10/6/2022

Confidence Intervals

As mentioned in the previous lecture, our goal with confidence intervals is to surround a point estimate with a margin of error that represents our level of confidence, or how much accuracy there is in our estimate. Confidence intervals are built from a point estimate and a margin of error. In this course, we will express our confidence intervals (CIs) as mathematical intervals where the lower bound of the interval is

$$point\ estimate - margin\ of\ error$$

and the upper bound is

$$point\ estimate + margin\ of\ error$$

Our margins of error will differ for our level of confidence and the sampling distribution standard deviation or the standard error.

$$ME = score_{confidence} \times standard\ error$$

In most cases the score (sometimes called the critical value) will come from the Student-T distribution, or the normal distribution, however, in some cases, we may use other appropriate distributions. The standard error, as we saw in the last lecture, for common situations, we have specific formulae to fall back on; however, in some cases we may have to rely on estimations from simulations (as we saw in the lab), to obtain standard errors for sampling distributions for less common statistics.

Most textbooks start their discussion of confidence intervals with a special case, the case where we are certain that the population we are drawing samples from is normally distributed, the sample size is sufficiently large, and the population standard deviation is known. We will follow suit here, but it should be emphasized that this is a special case, that we will not often encounter in the real world. For instance, we don't normally have a population standard deviation. We will deal with that case later, but for now it's important to note that we start here because it is the easiest case to prove (if we were doing not), not because it is the most common scenario.

Consider the following example:
A sample of 45 women have their heights measured and their mean height is found to be 64.3 inches. It is known from a large CDC study that the mean height of women has a population standard deviation of 3.1 inches. Construct a 95% confidence interval for the population mean.

Heights of people are indeed normally distributed. The population standard deviation is provided. And the sample size is greater than our rule of thumb, so we deem that to be sufficiently large. So this scenario satisfies our special condition. In this case the standard error we can use from the sampling distribution of the mean is $SE = \frac{\sigma}{\sqrt{n}}$. The confidence level of 95% is the standard confidence level we should assume if not level is specified. This value is going to get us the score we use as a multiplier for our standard error to obtain the margin of error. The score is essentially the z-score that $\pm z_{\alpha/2}$ contains that percent of the sampling distribution.

Think of the confidence level as saying that we want to have 95% probability that the true population mean is going to fall within the bounds we give. Thus, we say that we are 95% confident that the true mean is in our interval.

Confidence is given the notation $1-\alpha$. Thus, the notation in the value $\pm z_{\alpha/2}$. If we are 95% confident, then we are potentially leaving out a 5% chance that our interval will not contain the true value of our parameter. Since the normal curve is symmetric, we are leaving out 2.5% on each side (half of 5% each). So we want the z-score from the standard normal distribution for which $P(Z \leq z) = 0.025$. The reciprocal value on the right side would be $P(Z \geq z) = 0.025$ or equivalently that $P(Z \leq z) = 0.975$. While the notation technically uses the smaller value, it is negative, and when using the value in formulas, we prefer the positive value. Expressing that precisely would tend to make the notation more unwieldy, but it's worth keeping in mind what is meant here.

We may recall from the Empirical Rule that the value should be about 2. It turns out it's 1.96. Thus, we find our margin of error to be

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96\left(\frac{3.1}{\sqrt{45}}\right) \approx \ 0.9058$$

Our confidence interval is then

$$(64.3-0.91, 64.3+0.91) = (63.39, 65.21)$$

We interpret this interval by saying that we are 95% confident that the true mean height of women is between 63.39 inches and 65.21 inches.

If we are estimating a mean and one of these conditions fails: we are not sure the underlying distribution is normal, we do not have a sufficiently large sample size, or we are estimating the confidence interval from the sample standard deviation and not the population standard deviation, then we adjust our procedure to use a score from the Student-T distribution with $n-1$ degrees of freedom, instead of the standard normal distribution.

If the sample size is large enough, there will not be much difference in the resulting confidence interval, but a small sample size can increase the amount of variability, and estimating from another estimate (of the sample standard deviation), that, too, can increase the potential for variability that can be captured by the t-distribution.

In the example we did above on women's heights, the score from the t-distribution would be $t_{\alpha/2,44} = 2.01$ instead of 1.96 (the subscript must include both the confidence level and the degrees of freedom). This would make our margin of error 0.9289, which would make our confidence interval just a bit wider. When the same size is small, however, the difference will be much larger and more noticeable.

For one-sample proportions, we also discussed the normal approximation in the last lecture and the resulting standard error.

Suppose a candidate in an election is polling at 37.8% in a sample of 1400 residents. What is the 95% confidence interval for the true proportion of their support?

