# Lecture 18 (Ch. 7)

$\boxed{\sigma_x ?}$

Last time we built the CI for $\mu_x$ (pop. mean): $\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$ ← approx with $s_x$

and the CI for $\pi_x$ (pop. prop.): $p \pm z^* \sqrt{\frac{p(1-p)}{n}}$

Let's take care of the business of $\sigma_x$

Consider the (1-sample, 2 sided) C.I. for $\mu_x$: $\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$

We derived it from $z \equiv \frac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}} \sim N(0,1)$.

In practice, however, the CI is computed as $\bar{x} \pm z^* \frac{s_x}{\sqrt{n}}$

So, it's natural to ask what is the dist. of $\frac{\bar{x} - \mu_x}{s_x/\sqrt{n}}$.

In fact, upon a little thinking you can see that it cannot have a normal dist. For example, ask yourself which of the following has the "wider" sampling distr?

$$\underset{\text{r.v.}}{z} = \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} \ (\sigma_x \text{ fixed}) \quad \text{or} \quad t = \frac{\bar{X} - \mu_x}{s_x/\sqrt{n}}$$

This one has a wider dist. because it has 2 sources of variability, $\bar{x}$, $s_x$.

An English statistician (Gosset) worked out the distr. of t:

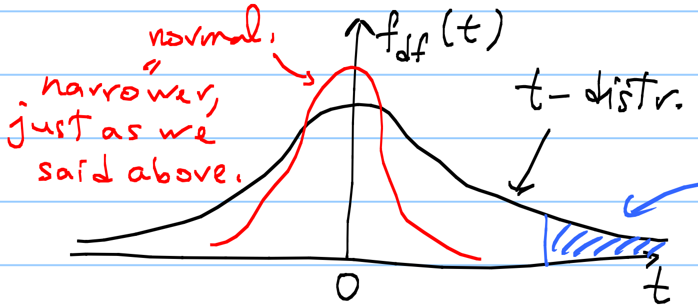
$z \sim$ Normal $(0,1)$

$t \sim$ t-distribution with "df" degrees of freedom ← param. of t-distr., like $\sigma^2$ of Normal.

$$f_{df}(t) = \frac{\Gamma\left(\frac{1}{2}(df+1)\right)}{\sqrt{\pi(df)}\,\Gamma\left(\frac{1}{2}df\right)\sqrt{\left(1+\frac{t^2}{df}\right)^{df+1}}}$$

This is just FYI. As far as you are concerned, the t-distr. is just another Table

Table VI ← 6 not 4!




if $df \to \infty$, then $t \to z$.

Table VI (6) gives Right areas.

Thm (Student's $t$)

any size, small or large.

For a sample of size $n$, from a Normal pop.

$t = \dfrac{\bar{x} - \mu_x}{s_x/\sqrt{n}}$ has a $t$-dist. with $df = n-1$

[As $n \to \infty$, $df \to \infty$, $\therefore t \to z$

[Analogous to $z = \dfrac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}}$ has a normal distr. with $\mu = 0, \sigma = 1$.]

If the pop. is not Normal, we don't know the distr. of $t$.
As a result of this, everything we do based on $t$ requires the distr. of the population to be Normal.

This is a restriction that does not effect the $z$-interval.
But for $t$, pop. should be Normal.
(or is assumed to be)

Now we can build a C.I. for $\mu_x$ based on the $t$-dist:

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

"self-evident fact"

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}} \Rightarrow \cdots \Rightarrow \cdots < \mu_x < \cdots$$

$\therefore$ C.I. for $\mu_x$ : $\bar{x} \pm t^* \dfrac{s_x}{\sqrt{n}}$ with $df = n-1$

Derive from Table VI (6).

This interval is also known as a "small sample C.I." or a $t$-interval.

see below for why "small".

Example: Sample of size 10 from a Normal pop, yields $\bar{x}=20$, $s=2$.

