Lecture 17 (Ch. 7)

a number determined by conf. level. e.g. 95%

Last time: C.I. for μ_x : $\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$ ← Approximate with sample std. dev. s, (for now).

There are 2 types of CIs (obs, random), with different interpretations:

1) We are 95% confident that μ_x is in the observed C.I $\bar{x}_{obs} \pm z^* \frac{\sigma_x}{\sqrt{n}}$

2) There is 95% prob. that a random CI, $\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$, covers μ_x.

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

⇒ There is random vs. observed vs. fixed pop. params.
↳ $\bar{x}$, CI, ... ↳ $\bar{x}_{obs}$, obs. CI, ... ↳ μ_x, σ_x, ...

E.g.
There are 3 means : $\bar{x}$, $\bar{x}_{obs}$, μ_x
There is random (e.g. $\bar{x}$) vs. not (e.g. $\bar{x}_{obs}, \mu_x$)

⇒ There is also random CI vs. observed CI

⇒ Then, there is confidence vs. probability

↳ for pop. params e.g. $\mu_x, \sigma_x, \ldots$ — $pr(\mu_x > 3)$ ✗ — C.I. for μ_x ✓

↳ for random thing e.g. $\bar{x}$ — $pr(\bar{x} > 3)$ ✓ — C.I. for $\bar{x}$ ✗

⇒ In the example from last lect, 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 σ) 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 μ_x to be within some range $\pm B$ (for Bound). Then

for now, approximate with s.

$$\frac{z^* \sigma_x}{\sqrt{n}} = B \Rightarrow n_{min} = \left(\frac{z^* \sigma_x}{B} \right)^2$$

Note That B is different from conf. level, or z^*. It has the dimensions of μ_x itself.

example (fish from last lect)

B units of $\bar{x}$

What min. sample size is required for a margin of error of 0.2 $\frac{\mu g}{g}$?

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

$$\simeq \left(\frac{2.575 \ (1.71)}{.2} \right)^2 = 485 \text{ type II Fish.} \quad \text{" " } 61$$

Note: If you have no sample to provide an estimate of σ_x, Then you guess it! It's not hard. For example, if we're dealing with people's height, Then $\sigma_x \sim$ 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 C.I. for π_x we need The sampling distr. of p, i.e. The sample proportion.

To build a C.I. for μ_x, we need The sampling distr. of $\bar{x}$, i.e. The sample mean.

In a hw, you show That even w/o knowing The sampl. dist. of p,

$$\mu_p \equiv E[p] = \underbrace{\pi_x}_{\text{pop. prop}}$$

$$\sigma_p \equiv \sqrt{V[p]} = \sqrt{\frac{\pi_x(1-\pi_x)}{n}}$$

Note resemblance to $\sigma_x/\sqrt{n}$, where $\sigma_x = \sqrt{\pi_x(1-\pi_x)}$ = std. dev. of Bernoulli!

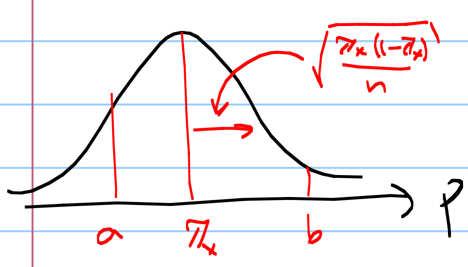
$$\mu_{\bar{x}} \equiv E[\bar{x}] = \mu_x$$

$$\sigma_{\bar{x}} \equiv \sqrt{V[\bar{x}]} = \frac{\sigma_x}{\sqrt{n}}$$

CLT: $p \sim N\left(\mu = \mu_p = \pi_x,\ \sigma = \sigma_p = \sqrt{\frac{\pi_x(1-\pi_x)}{n}}\right)$ | $\bar{x} \sim N\left(\mu = \mu_{\bar{x}} = \mu_x,\ \sigma = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$

So, again, we can find The prob That a random sample prop is ...:

$$\text{prob}(a < p < b) = \text{prob}\left(\frac{a-\mu_p}{\sigma_p} < \underbrace{\frac{p-\mu_p}{\sigma_p}}_{N(0,1)} < \frac{b-\mu_p}{\sigma_p}\right)$$



$$= \text{prob}\left(\frac{a-\pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}} < z < \cdots\right)$$

= table I.

Now that we know the sampling distr. of p, we can build CI for π_x.

