Understanding Confidence Intervals
Authors: Ian Castro & Gregory Du

Before you read this guide, please read Section 13 of the Data 8 textbook, Inferential Thinking. In particular, Section 13.4 focuses on hypothesis testing with confidence intervals. This guide is a supplement to that section.

The Basics of Confidence Intervals

Confidence intervals can be a little tricky and quite nuanced, so let's carefully investigate confidence intervals to get a better understanding of how they work, and what they mean. Note that this guide will not go over the process for generating confidence intervals; section 13 of the textbook is a wonderful resource to understand how we generate confidence intervals in the first place. I want to cover some of the traditionally tricky aspects of interpreting confidence intervals that can trip us up sometimes.

What is "Confidence"

We throw around the word confidence a lot so let's be precise in defining what exactly we mean when we say we're generating a 95% confidence interval. 95% refers to our confidence in the ***process*** we used to generate confidence intervals. Put another way we are 95% confident that the process we used will generate a confidence interval that captures the population parameter. As a consequence, we would expect 95% of all of the confidence intervals we generate to capture the true population parameter. If we generated 1000 95% confidence intervals, we'd expect 95% of those intervals (or 950 intervals) to contain the population parameter we're estimating.

What Does the Confidence Interval Tell Us (And What Doesn't it Tell Us)?

A confidence interval is simply an *informed guess* of the range some parameter may be in, and our *confidence level* tells us about how often our process of estimation is correct. Ex. If we generate a 95% confidence interval of average income amongst SF Bay residents to be [70000, 150000], we're providing an *informed guess* of the average income amongst SF Bay residents (our parameter), and we'd expect our process of estimation to generate an interval containing the true average income about 95% of the time. With that in mind, let's go through a list of possible interpretations of confidence intervals, and determine whether we can draw such a conclusion.

For the following examples, we're attempting to estimate the average amount of money in dollars spent by adults during the holiday season in the United States. We've generated our 95% confidence interval to be [1549, 1640].

Are the following interpretations correct?

Our estimate for the average expenditure of adults in the US during the holiday season is between $1549 and $1640, and we are 95% confident in the process used to generate this interval.

Correct. This is precisely the definition of a confidence interval.

95% of adults in the United States spend between $1549 and $1640 during the holiday season.

Incorrect. This is a *very* common misinterpretation about confidence intervals. Remember our confidence interval is estimating the *average expenditure of adults*, but makes absolutely **NO** claim about what percentage of adults actually spend between $1549 and $1640. If we take a quick step back and evaluate this interpretation you should be able to see that it doesn't really make sense. Do we actually expect 95% of all adults in the United States (around 200 million people!) to all somehow spend between $1549 and $1640? This would mean almost 200 million people all spend within $91 dollars of each other, which definitely seems wrong.

There is a 95% chance that the true average expenditure by adults in the US is between $1549 and $1640.

Incorrect. This is another common misinterpretation of confidence intervals. Once a confidence interval has been set, it no longer makes sense to talk about the probability that a value falls within that range. Think about it this way. There is some *true average expenditure* by adults in the US during the holiday season. This is a definite value; it's a parameter of our population. We might not know what this value is (in fact, we probably don't), but we know that this value exists. Either this unknown value falls in the range [1549, 1640], or it falls outside the range. There's no other option! Once the confidence interval has been determined, it will either contain the population parameter, or it won't so we cannot say there's some probability that it will contain the parameter. See the ***coin flip analogy*** section below if this topic is a little confusing.

Using the same process to generate another 95% confidence interval, there is a 95% chance that the next confidence interval I generate will contain the true average expenditure by adults in the US.

Correct. Well wait a minute. How can this be correct if the previous statement was incorrect? This is a very subtle difference, but notice that in this example, there **is no defined confidence interval**. Remember from our definition of the confidence interval above, we have 95% confidence in the process used to generate confidence intervals, which means we would expect 95% of all confidence intervals we generate using this process to contain the true average expenditure of adults in the US during the holidays. This means before we've taken a random sample and created a confidence interval from this sample, there's a 95% chance that the confidence interval we create will contain the true population parameter. Since there is not a concrete confidence interval, we can still speak about probability, since we're discussing ***the probability associated with our process*** of generating confidence intervals, **not** the probability associated with a single well-defined interval.

We are 95% confident that the average expenditure of adults in the US during the holiday season is between $1549 and $1640.

Correct. This is fine because we aren't talking about the **chance** associated with an interval (which we've already discussed as being incorrect), rather we're describing our confidence in an interval, and the word "confidence" is a flag that indicates we are talking about our confidence in the process used to generate the interval. If we're 95% confident in the process we used to generate confidence intervals (we expect 95% of the confidence intervals we generate to contain the population parameter) it's fine to say we're 95% confident in our generated interval. I do want to caution people from using this interpretation too liberally. It is very easy to conflate this with the 95% chance statement described above, and it can feel like these two are very similar. Be very careful in interpreting confidence intervals this way so as to not accidentally switch to discussing probabilities.

