Lecture 18 (Ch. 7 - end)

Unknown σ_x

Go over The examples in last lecture!

Consider The 1-sample, 2 sided C.I. for μ_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.

To see That $\frac{\bar{x}-\mu_x}{s_x/\sqrt{n}}$ is not normal, ask yourself

which of The following has the "wider" sampling distr?

r.v.

$$z = \frac{\bar{X} - \mu_x}{\sigma_x/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{X} - \mu_x}{S_x/\sqrt{n}}$$

fixed

This one is "wider" because it has 2 sources of variability: $\bar{x}, s_x$

An English statistician working for an Irish Beer company figured it out:

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

$t \sim$ t-distribution with "df" degrees of freedom

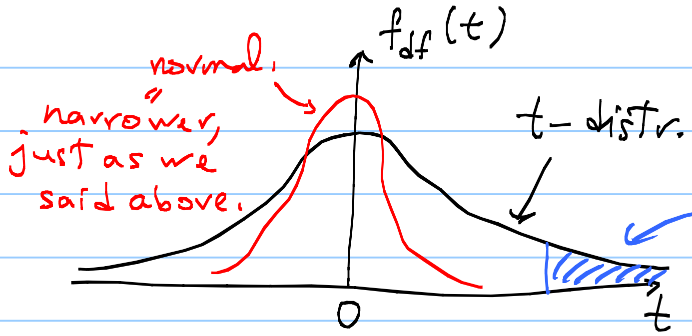
param. of t-distr., like σ^2 of Normal.

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

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)

For a sample of size n, from a Normal pop. any size, small or large.

$t = \dfrac{\bar{x}-\mu_x}{s_x/\sqrt{n}}$ 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 compute a C.I. for μ_x based on the t-dist:

$$\text{prob}(-t^* < t < t^*) = \text{conf. level}$$ "self-evident fact"

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

$\vdots$

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

Either derive it from Table VI (6), or look it up in Table IV (4), just like z^*.

This interval is also known as a "small sample C.I." (See next page).

Example: Sample of 16, from a Normal pop, yields $\bar{x} = 10$, $s = 2$

We are 95% confident that μ_x is in $10 \pm 2.13\left(\frac{2}{\sqrt{16}}\right)$

I.e. $[8.9, 11.1]$

$df = 16 - 1 = 15$

Table IV.

Note that this is wider than the z-interval:

$$10 \pm 1.96\left(\frac{2}{\sqrt{16}}\right) = [9.02, 10.98]$$

Remember that the C.I is made so that some percentage of them would cover the pop. param. In this case 95% of the intervals with $t^* = 2.13$ would do the job.

sometimes called t-intervals.

The one with $z^* = 1.96$ is narrower $\Rightarrow$ covers μ_x less than 95% of the time.

sometimes called z-interval.

The ... ± ... formulas for t-intervals are the same as those for z-intervals, because they are both derived from "self-evident facts."

$pr(-z^* < z < z^*) = $ conf. level $\qquad pr(-t^* < t < t^*) = $ conf. level

The diff. is that the t-interval has the df to find.

So, for example, the 2-sample t-interval for $\mu_1 - \mu_2$ is

$$\Rightarrow \quad (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

note s_1^2, s_2^2, not σ_1^2, σ_2^2

But what about the df = ?

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

← Welch's formula hard to show!

Then from table VI (6) or IV(4) we get t^*, and proceed.

$\Rightarrow$ And don't forget t^* still depends on 1-sided or 2-sided CI.

Note that the basic difference between the z-interval and the t-interval is in whether or not we know σ_x or not, respectively. So, the z-interval often appears under the header "Known σ_x", and the t-interval is under the header "Unknown σ_x." But these 2 intervals are also called "large-sample CI" and "small-sample CI", respectively, because if the sample is large, then s_x is going to be a very good approximation of σ_x; so, we can use $\bar{x} \pm z^* s_x/\sqrt{n}$. When the sample is small, the s_x is not a good approx. of σ_x, and so, we use $\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$.

