Lecture 21 (Ch. 7)

So far, we've done 1- and 2-sided C.I for μ_x.

2-sided $\quad \bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$ $\quad$ different z^*

Lower conf bound: $\bar{x} - z^* \frac{\sigma_x}{\sqrt{n}}$ $\quad$ upper: $\bar{x} + z^* \frac{\sigma_x}{\sqrt{n}}$

What about C.I. for pop. proportion π_x?

To build the C.I for π_x, we need the sampl. distr. of p, the sample proportion. [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 p,

$\mu_p \equiv E[p] = \pi_x$ $\quad$ pop. proportion.

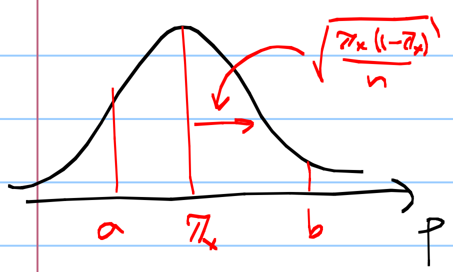
$\sigma_p \equiv \sqrt{V[p]} = \sqrt{\frac{\pi_x(1-\pi_x)}{n}}$ $\quad$ Note resemblance to $\sigma_x/\sqrt{n}$, where $\sigma_x = \sqrt{\pi_x(1-\pi_x)}$ [Binomial with $n=1$, also called Bernoulli(π).

CLT: $p \sim N\left(\mu = \mu_p = \pi_x,\ \sigma = \sigma_p = \sqrt{\frac{\pi_x(1-\pi_x)}{n}}\right)$

If $n\pi$, $n(1-\pi)$ are "large" (say > 5).

Therefore, we can again compute the prob that the sample prop. is between 2 numbers (or $<$ or $>$ ...):

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



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

$$= \text{table I}.$$

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

→ 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.

$< \pi_x <$

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

Good News: If n = large (> 30), then $\boxed{p \pm z^* \sqrt{\frac{p(1-p)}{n}}}$

simple eqn

Finally, 1-sided C.I. affects z^* only.

π_x denotes the proportion (e.g. of "goods") in pop. In the coin-tossing analog π_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 π_x is the same π that appeared in binomial. Now, you know how to make a confidence interval for it!

Example: A past survey from 390.   Sample props, p

Lab is good	: 17	21.25% ~ .21	= 17/80
" " bad	: 48	60.00% ~ .60	= 48/80
no opinion	: 15	18.75% ~ .19	= 15/80
	80		

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. Let's build the 95% C.I. for the various pop. proportions:

In this case the population of students consists of 3 categories: "Lab is Good", "Lab is Bad", "No opinion." So, we have 3 proportions for which to build CI's:

π_1: True prop. of students who think Lab is Good

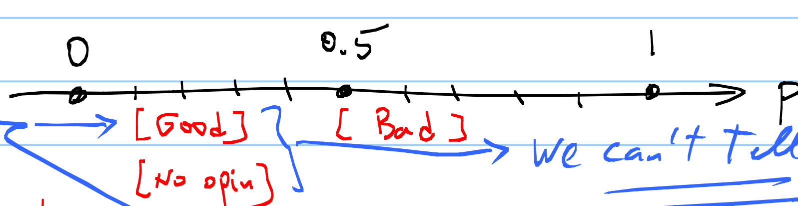
π_2: " " " " " " Bad

π_3: " " " " have no opinion.

[Note that these 3 props are NOT indep., because $\pi_1+\pi_2+\pi_3=1$
This will become important for CIs we will compute later.

Good	Bad	No opinion
$0.21 \pm 1.96\sqrt{\frac{.21(1-.21)}{80}}$ ← NOT 17!	$0.60 \pm 1.96\sqrt{\frac{.6(1-.6)}{80}}$	$0.19 \pm 1.96\sqrt{\frac{.19(1-.19)}{80}}$
0.21 ± 0.09	0.6 ± 0.11	0.19 ± 0.09
[.12, .3]	[.49, .71]	[.1, .28]

We are 95% confident that the proportion of students who like the Lab is here.