The Coin Flip Analogy:

Here's the coin flip analogy. This was the way confidence intervals were explained to me, which has really helped me to understand the nuanced difference between the 2 statements provided above. Suppose I have with me a fair coin.

Before I flip my coin, I ask you what's the probability the coin will land heads, and what's the probability it will land tails. You'd probably say there's a 50% chance that I'll get heads, and similarly, a 50% chance that I'd get tails, and you'd be correct!

Now I flip the coin, and it lands. Can I still say that there's a 50% chance that the coin is heads? **No, because the coin has already landed!** Either the coin landed heads, or it didn't. The outcome is already determined here, so it doesn't make sense to say there's a 50% chance that it's heads anymore.

To summarize: before flipping the coin or before an outcome has been determined, we can make claims about the probability of outcomes of our process, but once an outcome has been determined, it doesn't make sense to assign probabilities to the outcomes anymore.

Let's connect this back to confidence intervals. Before I generate my 95% confidence interval is it correct to say there's a 95% chance that the confidence interval my process generates will contain the true population parameter? Sure, this is exactly like before I flip my coin. I don't have a concrete confidence interval, and I'm making a claim about the probabilities of the outcome of my process. Once I generate a confidence interval is it correct to say there's a 95% chance that the created confidence interval contains the parameter? No not anymore. This is analogous to after the coin has landed. The interval has already been determined. It either captures the population parameter or it doesn't. There is no longer any chance involved.

Confidence Intervals and Hypothesis Testing

Although confidence intervals are used to estimate parameters and quantify uncertainty, we can also use them to perform hypothesis tests. This process follows a 4-step process similar to what we have seen with one-sample hypothesis testing and A/B testing.

Steps 1 and 2: Choosing your Hypothesis and Test Statistic

Remember that in hypothesis testing, we have a null hypothesis, which is some defined and "testable" model (or view of how the world works), as well as an alternative hypothesis, which is some model other than the null hypothesis.

Confidence interval tests are no different. Our null hypothesis will be some specified value for the population parameter, while the alternative hypothesis will be the opposite -- that the population parameter is not that value. The important distinction here, compared to hypothesis tests we have seen in the past, is that we come into this test with an "expectation" of what the parameter should be, rather than a null model that shows no pattern or is due to random chance.

In this guide, let's work with some exam score data for a computer science course here at Cal. This class has two midterms. We randomly sampled 30 students, the size of a typical lab section. From this sample, we want to figure out: did these students in this class do the same on Midterm 1 and Midterm 2, or is there a difference between their performance on the exams? We are going to measure this by comparing the averages of Midterm 1 versus Midterm 2.

```
Our statistic: (Average % score for MT2) - (Average % score for MT1)

Our null: The average percent scores for MT1 and MT2 are the same. In
other words, the true difference between MT1 and MT2 is 0.

Our alternative: The true difference between MT1 and MT2 is not 0.
```

Notice how, in this null estimate, we chose 0. Data scientists will often choose a null estimate that shows no opinion (0.5 or 50%), no relationship (a slope of 0), or no result (a difference of means of 0), but this is not a requirement; just choose a reasonable value that you could expect, given the data. In this example, we decided to use a difference of average scores (in percentages), since that is what the data is measuring. However, depending on your dataset, you will often see statistics such as the median, mean, or proportions -- it primarily depends on the value you are trying to estimate and if the bootstrap will create a symmetric distribution of resampled statistics.

Step 3: Performing the Bootstrap and Creating an Interval

To perform the test, we will compute a bootstrap with our sample and generate a confidence interval. This is the exact same as generating a confidence interval with the goal of estimating a

parameter. In other words, all we need to do is (1) create an array to store the resampled proportions, (2) resample from the original sample, with replacement and the same sample size, many times, and (3) calculate the statistic for each resample, saving it in the array.

