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.