→ CLT ⇒ If n = large, then $p \sim N\left(\pi_x, \sqrt{\frac{\pi_x(1-\pi_x)}{n}}\right)$

→ What, then, has a std. normal dist? $z = \frac{p - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}}$

→ Start with self-evident fact

[Recall $\text{prob}\left(-1.96 < \frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}} \; 1.96\right) = 0.95$

$< \mu_x <$ ⇒ 95% C.I. for μ_x.]

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

quadratic eqn in π_x. This is why the C.I. for π_x is a messy eqn.

$$< \pi_x <$$

$$\boxed{\text{C.I. for } \pi_x : \frac{1}{1+\frac{z^{*2}}{n}}\left[\left(p + \frac{z^{*2}}{2n}\right) \pm z^* \sqrt{\frac{p(1-p)}{n} + \frac{z^{*2}}{4n^2}}\right]}$$

Same 2 interpretations as before. Basically, any π in this CI is consistent with data/observations. Note: $0 < CI < 1$, as it should be for a proportion CI.

FYI So, we can't use this CI to test if $\pi = 0$ or $\pi = 1$ are consistent with data/obs.

A simple(r) eqn: If n = large, then $\boxed{p \pm z^* \sqrt{\frac{p(1-p)}{n}}}$ We'll use this one!

FYI The 1-sided CIs are obtained by simply changing z^*

π_x?

The pi_x denotes the true proportion (say, of girls) in population. In the coin-tossing analog it's the prob of a Head on a given toss. Note that this is all perfectly consistent, because the prob. of drawing a girl out of the population (ie prob of Head on a toss) is equal to the proportion of girls in the population.
Also, this pi_x is the same pi that appears in the binomial distribution. Back when we derived the binomial, the value of pi was simple given to us (eg 0.005 in an example). Now, you know how to make a confidence interval for it, too.

Example: A past survey from 390.

$$\begin{cases} \text{Lab is good} & : 17 \\ \text{'' '' bad} & : 48 \\ \text{no opinion} & : \underline{15} \\ & \ 80 \end{cases}$$

[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 categ. (The multiple-category case will be done later). So, let's consider

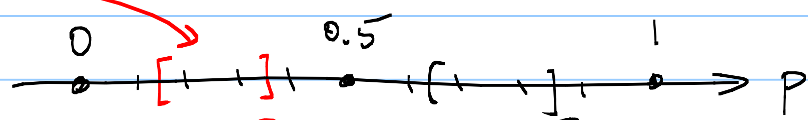
$$\begin{cases} \text{Lab is good} & : 17 \\ \text{'' '' bad} & : \underline{48} \\ & \ 65 \end{cases}$$

The sample proportion of students who like lab, p, is $p = \frac{17}{65} = .262$

Let π_x = 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 π_x: $p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = .262 \pm 1.96\sqrt{\frac{.262(1-.262)}{65}}$ ← Not 17

$= .262 \pm 0.107 = [0.16, 0.37]$



1) We are 95% Confident That π_x is in here (any π_x in here is consistent with data)

2) There is a 95% prob. That a random C.I will cover π_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: $1-(\text{CI for } \pi_x) = [0.63, 0.84]$

hw_lect17_1

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

hw-lect17-2

There are several ways of proving $E[p]=\pi$, $V[p]=\sqrt{\frac{\pi(1-\pi)}{n}}$, (dropping the subscript x, just for convenience). One way is to use a result which we have already derived, ie. $E[\bar{x}]=\mu_x$, $V[\bar{x}]=\frac{\sigma_x^2}{n}$. This result holds even if the x_i are 0 or 1. So, first,

a) Consider a sample of size n from a Bernoulli distribution, ie. n zeros and 1's, and show that the sample mean ($\bar{x}$) is equal to the sample proportion of 1's. Hint: if a sample of size n has n_0 0's and n_1 1's, then the sample prop. p is $\frac{n_1}{n}$.

So, at this point, it follows that $E[p]=E[\bar{x}]$, $V[p]=V[\bar{x}]$. But we already know that $E[\bar{x}]=\mu_x$ and $V[\bar{x}]=\sigma_x^2/n$, where μ_x and σ_x^2 are the dist. mean and dist. var. of variable taking only 0, 1 values, ie. a Bernoulli random variable. So,

b) For $x \sim$ Bernoulli(π), find μ_x and σ_x^2 starting from the definition of $E[x]$ and $V[x]$ from Ch. 2.

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