We can see this in action with our code for the polling data (with the data stored in a table called "exam_scores").

```
resampled_diffs = make_array()
reps = 1000
for i in np.arange(reps):
     one_resample = exam_scores.sample() # with replacement and same
                                         sample size, 30
     one_diff = np.average(one_resample.column("MT2%")) -
                np.average(one_resample.column("MT1%"))
     resampled_diffs = np.append(resampled_diffs,
     one_diff)
```

Step 4: Making the Conclusion

Now that we have our array of resampled statistics, we will need to calculate a confidence interval. In this case, we will use a level of 95% confidence. We can do this by using the percentile function, where the bounds are determined by this calculation:

```
Lower bound = 0 + (100 - 95%) / 2 = 2.5%ile
Upper bound = 100 - (100 - 95%) / 2 = 97.5%ile
```

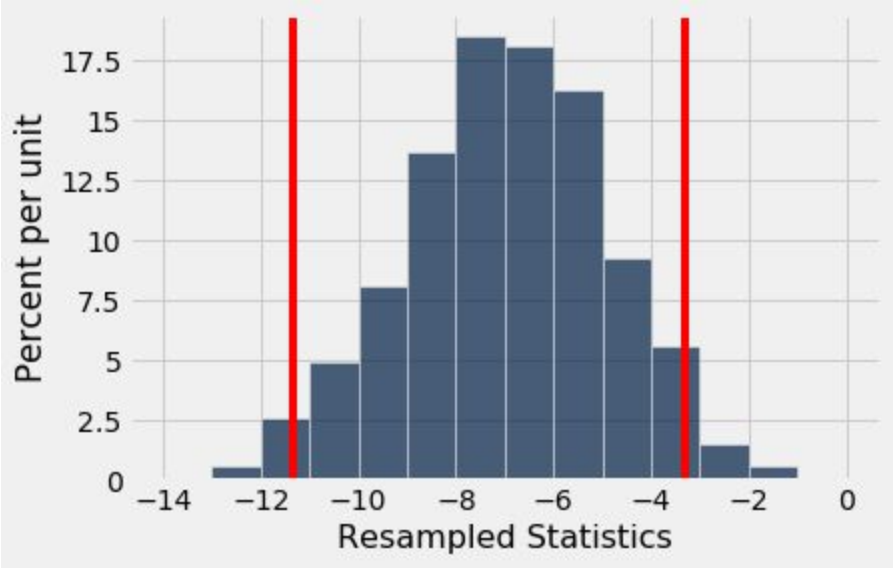
We do this so the interval includes 95% of resampled statistics (97.5 - 2.5 = 95%) and excludes the same amount of data (5%) on both the left and right sides of the interval (2.5% on the left and 2.5% on the right).

In terms of code, this is how we do it:

```
lower = percentile(2.5, resampled_diffs)
upper = percentile(97.5, resampled_diffs)
ci = make_array(lower, upper)
ci
```

When we did the test, we received the interval `array([-11.35, -3.32])`. You can see the distribution below:

```
Table().with_column("Resampled Statistics", resampled_diffs).hist()
```



So, what do we do with this information? We can make a few conclusions:

1) Our estimated interval of the true difference in midterm scores is [-11.35% -3.32%], and we are 95% confident in the process used to generate the interval. Put another way, given this specific interval, we are 95% confident** that the true difference in midterm scores is between -11.35% and -3.32%.
2) We would reject the null hypothesis that the true difference in midterm scores for the class is 0. In other words, students did not have the same performance on both exams, but rather did a bit worse on Midterm 2. Maybe it was a lot harder?

Note that although we are 95% **confident in the process, the ***probability*** that the population parameter is included in this specific, defined interval is either 0% or 100%.

How did we make this conclusion and how does it relate to p-values and p-value cutoffs? Well, remember that the definition of confidence (using a 95% level of confidence) states:

“We are confident that this process will produce a confidence interval that captures the true population parameter 95% of the time, on average.”

In other words, if we generated a new, representative sample and used the code above to generate a confidence interval and repeated that process many, many times, 95% of the intervals we generate will contain the population percentage. We are going to assume that [-11.34%, -3.32%] is one of those intervals that correctly captured the population parameter, and because our null value (0%) is not included in that interval, we will reject the null.

This connects back to the idea of the p-value cutoff and p-values. Recall that the p-value definition is “the probability of getting a result as or more extreme than the observed statistic under the null hypothesis”. The cut-off, in relation, is the probability that we accidentally reject the null hypothesis -- even though we should have failed to reject it (because the p-value was low, and the probability of the observed statistic occurring was not likely).

With 95% confidence, there is still a 5% chance that we get a “bad” interval that does not contain the population parameter. It’s unlikely, but it could still occur. In that case, if we did create one of those “bad” intervals, we may accidentally reject the null hypothesis -- which is the same idea as the p-value cutoff. As a result, we can make the following conclusion: the significance level, or the p-value cutoff, for a confidence interval test is going to be:

```
significance = 100 - (our % confidence)
```

With this specific example, our cut-off is 5%, since 100-95% is 5.

If you’re interested, the true population parameter (using the data from the full class) is -8.19%. Our conclusion was correct! However, note that most of the time, we do not actually have population data -- it isn’t too useful to do inferential statistics when you have all of the information -- so we usually cannot conclude if our interval did contain the parameter or not.