Lecture 6 - Statistical Hypothesis Testing

Will Calandra

Statistical Hypothesis Testing

In this lecture, we'll cover the basics of supporting your results empirically by using what we call statistical hypothesis testing. If you've ever heard the term "statistically significant," this is what we are talking about! Statistical hypothesis testing allows us to take our data (ex: the means in a dataset) and determine how extreme the phenomena we observe actually are!

Some definitions

- Statistical Hypothesis: a broader assumption/claim made by the researcher about the population of interest
    - Ex: A principal at a school thinks her students score an average of 7/10 in exams.
- Null hypothesis: The assumption/claim about the larger population
    - Ex: The mean score of students at this school is 7/10.
- Alternative hypothesis: Some proposition made by the researcher to test against the null hypothesis. This is "valid" if there is evidence in favor of the alternative hypothesis from observed data.
    - Ex: The mean score of students at this school is not 7/10.
- P-Value: The probability of finding the observed or more extreme results when the null hypothesis is true
    - Ex: From my sample of 30 students, there is a 4.2% chance that I get my observed mean of 6.4/10 or something more extreme (further from 7/10) if the actual mean score of students at this school is 7/10. My p-value is 0.042 in this example.
- Level of significance: This is how confident you want to be in your results, or basically what p-value you are comfortable with as the benchmark for rejecting the null hypothesis. Convention is 0.05 (people think this is arbitrary and it is, but we are generally comfortable with this threshold. A p-value threshold selection should depend on your tolerance for error types, which are discussed below).
    - Ex: I conduct the test of the school's mean test scores at a significance level of 0.05. So if my p-value is less than 0.05, I reject the null hypothesis and say I have significant evidence that the mean test scores of students at this school is not actually 7/10.
- Type I error: This comes right from the significance level; it's the probability that the null hypothesis is true when we actually rejected it from our test. It is equal to the significance level.
    - Ex: In my test above, I have a 5% chance of the null being true when I rejected it. So there is a 5% chance that the actual mean test scores of students at this school is 7/10 when I reject the claim (or, more precisely, the p-value is the % chance, or 4.2%. By declaring the 5% threshold, that is the zone where we could encounter a type I error).
- Type II error: The case in which you accepted (or failed to reject) a false null hypothesis. The probability of not making this mistake is known as the power of a test, which involves some calculations.
    - Ex: Say my p-value above was greater than 0.05, so I failed to reject the null hypothesis even though the mean test scores are actually not equal to 7/10.

Conditions

We can't conduct statistical hypothesis tests in every case. Each test has its separate conditions (you'll see why), but generally here are some conditions to think about when performing these tests:

- If you sample, it must be random. This makes sure there are no confounding variables in your sample data and you are not cherry picking for results. Again, some uncertainty needs to be there.
- Your observations should be independent or have no effect on subsequent observations. It is often difficult to control for this, but we try our best in our experimental setup!
- The data must come from a normal distribution or a solid sample size (generally, by the CLT, over 30).

Why sample? It allows our statistical computations and assumptions to be valid, and IRL, sampling is more efficient because it is virtually impossible (and costly) to get the data for the entire population of interest.

Notes

^These are a lot of definitions and conditions, so it is easier and more intuitive to learn about this as we go through some exercises.

General workflow for hypothesis testing:

- Check conditions for the test
- State the hypothesis, pick a significance level
- Analyze sample data
- Interpret results

One Sample T-test

Now that we've covered the basics and talked conceptually, we can start with some exercises. There are many tests R can conduct (we'll go through some), but the first one is called a one sample t-test. It takes one sample, calculates the mean value, and evaluates it against a hypothesized value.

