Lecture 15 (Ch. 7)

Review:
The transition from the 1st half of the course (descriptive stats) to the latter half (inference) is based on the concept of the (empirical) sampling distr. and the CLT.

The thought experiment of taking multiple samples (of size n) from a dist./pop., computing some sample statistic (e.g. $\bar{x}$) for each, and then looking at their hist (technically, distr.).

Then, we can compute probs, e.g.,

$$P(a < \bar{x} < b) = P\left( \frac{a - \mu_x}{\sigma_x/\sqrt{n}} < \frac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}} < \frac{b - \mu_x}{\sigma_x/\sqrt{n}} \right) = \text{Table I.}$$

Two types of probs are useful in statistics/data analysis:

- $P(V < \bar{x} < V_{obs}) \Rightarrow p$-value (Ch. 8)
- $P(-1.96 < Z < 1.96) = 0.95 \Rightarrow$ confidence int. (Ch. 7)

Recall our notation:

- $\bar{x}$ (sample mean) is a point estimate of $\mu_x$ (pop. mean)
- $s$ (n std. dev.)
- $\hat{p}$ (n prop.)
- $\sigma_x$ (n std. dev.)
- $\hat{p}$ (n prop.)
- $n$ (n size) is not related to pop. size, $\Rightarrow$ for us = $\infty$
- $\mu_x$ = mean of the sampling dist. of $\bar{x}$
- $\sigma_{\bar{x}}$ = std. dev.
The 1st way to build a confidence Interval (CI) for \( \mu_x \):

The procedure is to start with \( P(x < z < b) = \text{blah} \), with specific values of \( a, b, \) and \( \text{blah} \). E.g.

\[
\Pr(-1.96 < z < 1.96) = 0.95
\]

\[
\Pr(\frac{x - \mu_x}{\sigma_x} < z < 1.96) = 0.95
\]

\[
\Pr(\frac{x - \mu_x}{\sigma_x} < x - \mu_x < 1.96 \frac{\sigma_x}{\sqrt{n}}) = 0.95
\]

\[
\frac{x - 1.96 \frac{\sigma_x}{\sqrt{n}}}{\sqrt{n}} < \mu_x < \frac{x + 1.96 \frac{\sigma_x}{\sqrt{n}}}{\sqrt{n}}
\]

\[
\therefore \text{95\% CI for } \mu_x : \frac{x \pm 1.96 \frac{\sigma_x}{\sqrt{n}}}{\text{pop. mean}}
\]

This is a random CI, because \( x \) is random (how else would it have a sampling dist.?!) The (observed) 95\% CI for \( \mu_x \) is \( \bar{x}_{\text{obs}} \pm 1.96 \frac{\sigma_x}{\sqrt{n}} \)

1 Interpretation: We are 95\% Confident That \( \mu_x \) is in here.

2nd " : Below.

Often we forget saying "observed". It's up to you to find out if we're talking about a random CI or the observed CI.
A sample of size 25 yields \( \bar{x} = 3, s = 1.5 \).

Last time: Suppose pop is normal (\( \mu_x = 2, \sigma_x = 1 \)).

What's the prob of getting an even larger sample mean?

This time: What's the (observed) 95% C.I. for \( \mu_x \)?

**First attempt:** We don't use/know \( \mu_x = 2 \), but we know/use \( \sigma_x = 1 \).

(>observed) 95% C.I. for \( \mu_x \):

\[
\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}
\]

\[
3 \pm 1.96 \frac{1.5}{\sqrt{25}} = 3 \pm 0.392 = [2.6, 3.4]
\]

**Second attempt:** We don't use/know either \( \mu_x \) or \( \sigma_x \).

(Approximate \( \sigma_x \) with sample std. dev.)

We will improve on

\[
3 \pm 1.96 \frac{1.5}{\sqrt{25}} = 3 \pm 0.588 = [2.4, 3.6]
\]

This, later, when we get to \( t \)-dist.

Eitherway, here is one (of 2) interpretation of the C.I.:

We are 95% **confident** that the true mean is in here.

It's important to note that the word probability does not appear in this interpretation, even though we started with a probability.

It happened in the last step of the derivation of the C.I. for \( \mu_x \), when I dropped the \( p \). That is because \( p \) of \( \mu_x \) does not exist, because \( \mu_x \) is a fixed population value, not random.
But, there is a way of squeezing "probability" into the conclusions, but it has to pertain to the random CI.

We are 95% confident that the pop. mean is in the interval \( \bar{x} \pm 1.96 \frac{\sigma_x}{\sqrt{n}} \).

Equivalent interpretations of CI:

- There is a 95% prob. that a random sample will yield a CI. \( (\bar{x} \pm 1.96 \frac{\sigma_x}{\sqrt{n}}) \) that covers \( \mu_x \).

