Prob & Stats - Section 7.1: Confidence Intervals

The whole point of statistical inference is STILL to see what statistics tell us about parameters.

Statistics are numerical characteristics of random samples.
Parameters are numerical characteristics of populations.

We saw last time that our main parameters of interest have certain statistics associated with them that we consider to be "good" estimators, and we defined "good" to mean "unbiased."

Our interest today is in taking those unbiased estimators and identifying an interval around them that we can be reasonably sure holds some desirable properties. In fact, let's define right away:

The interval [a,b] is called a 95% confidence interval for parameter θ if
$P(a < \theta < b) = 0.95$.
(There will be a little more to it than this. And it does not have to be 95%. It could be any amount we desire, but 95% is the most common, followed by 99% and 98%, and maybe 90%. Through this discussion I'll use 95% each time, just for convenience.)

It is important to observe in the statement $P(a < \theta < b)$ that θ is a parameter, and therefore not a probabilistic thing. In this statement, the a & b are statistics, so the probability is actually the probability a & b hold the stated relationship to θ.

Here's what's going to happen in the background. We're going to take a random sample of size n from some population, we're going to calculate a & b (from some formula we still need to find). Then we take another random sample of size n and calculate again. We keep doing this over and over, each time constructing the interval [a,b]. Those intervals will be different, because they are based upon different random samples. If we look at a number line, somewhere along that number line will be the actual value of our parameter θ. And also along the number line will be all of these theoretical intervals we've found (or could find). We're going to set up these intervals so that 95% of the time the interval will actually surround θ, but 5% of the time it will miss. And there is our interpretation of a 95% CI:

Interpretation: To say that [a,b] is a 95% CI for θ means that we are 95% sure that the interval [a,b] surrounds θ.

In my classes, I usually try not to teach you wrong things. But here I'm going to teach you something wrong. The reason is, some statistics books and classes use a different interpretation of the CI. I'm not saying they are wrong, but they are wrong. But if you want to communicate with someone who has learned stats the other way, you need to know what they have learned. And we don't condemn them for not knowing the true mathematical interpretation of the CI. We just remain smug in our enhanced understanding of the topic.

The wrong interpretation: To say that [a,b] is a 95% CI for θ means that we are 95% sure that θ falls into the interval [a,b].

Can you see the difference between the two interpretations? One of them gives action to the interval (the interval surrounds), while the other one gives action to the parameter (θ falls). But θ can't fall one place or another. It's a parameter, and it is fixed to one place. We can't really talk about it falling one place or another. That would be trying to make it a probabilistic object, which (as we earlier discussed) it is not.

You should learn both interpretations, because we want to know the correct interpretation, but we want to be able to communicate with others who have learned an incorrect interpretation.

Okay, let's figure out what a & b must be in the case where we are trying to estimate the mean μ of a normal population that has known std dev of σ. You won't have to reconstruct this for me.

$a < \mu < b$	our starting interval
$a - \overline{X} < \mu - \overline{X} < b - \overline{X}$	subtract through by $\overline{X}$
$\overline{X} - b < \overline{X} - \mu < \overline{X} - a$	run through with a minus, and rearrange
$\frac{\overline{X}-b}{\sigma/\sqrt{n}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{\overline{X}-a}{\sigma/\sqrt{n}}$	divide through by $\sigma/\sqrt{n}$

That middle part is a Z, and the outside parts are just numbers, so now we're looking for two numbers, say α & β, such that $P(\alpha < Z < \beta) = 0.95$.

Now, if we want this CI to estimate our parameter μ, it probably makes sense that we want it to be as narrow as possible. Without too much trouble we could show that this interval is narrowest if they are symmetric around zero, so that means $\alpha = -\beta$. (Continued on the next page.)

So we need to find a z-score β such that $P(-\beta < Z < \beta) = 0.95$, but I'm going to just call it z, instead of β.

We need to find z such that $P(-z < Z < z) = 0.95$. We will find that actual value of z in a moment. Let's get the formula first.

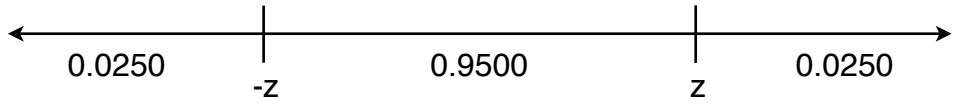
We have $\frac{\overline{X}-b}{\sigma/\sqrt{n}} = -z$ and $\frac{\overline{X}-a}{\sigma/\sqrt{n}} = z$, which rearrange to

$a = \overline{X} - z^* \frac{\sigma}{\sqrt{n}}$ and $b = \overline{X} + z^* \frac{\sigma}{\sqrt{n}}$

And there's our formula (with some adjustments coming later):

$$\overline{X} \pm z^* \frac{\sigma}{\sqrt{n}}$$

How do we find the z-value? The z & -z values are supposed to trap 0.9500 area between them (since it's a 95% CI). That means there is 0.0500 area outside the interval, but it's split evenly between above & below the interval (because of the symmetry). See the number line below.



So our upper z-value exceeds an area of 0.9750. We go to the z-chart, look up the area 0.9750 inside the chart, and read off the z-score from the row & column, getting z = 1.96. (It just so happens that 0.9750 is exactly in the chart.) Now we could do the same for -z, but it's called -z for a reason: it's the negative of the other z. So -z = -1.96.

If we were using another confidence level than 95%, then we would have to adjust our areas on the number line and look up a different z-value.

<u>Before we do an example, some details about our problem:</u>
a) As we've seen before, $\overline{X}$ doesn't always have a normal distribution (ND). So we will have to observe if either the population was normal or the sample size was 30 or more. If neither of those are true, then we will have to "Assume Pop ND."
b) We will not always know what σ is. In fact, knowing σ (in practice) is quite rare. If we don't know σ, we will have to use s instead, but that will change the theoretical distribution from a z-distribution to a t-distribution. We will look at that later.

Example 1: Suppose we know that IQs at ENC are ND(μ, 15). We wish to estimate μ, so we take a random sample of size n=16, and we observe $\bar{X}$ = 110. Find a 95% confidence interval for μ.

Solution: 95% CI for μ. We observe that σ is given, so we use the z version of the CI. And we notice that the pop was ND, so there is no need to make any assumption.

$$\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}} \Rightarrow 110 \pm 1.96^* \frac{15}{\sqrt{16}} \Rightarrow 110 \pm 1.96^*3.75 \Rightarrow 110 \pm 7.35$$

So the confidence interval is [102.65, 117.35]

If we want to interpret this, we say:
We are 95% sure that this interval surrounds the true value of μ.
We also could think about it this way:
If we had to identify which values of μ are reasonable and which are unreasonable, we would choose those inside the interval as reasonable, and those outside the interval as unreasonable. So if we were wondering if the true mean
---

Another comment: That part of the formula after the ± is called the Margin of Error. Perhaps you've heard of that before. When you see survey or polling data in the news, they frequently tell you what the sample mean $\bar{X}$ is, as well as the margin of error. From that you could construct the CI. Also, they almost always use 95% when finding their margin of error.

<u>For you to try (answer on the next page)</u>
One of these problems fits with our formula, and one does not. Determine which one fits, and then find the confidence interval for that one.

Example 2A. In order to monitor the ecological health of the Florida Everglades, various measurements are recorded at different times. The bottom temperatures are recorded at the Garfield Bight station and the mean of 30.4 deg C is obtained for 61 temperatures recorded on 61 different days. Assuming that σ = 1.7 deg C, find a 95% CI for the true mean of the temperatures.

Example 2B. In a sample of seven cars, each car was tested for nitrogen-oxide emissions (in grams per mile) and the following results were obtained:
0.06, 0.11, 0.16, 0.15, 0.14, 0.08, 0.15.
Construct a 95% CI of the mean NO-emissions for all cars. For your convenience, the sample yielded a mean of 0.1214 and a std dev of 0.0389.

Answer to previous page: Example 2B does not apply, because we do not know σ.

Answer to Example 2A:

95% CI for μ
σ known, so use $\bar{X} \pm z^*\sigma/\sqrt{n}$

$30.4 \pm 1.96*1.7/\sqrt{61}$

30.4 ± 0.4266, so 30.4 ± 0.43

[29.97, 30.83]

Final note: We saw earlier that, if the confidence level is 95%, then the z-value is 1.96. To help you with your homework, let's be a little more precise and inclusive. Since you won't need to memorize these values, you might make sure you can see how they came out of the z-chart.

If we are using z in our CI, then we get the following values:

99% CI	z = 2.575
98% CI	z = 2.32
95% CI	z = 1.96
90% CI	z = 1.645

Homework for this section:

Section 7.1, # 2, 4abc
Find your answers to these, then either:
a) email those answers to me. No need to include any work.
or b) scan/photograph your hw page and email it to me.