We can't tell if there is a diff.

important (see examples, below)

This page FYI

We skipped sec. 3.6, but it has one result that we keep using:

constant

$E[ax] = a\,E[x] \qquad V[ax] = a^2\,V[x]$

$E[x \pm y] = E[x] \pm E[y]$

always plus.

x and y indep.

$V[x \pm y] = V[x] + V[y] + 0$

I'm going to include a derivation of one of these, but only FYI

Pf: with 2 (instead of 1) variable, E[ ] is defined as:

$E[x] = \sum_x x\, p(x) \implies E[x] = \sum_x \sum_y x\, p(x,y)$

where $\sum_x \sum_y p(x,y) = 1$ [defn. of density]

and $\sum_x p(x,y) = p(y), \quad \sum_y p(x,y) = p(x)$

Then it's easy to derive the above results. E.g.

$$
\begin{aligned}
E[x \pm y] &= \sum_x \sum_y (x \pm y)\, p(x,y) \\
&= \sum_x \sum_y x\, p(x,y) \pm \sum_x \sum_y y\, p(x,y) \\
&= \sum_x x \underbrace{\sum_y p(x,y)}_{p(x)} \pm \sum_y y \underbrace{\sum_x p(x,y)}_{p(y)} \\
&= E[x] \pm E[y].
\end{aligned}
$$

FYI

Now, something "new":

What we have so far is 1-Sample C.I. (1 sided and 2-sided) for μ_x and π_x.

In some situations, all we really care about is some kind of comparison between 2 means or 2 proportions.

For example, is The mean CPU speed of Mac computers higher Than That of Dell computers? or is The proportion of people who have iphones different from The proportion of people who have Samsung phones?

Better way is to build a C.I. for The difference.

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

Dropping The x for simplicity.

These are called 2-sample C.I. ← 2 populations.

Analogous to

Q Whose sampling distr. do we need?

$(\bar{x}_1 - \bar{x}_2)$ or $(p_1 - p_2)$ — $\bar{x}$, p

Q What are Their E[ ] and V[ ]

$E[\bar{x}_1 \pm \bar{x}_2] = E[\bar{x}_1] \pm E[\bar{x}_2] = \mu_1 \pm \mu_2.$ — $E[\bar{x}] = \mu_x$

$V[\bar{x}_1 \pm \bar{x}_2] = V[\bar{x}_1] + V[\bar{x}_2] - 0 = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}.$ — $V[\bar{x}] = \dfrac{\sigma^2}{n}$

indep. $\bar{x}_1, \bar{x}_2$.

CLT $\Rightarrow$ sampling distr. of $(\bar{x}_1 - \bar{x}_2)$ is Normal with above params.

Even if pop is non-normal, but then must have n_1, n_2 = large

<u>Q</u> What's the quantity that has std. Normal?

<u>A</u> $$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ [analogous to $\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$]

Self-evident fact:

1.96, 95%

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

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\cdots}}$$

$$\cdots < \mu_1 - \mu_2 < \cdots$$

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

if samples = indep.

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

Setting this to B, will give minimum sample size.

Interpretation: Same as before

again, for now, approximate σ_1, σ_2 with s_1, s_2.

Similarly,

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

7.25 in book

if samples = indep.

$$\underbrace{(p_1 - p_2)}_{\text{obs.}} \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Q1: Does it make sense to talk about a 1-sided C.I. for $\pi_1 - \pi_2$?

A) Nope B) Yup C) Only when $\pi_1 - \pi_2 > 0$

E.g. We want to know if $\pi_1 > \pi_2$

Then build lower Conf. Bound. for $\pi_1 - \pi_2$

or " upper " " for $\pi_2 - \pi_1$

Example: Here is another data set:

	pop. 1 Spring quarter	pop. 2 Winter quarter
Lab is good :	17 (.21)	10 (0.10)
" Bad :	48 (.60) = p_1	55 (0.55) = p_2
No opinion :	15 (.19)	35 (0.35)
	80	100

According to sample, the prop. of students who do NOT like Lab in Spring (0.6) is very close to that in Winter (0.55). Does this data provide sufficient evidence to claim that the 2 props in the populations are different? at 95% conf. level.

