Section 6
Estimating a Proportion

6.1 – Deriving a Confidence Interval

In the last two sections, we talked about one method of statistical inference: hypothesis testing. Remember that the goal of inference is to use our *sample* data to infer what our *population* might look like. Hypothesis testing aims to determine if our data could have been reasonably generated from a population with a certain assumption or hypothesis.

There may be times though when we collect data and don't have a reasonable hypothesis to place upon a theoretical population. But what if we still want to use our data to infer what the population would look like? We can do this through estimation – that is, trying to determine what the value of a parameter might be based upon sample data.

Constructing an interval

How do we go about estimating a parameter? This is something we already did during the kissing the "right" way activity. Let's revisit this example now.

> *Example*: A German bio-psychologist, Onur Gunturken, was curious whether the human tendency for right-handedness and right-eye dominance manifested in other situations. He investigated whether kissing couples would turn their heads to the right or the left. He and his researchers observed couples in public places across the US, Germany, and Turkey, and found that of the 124 couples observed, 80 of them turned their heads to the right. What is the best estimate for the percentage of *all* couples that lean their head to the right?

We can thus use sample data to estimate what our population might look like. However, there is a great deal of sample to sample variation just by random chance of who you select for a survey. Thus, estimates we obtain from samples are naturally going to be wrong. Thus, we ideally want to get a wider _____ to estimate a parameter. To better understand this variation, we can use information we know about the distribution of a proportion. Recall that for a binomial random variable X with parameters n and p, we learned that

$$E(X) = np \qquad \text{Var}(X) = np(1 - p)$$

If we want to redefine this binomial variable X as a proportion, we can simply take this count and divide it by the sample size, n.

$$E(\hat{p}) = E\left(\frac{X}{n}\right) =$$

$$\text{Var}(\hat{p}) = Var\left(\frac{X}{n}\right) =$$

Because we don't know the population proportion p, the next best thing we can use to estimate these values is to use the sample proportion, $\hat{p}$.

From simulating various scenarios based on a percentage in TinkerPlots, we could notice that the shapes of these binomial distributions were fairly bell shaped. We'll formalize this idea next section, but for now, let's assume that we can assume that the distribution of possible sample proportions, $\hat{p}$, is normal. Let's start with a basic fact about the normal distribution:

$$P(-1.96 < Z < 1.96) = 0.95$$

Let's now standardize the sample proportion $\hat{p}$ so that it also follows a standard normal distribution and put it into this inequality:

$$-1.96 < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < 1.96$$

We call these upper and lower bounds a *95% confidence interval for p*. In practice, we typically write out confidence intervals like we would any open interval, using parentheses separated by a comma. While 95% is a fairly common level we could pick for confidence, you could pick a different confidence level and find the appropriate values from a standard normal distribution to get a confidence interval for that confidence level. Thus, we would generically write out the interval as shown below:

Note that the approximation to the normal distribution does not always hold – if the sample size is small or the proportion is very large or small, the distribution will not look very normal and this will be a bad approximation. Thus, we only use this interval when the following sample size conditions are met:

One thing we can notice about the structure of this confidence interval is that the distance we go out on either side of $\hat{p}$ is the same. We call this distance from $\hat{p}$ to the ends of the confidence interval the _____.

You've probably seen this term used before in election contexts as well – this helps to account for how wrong you might be in a political poll due to sampling variability. But remember that this assumes a perfectly random sample was taken with no non-response or bias issues, and political polls come with plenty of this. This NYTimes article explains many other sources of error introduced in political polling due to the imperfect nature of sampling.

With that said, let's assume a perfectly random sample in our kissing example for calculating a confidence interval:

> *Example*: A German bio-psychologist, Onur Gunturken, was curious whether the human tendency for right-handedness and right-eye dominance manifested in other situations. He investigated whether kissing couples would turn their heads to the right or the left. He and his researchers observed couples in public places across the US, Germany, and Turkey, and found that of the 124 couples observed, 80 of them turned their heads to the right. Find a 95% confidence interval for p, the population proportion of couples that lean their head to the right. Interpret this interval and determine the margin of error.

