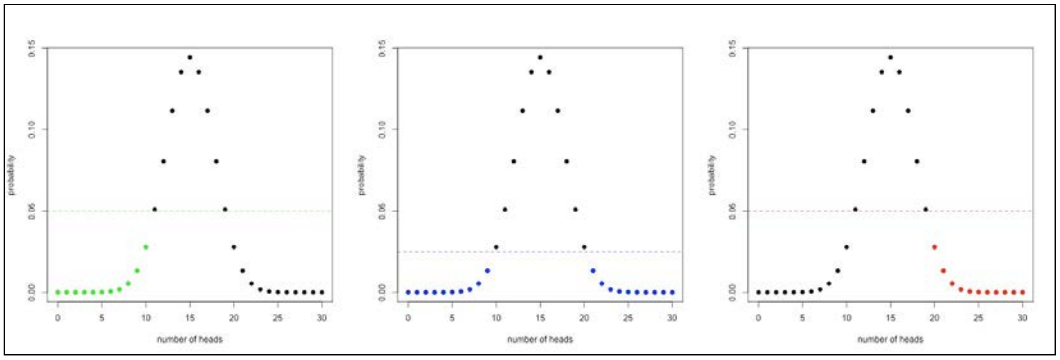

Figure 6.2: The three panels, from left to right, depict the critical regions of the left ("lesser") one-tailed, two-tailed, and right ("greater") alternative hypotheses. The dashed horizontal line is meant to show that, for the two-tailed tests, the critical region starts below p=.025, because it is being split between two tails. For the one-tailed tests, the critical region is below the dashed horizontal line at p=.05.

As you can see from the figure, for the directional alternative hypothesis that heads has a probability less than .5, 10 heads is now included in the green critical region.

We can use the binom.test function to test this directional hypothesis, too. All we have to do is specify the optional parameter alternative and set its value to "less" (its default is "two.sided" for a two-tailed test).

```
> binom.test(10,30, alternative="less")

        Exact binomial test

data:  10 and 30
number of successes = 10, number of trials = 30, p-value = 0.04937
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4994387
sample estimates:
probability of success
             0.3333333
```

If we wanted to test the directional hypothesis that the probability of heads was greater than .5, we would use alternative="greater".

Take note of the fact that the p-value is now less than .05. In fact, it is identical to the probability we got from the pbinom function.

When things go wrong

Certainty is a card rarely used in the deck of a data analyst. Since we make judgments and inferences based on probabilities, mistakes happen. In particular, there are two types of mistakes that are possible in NHST: *Type I errors* and *Type II errors*.

- A Type I error is when a hypothesis test concludes that there is an effect (rejects the null hypothesis) when, in reality, no such effect exists
- A Type II error occurs when we fail to detect a real effect in the world and fail to reject the null hypothesis even if it is false

Check the following table for errors encountered in the coin example:

Coin type	**Failure to reject null hypothesis (conclude no detectable effect)**	**Reject the null hypothesis (conclude that there is an effect)**
Coin is fair	Correct positive identification	Type I error (false positive)
Coin is unfair	Type II error (false negative)	Correct identification

In the criminal justice system, Type I errors are considered especially heinous. Legal theorist William Blackstone is famous for his quote: *it is better that ten guilty persons escape than one innocent suffer*. This is why the court instructs jurors (in the United States, at least) to only convict the defendant if the jury believes the defendant is guilty beyond a reasonable doubt. The consequence is that if the jury favors the hypothesis that the defendant is guilty, but only by a little bit, the jury must give the defendant the benefit of the doubt and acquit.

This line of reasoning holds for hypothesis testing as well. Science would be in a sorry state if we accepted alternative hypotheses on rather flimsy evidence willy-nilly; it is better that we err on the side of caution when making claims about the world, even if that means that we make fewer discoveries of honest-to-goodness, real-world phenomena because our statistical tests failed to reach significance.

This sentiment underlies that decision to use an alpha level like .05. An alpha level of .05 means that we will only commit a Type I error (false positive) 5% of the time. If the alpha level were higher, we would make fewer Type II errors, but at the cost of making more Type I errors, which are more dangerous in most circumstances.

There is a similar metric to the alpha level, and it is called the *beta level* (β level). The beta level is the probability that we would fail to reject the null hypothesis if the alternative hypothesis were true. In other words, it is the probability of making a Type II error.

The complement of the beta level, 1 minus the beta level, is the probability of correctly detecting a true effect if one exists. This is called *power*. This varies from test to test. Computing the power of a test, a technique called power analysis, is a topic beyond the scope of this book. For our purposes, it will suffice to say that it depends on the type of test being performed, the sample size being used, and on the size of the effect that is being tested (*the effect size*). Greater effects, like the average difference in height between women and men, are far easier to detect than small effects, like the average difference in the length of earthworms in Carlisle and in Birmingham. Statisticians like to aim for a power of at least 80% (a beta level of .2). A test that doesn't reach this level of power (because of a small sample size or small effect size, and so on) is said to be underpowered.

A warning about significance

It's perhaps regrettable that we use the term *significance* in relation to null-hypothesis testing. When the term was first used to describe hypothesis tests, the word significance was chosen because it signified something. As I wrote this chapter, I checked the thesaurus for the word *significant,* and it indicated that synonyms include notable, worthy of attention, and important. This is misleading in that it is not equivalent to its intended, vestigial meaning. One thing that really confuses people is that they think statistical significance is of great importance in and of itself. This is sadly untrue; there are a few ways to achieve statistical significance without discovering anything of significance, in the colloquial sense.

As we'll see later in the chapter, one way to achieve non-significant statistical significance is by using a very large sample size. Very small differences, that make little to no difference in the real world, will nevertheless be considered statistically significant if there is a large enough sample size.

