Lecture 16 (Ch. 7)

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

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

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

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

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

$$\frac{z^* \sigma_x}{B} = \frac{1}{\sqrt{n}} \Rightarrow n = \left( \frac{z^* \sigma_x}{B} \right)^2$$

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

What min. sample size is required for a margin of error of 0.03 $\bar{x}$?

$$n = \left( \frac{z^* \sigma_x}{B} \right)^2 \approx \left( \frac{1.96 (1.27)}{0.03} \right)^2 = 6,885 \text{ type I Fish.}$$

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 $\sigma \approx$ a few inches.
So far: 2-sided (and 1-sided (UCB or LCB)) for \( \mu_x \) population \( \mu \).

As you may have noticed, there are some pop. 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 those pertain to a continuous random variable (\( x \)). But we also care about situations where the population consists of a 2-level categorical random variable, e.g. gender. Then, we care about estimating the true/pop. proportion of one of the 2 levels (say, girls).

That proportion is denoted \( \pi_x \). The point estimate for \( \pi_x \) is \( \hat{p} \), the sample proportion (of, say girls). Here's the interval estimate.

To build the CI for \( \pi_x \), we need the sampl. dist. of \( \hat{p} \), the sample proportion. \([\text{Recall } \bar{x} \sim N(\mu_x, \frac{\sigma_x}{\sqrt{n}})]\).

In a hw, you show that even w/o knowing the sampl. dist. of \( \hat{p} \),

\[
\hat{p} \equiv E[\hat{p}] = \pi_x \leftarrow \text{pop. proportion.}
\]

\[
\sigma_p = \sqrt{\text{Var}([\hat{p}])] = \sqrt{\frac{\pi_x(1-\pi_x)}{n}} \leftarrow \text{Note resemblance to } \frac{\sigma_x}{\sqrt{n}},
\]

CLT: \( \hat{p} \sim N(\mu_p = \pi_x, \sigma = \sigma_p = \sqrt{\frac{\pi_x(1-\pi_x)}{n}}) \) \([\text{If } n \pi_x, n(1-\pi_x) \text{ are } \text{large} (\text{say } > 5)]\).

So, again, we can find the prob. that a random sample prop. is \( \hat{p} \) is:

\[
\text{Prob}(\alpha < \hat{p} < b) = \text{Prob}(\frac{\alpha - \mu_p}{\sigma_p} < \frac{\hat{p} - \mu_p}{\sigma_p} < \frac{b - \mu_p}{\sigma_p})
\]

\[
\Rightarrow \text{Prob}(\frac{\alpha - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}} < z < \frac{b - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}}) = \text{table II}.
\]
Now that we know the sampling distr. of \( p \), we can build CI for \( \pi \).

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

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

Start with self-evident fact

\[
\begin{align*}
\text{Recall } & \quad \text{prob} \left( -1.96 < \frac{\hat{p} - \pi_x}{\frac{\pi_x(1-\pi_x)}{\sqrt{n}}} < 1.96 \right) = 0.95 \\
\text{Let } & \quad \hat{\pi}_x < \Rightarrow 95\% \text{ CI for } \pi_x.
\end{align*}
\]

\[
\begin{align*}
\text{prob} \left( -z^* < \frac{p - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}} < z^* \right) = \text{conf. level} \\
\pi_x < \Rightarrow \text{why the CI for } \pi_x \text{ is a messy eqn.}
\end{align*}
\]

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

\[
\frac{1}{1 + \frac{z^*^2}{n}} \left( p + \frac{z^*}{2n} \right) \pm z^* \sqrt{\frac{p(1-p)}{n} + \frac{z^*^2}{4n^2}}
\]

Same 2 interpretations as before. Basically, any \( \pi \) in this CI is consistent with data/observations. Note: \( \pi \notin CI \), 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 \( n \text{ is large} \), then

\[
\frac{p \pm z^* \sqrt{\frac{p(1-p)}{n}}}{n}\]

We’ll use this one!

(FYI) The 1-sided CIs 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.

\[\begin{align*}
\text{Lab is good} & : 17 \\
\text{bad} & : 48 \\
\text{no opinion} & : 15 \\
\text{Total} & : 80
\end{align*}\]

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 cats. (The multiple-category case will be done later). So, let’s consider

\[\begin{align*}
\text{Lab is good} & : 17 \\
\text{bad} & : 48 \\
\text{Total} & : 65
\end{align*}\]

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

Let \( \hat{\pi}_x \) = True/actual prop. of students who like lab.

(The True prop. of students who don’t like the lab is \( (1 - \hat{\pi}) \).)