Finding an interval in R

We've already learned how to calculate a confidence interval in R, we just didn't use that output previously! Using the `binom.test` function produces this interval as long as the default "`two.sided`" option is given for the alternative. Thus, to use this function for a confidence interval, you can use the following code:

```
binom.test(x, n, conf.level=0.95)
```

Important note: the `binom.test` function finds the confidence interval using a binomial distribution rather than a normal approximation. Thus, it is a bit more accurate than the equation we derived ourselves; however, because binomial distributions are not perfectly symmetric (unless $p = 0.5$), the binomial interval will not be centered at $\hat{p}$ like it is in the equation. Unless a specific method is specified, both the equation and R methods are valid for computing a confidence interval on homework/exams.

Note that I didn't specify a value for the null proportion – since we are not conducting a hypothesis test here, this value is not relevant, and it defaults to the value of 0.5 if not specified. What would happen if we tried different levels of confidence?

Example: Find a 90% confidence interval for the previous kissing example.

Example: Find a 99% confidence interval for the previous kissing example.

6.2 – Additional Estimation Topics

Connecting confidence intervals and hypothesis testing

The reason that we see confidence intervals provided with our testing output is because there is actually an intimate relationship between hypothesis testing and confidence intervals:

- We can conduct a _____ hypothesis test with significance level α by computing a _____% confidence interval.
- Reject H_0 if _____ is not contained in your interval.
- Fail to reject H_0 if _____ is contained in the interval.

Conceptually, remember that confidence intervals provide a range of reasonable values for a parameter at some level of confidence. Hypothesis testing in some way is the exact opposite of that idea, as it's testing to see if some value for a parameter (the null) is or isn't reasonable, and deems it unreasonable if the *p*-value reaches a pre-determined threshold. It turns out that the threshold of the significance level that we use is nicely tied to the confidence level!

> *Example:* Using the confidence interval generated in the R output for the kissing example, draw a conclusion for a hypothesis test with hypotheses H_0: $p = 0.5$, H_a: $p \neq 0.5$.

Bootstrap confidence intervals

When we did this kissing example for our activity earlier, we didn't use a formulaic approach to find the confidence interval – instead, we used the bootstrapping technique. This technique runs under the assumption that our sample is a good representation of our population (which should be reasonable if it is randomly sampled) and uses that sample as an approximation for what the population looks like. Thus, for the kissing example, if one of our 124 couples leaned to the right, we are saying that this couple represents 1/124 of the couples in our population. In total, this means that when we sample from the population, there is an 80/124 (64.5%) chance that we pick a right-kissing couple.

To determine a confidence interval under this assumption, we simply re-sample from this new population many times, and see what the middle 95% of our new samples is. Just as we did for hypothesis testing, we can use the same simulation approach with sampling and a for loop. To sample from our new population, we need to first create a vector to re-sample from. Let's let the value 1 represent right-kissers, and the value of 0 represent the left-kissers.

```
data = c(rep(1, 80), rep(0, 44))
```

Here, I'm using the `rep()` function again to repeat the same value multiple times rather than type out a vector of 80 ones and 44 zeros. Using the `c()` function on both of these vectors joins them together in one vector.

Now, we want to sample from this vector to get a new bootstrapped sample. We can use the `sample()` function again to do this, making sure to have sampling with replacement turned on.

```
sample(data, 124, replace=TRUE)
```

This produces a new vector of 1s and 0s to represent right and left kissing. To summarize this into the percentage of right-kissers, we can simply just take the mean of this vector, as finding the percentage would be just adding up the 1's and dividing by the sample size, 124.

```
mean(sample(data, 124, replace=TRUE))
```

Thus, with this setup, we can now use the approach of having a placeholder vector for percentages and then run this through a for loop many times to build up our simulation.