For this reason, many people make the distinction between statistical significance and practical significance or clinical relevance. Many hold the view that hypothesis testing should only be used to answer the question *is there an effect?* or *is there a discernable difference?*, and that the follow-up questions *is it important?* or *does it make a real difference?* should be addressed separately. I subscribe to this point of view.

To answer the follow-up questions, many use effect sizes, which, as we know, capture the magnitude of an effect in the real world. We will see an example of determining the effect size in a test later in this chapter.

A warning about p-values

P-values are, by far, the most talked about metric in NHST. P-values are also notorious for lending themselves to misinterpretation. Of the many criticisms of NHST (of which there are many, in spite of its ubiquity), the misinterpretation of p-values ranks highly. The following are two of the most common misinterpretations:

1. A p-value is the probability that the null hypothesis is true. This is not the case. Someone misinterpreting the p-value from our first binomial test might conclude that the chances of the coin being fair are around 10%. This is false. The p-value does not tell us the probability of the *hypothesis' truth or falsity*. In fact, the test assumes that the null hypothesis is correct. It tells us the proportion of trials for which we would receive a result as extreme or more extreme than the one we did if the null hypothesis was correct. I'm ashamed to admit it, but I made this mistake during my first college introductory statistics class. In my final project for the class, after weeks of collecting data, I found my p-value had not passed the barrier of significance—it was something like .07. I asked my professor if, after the fact, I could change my alpha level to .1 so my results would be positive. In my request, I appealed to the fact that it was still more probable than not that my alternative hypothesis was correct—after all, if my p-value was .07, then there was a 93% chance that the alternative hypothesis was correct. He smiled and told me to read the relevant chapter of our text again. I appreciate him for his patience and restraint in not smacking me right in the head for making such a stupid mistake. Don't be like me.
2. A p-value is a measure of the size of an effect. This is also incorrect, but its *wrongness* is more subtle than the first misconception. In research papers, it is common to attach phrases like *highly significant* and *very highly significant* to p-values that are much smaller than .05 (like .01 and .001). It is common to interpret p-values such as these, and statements such as these, as signaling a bigger effect than p-values that are only modestly less than .05. This is a mistake; this is conflating statistical significance with practical significance. In the previous section, we explained that you can achieve significant p-values (sometimes *very highly significant* ones) for an effect that is, for all intents and purposes, small and unimportant. We will see a very salient example of this later in this chapter.

Testing the mean of one sample

An illustrative and fairly common statistical hypothesis test is the *one sample t-test*. You use it when you have one sample and you want to test whether that sample likely came from a population by comparing the mean against the known population mean. For this test to work, you have to know the population mean.

In this example, we'll be using R's built-in `precip` data set that contains precipitation data from 70 US cities.

```
> head(precip)
Mobile    Juneau    Phoenix   Little Rock   Los Angeles   Sacramento
 67.0      54.7      7.0         48.5           14.0        17.2
```

Don't be fooled by the fact that there are city names in there—this is a regular old vector - it's just that the elements are labeled. We can directly take the mean of this vector, just like a normal one.

```
> is.vector(precip)
[1] TRUE
> mean(precip)
[1] 34.88571
```

Let's pretend that we, somehow, know the mean precipitation of the rest of the world—is the US' precipitation significantly different to the rest of the world's precipitation?

Remember, in the last chapter, I said that the sampling distribution of sample means for sample sizes under 30 were best approximated by using a t-distribution. Well, this test is called a *t-test*, because in order to decide whether our samples' mean is consistent with the population whose mean we are testing against, we need to see where our mean falls in relation to the sampling distribution of population means. If this is confusing, reread the relevant section from the previous chapter.

In order to use the t-test in general cases—regardless of the scale—instead of working with the sampling distribution of sample means, we work with the sampling distribution of the t-statistic.

Remember z-scores from *Chapter 3*, *Describing Relationships*? The t-statistic is like a z-score in that it is a scale-less measure of distance from some mean. In the case of the t-statistic, though, we divide by the standard error instead of the standard deviation (because the standard deviation of the population is unknown). Since the t-statistic is *standardized*, any population, with any mean, using any scale, will have a sampling distribution of the t-statistic that is exactly the same (at the same sample size, of course).

The equation to compute the t-statistic is this:

$$t = \frac{\overline{x} - \mu}{s / \sqrt{N}}$$

where $\overline{x}$ is the sample mean, μ is the population mean, s is the sample' standard deviation, and N is the sample size.

Let's see for ourselves what the sampling distribution of the t-statistic looks like by taking 10,000 samples of size 70 (the same size as our `precip` data set) and plotting the results:

```
# function to compute t-statistic
t.statistic <- function(thesample, thepopulation){
  numerator <- mean(thesample) - mean(thepopulation)
  denominator <- sd(thesample) / sqrt(length(thesample))
  t.stat <- numerator / denominator
  return(t.stat)
}

# make the pretend population normally distributed
# with a mean of 38
population.precipitation <- rnorm(100000, mean=38)
t.stats <- numeric(10000)
for(i in 1:10000){
  a.sample <- sample(population.precipitation, 70)
  t.stats[i] <- t.statistic(a.sample, population.precipitation)
}

# plot
library(ggplot2)
tmpdata <- data.frame(vals=t.stats)
qplot(vals, data=tmpdata, geom="histogram",
      color=I("white"),
      xlab="sampling distribution of t-statistic",
      ylab="frequency")
```