We are 70% confident that $\mu_x$ is in

$20 \pm 1.1 \left(\frac{2}{\sqrt{10}}\right) = [19.30, 20.70]$

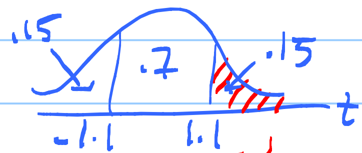
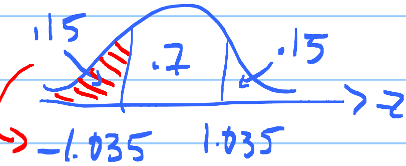



table VI (below)

Note: This is wider than the z-interval:

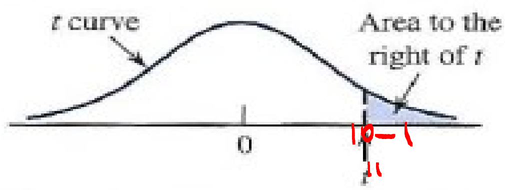
$20 \pm 1.035 \left(\frac{2}{\sqrt{10}}\right) = [19.35, 20.65]$

Table I




Recall that a 95% CI is designed to cover the pop. param. 95% of the time. The "t-interval" (with $t^* = 2.13$) has that property. The "z-interval" (with $z^* = 1.96$) is narrower, and so it covers $\mu_x$ less than 95% of the time.

**Table VI** Tail areas for *t* curves




| t \ df | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 |
| 0.1 | .468 | .465 | .463 | .463 | .462 | .462 | .462 | .461 | .461 |
| 0.2 | .437 | .430 | .427 | .426 | .425 | .424 | .424 | .423 | .423 |
| 0.3 | .407 | .396 | .392 | .390 | .388 | .387 | .386 | .386 | .386 |
| 0.4 | .379 | .364 | .358 | .355 | .353 | .352 | .351 | .350 | .349 |
| 0.5 | .352 | .333 | .326 | .322 | .319 | .317 | .316 | .315 | .315 |
| 0.6 | .328 | .305 | .295 | .290 | .287 | .285 | .284 | .283 | .282 |
| 0.7 | .306 | .278 | .267 | .261 | .258 | .255 | .253 | .252 | .251 |
| 0.8 | .285 | .254 | .241 | .234 | .230 | .227 | .225 | .223 | .222 |
| 0.9 | .267 | .232 | .217 | .210 | .205 | .201 | .199 | .197 | .196 |

z vs. t <=> known vs. unknown sigma <=> large sample vs. small sample

Note that the basic difference between The z-interval and the t-interval is in whether we know sigma, or not, respectively. So, in books the t-interval often appears under the header "Known sigma," and the t-interval is under the header "Unknown sigma." But often these 2 intervals are also called "large-sample CI," and "small-sample CI," respectively. The reason for that naming is that if the sample is large, then the sample std dev s is going to be a very good approximation of sigma, and so, we can use our CI formula with s instead of sigma. When the sample is small, then s is not a good approximation of sigma, and so, we use the t-based CI.

So far, in all of our examples, 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. $\mu_1 - \mu_2$ or $\pi_1 - \pi_2$

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 CI's for $\mu_1$ or $\pi_1$ and $\mu_2$ $\pi_2$, seperately, and compare.

But, 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.

---

$\Rightarrow$ 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 $= \pi_1$
from " " of Mac " " girls? $= \pi_2$

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

$\Rightarrow$ Here is an incorrect example:

Is the proportion of people who use Macs different $= \pi$
from " " " " " " other computers? $= 1-\pi$

The 2 props in this example are constrained: prop(Macs) + prop(other) = 1

So, it's like the lab example, above; there is only 1 indep. prop.

Q: To build a CI for $\mu_1 - \mu_2$ or $\pi_1 - \pi_2$, whose sampling distr. do we need?

A: $(\bar{x}_1 - \bar{x}_2)$ or $(p_1 - p_2)$

The E and V of the sampling dist. are

$$E[\bar{x}_1 - \bar{x}_2] = E[\bar{x}_1] - E[\bar{x}_2] = \mu_1 - \mu_2 .$$

$$V[\bar{x}_1 - \bar{x}_2] = V[\bar{x}_1] + V[\bar{x}_2] - \underset{\bar{x}_1, \bar{x}_2 \text{ indep.}}{0} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} .$$

$\Rightarrow$ CLT: $\bar{x}_1 - \bar{x}_2 \sim N( \ , \sqrt{\ } )$

Q What's the quantity that has a z (t) dist?

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \text{t-distr. with}$$
df = Welch (see below)

Self-evident fact:

$pr(-z^* < z < z^*) = \text{Conf. level}$ — solve for $\mu_1 - \mu_2$

$pr(-t^* < t < t^*) = \text{Conf. level}$ — Solve for $\mu_1 - \mu_2$

CI for $\mu_1 - \mu_2$:

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

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
df = Welch.

