4.14: Estimating a confidence interval

> *Statistics means never having to say you're certain – Unknown origin*

Up to this point in this chapter, we've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to **quantify** the amount of uncertainty in our estimate. It's not enough to be able guess that the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ. I've just finished running my study that has N participants, and the mean IQ among those participants is $\bar{X}$. We know from our discussion of the central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. To be more precise, we can use the **qnorm()** function to compute the 2.5th and 97.5th percentiles of the normal distribution

```
qnorm( p = c(.025, .975) ) [1] -1.959964 1.959964
```

Okay, so I lied earlier on. The more correct answer is that a 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean.

Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean $\bar{X}$ that we have actually observed lies within 1.96 standard errors of the population mean. Oof, that is a lot of mathy talk there. We'll clear it up, don't worry.

Mathematically, we write this as:

$$\mu - (1.96 \times \text{SEM}) \ \leq \ \bar{X} \ \leq \ \mu + (1.96 \times \text{SEM})$$

where the SEM is equal to $\sigma / \sqrt{N}$, and we can be 95% confident that this is true.

However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean, given that we know what the population parameters are. What we **want** is to have this work the other way around: we want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

$$\bar{X} - (1.96 \times \text{SEM}) \ \leq \ \mu \ \leq \ \bar{X} + (1.96 \times \text{SEM})$$

What this is telling is is that the range of values has a 95% probability of containing the population mean μ. We refer to this range as a **95% confidence interval**, denoted CI_{95}. In short, as long as N is sufficiently large – large enough for us to believe that the sampling distribution of the mean is normal – then we can write this as our formula for the 95% confidence interval:

$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}} \right)$$

Of course, there's nothing special about the number 1.96: it just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I could have used the **qnorm()** function to calculate the 15th and 85th quantiles:

```
qnorm( p = c(.15, .85) ) [1] -1.036433 1.036433
```

and so the formula for CI_{70} would be the same as the formula for CI_{95} except that we'd use 1.04 as our magic number rather than 1.96.

A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation σ.

Yet, before we stressed the fact that we don't actually **know** the true population parameters. Because we don't know the true value of σ, we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the quantiles of the t-distribution rather than the normal distribution to calculate our magic number; and the answer depends on the sample size. Plus, we haven't really talked about the t distribution yet.

When we use the t distribution instead of the normal distribution, we get bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation $\hat{\sigma}$ might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like… and this uncertainty ends up getting reflected in a wider confidence interval.