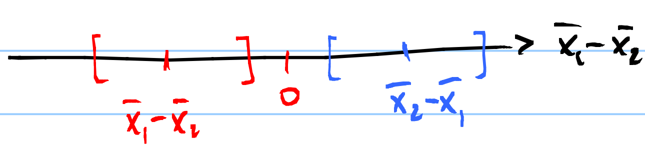
Q1: Suppose Joe computes a C.I. for $\mu_1 - \mu_2$, but Jane computes a CI for $\mu_2 - \mu_1$. So, they are wondering if they need to re-calculate.

a) The 2 CIs will be identical

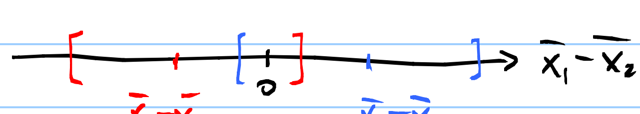
b) The 2 CIs will be different, but the "corollary" (i.e. the simple answer to the question Are the 2 means different?) will be the same.

c) The 2 CIs will be different, and the "corollary" is diff. too.

d) There is no relation between the 2 CIs.



$\Rightarrow$ diff. CI, same corollary



$\Rightarrow$ diff C.I, same corollary

Paired data

Recall that we required the 2 samples (in a 2-sample problem) to be independent. It happened when we wrote

$V[\bar{x}_1 - \bar{x}_2] = V[\bar{x}_1] + V[\bar{x}_2] + 0 \leftarrow \quad = \sigma_1^2/n_1 + \sigma_2^2/n_2$

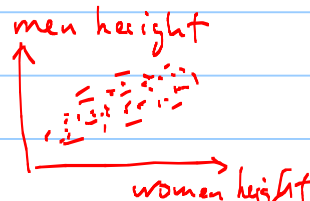
But there exist problems where the 2 samples are not independent.

E.g. 1: suppose you want to see if the mean of height is different for men and women.

If you take 100 men and 100 women, randomly, then you can claim the 2 samples are independent. But if your data comes from married couples, then they are not independent.

Such data are called "paired".

You can usually see/test this by looking at:



E.g. 2: IQ before and after some pill.

How do we build a C.I. for $\mu_1 - \mu_2$ from paired data?

1) Figure out/estimate the 0 term in $V[\bar{x}_1 - \bar{x}_2]$ Too hard!

2) Simpler way: "Make a new column"

	IQ before "x_1"	IQ after "x_2"	$d = x_1 - x_2$
person 1	$\cdot$	$\cdot$	$\cdot$
person 2	$\cdot$	$\cdot$	$\cdot$
$\vdots$	$\cdot$	$\cdot$	$\cdot$
			$\bar{d}$, s_d

n

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

$\bar{d} \pm t^* \frac{s_d}{\sqrt{n}}$, $df = n-1$

Depends on 1-sided or 2-sided.

[The Math is trivial! Determining paired vs. not is NOT trivial. Paired vs. NOT should be the first question you ask yourself.

Example Consider The fish example again. The data

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

was collected by catching The fish (both types) from some lake. This time, suppose we want to know if $\mu_1 > \mu_2$, where

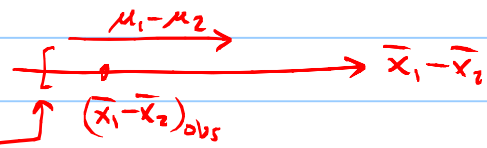
$\mu_1 =$ pop. mean zinc in Type I
$\mu_2 =$ " " II

} Important to define (The pop. parameters) clearly.

