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. For me personally, the 68-95-99.7 rule is not something I remember day-to-day in the long-term, and even when I do manage to commit it to short-term memory, I would remember "2 standard deviations" rather than "1.96 standard deviations." Memorizing the more precise number 1.96 feels like unnecessary brain clutter to me! ☺

> 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 their 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.

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?

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?"