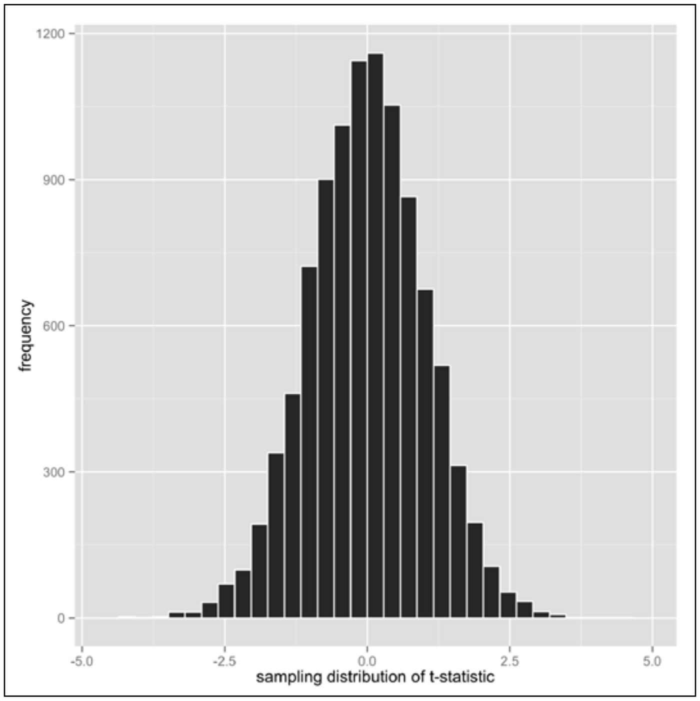

Figure 6.3: The sampling distribution of the t-statistic

Ah, there's that familiar shape again!

Fortunately, the sampling distribution of the `t-statistic` is well known, so we don't have to create our own. In fact, the sampling distribution for many test statistics are well known, so we won't be running our own simulations of them anymore. Lucky us!

Okay, so how does our sample's t-statistic compare to the t-distribution? Our t-statistic, using our function from the last code-snippet, is:

```
> t.statistic(precip, population.precipitation)
[1] -1.901225
```

Though, you can work this out for yourself easily.

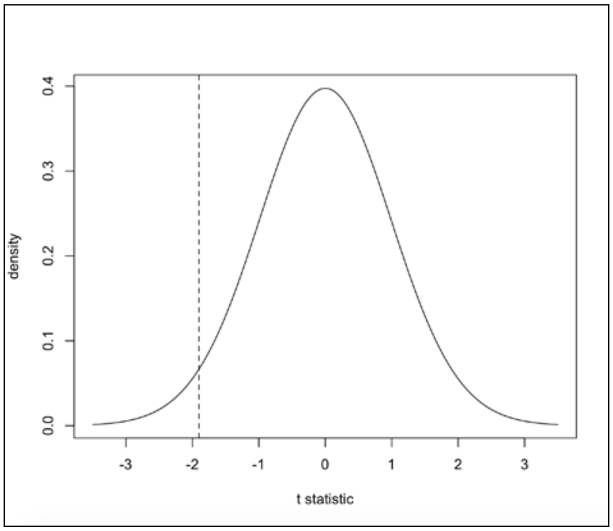

Figure 6.4: The t-distribution with 69 degrees of freedom. The t-statistic of our sample is shown as the dashed line

Hmm, it looks like a pretty unlikely occurrence to me, but is it statistically significant? First, let's formally define our hypotheses:

- H0 = the average (mean) precipitation in the US is equal to the known average precipitation in the rest of the world
- H1 = the average (mean) precipitation in the US is different than the known average precipitation in the rest of the world

Then, we prespecify an alpha level of .05, as is customary.

Since our hypothesis is non-directional (we only hypothesize that the precipitation in the US is different than the world, not less or more), we define our critical region to cover 5% of the area on each side of the curve.

```
> qt(.025, df=69)
[1] -1.994945
> # the critical region is less than -1.995 and more than +1.995
```

What does it look like now?

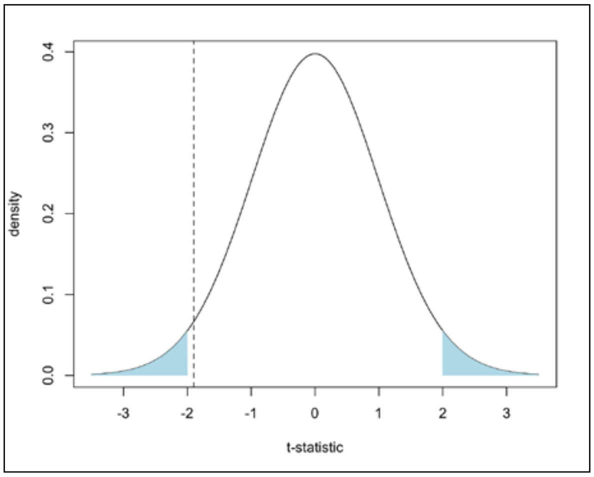

Figure 6.5: The previous figure with the critical region for non-directional hypothesis highlighted

Oh, too bad! It looks like our sample mean falls just out of the critical region. So, we fail to reject the null hypothesis.

The cruel truth if we, for some reason, hypothesized that the US precipitation was *less* than the average world precipitation is:

- H0 = mean US precipitation >= mean world precipitation
- H1 = mean US precipitation < mean world precipitation

We would have achieved significance at `alpha = .05`.