Let's set up our data:

```
## Get sample data
set.seed(30)
x <- as.data.frame(rnorm(100, mean = 6.4, sd = 1))
```

We define our sample data (assumptions taken care of), x, as a normal distribution with a mean of 6.4 and a standard deviation of 1. Say this is our sample from the school - is it fair to say that the mean scores of the school are equal to 7?

```
## Run one sample t-test against mean 7
t.test(x, mu = 7)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -6.3811, df = 99, p-value = 5.684e-09
## alternative hypothesis: true mean is not equal to 7
## 95 percent confidence interval:
##  6.114636 6.534644
## sample estimates:
## mean of x
##   6.32464
```

Look at this nice output! R already set up our hypotheses (alternative is mean is not equal to 7, null being equal to 7). R gives us a 95% confidence interval of the true mean based on this sample, it also gives us the observed sample mean, and we get the corresponding p-value! Note that the t-score and the corresponding degrees of freedom (df) are given. That's how the p-value is calculated, from the t-distribution (basically a normal curve with wider tails so it accounts for more wonky phenomena).

We can also conduct this t-test to give it a direction. What if the principal of this school wanted to claim that their school's test scores are better than a rival school's test scores of 6.8? How do our hypotheses change? We say the null hypothesis is that mean test scores are 6.8 (can never have a null hypothesis with direction), but the alternative hypothesis is that the school's mean test scores are greater than 6.8. We conduct this below:

```
## One sample t-test, alternative as greater than
t.test(x, mu = 6.8, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  x
## t = -4.4914, df = 99, p-value = 1
## alternative hypothesis: true mean is greater than 6.8
## 95 percent confidence interval:
##  6.148909      Inf
## sample estimates:
## mean of x
##   6.32464
```

That p-value is astronomical, so we don't have evidence to support our claim of greater test scores. Notice that 'inf' on the confidence interval - that's normal, because we are really only testing the lower bound here. This p-value is high, so let's flip around the signs to see if this principal's test scores are actually LESS than the rival school.

```
## One sample t-test, alternative as less than
t.test(x, mu = 6.8, alternative = 'less')
```

```
##
##  One Sample t-test
##
## data:  x
## t = -4.4914, df = 99, p-value = 9.59e-06
## alternative hypothesis: true mean is less than 6.8
## 95 percent confidence interval:
##      -Inf 6.500371
## sample estimates:
## mean of x
##   6.32464
```