π_1 = prop. of students in pop. who don't like Lab, in Spring
π_2 = " " Winter

So, we need a 2-sided 95% C.I. for $\pi_1 - \pi_2$:

$$(p_1 - p_2) \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$(0.6 - 0.55) \pm 1.96 \sqrt{\frac{0.6(0.4)}{80} + \frac{0.55(0.45)}{100}} = (-0.1, 0.2)$$

Interpretation: 1) We are 95% confident that $\pi_1 - \pi_2$ is in .

Corollary: zero is included in the interval.

Correct Conclusion: So, we cannot conclude anything about the relative size of π_1 and π_2!

Note that I did NOT say

Incorrect Concl. → "we can conclude that π_1 and π_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 "Attenders" and "Non-attenders".

		n	$\bar{x}$	s
1	Non-attend	30	11.8	3.32
2	Attend	82	13.25	3.04

} Data

Sample suggests that mean of Attend is higher than Non-attend.

Is this true for the population (ie. all 390 courses)? 95% Conf. level

$\bar{x}_2 > \bar{x}_1$ ie. $\bar{x}_2 - \bar{x}_1 > 0$. $\mu_2 - \mu_1 > 0$?

Important

μ_1 = mean of test 1 for Non-attend students

μ_2 = " " Attend students.

The hardest part of these problems is determining what type of CI to compute.

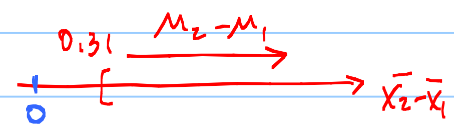
We need to build the LOWER conf. bound for $\mu_2 - \mu_1$:

$$(\bar{x}_2 - \bar{x}_1) - 1.645\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(13.25 - 11.8) - 1.645\sqrt{\frac{(3.32)^2}{30} + \frac{(3.04)^2}{82}} = 1.45 - 1.645(.693)$$

Important

$$1.45 - 1.14 = 0.31 \implies$$



Interpretation: We are 95% confident that $\mu_2 - \mu_1 > 0.31$

ie. that μ_2 exceeds μ_1 by at least* 0.31.

Corollary: Zero is not included in that interval. So there is evidence that attending students have a higher pop. mean than Non-attend.

* Note: "μ_2 exceeds μ_1 by 13" translates to $\mu_2 = \mu_1 + 13$

Example: Lab (Good/Bad), but by quarter.

	pop. 1 Spring quarter	pop. 2 Winter quarter
Lab is good :	17 $(.21) = p_1$	10 $(0.10) = p_2$
" Bad :	48 (.60)	55 (0.55)
No opinion :	15 (.19)	35 (0.35)
	80	100

According to the sample, the prop. of students who like the lab is higher in Spring than Winter. Is this true for the population?

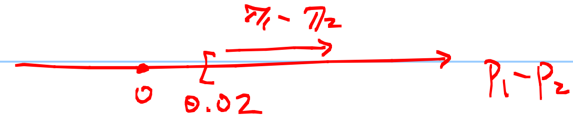
i.e. $p_1 > p_2$, ie. $p_1 - p_2 > 0$ $\longrightarrow$ $\pi_1 - \pi_2 > 0$?

Let π_1 = prop. of students in pop. who like the Lab in Spring.
π_2 = " " Winter } Important

We need to build the LOWER 95% conf. bound for $\pi_1 - \pi_2$:

$$(p_1 - p_2) - 1.645\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$(.21 - .10) - 1.645\sqrt{\frac{.21(0.79)}{80} + \frac{.1(.9)}{100}} = 0.11 - 0.09 = 0.02$$



Important.