Analogous to C.I. for $\mu_x$:

$\bar{x}$

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

$V[\bar{x}] = \frac{\sigma_x^2}{n}$

$\bar{x} \sim N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

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

$t = \frac{\bar{x} - \mu_x}{s_x / \sqrt{n}} \sim t$, df = n-1

$-z^* < \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} < z^*$

$\cdots < \mu_x < \cdots$

$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

$\left\{ \begin{array}{l} \bar{x} \pm t^* \frac{s}{\sqrt{n}} \\ df = n-1 \end{array} \right.$

Interpretation: Same as before. $\left\{ \begin{array}{l} \text{random CI} \Rightarrow 1^{st} \text{ interp.} \\ \text{observed CI} \Rightarrow 2^{nd} \text{ interp.} \end{array} \right.$

---

It's not important, but the df for 2-sample CI is given by Welch:

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

That was CI for $\mu_1-\mu_2$. The analog for $\pi_1-\pi_2$ is:

See 7-27

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

t-based CI for props does not exist.

↓↓↓ Go over all These examples ↓↓↓ I'll go over Them next time, too.

Example: Here is another data set:

| | pop. 1 winter quarter | pop. 2 Spring quarter |
|---|---|---|
| Lab is good: | 10 (.152) | 17 (.262) |
| " Bad: | 56 (.848) = $p_1$ | 48 (.738) = $p_2$ |
| | 66 | 65 |

Does data provide sufficient evidence to claim That The proportion of "bads" in The 2 populations are different? 90% Conf. level

$\pi_1$ = prop. of students in pop. who don't like Lab, in Winter

$\pi_2$ = " " " in Spring

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

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

$$(.848-.738) \pm 1.645\sqrt{\frac{.848(1-.848)}{66}+\frac{.738(1-.738)}{65}} = .11 \pm .115 = (-.005, .225)$$

Interpretation: 1) We are 90% Confident That $\pi_1-\pi_2$ is in ↑. 2) ...

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 $\pi_1-\pi_2$.

Incorrect Concl. → $\pi_1$ and $\pi_2$ are same. Very big error!

**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 |

} sample

Important

$\mu_1$ = mean of test1 for Non-attend students who have ever taken 390. (population)

$\mu_2$ = " " Attend students " " " " ".

Is there evidence from data that $\mu_1$ and $\mu_2$ are different?

We need to build the 2-sample (2-sided) CI for $\mu_2 - \mu_1$:

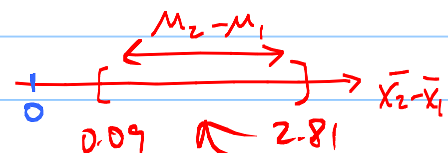
$$(\bar{x}_2 - \bar{x}_1) \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

95% — I'm using $z^*$ here, just to focus on the interp. of CI's.

$$(13.25 - 11.8) \pm 1.96 \sqrt{\frac{(3.32)^2}{30} + \frac{(3.04)^2}{82}} = 1.45 \pm 1.96\ (.693)$$

Important

$$1.45 \pm 1.36 = (0.09, 2.81) \Rightarrow$$




Interpretation: We are 95% confident that $(\mu_2 - \mu_1)$ is in here.

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.

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

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

$\mu_1$ = pop. mean zinc in Type I
$\mu_2$ = " " II

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

This time, let's use $t^*$ to get some practice.

df Welch $= \cdots = 110.32 \Rightarrow$
not important.

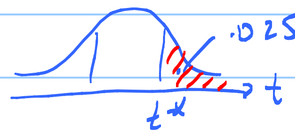



Table VI $\Rightarrow t^* \approx 2.0$
R $\Rightarrow$ qt(.025, df) = 1.98

95% C.I. for $\mu_1 - \mu_2$ : $(9.15 - 3.08) \pm 1.98 \sqrt{\frac{(1.27)^2}{56} + \frac{(1.71)^2}{61}}$

$6.07 \pm 0.55 = [5.52, 6.62]$

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: 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.

FYI

Because the C.I. is entirely to the right of $\phi$, there is evidence that $\mu_1 > \mu_2$, but not with 95% conf.
The appropriate test of whether $\mu_1 > \mu_2$ requires building the lower conf. bound (LCB) for $\mu_1 - \mu_2$.

hw_lect18_1 use t

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 mu1 denote the true mean of y when x = (first lelvel of x), and mu2 denote the true mean of y when x= (2nd level of x).

a) compute a 2-sided, 95% C.I. for mu1-mu2.

b) Is there evidence from data that mu1 and mu2 are different?

hw_lect18_2

Let pi_1 denote the true proportion of defective bridges in the USA, and pi_2 .... in Canada.

A sample of n1=80, and n2=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 pi_1 is larger than pi_2?

skip part b, because it requires 1-sided CIs, which we are skipping this quarter.