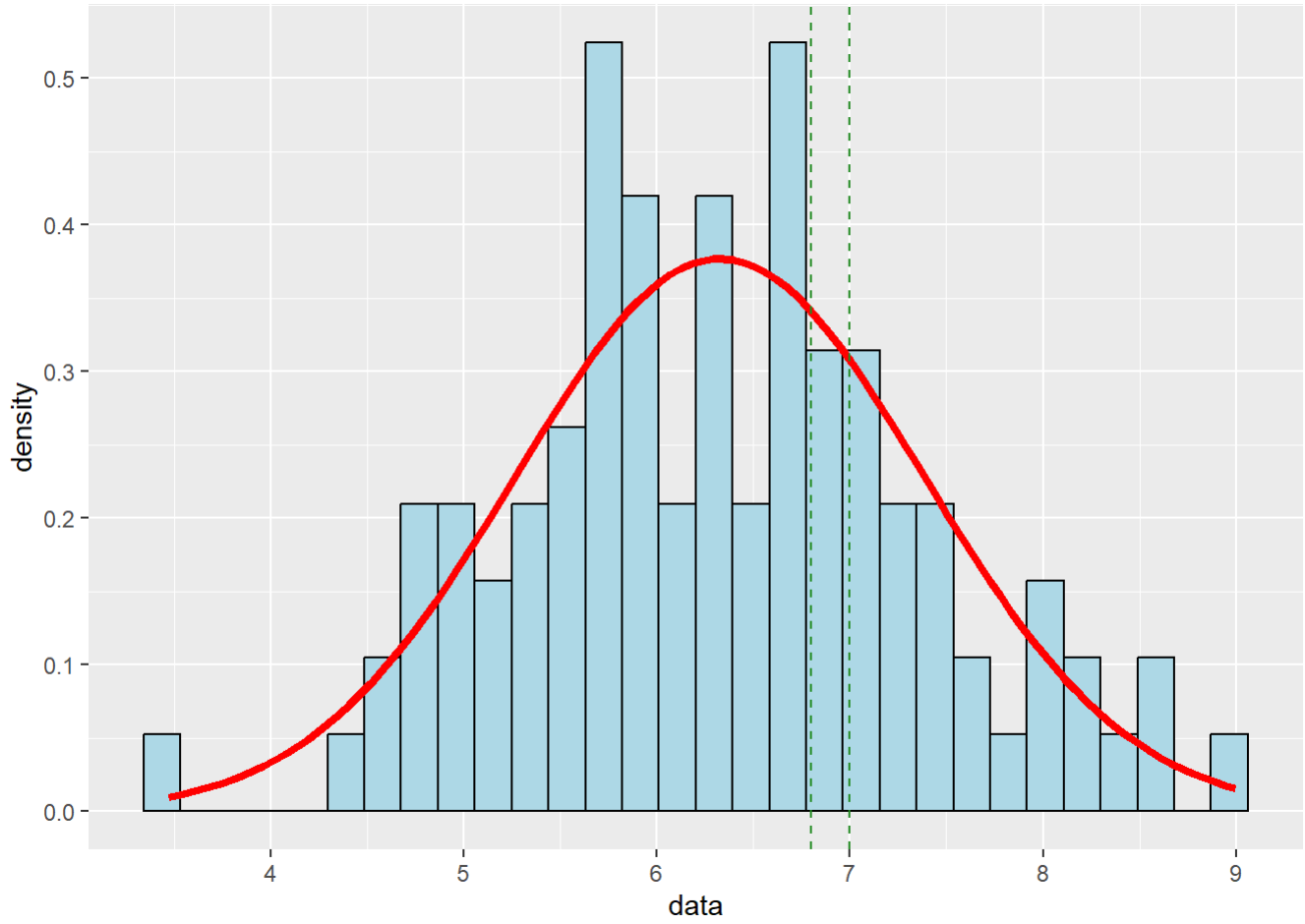
Uh oh, it's significant that the principal school's test scores are lower than the rival school. Ouch.

If we put our data viz skills to the test, you can see why we get the results that we do…

```
## Visualize distributions
library(ggplot2)
colnames(x) <- "data"

ggplot(x, aes(x=data)) + geom_histogram(aes(y = after_stat(density)), color = 'black', fill = 'l
ightblue') +
  stat_function(fun = dnorm, args = list(mean = mean(x$data), sd = sd(x$data)), lwd = 1.5, color
= 'red') +
  geom_vline(xintercept = 6.8, linetype = "dashed", color = "forestgreen") +
  geom_vline(xintercept = 7, linetype = "dashed", color = "forestgreen")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The dashed lines represent the hypothesized values (null hypotheses), the histogram is our sample data, and the curve is the distribution in which our data came from (usually it's an estimate we build from the sample data). You can see why we made the conclusions we did before!

Two Sample T-test

In the above example, we already had the average of the rival school. But let's say we didn't! Then we would have two samples in which we'd need to compare.

```
## Two-sample t-test
# Generate data
set.seed(30)
y <- as.data.frame(rnorm(100, mean = 6.8, sd = 0.5))
```

In a two-sample t-test, the null hypothesis is that the difference in the two population means is 0 (or they are equal), and the alternative hypothesis is that the difference is not 0 (or the means are not equal). Our test will account for differences in variance between the two groups, so our assumptions pass. Let's let R tell us the answer here!