Interpretation: 1) We are 95% confident that the true difference $\pi_1 - \pi_2$ is greater than 0.02 (or π_1 exceeds π_2 by at atleast 0.02)

2) There is 95% prob. that a random lower conf. bound will be lower than $\pi_1 - \pi_2$.

Corollary : zero is not included in that interval.
So, $\pi_1 > \pi_2$ (but $\boxed{\text{NOT}}$ with 95% confidence.) how below

Example Back to problem 7.12 (from previous lecture)

Concentration of zinc in 2 types of fish.

	n	$\bar{x}$	s
Type I	56	9.15	1.27
Type II	61	3.08	1.71

The sample says that the 2 types of fish have different mean zinc levels. Is this true of the pop. means?

μ_1 = pop. mean zinc in Type I
μ_2 = " " II

Important to define μ_1, μ_2 (the pop. parameters) clearly.

95% C.I. for $\mu_1 - \mu_2$: $(9.15 - 3.08) \pm 1.96 \sqrt{\frac{(1.27)^2}{56} + \frac{(1.71)^2}{61}}$

$$6.07 \pm 0.54$$

$$[5.53, 6.61]$$

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 $\mu_1 \neq \mu_2$.

Note: Because the C.I. is entirely to the right of ϕ, there is evidence that $\mu_1 > \mu_2$, but not with 95% conf.
A more appropriate test of whether $\mu_1 > \mu_2$ requires building the lower conf. bound for $\mu_1 - \mu_2$.

Summary:

→ (Large sample), 2- and 1-sided C.I. for μ_x and π_x

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}} \quad , \quad p \pm z^* \sqrt{\frac{p(1-p)}{n}}$$

also called 1-sample z-intervals.

→ 2-sample z-intervals are for comparing 2 (independent) Samples taken from 2 populations:

C.I. for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm \boxed{z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

C.I. for $\pi_1 - \pi_2$: $(p_1 - p_2) \pm \boxed{z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$

setting this to B, gives n_{min}, like before.

→ And both of These 2-sample intervals come in a 1-sided and 2-sided variety, paired and unpaired.

upper or lower

→ All C.Is should be interpreted in atleast 1 of 2 ways we have discussed. These interpretations say something about ~~reliability~~ precision. A wider C.I. is less ~~reliable~~ precise, and vice versa.

→ Each interpretation of a C.I. is also accompanied by what I call a corollary, where the C.I is used to make a decision.

Study all The examples here carefully, even if I don't cover Them in class.

hw_lect21_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 at most 2.5% of these screws suffer from this defect? Explain your reasoning for what is the most appropriate type of interval (2-sided, lower or upper conf bound), and the conclusion that follows from it. You may use the "simple formula" appropriately revised .

hw_lect21_2

Let p, the sample proportion of girls, be written as $p = \frac{n_g}{n}$, where n = sample size, and n_g = number of girls in n. Show that

$E[p] = \pi_x \quad V[p] = \pi_x(1-\pi_x)/n$ where π_x = prop. of girls in the pop.

Do not use sums of 0's and 1's, like the book does.
Instead repeat the way we derived $E[\bar{x}]$ and $V[\bar{x}]$, above but now keeping in mind that something in this problem is actually binomial. Hint: find out what's binomial, first.

hw_lect21_3

In one example above, we tested $\pi_1 - \pi_2$, where

π_1 = prop. of students in pop. who like the Lab in Spring.
π_2 = " " Winter.

We found that we are 95% Confident that $\pi_1 - \pi_2 > 0.2$. That does suggest that $\pi_1 - \pi_2 > 0$ (i.e. $\pi_1 > \pi_2$), but not with 95% Confidence. Determine the conf. level at which $\pi_1 - \pi_2 > 0$.

hw_lect21_4

Let π_1 be the true proportion of defective bridges in the USA.
" π_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 π_1 and π_2 are different?

b) " " " " π_1 is greater than π_2?