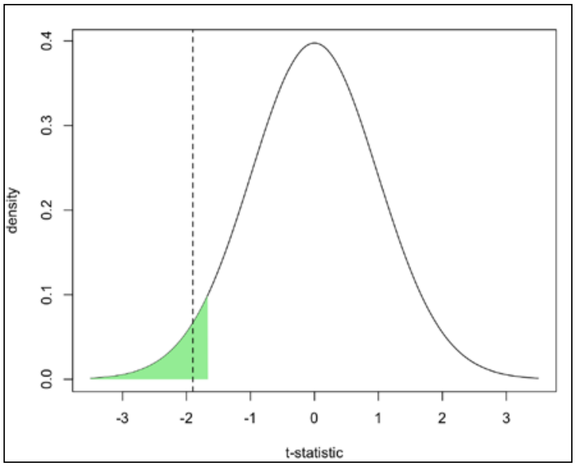

Figure 6.6: Figure 6.4 with directional critical region highlighted

Of course, we have no reason to think that US precipitation was less or more than the world's average. And to change our hypothesis now would be cheating. You're not a cheater, are you?

Now that we know what we're doing, we won't be manually calculating our test statistics anymore; we'll just be using the test functions that R provides.

Let's use the function that R provides now. The one sample t-test can be performed by the `t.test` function. In its most basic form, it takes a vector of sample observations as its first argument and the population mean as its second argument..

```
> t.test(precip, mu=38)

          One Sample t-test

  data:  precip
  t = -1.901, df = 69, p-value = 0.06148
  alternative hypothesis: true mean is not equal to 38
  95 percent confidence interval:
   31.61748 38.15395
  sample estimates:
  mean of x
   34.88571
```

Among other things, this test tells us that the t-statistic is 1.9 (just like we calculated ourselves), the degrees of freedom were 69 (the sample size minus 1), and the p-value, which is 0.06148. Like our plot with the two-tailed critical regions showed, this p-value is greater than our prespecified alpha level of 0.05. We fail to reject the null hypothesis.

Just for kicks, let's run the *one-tailed hypothesis test*:

```
> t.test(precip, mu=38, alternative="less")

        One Sample t-test

data:  precip
t = -1.901, df = 69, p-value = 0.03074
alternative hypothesis: true mean is less than 38
95 percent confidence interval:
     -Inf 37.61708
sample estimates:
mean of x
 34.88571
```

Now our p-value is < .05. C'est la vie.

Note that the R output indicates that the alternative hypothesis which is the true mean is less than 38—compare this with the last t-test output.

Assumptions of the one sample t-test

There are two main assumptions of the one sample t-test:

- The data are sampled from a normal distribution. This actually has more to do with the sampling distribution of sample means being approximately normal than the actual population. As we know, the sampling distribution of sample means for sufficiently large sample sizes will always be normally distributed, even if the population is not. In reality, this assumption can be violated somewhat, and the results will be valid, especially for sample sizes of over 30. We have nothing to worry about here. Usually, people check this assumption by plotting the sample means and making sure it's kind-of normal, though there are more formal ways of doing this, which we will see later. If the assumption of normality is in question, we may want to use an alternative test, like a *non-parametric test*; we'll see some examples at the end of this chapter.
- Independence of samples: Had we tested whether the US precipitation likely came from the population of the entire world's precipitation, we would have been violating this assumption. Why? Because we know that the US is a member of the set (it is indeed 'in the world'), so of course it was drawn from that population. This is why we tested whether the US precipitation was on par with the rest of the world's precipitation. In other examples of the one sample t-tests, this assumption basically requires that the sample be random.

Testing two means

An even more common hypothesis test is the independent samples t-test. You would use this to check the equality of two samples' means. Concretely, an example of using this test would be if you have an experiment where you are testing to see if a new drug lowers blood pressure. You would give one group a placebo and the other group the real medication. If the mean improvement in blood pressure was significantly greater than the improvement with the placebo, you might infer that the blood pressure medication works. Outside of more academic uses, web companies use this test all the time to test the effectiveness of, for example, different internet ad campaigns; they expose random users to either one of two types of ads and test if one is more effective than the other. In web-business parlance, this is called an A-B test, but that's just business-ese for *controlled experiment*.

The term *independent* means that the two samples are separate, and that data from one sample doesn't affect data in the other. For example, if instead of having two different groups in the blood pressure trial, we used the same participants to test both the conditions (randomizing the order we administer the placebo and the real medication), we would violate independence.

The dataset we will be using for this is the `mtcars` dataset that we first met in *Chapter 2, The Shape of Data* and saw again in *Chapter 3, Describing Relationships*. Specifically, we are going to test the hypothesis that the mileage is better for manual cars than it is for cars with automatic transmission. Let's compare the means and produce a boxplot:

```
> mean(mtcars$mpg[mtcars$am==0])
[1] 17.14737
> mean(mtcars$mpg[mtcars$am==1])
[1] 24.39231
>
> mtcars.copy <- mtcars
> # make new column with better labels
> mtcars.copy$transmission <- ifelse(mtcars$am==0,
                                      "auto", "manual")
> mtcars.copy$transmission <- factor(mtcars.copy$transmission)
> qplot(transmission, mpg, data=mtcars.copy,
+       geom="boxplot", fill=transmission) +
+   # no legend
+   guides(fill=FALSE)
```

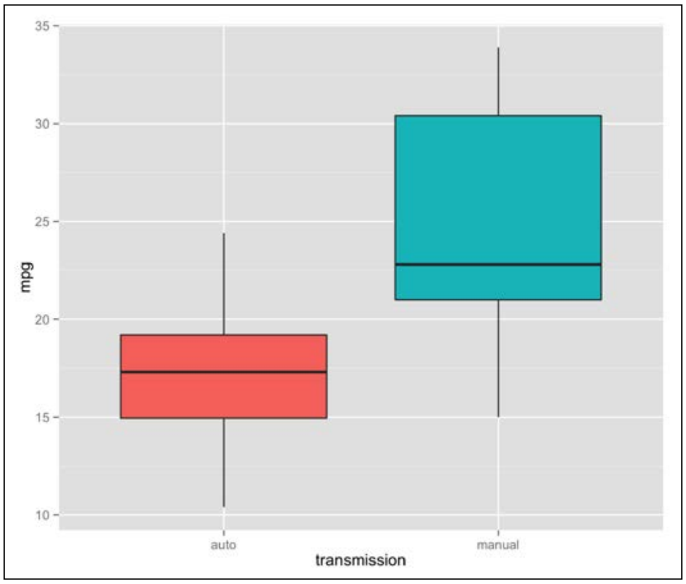

Figure 6.7: Boxplot of the miles per gallon ratings for automatic cars and cars with manual transmission