```
# Test for differences
t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -3.6988, df = 145.59, p-value = 0.0003063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6715453 -0.2038145
## sample estimates:
## mean of x mean of y
##   6.32464   6.76232
```

We do have evidence that the means between the two schools' test scores are significant. But in what direction? Conducting that test below:

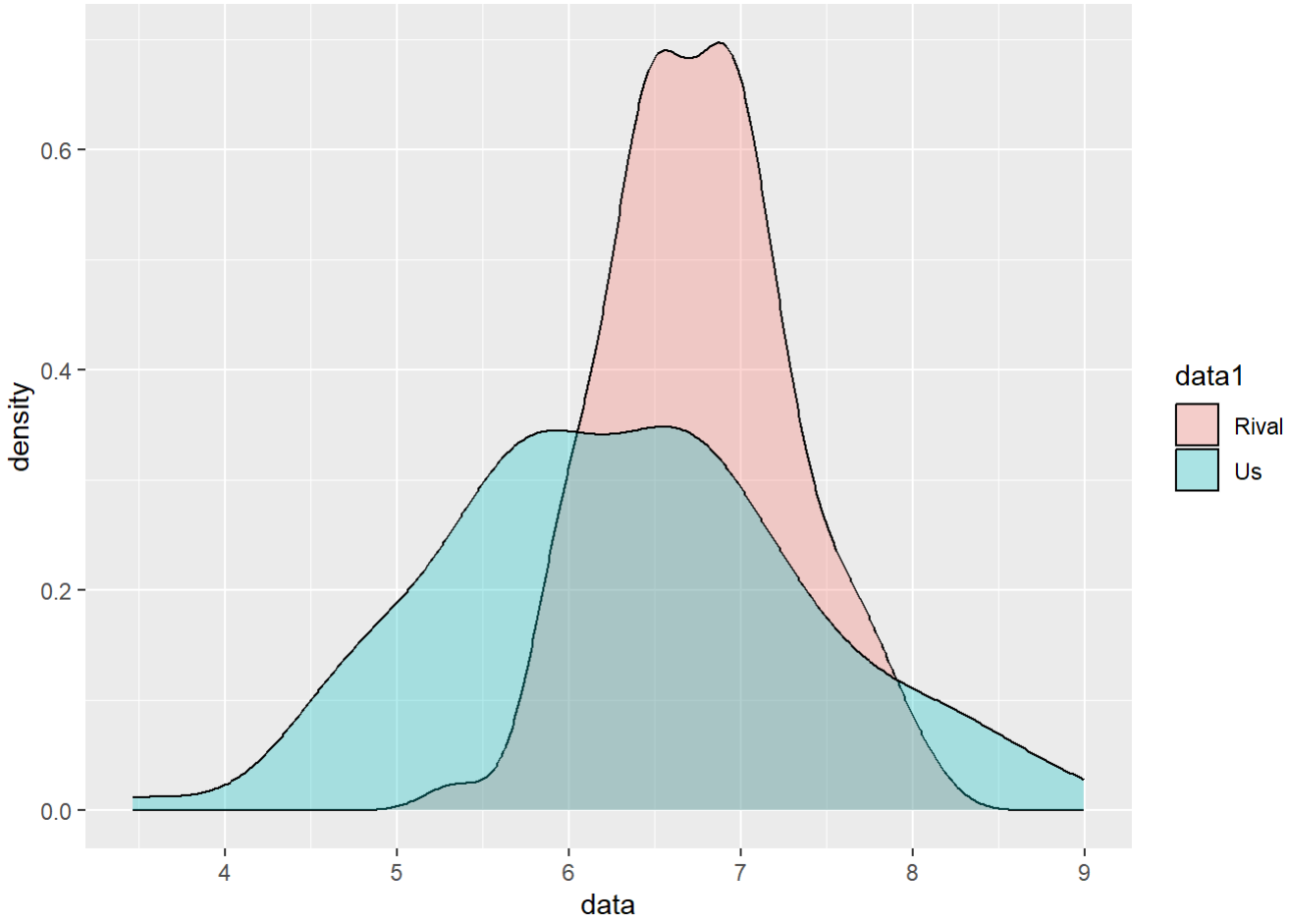
```
## Two sample t-test, one-tailed
t.test(x,y, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -3.6988, df = 145.59, p-value = 0.0001531
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2417989
## sample estimates:
## mean of x mean of y
##   6.32464   6.76232
```

We do have enough evidence to say that the rival school's mean test scores are greater than the principal's school. This is an example of the difference between what we call a one-tailed (direction) and two-tailed (no direction, just testing equality) test.