Look at the derivation of CI; this is obvious.

\[ \begin{align*}
\text{Sample 2} & \quad [x, \bar{x}] \\
\text{Sample 1} & \quad [x, \bar{x}] \\
\mu_x & \quad \text{pop. mean}
\end{align*} \]

\[ \bar{x} \pm 1.96 \frac{\sigma_x}{\sqrt{n}} \]

\[ \Rightarrow 95\% \text{ of these intervals cover } \mu_x. \]

\[ \Rightarrow \text{I.e. the prob. that a random C.I. } (\bar{x} \pm 1.96 \frac{\sigma_x}{\sqrt{n}}) \text{ will include } \mu_x \text{ is } 0.95. \]

\[ \Rightarrow \text{If you want to say something directly about } \mu_x, \text{ use "confidence" not prob.} \]

\[ \{
\begin{align*}
\text{C.I.s are all about coverage;} \\
\text{a 95\% C.I. for } \mu_x \text{ is designed to cover } \mu_x \text{ in 95\% of samples.}
\end{align*}
\]

\[ \text{Above Example Data} \Rightarrow 95\% \text{ observed C.I. } : [2.6, 3.4] \]

Conclusion(s):
1) We are 95\% confident that \( \mu_x \) is in [2.6, 3.4].
2) There is a 95\% prob. that a random CI covers \( \mu_x \).

No mention of anything observed.
What about other confidence levels (≠ 0.95)?

E.g. 99% conf. level: \( \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \) \( \sim \) C.I. for \( \mu \): \( \bar{x} \pm 2.575 \frac{\sigma}{\sqrt{n}} \)

In general: C.I. for \( \mu \): \( \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \)

where \( z^* = 1.645, 1.96, 2.575, \ldots \)

for conf. level = 90%, 95%, 99%, \ldots = 1 - \alpha

\( \alpha \)-level = 0.1, 0.05, 0.01, \ldots

you can either "derive" these \( z^* \) values from Table I

or look them up on the last line of Table IV.
Example problem 7.12

Concentration of zinc in 2 types of fish

\[ \bar{x} \pm \frac{s}{\sqrt{n}} \]

Type 1 5.6 9.15 1.27 \{ sample data \}
Type 2 6.1 3.08 1.71

What's the true/pop. mean for Type 1 fish, at 95% conf. level?

In the old days, all we could write was:

Type 1

\[ \bar{x} \pm \frac{s}{\sqrt{n}} \]

\[ 9.15 \pm 1.96 \frac{1.27}{\sqrt{56}} \]

9.15 \pm 0.333

(8.82, 9.48)

But now, we have an interpretation:

\[ \bar{x} \pm 2.575 \frac{s}{\sqrt{n}} \]

\[ 3.08 \pm 2.575 \frac{1.71}{\sqrt{61}} \]

3.08 \pm 0.564

(2.52, 3.64)

To go with our CI formula:

**Important**

- We are 95% confident that the true pop. mean of zinc concentration for Type 1 fish is between 8.8 and 9.5.
- There is a 95% prob. that a random sample will yield a CI that covers the true mean of zinc concentration.

Note that the 2nd interpretation makes no reference to the observed CI (8.8, 9.5) at all.

**Note:** CI for \( \mu_x \) of Type 2 fish is wider (i.e. our estimate for \( \mu_x \) is less reliable/precise). Why?

- The conf. level is higher.
- Sample std. dev. (s) is larger.
- Even though \( n \) is larger (which shrinks the CI), the increase in \( n \) is not enough to compensate for the increase in conf. level and \( s \).
The math is trivial. It’s the correct jargon & interpretations that are tricky.

There is random vs. observed vs. fixed pop. params.
\( \bar{x}, CI, \ldots \) vs. \( \bar{x}_{obs}, obs. CI, \ldots \) vs. \( \mu, \sigma, \ldots \)

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

There is also random CI vs. observed CI

Then, there is confidence vs. probability
for pop. params vs. for random thing
\( \Pr(\mu > 3) \times \) vs. \( \Pr(\bar{x} > 3) \checkmark \)

C.I. for \( \mu \) \( \checkmark \) C.I. for \( \bar{x} \) \( \times \)

We have reached our goal of being able to say something about a pop. from a sample!

Celebrate a little! But there is more (a lot more).
A sample of size 25 yields $\bar{x}_{\text{obs}} = 3$, $s_{\text{obs}} = 1.5$. Suppose we know $\sigma_x = 1$.

Then, as shown in the example, we can be 95% confident that $\mu_x$ is in the interval $(2.6, 3.4)$.

a) Can you find $p(2.6 < \bar{x} < 3.4)$? If yes, find it. If not, why not?

b) Show $p(2.6 < \bar{x} < 3.4) = p(z_{\text{obs}} - 1.96 < z < z_{\text{obs}} + 1.96)$, where

$$z_{\text{obs}} = \frac{\bar{x}_{\text{obs}} - \mu_x}{s_x / \sqrt{n}}.$$  

Hint: $2.6 = \bar{x}_{\text{obs}} - 1.96 \frac{s_x}{\sqrt{n}}$ from our data.

c) If $\mu_x = \bar{x}_{\text{obs}}$, what is the numerical value of $p(2.6 < \bar{x} < 3.4)$?

d) If $\mu_x = 2$, what is the numerical value of $\mu_x$?

e) If $\mu_x = 2.8$ and $s_x = 0.608$, $\mu_x$?

Suppose $s_x / \sqrt{n} = 1$. Find the possible $\bar{x}_{\text{obs}}$ values such that the 95% obs. CI for $\mu_x$ allows for $\mu_x$ to be zero.