95% C.I. for \( \hat{\pi}_x \): \( \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.262 \pm 1.96 \sqrt{\frac{0.262(1-0.262)}{65}} \approx 0.262 \pm 0.107 = [0.155, 0.369] \)

1) We are 95% confident that \( \hat{\pi}_x \) is in here (any \( \pi \) in here is consistent with data)
2) There is a 95% prob. that a random C.I will cover \( \hat{\pi}_x \).
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 \( (1-\pi) \) is: \( 1 - \text{(CI for } \pi) = [0.63, 0.84] \)
In all of the above problems, we have been dealing with the C.I. for a single $\mu_x$ or a single $\pi_x$. But there are times when all we care about is some kind of comparison between 2 $\mu$'s or between 2 $\pi$'s, e.g. $\mu_1$ vs $\mu_2$ or $\pi_1$ vs $\pi_2$. Note that I’m dropping the $x$ subscript to keep notation simple.

For example, here is a question pertaining to 2 means:

Is the mean CPU speed of Mac computers $= \mu_1$ different from that of Dell computers $= \mu_2$?

We could build C.I.'s for $\mu_1$ and $\mu_2$, separately, and compare.

But, better way is to build a C.I. for the difference:

C.I. for $\mu_1$ vs $\mu_2$ or for $\pi_1$ vs $\pi_2$.

Dropping the x for simplicity.

These are called 2-sample C.I.

2-Sample problems involving 2 means are easy to recognize. Examples involving 2 props are more tricky. Here is a correct one:

Is the prop. of Mac users among boys different from that of Mac users among girls?

Note $\pi_1 + \pi_2 \neq 1$. I.e. $\pi_1, \pi_2$ are 2 different props.

Here is an incorrect example:

Is the proportion of people who use Macs different from that of other computers?

The 2 props in this example are constrained: $\text{prop(Macs)} + \text{prop(other)} = 1$.

So, it’s like the lab example (last lect): There is only 1 independent prop.
To build a CI for $\mu_1 - \mu_2$ or $\pi_1 - \pi_2$, whose sampling dist. do we need? 

$$(\bar{x}_1 - \bar{x}_2) \quad or \quad (\hat{p}_1 - \hat{p}_2)$$

The E and V of the sampling dist. are

$$\begin{align*}
E[\bar{x}_1 - \bar{x}_2] &= E[\bar{x}_1] - E[\bar{x}_2] = \mu_1 - \mu_2, \\
V[\bar{x}_1 - \bar{x}_2] &= V[\bar{x}_1] + V[\bar{x}_2] - 0 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.
\end{align*}$$

$\bar{x} \sim N(\mu, \frac{\sigma}{n})$

$\bar{x} \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$

$\bar{x} \sim N(\mu, \frac{\sigma^2}{\sqrt{n}})$

$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$

$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$

Self-evident fact: 1.96, 95%

$$\text{prob}(-z < z < z^*) = \text{Conf. level}$$

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < \mu_1 - \mu_2 < \frac{\bar{x}_1 - \bar{x}_2 + (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Setting this to $B$, will give minimum sample size.

C.I. for $\mu_1 - \mu_2$:

if samples = indep. 

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{obs}$$

Interpretation: Same as before.

Similarly,

C.I. for $\pi_1 - \pi_2$:

if samples = indep. 

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

again, for now, approximate $\sigma_1, \sigma_2$ with $s_1, s_2$. 

$$s = \frac{(x_1 + x_2)}{2}$$
Example: Here is another data set:

<table>
<thead>
<tr>
<th>Winter quarter</th>
<th>Spring quarter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab is good:</td>
<td>10 (0.152)</td>
</tr>
<tr>
<td>Bad:</td>
<td>56 (0.848)</td>
</tr>
<tr>
<td></td>
<td>66</td>
</tr>
</tbody>
</table>

Again, the hard part is figuring out what to ask! Let's focus on the "bads." Then, we can ask:

Does data provide sufficient evidence to claim that the proportion of "bads" in the 2 populations are different? (90% Conf. Int.)

\( \pi_1 = \text{prop. of students in pop. who don't like lab in Winter} \)
\( \pi_2 = \text{prop. of students in pop. who don't like lab in Spring} \)

So, we need a 2-sided 90% C.I. for \( \pi_1 - \pi_2 \):

\[
(\hat{\pi}_1 - \hat{\pi}_2) \pm Z^* \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}
\]

\[
(0.848 - 0.738) \pm 1.645 \sqrt{\frac{0.848(1-0.848)}{66} + \frac{0.738(1-0.738)}{65}} = 0.11 \pm 0.115
\]

\[
(-0.005, 0.225)
\]

Interpretation: 1) We are 90% confident that \( \pi_1 - \pi_2 \) is in \((-0.005, 0.225)\).

Corollary: Zero is included in the interval.

Correct Conclusion: Cannot conclude that \( \pi_1 \) and \( \pi_2 \) are different. "..." "..." "..." "..." equal.

Sure, you are thinking that it is possible that they are equal. But the data provide no evidence for it!

The data provide no evidence that they are different either! Basically, we cannot conclude anything about their relative size.

Incorrect Concl.: "we can conclude that \( \pi_1 \) and \( \pi_2 \) are equal." (Very Important Distinction.)
Example: 82 students have picked up their test, but 30 have not, even 1 week after the test was returned. Call these 2 groups “Attendees” and “Non-attenders.”