```
# Visualize distributions
colnames(y) <- "data"
Schools <- rbind(x,y)
Schools$data1 <- ifelse(seq.int(nrow(Schools)) < 101, "Us", "Rival")
Schools$dataX <- ifelse(Schools$data1 == "Us", Schools$data, NA)
Schools$dataY <- ifelse(Schools$data1 == "Rival", Schools$data, NA)

ggplot(data = Schools, aes(x = data, group = data1, fill = data1)) + geom_density(alpha = 0.3)
```



A visual aid helps us see the distributions of the two schools. We can see that generally, most of the red curve is to the right of the blue curve, and our test confirms this result empirically. You can also tell the means are different, which confirms our result from the first test.

Proportion Testing

Proportion testing takes the same idea as t-testing but works with the normal distribution (z) rather than the t distribution (t). This is because the standard deviation of the proportion is a function of itself, so it's a bit more reliable of an estimate where we don't need wide tails.

For our example, we are going to simulate a coin flipping game, where we have two teams flip a coin 100 times. Whichever team accumulates more "heads" flips in the game wins! Assumptions will pass if we have a binary outcome, enough sample, and independence. We do!

Let’s set up our data. We’ll randomize results by “sampling” a coin with its respective probabilities of heads and tails. However, notice something: team 2 is cheating! They are using a weighted coin that is more likely to land on heads than tails. Let’s see if we can catch them using our hypothesis testing! Code below sets up our data:

```
## Coin initialization
coin <- c("heads", "tails")
n <- 100

## Team 1, fair coin
Prob1 <- c(0.5,0.5)
set.seed(30)
samps1 <- as.data.frame(table(sample(coin, n, replace = TRUE, prob = Prob1)))

## Team 2, weighted coin
Prob2 <- c(0.6,0.4)
set.seed(30)
samps2 <- as.data.frame(table(sample(coin, n, replace = TRUE, prob = Prob2)))

## See data
samps1
```

```
##    Var1 Freq
## 1 heads   39
## 2 tails   61
```

```
samps2
```

```
##    Var1 Freq
## 1 heads   68
## 2 tails   32
```

R needs a summary table of our frequencies so we can place the proper inputs into its prop.test() function. Done here:

```
# Compare
summary <- cbind(samps1, samps2)
summary <- summary[-c(3)]
summary
```

```
##    Var1 Freq Freq.1
## 1 heads   39     68
## 2 tails   61     32
```

Whoa, that’s a big win for team 2! Let’s evaluate how extreme these results are - let’s call our two-proportion test!

```
# Significant win?
prop.test(x = c(39,68), n = c(100,100), alternative = "less", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(39, 68) out of c(100, 100)
## X-squared = 15.757, df = 1, p-value = 3.601e-05
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.1689876
## sample estimates:
## prop 1 prop 2
##   0.39   0.68
```

The p-value of this test is very low, so we did witness a statistically significant event. We can also use the confidence interval to show that the difference in proportions is quite large. This means that it is highly unlikely that the coins were fair for both sides. Let's take a look at team 2's results to see if they used a weighted coin:

```
# Cheated? One Sample proportion
prop.test(x = 68, n = 100, p = 0.5, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  68 out of 100, null probability 0.5
## X-squared = 12.25, df = 1, p-value = 0.0002326
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5942311 1.0000000
## sample estimates:
##    p
## 0.68
```

Notice the difference in the syntax for a one-sample proportion test (one value for x, n, and inclusion of p - hypothesized value). We see that we have convincing evidence that team 2 did something funky with the coin. Uh oh.

Chi-Sq Test for Association + Goodness of Fit Test

The last set of tests we will cover involve the Chi-Sq distribution. The Chi-Sq distribution comes from a sum of independent squared random normal variables - it really has no parameters other than degrees of freedom (df), so it is useful for many statistical analyses.

We can use Chi-Sq tests for categorical variables. We will first use the test for association to determine if there is an association between variables (null is if the variables are not associated, alternative is that the variables are associated). We will then use the goodness of fit test to determine how closely our data resembles some real-world benchmark (null is if observed values are not different than expected values, alternative is if observed values are different than expected).

For the Chi-Sq test for association, we need a contingency table, in that the categories for one variable are in the rows and the categories for another variable are in the columns. Contingency tables are also referred to as "cross-tabs" or "two-way tables," which you may have seen before.

In our example, we'll use the iris dataset to see if there is first an association of sepal length size to species type. We will then test it against an expected distribution of sepal length size to check results.

First, the data and a visual aid:

Source: https://statsandr.com/blog/chi-square-test-of-independence-in-r/ (https://statsandr.com/blog/chi-square-test-of-independence-in-r/)

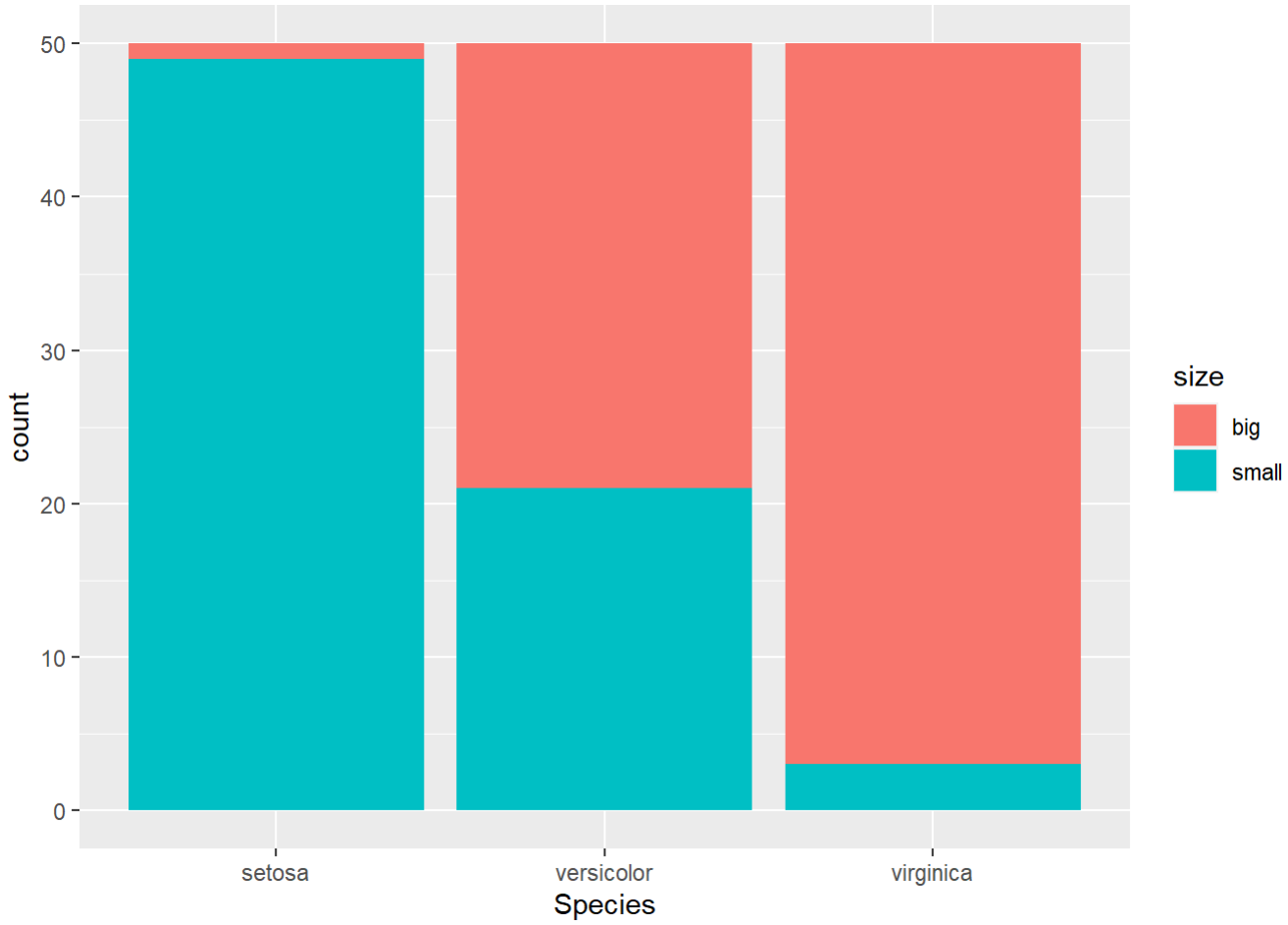
```
## Chi-sq test for association
data <- iris