Hmm, looks different… but let's check that hypothesis formally. Our hypotheses are:

- H0 = mean of sample1 - mean of sample2 >= 0
- H1 = mean of sample1 - mean of sample2 < 0

To do this, we use the `t.test` function, too; only this time, we provide two vectors: one for each sample. We also specify our directional hypothesis in the same way:

```
> automatic.mpgs <- mtcars$mpg[mtcars$am==0]
> manual.mpgs <- mtcars$mpg[mtcars$am==1]
> t.test(automatic.mpgs, manual.mpgs, alternative="less")

        Welch Two Sample t-test

data:  automatic.mpgs and manual.mpgs
t = -3.7671, df = 18.332, p-value = 0.0006868
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.913256
sample estimates:
mean of x mean of y
 17.14737  24.39231
```

```
p < .05. Yipee!
```

There is an easier way to use the t-test for independent samples that doesn't require us to make two vectors.

```
> t.test(mpg ~ am, data=mtcars, alternative="less")
```

This reads, roughly, perform a t-test of the `mpg` column grouping by the `am` column in the data frame `mtcars`. Confirm for yourself that these incantations are equivalent.

Don't be fooled!

Remember when I said that statistical significance was not synonymous with *important* and that we can use very large sample sizes to achieve statistical significance without any *clinical* relevance? Check this snippet out:

```
> set.seed(16)
> t.test(rnorm(1000000,mean=10), rnorm(1000000, mean=10))

        Welch Two Sample t-test

data:  rnorm(1e+06, mean = 10) and rnorm(1e+06, mean = 10)
```

```
t = -2.1466, df = 1999998, p-value = 0.03183
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0058104638 -0.0002640601
sample estimates:
mean of x mean of y
 9.997916 10.000954
```

Here, two vectors of one million normal deviates each are created with a mean of 10. When we use a t-test on these two vectors, it should indicate that the two vectors' means are not significantly different, right?

Well, we got a p-value of less that .05—why? If you look carefully at the last line of the R output, you might see why; the mean of the first vector is `9.997916`, and the mean of the second vector is `10.000954`. This tiny difference, a meagre .003, is enough to tip the scale into *significant* territory. However, I can think of very few applications of statistics where .003 of anything is noteworthy even though it is, technically, statistically significant.

The larger point is that the t-test tests for equality of means, and if the means aren't exactly the same in the population, the t-test will, with enough power, detect this. Not all tiny differences in population means are important, though, so it is important to frame the results of a t-test and the p-value in context.

As mentioned earlier in the chapter, a salient strategy for putting the differences in context is to use an effect size. The effect size commonly used in association with the t-test is *Cohen's d*. Cohen's d is, conceptually, pretty simple: it is a ratio of the variance explained by the "effect" and the variance in the data itself. Concretely, Cohen's d is the difference in means divided by the sample standard deviation. A high d indicates that there is a big effect (difference in means) relative to the internal variability of the data.

I mentioned that to calculate d, you have to divide the difference in means by the sample standard deviation—but which one? Although Cohen's d is conceptually straightforward (even elegant!), it is also sometimes a pain to calculate by hand, because the sample standard deviation from both samples has to be *pooled*. Fortunately, there's an R package that let's us calculate Cohen's d—and other effect size metrics, to boot, quite easily. Let's use it on the auto vs. manual transmission example:

```
> install.packages("effsize")
> library(effsize)
> cohen.d(automatic.mpgs, manual.mpgs)
```

```
Cohen's d

d estimate: -1.477947 (large)
95 percent confidence interval:
       inf        sup
-2.3372176 -0.6186766
```

Cohen's d is -1.478, which is considered a very large effect size. The `cohen.d` function even tells you this by using canned interpretations of effect sizes. If you try this with the two million element vectors from above, the `cohen.d` function will indicate that the *effect* was negligible.

Although these canned interpretations were on target these two times, make sure you evaluate your own effect sizes in context.

Assumptions of the independent samples t-test

Homogeneity of variance (or homoscedasticity - a scary sounding word), in this case, simply means that the variance in the miles per gallon of the automatic cars is the same as the variance in miles per gallon of the manual cars. In reality, this assumption can be violated as long as you use a *Welch's T-test* like we did, instead of the *Student's T-test*. You can still use the Student's T-test with the `t.test` function, like by specifying the optional parameter `var.equal=TRUE`. You can test for this formally using `var.test` or `leveneTest` from the `car` package. If you are sure that the assumption of homoscedasticity is not violated, you may want to do this because it is a more powerful test (fewer Type II errors). Nevertheless, I usually use Welch's T-test to be on the safe side. Also, always use Welch's test if the two samples' sizes are different.

- The sampling distribution of the sample means is approximately normal: Again, with a large enough sample size, it always is. We don't have a terribly large sample size here, but in reality, this formulation of the t-test works even if this assumption is violated a little. We will see alternatives in due time.
- Independence: Like I mentioned earlier, since the samples contain completely different cars, we're okay on this front. For tests that, for example, use the same participants for both conditions, you would use a *Dependent Samples T-test* or *Paired Samples T-test* , which we will not discuss in this book. If you are interested in running one of these tests after some research, use `t.test(<vector1>, <vector2>, paired=TRUE).`

Testing more than two means

Another really common situation requires testing whether three or more means are significantly discrepant. We would find ourselves in this situation if we had three experimental conditions in the blood pressure trial: one groups gets a placebo, one group gets a low dose of the real medication, and one groups gets a high dose of the real medication.

Hmm, for cases like these, why don't we just do a series of t-tests? For example, we can test the directional alternative hypotheses:

- The low dose of blood pressure medication lowers BP significantly more than the placebo
- The high dose of blood pressure medication lowers BP significantly more than the low dose

Well, it turns out that doing this first is pretty dangerous business, and the logic goes like this: if our alpha level is 0.05, then the chances of making a Type I error for one test is 0.05; if we perform two tests, then our chances of making a Type I error is suddenly .09025 (near 10%). By the time we perform 10 tests at that alpha level, the chances of us having making a Type I error is 40%. This is called the multiple testing problem or multiple comparisons problem.

To circumvent this problem, in the case of testing three or more means, we use a technique called Analysis of Variance, or ANOVA. A significant result from an ANOVA leads to the inference that at least one of the means is significantly discrepant from one of the other means; it does not lend itself to the inference that all the means are significantly different. This is an example of an *omnibus* test, because it is a global test that doesn't tell you exactly where the differences are, just that there are differences.

You might be wondering why a test of equality of means has a name called **Analysis of Variance**; it's because it does this by comparing the variance between cases to the variance within cases. The general intuition behind an ANOVA is that the higher the ratio of variance between the different groups than within the different groups, the less likely that the different groups were sampled from the same population. This ratio is called an *F ratio*.

For our demonstration of the simplest species of ANOVA (the one-way ANOVA), we are going to be using the `WeightLoss` dataset from the car package. If you don't have the `car` package, install it.

```
> library(car)
> head(WeightLoss)
```

```
     group wl1 wl2 wl3 se1 se2 se3
1 Control   4   3   3  14  13  15
2 Control   4   4   3  13  14  17
3 Control   4   3   1  17  12  16
4 Control   3   2   1  11  11  12
5 Control   5   3   2  16  15  14
6 Control   6   5   4  17  18  18
>
> table(WeightLoss$group)

Control    Diet  DietEx
     12      12      10
```

The `WeightLoss` dataset contains pounds lost and self esteem measurements for three weeks for three different groups: a control group, one group just on a diet, and one group that dieted and exercised. We will be testing the hypothesis that the means of the weight loss at week 2 are not all equal:

- H0 = the mean weight loss at week 2 between the control, diet group, and diet and exercise group are equal
- H1 = at least two of the means of weight loss at week 2 between the control, diet group, and diet and exercise group are not equal

Before the test, let's check out a box plot of the means:

```
> qplot(group, wl2, data=WeightLoss, geom="boxplot", fill=group)
```

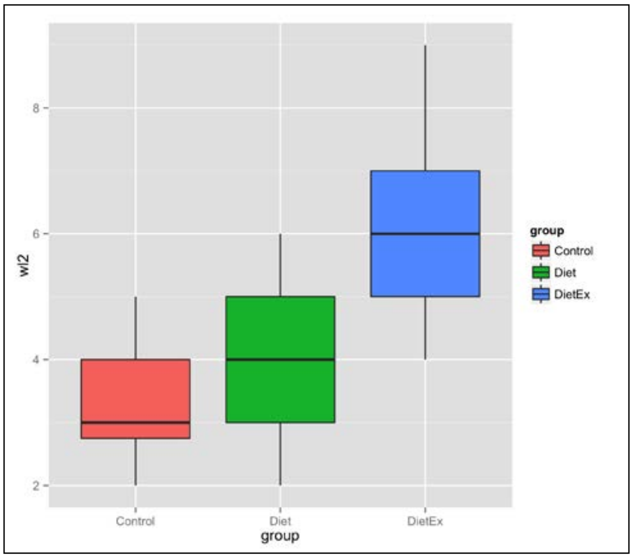

Figure 6.8: Boxplot of weight lost in week 2 of trial for three groups: control, diet, and diet & exercise

Now for the ANOVA…

```
> the.anova <- aov(wl2 ~ group, data=WeightLoss)
> summary(the.anova)
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2  45.28  22.641   13.37 6.49e-05 ***
Residuals   31  52.48   1.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Oh, snap! The p-value `(Pr(>F))` is `6.49e-05`, which is .000065 if you haven't read scientific notation yet.

As I said before, this just means that at least one of the comparisons between means was significant—there are four ways that this could occur:

- The means of *diet* and *diet and exercise* are different
- The means of *diet* and *control* are different
- The means of *control* and *diet and exercise* are different
- The means of *control*, *diet*, and *diet and exercise* are all different

In order to investigate further, we perform a *post-hoc test*. Quite often, the post-hoc test that analysts perform is a suite of t-tests comparing each pair of means (*pairwise t-tests*).

But wait, didn't I say that was dangerous business? I did, but it's different now:

- We have already performed an honest-to-goodness omnibus test at the alpha level of our choosing. Only after we achieve significance do we perform pairwise t-tests.
- We correct for the problem of multiple comparisons

The easiest multiple comparison correcting procedure to understand is *Bonferroni correction*. In its simplest version, it simply changes the alpha value by dividing it by the number of tests being performed. It is considered the most conservative of all the multiple comparison correction methods. In fact, many consider it too conservative and I'm inclined to agree. Instead, I suggest using a correcting procedure called *Holm-Bonferroni correction*. R uses this by default.

```
> pairwise.t.test(WeightLoss$wl2, as.vector(WeightLoss$group))

        Pairwise comparisons using t tests with pooled SD
```

```
data:  WeightLoss$wl2 and as.vector(WeightLoss$group)

       Control Diet
Diet   0.28059 -
DietEx 7.1e-05 0.00091

P value adjustment method: holm
```

This output indicates that the difference in means between the *Diet* and *Diet and exercise* groups is `p < .001`. Additionally, it indicates that the difference between *Diet and exercise* and *Control* is `p < .0001` (look at the cell where it says `7.1e-05`). The p-value of the comparison of just diet and the control is .28, so we fail to reject the hypothesis that they have the same mean.

Assumptions of ANOVA

The standard one-way ANOVA makes three main assumptions:

- The observations are independent
- The distribution of the residuals (the distances between the values within the groups to their respective means) is approximately normal
- Homogeneity of variance: If you suspect that this assumption is violated, you can use R's `oneway.test` instead

Testing independence of proportions

Remember the University of California Berkeley dataset that we first saw when discussing the relationship between two categorical variables in *Chapter 3, Describing Relationships*. Recall that UCB was sued because it appeared as though the admissions department showed preferential treatment to male applicants. Also recall that we used cross-tabulation to compare the proportion of admissions across categories.

If admission rates were, say 10%, you would expect about one out of every ten applicants to be accepted regardless of gender. If this is the case—that gender has no bearing on the proportion of admits—then gender is independent.

Small deviations from this 10% proportion are, of course, to be expected in the real world and not necessarily indicative of a sexist admissions machine. However, if a test of independence of proportions is significant, that indicates that a deviation as extreme as the one we observed is very unlikely to occur if the variable were truly independent.

A test statistic that captures divergence from an idealized, perfectly independent cross tabulation is the *chi-squared statistic* χ^2 statistic), and its sampling distribution is known as a *chi-square distribution*. If our chi-square statistic falls into the critical region of the chi-square distribution with the appropriate degrees of freedom, then we reject the hypothesis that gender is an independent factor in admissions.

Let's perform one of these chi-square tests on the whole UCB Admissions dataset.

```
> # The chi-square test function takes a cross-tabulation
> # which UCBAdmissions already is. I am converting it from
> # and back so that you, dear reader, can learn how to do
> # this with other data that isn't already in cross-tabulation
> # form
> ucba <- as.data.frame(UCBAdmissions)
> head(ucba)
     Admit Gender Dept Freq
1 Admitted   Male    A  512
2 Rejected   Male    A  313
3 Admitted Female    A   89
4 Rejected Female    A   19
5 Admitted   Male    B  353
6 Rejected   Male    B  207
>
> # create cross-tabulation
> cross.tab <- xtabs(Freq ~ Gender+Admit, data=ucba)
>
> chisq.test(cross.tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  cross.tab
X-squared = 91.6096, df = 1, p-value < 2.2e-16
```

The proportions are almost certainly not independent (`p < .0001`). Before you conclude that the admissions department is sexist, remember *Simpson's Paradox*? If you don't, reread the relevant section in *Chapter 3, Describing Relationships*.

Since the chi-square independence of proportion test can be (and is often used) to compare a whole mess of proportions, it's sometimes referred to an omnibus test, just like the ANOVA. It doesn't tell us what proportions are significantly discrepant, only that some proportions are.

What if my assumptions are unfounded?

The t-test and ANOVA are both considered *parametric statistical tests*. The word *parametric* is used in different contexts to signal different things but, essentially, it means that these tests make certain assumptions about the parameters of the population distributions from which the samples are drawn. When these assumptions are met (with varying degrees of tolerance to violation), the inferences are accurate, powerful (in the statistical sense), and are usually quick to calculate. When those parametric assumptions are violated, though, parametric tests can often lead to inaccurate results.

We've spoken about two main assumptions in this chapter: *normality* and *homogeneity of variance*. I mentioned that, even though you can test for homogeneity of variance with the `leveneTest` function from the `car` package, the default `t.test` in R removes this restriction. I also mentioned that you could use the `oneway.test` function in lieu of `aov` if you don't have to have to adhere to this assumption when performing an ANOVA. Due to these affordances, I'll just focus on the assumption of normality from now on.

In a t-test, the assumption that the sample is an approximately normal distribution can be visually verified, to a certain extent. The naïve way is to simply make a histogram of the data. A more proper approach is to use a **QQ-plot** (**quantile-quantile plot**). You can view a QQ-plot in R by using the `qqPlot` function from the `car` package. Let's use it to evaluate the normality of the miles per gallon vector in `mtcars`.

```
> library(car)
> qqPlot(mtcars$mpg)
```

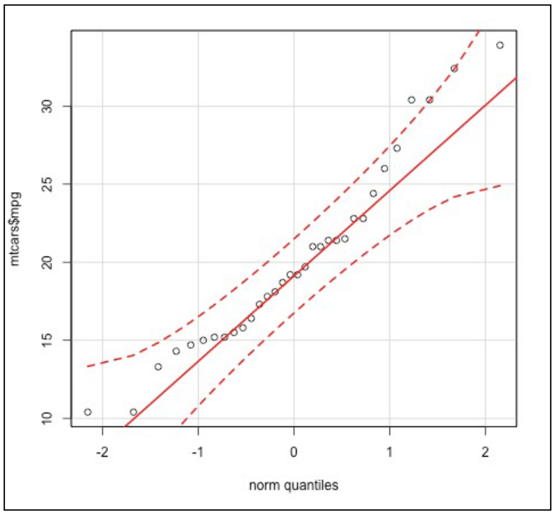

Figure 6.9: A QQ-plot of the mile per gallon vector in mtcars

A QQ-plot can actually be used to compare any sample from any theoretical distribution, but it is most often associated with the normal distribution. The plot depicts the quantiles of the sample and the quantiles of the normal distribution against each other. If the sample were perfectly normal, the points would fall on the solid red diagonal line—its divergence from this line signals a divergence from normality. Even though it is clear that the quantiles for mpg don't precisely comport with the quantiles of the normal distribution, its divergence is relatively minor.

The most powerful method for evaluating adherence to the assumption of normality is to use a *statistical test*. We are going to use the *Shapiro-Wilk test*, because it's my favorite, though there are a few others.