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>x̄</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-attend</td>
<td>30</td>
<td>11.8</td>
<td>3.32</td>
</tr>
<tr>
<td>Attend</td>
<td>82</td>
<td>13.25</td>
<td>3.04</td>
</tr>
</tbody>
</table>

For now, approximate with $\sigma_1^2, \sigma_2^2$.

$(13.25 - 11.8) \pm 1.96 \sqrt{\frac{(3.32)^2}{30} + \frac{(3.04)^2}{82}} = 1.45^+ - 1.96 (-2.81)$

$1.45 \pm 1.36 = (0.09, 2.81) \Rightarrow M_2 - M_1 \rightarrow \bar{x}_2 - \bar{x}_1$

Important: We are 95% confident that $(M_2 - M_1)$ is in here.

Interpretation: 0.09

Corollary: Zero is not included in that interval, so there is evidence that there is a difference between the mean of attending and non-attending students, with 95% confidence.

FYI: In fact, because the entire CI is to the right of zero, we can say that attending students have a higher mean. However, this conclusion is not true with 95% confidence, but a slightly higher confidence. If we are really interested in whether one mean is larger (or smaller) than another mean, then we should build 1-sided UCB or LCB.
Example: Back to the fish example:

Concentration of zinc in 2 types of fish.

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>$\bar{x}$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>54</td>
<td>9.15</td>
<td>1.27</td>
</tr>
<tr>
<td>Type II</td>
<td>61</td>
<td>3.08</td>
<td>1.71</td>
</tr>
</tbody>
</table>

Suppose we ask: Are the true/pop. means different?

$\mu_1 =$ pop. mean zinc in Type I

$\mu_2 =$ pop. mean zinc in Type II

Important to define $\mu_1, \mu_2$ (the pop. parameters) clearly.

95% C.I. for $\mu_1 - \mu_2$:

\[
(9.15 - 3.08) \pm 1.96 \sqrt{\frac{(1.27)^2}{54} + \frac{(1.71)^2}{61}}
\]

\[
6.07 \pm 0.54
\]

\[
[5.53, 6.61]
\]

Important

Interpretation:
1) We are 95% confident that $\mu_1 - \mu_2$ is in
2) There is 95% prob. that a random C.I. will include $\mu_1 - \mu_2$.

Corollary: The number zero is not included in the C.I.

So, there is evidence that $\mu_1 \neq \mu_2$.

Because the C.I. is entirely to the right of 0, there is evidence that $\mu_1 > \mu_2$, but not with 95% confidence.

The appropriate test of whether $\mu_1 > \mu_2$ requires building the lower conf. bound (LCB) for $\mu_1 - \mu_2$.

Note: The qualitative comparison of boxplots that we learned to do in Ch. 1.2 is now more quantitative. The only subjectivity is in the choice of the conf. level.
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.

For the data you collected, consider one of the continuous variables (call it y), and one of the categorical/discrete variables (call it x). Let \( \mu_1 \) denote the true mean of y when \( x = \) (first level of x), and \( \mu_2 \) denote the true mean of y when \( x = \) (2nd level of x).

a) compute a 2-sided, 95% C.I. for \( \mu_1 - \mu_2 \).

b) Is there evidence from data that \( \mu_1 \) and \( \mu_2 \) are different?

Let \( p_1 \) denote the true proportion of defective bridges in the USA, and \( p_2 \) .... in Canada.

A sample of \( n_1 = 80 \), and \( n_2 = 50 \) bridges from the two countries, respectively, is taken, and it is found that 21% of the bridges in the USA, and 10% of the bridges in Canada are defective. At 95% confidence level

a) Is there evidence that the true proportions are different?

b) Is there evidence that \( p_1 \) is larger than \( p_2 \)?

There are several ways of proving \( \mathbb{E}[p] = \pi \), \( \text{Var}[p] = \frac{\pi(1-\pi)}{n} \),

dropping the subscript \( \pi \) just for convenience. One way is to return to our derivation of \( \mathbb{E}[\bar{x}] = \mu_x \), \( \text{Var}[\bar{x}] = \frac{\sigma^2}{n} \), 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 \( \frac{n_1}{n} \).

b) Therefore, \( \mathbb{E}[\bar{x}] = \mathbb{E}[p] = \mu_x \), and \( \text{Var}[\bar{x}] = \text{Var}[p] = \frac{\sigma^2}{n} \). Finally, for \( x \sim \text{Bernoulli}(\pi) \), find \( \mu_x \) and \( \sigma^2 \) starting from the defn. of \( \mathbb{E}[x] \) and \( \text{Var}[x] \) from CH.2.

You are done, you will have proven \( \mathbb{E}[p] = \pi \), \( \text{Var}[p] = \frac{\pi(1-\pi)}{n} \),

using equations that we had proven before, i.e. \( \mathbb{E}[\bar{x}] = \mu_x \), \( \text{Var}[\bar{x}] = \frac{\sigma^2}{n} \).