Select Reading Question Responses (3/9)

> In Chapter 4, we learned that 95% of data should be within 2 standard deviations of the mean. In Chapter 5, it says that "95% of the data is within 1.96 standard deviations of the mean". Is the statement in Chapter 5 just a more precise formulation that is only needed for calculating confidence intervals or should we apply the rule of 1.96 standard deviations in other situations as well?

The 1.96 number is just more precise.

In any case, the 68-95-99.7 rule is not something I remember day-to-day in the long-term. Memorizing it feels like unnecessary brain clutter, and it's easier to just remember how to get a computer to calculate the number. ☺

> For the Central Limit Theorem, in order for a distribution to be normal, does the sample size have to be greater than 10? When the textbook mentions the "sufficiently large" part, is that what it means? What if the sample size is 9?

According to the version of the Central Limit Theorem that was in the reading due today, the sampling distribution for a proportion is roughly normal if the "success-failure condition" is satisfied, ie, if $np \geqslant 10$ and $n(1-p) \geqslant 10$.

If the sample size is 10 (or less), the success-failure condition cannot be satisfied! If $n = 10$, then $np \geqslant 10$ means $10p \geqslant 10$, which means that $p \geqslant 1$. But p is a probability, so that means that $p = 1$. But then $1 - p = 0$, so $n(1-p) = 10 \cdot 0 = 0$ and 0 is not greater than or equal to 10. In fact, one can prove that the smallest possible n for which the success-failure *might* be satisfied is $n = 20$ (but this only happens if $p = 0.5$). You could still have sample sizes much larger than 20 where the success-failure condition is still not satisfied. For example, if $n = 100$ but $p = 0.01$, then $np = 1 < 10$ so the success-failure condition is not satisfied.

Anyway, my point is that normality of the sampling distribution for a proportion depends not just on the size of n, but also on what p is.

> The book only highlighted 99% and 95% confidence levels as I assume these are the most common, but are there other more precise or less precise confidence levels. If so, why or what would they be used for?

> I am wondering when a 50% confidence interval would be more valuable 95%. I am struggling to see how it might make sense for a 50% CI to be preferred.

Why would we decide to use a smaller confidence interval (95% vs. 99%) if a slightly larger confidence interval may help us more accurately describe where the population proportion falls?

When would we want to calculate a 90% confidence interval as opposed to a 95% confidence interval or 99% confidence interval? Are there different general standards with different types of data or samples, or is it chosen on a case-by-case basis?

Why would you ever want to use a lower confidence level for a confidence interval?

It's true that you can be more certain that the parameter of interest lies in a 95% confidence interval than a 50% confidence interval. That makes it seem better, but the 95% confidence interval is also much *wider*!

For example, let's say I want to measure the proportion of cats that are orange tabbies. I can be 100% confident that the percentage of cats that are orange tabbies is between 0% and 100%, but this is a *useless* statement! In other words, I've "computed" a 100% confidence interval here, but the 100% confidence interval is completely pointless because it's so wide.

You might not need 100% confidence. You might not even need 99% confidence, or 95% confidence. Maybe you can get by for whatever application you have in mind with just 90% confidence, or even 50% confidence. You'll be less certain that you're right, but you'll be working with a narrower (ie, less useless) range. There isn't any clear "rule" that tells you when a 90% confidence interval might be more appropriate than a 95% one; this is just something you as a statistician have to make a judgment about. Bear in mind that more certainty comes at the cost of having wider intervals, and then ask yourself, "How certain do I really need to be in this situation?"

I have often encountered in other science classes that researchers are typically held to the 95 percent confidence level. Why is that particular figure so widely used?

The number is a little arbitrary, but you can find a discussion of this here on the website of the textbook. You'll also find frequent critiques of blindly applying 95% confidence. It's fine to use 95% confidence if you don't have a reason to use something else, but you should at least think through reasons for using something else first.

If a sample population is typically used to determine nation-wide and global estimates for a study, how accurate are those studies realistically? I know that it is unrealistic to assume that every individual on the planet or in a country would be willing to be involved in a scientific or statistical test, but what does this mean for the accuracy of tests claiming they know the statistics of such a large group? Should the results be trusted? When all of these studies are working off of somewhat small sample proportions relative to the national and global population, is it even useful to conduct these studies?

One of the cool things about statistics is that it's possible to quantify how much uncertainty is caused by sampling error. This is what the "sampling distribution" measures, and one thing we learn from this theoretical analysis is that sampling error doesn't cause a huge amount of uncertainty when one has moderate sample sizes, even if the population of interest is huge. This is very surprising, very cool, and very convenient.

That being said, we can only use these ideas to quantify the uncertainty caused by sampling error. We cannot use it to quantify uncertainty caused by biases, and there isn't a great quantitative way of accounting for sample biases. Even if you have a reasonably large sample size, if you're only sampling certain kinds of people and not other people and the type of people you're sampling are likely to respond to your study in a particular way and the type of people you're not sampling will respond in a different way, your data just won't generalize. Here is another place where lucid reasoning (rather than blind application of statistical tools) is required. You have to think through your data collection process and possible biases that might be introduced by your data collection process. There are usually clever ways of avoiding or at least minimizing these kinds of biases, but there aren't hard-and-fast rules that work in every situation.

In any case, the point is: if you've thought carefully about minimizing biases in data collection, all that's left in your data is uncertainty from sampling. That kind of uncertainty is both easily quantifiable and smaller than you would think.