We use the sample proportion as p in our confidence interval. If we want to be more conservative, the largest margin of error will occur at $p = 50\%$, but it's more common to use the sample proportion here. We can check out test to see that npq is over 300, so the distribution is sufficiently normal. We can use the 1.96 we found before.

$$ME = 1.96\sqrt{\frac{0.378(1-0.378)}{1400}} \approx 0.0254$$

Thus, the confidence interval is

$$(0.378 - 0.0254, 0.378 + 0.0254) = (0.3526, 0.4034)$$

We are 95% sure that the true support for the candidate in the population is between 35.26% and 40.34%.

It's worth noting that the standard error we are using here is an estimation. The Devore text [1] linked below gives a more precise formation for the margin of error for a sample proportion.

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}\hat{q}/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$$

Statistical programs that have built-in confidence interval calculator may use this more accurate version, but in cases where we have to calculate values by hand, we'll continue to use the simplified version.

We can back out of a confidence interval to find the point estimate and either the sample size or standard deviation if we have the other one.

Consider the confidence interval $(7.9, 10.5)$ which measures the alcohol content by percentage in wine at a particular winery. The point estimate is the midpoint of the interval. $\frac{10.5+7.9}{2} = 9.2$. The margin of error is the difference between the point estimate and one endpoint, or half the range. $\frac{10.5-7.9}{2} = 1.3$. If this is a 95% confidence interval, we can divide by 1.96 to estimate the standard error. $\frac{1.3}{1.96} = 0.66$. If we knew the sample size, we could use a t-interval to make a better estimate. The 0.66 is our estimate of the standard error. If we know our sample was taken from a sample of only 10 bottles, we would prefer that t-distribution estimate. Our $t_{\alpha/2,9} = 2.26$, so our standard error would be $\frac{1.3}{2.26} = 0.575$. And that is equal to $\frac{s}{\sqrt{10}} = 0.575$, making $s = 1.8$.

The calculation is a little messier if we have the standard deviation, since both t and the $\sqrt{n}$ both depend on n. If we have a proportion, we can use the center of the interval as our estimate for the proportion in the standard error so we'll be able to find the sample size as long as we know the confidence level.

By a similar means, we can create formulas for both of these situations to estimate the sample size needed to obtain a particular margin of error. The algebra is relatively straightforward so I will just report the results below.

For means:

$$n = \left(\frac{z^*\sigma}{E}\right)^2$$

For proportions:

$$n = p(1-p)\left(\frac{z^*}{E}\right)^2$$

When we are estimating the required sample size ahead of a study, we are estimating. We can't use the t-distribution, since that depends on the sample size, so we use the standard normal distribution. The z^* is the z-critical value desired. If we don't have σ we can use an estimated standard deviation s, often from a smaller, preliminary sample. E is the desired margin of error. In the proportion case, we usually estimate n with $p = 0.5$ as the largest possible standard error, but if we have prior evidence to suggest it is another value, we can use that. Remember to express both the margin of error and the proportion in decimal or fraction form. Once we find n, we always round up to the next whole value.

Consider if we want to estimate how large a sample we'd need in a public poll to determine who is ahead in a political race within 3% of the true value, with 95% confidence.

$$n = \frac{(0.5)(0.5)(1.96)^2}{(0.03)^2} = 1067.1\ldots \approx 1068$$

Sample sizes for proportions are usually much larger than sample sizes for means studies.

We can also consider confidence intervals when comparing two samples. The point estimate is the difference between the sample means (or proportions), and the standard error formula changes, but the general procedure remains the same. When we discuss two-sample hypothesis tests, we'll return to this subject in more detail.

It's also possible to construct one-sided confidence intervals. We may only be considered with values that are larger than some maximum or smaller than some minimum. For instance, in a rainfall forecast, we may be concerned with rainfall exceeding our forecast, but not if it ends up being lower than the forecasted amount. In such a case, we can then put all 5% on one side of the interval instead of split in half on both sides. This will change the critical value calculation to $P(Z \leq z) = 0.95$ or $P(Z \leq z) = 0.05$ depending on which side of our interval we want to be infinity. The critical value would be 1.64 in this case and we'd notate it z_α instead of $\alpha/2$. We would only have to calculate the one bound.

One-sided intervals are uncommon but it's worth knowing they do exist.

In all of these cases, we have mathematically well-established standard errors to fall back on. This is not always the case. In such situations we can simulate the distribution and drawing samples from it. If we wish to build a 95% confidence interval, we can estimate the bounds that contain 95% of our simulated

samples. We can use these methods for uncommon situations such as distributions of the sample median, or complex situations which are difficult to find mathematically compact formulas for. The lab looked at some of these situations, and we'll return to these ideas later in the course.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf