Lecture 17 (CH. 7)

Last time: C.I. for $\mu_x$ : $\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$

Approximate with
Sample std. dev. $s$,
(for now).

Interpretations (e.g., at 95% conf. level)

1) We are 95% confident that $\mu_x$ is in $\bar{x}_{\text{obs}} \pm 1.96 \frac{s_x}{\sqrt{n}}$.

2) There is 95% prob. That a random CI, $\bar{x} \pm 1.96 \frac{s_x}{\sqrt{n}}$, covers $\mu_x$.

WARNING!
The math is trivial.
It's the correct jargon & interpretations that are tricky.

$\Rightarrow$ There is random vs. observed vs. fixed pop. params.

$\Rightarrow$ There are 3 means: $\bar{x}$, $\bar{x}_{\text{obs}}$, $\mu_x$

There is random (e.g. $\bar{x}$) vs. not (e.g. $\bar{x}_{\text{obs}}$, $\mu_x$)

$\Rightarrow$ There is also random CI vs. observed CI

$\Rightarrow$ Then, there is confidence vs. probability

$\Rightarrow$ For pop. params

- $\mu$, $\sigma_x$, ...
- $P(\mu_x > 3)$
- C.I. for $\mu_x$

$\Rightarrow$ For random thing

- $\bar{x}$
- $P(\bar{x} > 3)$
- C.I. for $\bar{x}$

$\Rightarrow$ In the example from last lec., we also talked about how there are 3 things that affect the width of a CI

1) Conf. level (through $z^*$), 2) $s$ (approx. of o), 3) n
The formula for C.I. can be used to decide what minimum sample size is necessary, even before taking any sample!

But you need to specify what is meant by necessary.

For example, say, you want your estimate of \( \mu_x \) to be within some range \( \pm B \) (for Bound). Then:

\[
\frac{z^* \sigma_x}{B} = \sqrt{n} \implies n_{\text{min}} = \left( \frac{z^* \sigma_x}{B} \right)^2
\]

Note that \( B \) is different from conf. level, or \( z^* \). It has the dimensions of \( \mu_x \) itself.

(example (fish from last lect))

What min. sample size is required for a margin of error of 0.2 kg?

\[
n = \left( \frac{z^* \sigma_x}{B} \right)^2 = \left( \frac{1.96 (1.27)}{0.2} \right)^2 = 155\text{ type I Fish. instead of } 56
\]

\[
\approx \left( \frac{2.575 (1.70)}{0.2} \right)^2 = 485\text{ type II Fish. } \approx 61
\]

Note: If you have no sample to provide an estimate of \( \sigma_x \), then you guess it! It’s not hard. For example, if we’re dealing with people’s height, then \( \sigma_x \approx \text{a few inches} \).
So far, we have been talking about the (2-sided) CI for the population mean, $\mu_x$.

This quarter, we skip 1-sided intervals: Lower Conf. Bound (LCB) and Upper Conf. Bound (UCB).

There are some population parameters that we care about a lot; the pop. mean ($\mu_x$) is one of them, and the pop. std. dev. ($\sigma_x$) is another. Both of these pertain to a continuous random variable ($x$). But we also care about situations where the population consists of a categorical random variable. We will deal with the multi-level case later (when we learn about something called the chi-squared distribution). Here, let's focus on the 2-level case ($x=0,1$), e.g. healthy vs. sick person, safe vs. unsafe email. Then, we care about estimating the true/population proportions of the two levels, e.g. the true proportion of people who have covid19. It is sufficient to estimate only one of the two proportions because the proportion of the other level is just one minus the first proportion. Let's say we want to estimate the true proportion of $x=1$, and let $\pi_x$ denote that population proportion. Here we will build the (2-sided) CI for the population proportion, $\pi_x$.

To build a CI for $\pi_x$, we need:

1. The sampling distrib. of $p$, i.e. the sample proportion.
2. The sample size, $n$.

In a hw, you show that even w/o knowing the sampling dist. of $p$, we have:

\[
\pi_x = \frac{\sum_{i=1}^{n} I(x_i = 1)}{n}
\]

where $\pi_x = \frac{\sum_{i=1}^{n} (x_i = 1)}{n} = \text{pop. prop}$

\[
\pi_x \sim \text{Binomial}(n, \pi_x)
\]

Note resemblance to $\sigma / \sqrt{n}$,

where $\sigma_x = \sqrt{\pi_x(1-\pi_x)} = \text{std. dev. of Bernoulli}!

CLT: $p \sim \mathcal{N}(\pi = \pi_x, \sigma = \pi_x(1-\pi_x) / n)$

So, we can find the prob that a random sample prop is $\alpha$:

\[
\Pr(\alpha < p < b) = \Pr\left( \frac{\alpha - \pi}{\sigma_p} < \frac{p - \pi}{\sigma_p} < \frac{b - \pi}{\sigma_p} \right)
\]

\[
\frac{\alpha - \pi}{\sigma_p} \sim \mathcal{N}(0,1)
\]

\[
\frac{b - \pi}{\sigma_p} \sim \mathcal{N}(0,1)
\]

\[
\frac{\pi_x}{\sigma_x} = \frac{\pi_x}{\sqrt{\pi_x(1-\pi_x) / n}} \sim \mathcal{N}(0,1)
\]

Table I.
Now that we know the sampling distr. of \( p \), we can build CI for \( \pi_x \).

\[ \text{CLT} \Rightarrow \text{if } n = \text{large, then } p \sim N \left( \pi_x, \sqrt{\frac{\pi_x(1-\pi_x)}{n}} \right) \]

What, then, has a std. normal distr? \[ z = \frac{p - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}} \]

Start with self-evident fact

\[ \left[ \text{Recall } \Pr(-1.96 < \frac{\bar{x} - \mu_x}{\sigma/\sqrt{n}} < 1.96) = 0.95 \right. \]
\[ \left. < \mu_x < \Rightarrow 95\% \text{ C.I. for } \mu_x \right] \]

\[ \text{prob} \left( -z^* < \frac{p - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}} < z^* \right) = \text{conf. level} \]

\[ \frac{\pi_x}{n} < \text{why the C.I. for } \pi_x \text{ is a messy eqn.} \]

C.I. for \( \pi_x \):
\[ \left( p \pm z^* \sqrt{\frac{\pi_x(1-\pi_x)}{n}} \pm \frac{z^*^2}{2n} \right) \]

Say 2 interpretations as before. Basically, any \( \pi \) in this CI is consistent with data/observations. Note: 0 < CI < 1, as it should be for a proportion CI. So we can't use this CI to test if \( \pi = 0 \) or \( \pi = 1 \) are consistent with data/obs.

A Simple(v) eqn: If \( n = \text{large} \), Then \[ p \pm z^* \sqrt{\frac{p(1-p)}{n}} \]

We'll use this one!

F.Y.I The 1-sided C.I.s are obtained by simply changing \( z^* \)

\( \pi_x \) denotes the proportion (say, of girls) in pop. In the coin-tossing analogy \( \pi_x \) is the prob. of a head on a given toss. Note that this is all perfectly consistent, because the prob. of drawing a single "good" out of the population (i.e. prob of heads on a toss) is equal to the proportion of goods in pop. This \( \pi_x \) is the same \( \pi \) that appeared in binomial. Now, you know how to make a confidence interval for it!
Example: A past survey from 390:

- Lab is good: 17
- "bad": 48
- No opinion: 15

Only part of the class voted, but assuming that the voters are a random sample from the whole class, we can find the true proportion of students who like the lab, etc.

Our CI formulas pertain to a pop. of things with 2 catg. (The multiple-category case will be done later). So, let's consider

- Lab is good: 17
- "bad": 48

The sample proportion of students who like lab, \( \hat{p} \), is \( \hat{p} = \frac{17}{65} = 0.262 \)

Let \( \pi_x = \text{True/distr. prop. of students who like lab} \).

(The True prop. of students who don't like the lab is \( 1 - \pi_x \)).

95% C.I. for \( \pi_x \):

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.262 \pm 1.96 \sqrt{\frac{0.262(1-0.262)}{65}} = 0.262 \pm 0.107 = [0.16, 0.37]
\]

1) We are 95% confident that \( \pi_x \) is in here (any \( \pi_x \) in here is consistent with data)

2) There is a 95% prob. that a random C.I. will cover \( \pi_x \).

3) Corollary: (A simple, non-mathematical answer, "in English"): Students are generally unhappy with lab.

If the C.I. had covered 0.5, then we would say "We don't know!"

FYI: 95% C.I. for \( 1-\pi_x \) is: \( 1 - \text{(C.I. for } \pi_x \text{)} = [0.63, 0.84] \)
A sample of 2000 aluminum screws used in the assembly of electronic components was examined, and it was found that 44 of these screws stripped out during the assembly process. Does it appear that the true percentage of defective screws is (or is not) 2.5%? Explain your reasoning and the conclusion that follows from it. You may use the "simple formula" appropriately revised. Use 90% confidence level.

There are several ways of proving $E[p] = \mu$, $V[p] = \frac{\mu(1-\mu)}{n}$, (dropping the subscript $x$, just for convenience). One way is to return to our derivation of $E[\bar{x}] = \mu_x$, $V[\bar{x}] = \frac{\sigma_x^2}{n}$, and note that the derivation is correct even if $\bar{x}$ is the sample mean of $x_i$, where $x_i = 0$ or 1. So, first,

a) Show that for a sample of size $n$ taken from the Bernoulli dist., the sample mean is equal to the sample proportion. I.e., show $\bar{x} = p$. Hint: if a sample of size $n$ has $n_0$ zeros and $n_1$ 1's, then the sample proportion $p$ is $n_1/n$.

b) Therefore, $E[\bar{x}] = E[p] = \mu_x$, and $V[\bar{x}] = V[p] = \frac{\sigma_x^2}{n}$. Finally, for $x \sim \text{Bernoulli}(\mu)$, find $\mu_x$ and $\sigma_x^2$ starting from the defn. of $E[x]$ and $V[x]$ from Ch. 2.

Moral: When you are done, you will have proven $E[p] = \mu$, $V[p] = \frac{\mu(1-\mu)}{n}$ using equations that we had proven before, i.e. $E[\bar{x}] = \mu_x$, $V[\bar{x}] = \frac{\sigma_x^2}{n}$. 