data$size <- ifelse(data$Sepal.Length < median(data$Sepal.Length),
                    "small", "big"
)

# Generate contingency table
Table <- table(data$Species, data$size)
Table
```

```
##
##              big small
##   setosa       1    49
##   versicolor  29    21
##   virginica   47     3
```

```
# Visual Aid
ggplot(data) + aes(x = Species, fill = size) +
  geom_bar()
```



You can see the large differences between sepal length and species. This probably means there is an association between our two variables - sepal length and species. Let's see!

```
# Run chi-sq test
test <- chisq.test(Table)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  Table
## X-squared = 86.035, df = 2, p-value < 2.2e-16
```

```
test$expected
```

```
##
##                  big    small
##   setosa     25.66667 24.33333
##   versicolor 25.66667 24.33333
##   virginica  25.66667 24.33333
```

Notice how we do get a significant result. We can also save our test as an object and access its values; this is where I pull up the values in the case that there was no association. Notice how different our observed values are from these! We have evidence for an association between sepal length and species.

Using the Chi-Sq distribution, we can also test how closely the observed values follow some distribution. For instance, let's say we expected 1/6 of big sepal lengths coming from setosas, 1/3 coming from versicolors, and 1/2 coming from virginicas. We can test our observed values against the hypothetical with a goodness of fit test:

```
# Test against expected probability of big per species
big <- c(1, 29, 47)
test <- chisq.test(big, p = c(1/6, 1/3, 1/2))
test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  big
## X-squared = 13.221, df = 2, p-value = 0.001346
```

```
test$expected
```

```
## [1] 12.83333 25.66667 38.50000
```

Notice how we add the probability vector, and the expected values of the test change. We can see that there is a significant difference between our observed values versus the expected result! Note: condition for the GOF test is that all expected counts must be >5 (which they are here).

Notes

This concludes our lecture on statistical hypothesis testing; my goal was to give you an overview of how tests work and how they are relatively easy to run in R. Some general notes while you perform these analyses…

- Don't forget to check your conditions! We were fine for our datasets, but IRL, it is important that you validate that your conditions are satisfied before you run the tests.
- Use a visual aid as often as possible. This will help validate your results and look at your sample data for context.
- Confidence intervals: look them up. We discuss them briefly in lecture, but they are foundational to statistical tests and an alternative way to test hypotheses.
- Your answers are not certain! Interpret your results and know that your results are not definitive. The data you collect is inherently uncertain. Therefore, it is important you understand that tests do a good job of extracting meaning from uncertainty but do not eliminate it from your analysis.
- Some other statistical tests to check out…
    - One way ANOVA test - for a categorical variable interacting with an interval variable (numeric)
    - McNemar test - for binary outcomes - often used in UX (i.e. task completion rates)
    - Not really a test, but correlation matrix - this provides a useful snapshot of the relationships in your data