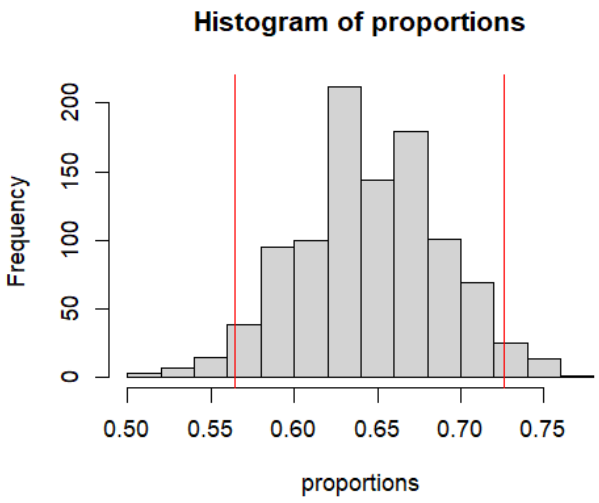
```
proportions = rep(0, 1000)
for (i in 1:1000) {
  proportions[i] = mean(sample(data, 124, replace=TRUE))
}
```

We now have a vector of helpers that tells us all of the different percentages of right-kissers we got in each bootstrapped sample. To create a 95% confidence interval now, we just need to find the middle 95%. This can be found by finding the 2.5th and 97.5th percentiles, which we can do with the `quantile()` function.

```
quants = quantile(proportions,
c(0.025, 0.975))
```

This produces an interval that is nearly consistent with the interval we found with the equations and with `binom.test`. Of course, since bootstrapping is based on simulated results, the interval may change slightly each time you run it. Increasing the number of repetitions from 1000 to a larger number will increase the consistency of your results! We can also visualize what this CI would look like on a histogram.

```
hist(proportions)
abline(v=quants[1], col="red")
abline(v=quants[2], col="red")
```



So how might you choose among these three methods: the normal approximation equations we derived, the confidence interval based on a binomial distribution, or the bootstrapping method? Let's start with the normal approximation: remember that this approximation is only valid when the sample size conditions mentioned above are met. So what happens when these aren't met?

> *Example*: Consider the previous helper/hinderer scenario where 14 of the 16 babies chose the helper toy. Compute a 95% confidence interval for the percentage of babies that choose the helper toy using each of the three methods.

Let's summarize these three methods and their limitations. The equation has obvious ones – the binomial distribution for this data is very skewed to the left, which shows up in our histogram of the bootstrap simulation. This gave us an obviously erroneous upper limit of the interval.

The bootstrap is improved, but it's also a bit problematic. It provides an upper limit of 1 or 100%, which we know is not the case in our population, as two babies chose the hindering toy. Thus, it doesn't do a great job with the boundaries (0 or 100%) or skewed distributions

The method used in `binom.test` is the most accurate, but it is a "conservative" method. Due to binomial being discrete in nature, there's often no way to get an exact middle 95% of a sampling distribution, as you have to either include or leave out an entire count/percentage. Thus, the method used to determine this confidence interval (known as the Clopper-Pearson method) actually is more confident than the percentage specified, meaning that it is wider than it actually needs to be. That being said, this is probably the most accurate method of the three generally.

Thus, since we know the root distribution for percentages/proportions in this scenario is based on a binomial distribution, methods that use this distribution are going to be the most accurate. However, there will be cases we see in the next section where the population distribution is unknown, and bootstrapping will relatively be much more useful!

6.3 – Additional Practice

Example: In the last section, the additional practice problem described a scenario about testing whether red-headed people are more commonly left-handed. The data collected in that study found that 40 out of 125 randomly sampled red-headed people were left-handed. Find a 95% confidence interval for the proportion of red-headed individuals that are left handed and give an interpretation of this interval.

Would a 90% confidence interval be wider, narrower, or the same width as the one you found above? Explain.