```
> shapiro.test(mtcars$mpg)

        Shapiro-Wilk normality test

data:  mtcars$mpg
W = 0.9476, p-value = 0.1229
```

This non-significant result indicates that the deviations from normality are not statistically significant.

For ANOVAs, the assumption of normality applies to the residuals, not the actual values of the data. After performing the ANOVA, we can check the normality of the residuals quite easily:

```
> # I'm repeating the set-up
> library(car)
> the.anova <- aov(wl2 ~ group, data=WeightLoss)
>
> shapiro.test(the.anova$residuals)

        Shapiro-Wilk normality test

data:  the.anova$residuals
W = 0.9694, p-value = 0.4444
```

We're in the clear!

But what if we do violate our parametric assumptions!? In cases like these, many analysts will fall back on using non-parametric tests.

Many statistical tests, including the t-test and ANOVA, have non-parametric alternatives. The appeal of these tests is, of course, that they are resistant to violations of parametric assumptions—that they are robust. The drawback is that these tests are usually less powerful than their parametric counterparts. In other words, they have a somewhat diminished capacity for detecting an effect if there truly is one to detect. For this reason, if you are going to use NHST, you should use the more powerful tests by default, and switch only if you're assumptions are violated.

The non-parametric alternative to the independent t-test is called the *Mann-Whitney U test*, though it is also known as the *Wilcoxon rank-sum test*. As you might expect by now, there is a function to perform this test in R. Let's use it on the auto vs. manual transmission example:

```
> wilcox.test(automatic.mpgs, manual.mpgs)

        Wilcoxon rank sum test with continuity correction

data:  automatic.mpgs and manual.mpgs
W = 42, p-value = 0.001871
alternative hypothesis: true location shift is not equal to 0
```

Simple!

The non-parametric alternative to the one-way ANOVA is called the *Kruskal-Wallis test*. Can you see where I'm going with this?

```
> kruskal.test(wl2 ~ group, data=WeightLoss)

        Kruskal-Wallis rank sum test

data:  wl2 by group
Kruskal-Wallis chi-squared = 14.7474, df = 2, p-value = 0.0006275
```

Super!

Exercises

Here are a few exercises for you to practise and revise the concepts learned in this chapter:

- Read about data-dredging and *p-hacking*. Why is it dangerous not to formulate a hypothesis, set an alpha level, and set a sample size *before* collecting data and analyzing results?

- Use the command `library(help="datasets")` to find a list of datasets that R has already built in. Pick a few interesting ones, and form a hypothesis about each one. Rigorously define your null and alternative hypotheses before you start. Test those hypotheses even if it means learning about other statistical tests.
- How might you quantify the effect size of a one-way ANOVA. Look up *eta-squared* if you get stuck.
- In ethics, the doctrine of moral relativism holds that there are no universal moral truths, and that moral judgments are dependent upon one's culture or period in history. How can moral progress (the abolition of slavery, fairer trading practices) be reconciled with a relativistic view of morality? If there is no objective moral paradigm, how can criticisms be lodged against the current views of morality? Why replace existing moral judgments with others if there is no standard to which to compare them to and, therefore, no reason to prefer one over the other.

Summary

We covered huge ground in this chapter. By now, you should be up to speed on some of the most common statistical tests. More importantly, you should have a solid grasp of the theory behind NHST and why it works. This knowledge is far more valuable than mechanically memorizing a list of statistical tests and clues for when to use each.

You learned that NHST has its origin in testing whether a weird lady's claims about tasting tea were true or not. The general procedure for NHST is to define your null and alternative hypotheses, define and calculate your test statistic, determine the shape and parameters of the sampling distribution of that test statistic, measure the probability that you would observe a test statistic as or more extreme than the one we observed (this is the p-value), and determine whether to reject or fail to reject the null hypothesis based on the whether the p-value was below or above the alpha level.

You then learned about one vs. two-tailed tests, Type I and Type II errors, and got some warnings about terminology and common NHST misconceptions.

Then, you learned a litany of statistical tests—we saw that the one sample t-test is used in scenarios where we want to determine if a sample's mean is significantly discrepant from some known population mean; we saw that independent samples t-tests are used to compare the means of two distinct samples against each other; we saw that we use one-way ANOVAs for testing multiple means, why it's inappropriate to just perform a bunch of t-tests, and some methods of controlling Type I error rate inflation. Finally, you learned how the chi-square test is used to check the independence of proportions.

We then directly applied what you learned to real, fun data and tested real, fun hypotheses. They were fun... right!?

Lastly, we discussed parametric assumptions, how to verify that they were met, and one option for circumventing their violation at the cost of power: non-parametric tests. We learned that the non-parametric alternative to the independent samples t-test is available in R as `wilcox.test,` and the non-parametric alternative to the one-way ANOVA is available in R using the `kruskal.test` function.

In the next chapter, we will also be discussing mechanisms for testing hypotheses, but this time, we will be using an attractive alternative to NHST based on the famous theorem by Reverend Thomas Bayes that you learned about in *Chapter 4*, *Probability*. You'll see how this other method of inference addresses some of the shortcomings (deserved or not) of NHST, and why it's gaining popularity in modern applied data analysis. See you there!