The appropriate "interval" is a lower conf. bound for $\mu_1 - \mu_2$:

$$(9.15 - 3.08) - 1.645\sqrt{\frac{(1.27)^2}{56} + \frac{(1.71)^2}{61}}$$

$$6.07 - 0.455 = 5.53$$

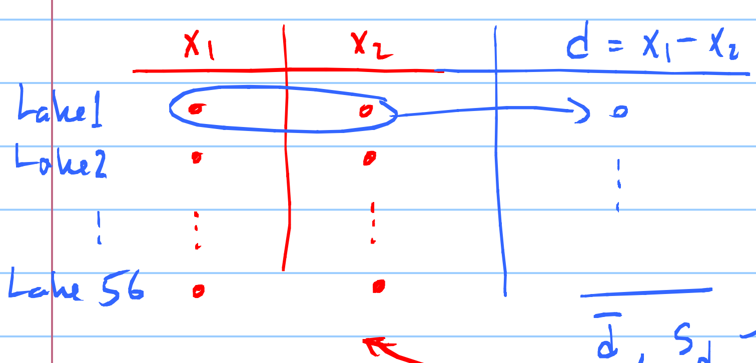


Conclusion: we are 95% confident That $\mu_1 > \mu_2 + 5.53$

Corollary: Yes, There is evidence That $\mu_2 > \mu_1$. [not with 95% conf.]

Now, suppose The way we collect The data is different. Suppose we catch a type I and a type II fish from one lake, and Then another pair of type I, type II from another lake, etc. from 56 lakes. Same question: is $\mu_1 > \mu_2$?

Now The data from The 2 populations are paired.



95% paired C.I:

$$\bar{d} - t^* \frac{s_d}{\sqrt{56}}$$

$df = n-1$

We don't have The actual data, so I can't compute This here. But it can be shown That if The data are paired, Then you'll get a number larger Than 5.53. In general paired CIs are narrower Than unpaired CIs If The data are truly paired. Narrower CI = better = more precise. That is The beauty of paired CI's! See how (below).

List of CIs:

z-based CI's for (If σ_x = known. If not, then n = large)

μ_x	π_x	$\mu_1 - \mu_2$	$\pi_1 - \pi_2$
$\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$	$p \pm z^* \sqrt{\frac{p(1-p)}{n}}$	$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$(p_1 - p_2) \pm z^* \sqrt{\frac{p_1(1-p_1)}{n} + \cdots}$

t-based CI's for (If σ_x = unknown. Must have pop = normal)

μ_x	π_x	$\mu_1 - \mu_2$	$\pi_1 - \pi_2$
$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ $df = n-1$	X	$z^* \to t^*$ df = Welch	X

use bootstrap (see lab)

These come in the 2-sided and 1-sided variety

Don't forget that we also saw C.I. for σ_x (hw), π_1/π_2 (7.29), ...

And on top of all that, you need to decide paired vs. unpaired

Let this be the first question you ask yourself!

hw-lect 18-1

In The last example, above, we have $n=16$ and so $df=n-1=15$.
One way to get t^* for The C.I. is from Table IV (4).
under The 2-sided 95% interval, for $df=15$,
you will find 2.131.

a) Now, use Table VI (6); what value of t^* do you get?

b) Now, suppose we are interested in building a 1-sided C.I for μ_x. According to Table IV (4), with $df=15$, and 95% confidence level, The value of t^* is 1.753. Again, what value of t^* do you get from Table VI (6)?

hw-lect 18-2

For the data collected in hw_lect1, consider one of the continuous variables (call it y), and one of the categorical 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 t-based, 2-sided, 95% C.I. for mu1-mu2.
b) Is there evidence from data that mu1 and mu2 are diffiererent?

hw-lect 18-3

Consider the following data on x1 and x2 which was collected in a paired design:
x1 = c(-0.27, -0.14, 1.61, 0.09, 0.00, 2.07, 0.56, -1.67, -0.51, -0.54)
x2 = c(-0.32, 0.20, 1.93, 0.54, 0.75, 1.77, 0.84, -0.29, -0.33, 0.17)
a) Compute a 2-sided, 95% CI for the difference between the two true means. You may use R to do simple claculations, but use the CI formulas derived in class. BTW, you can "test" that x1 and x2 are paired by looking at their scatterplot:
plot(x1,x2)   # I see a linear association

b) Provide one interpretation of the observed CI, AND state the conclusion in English, i.e., the "corollary."

c) Consider the following data, which is the same as above, except the cases in x2 have been randomly shuffled. Compute an appropriate 95% 2-sided CI.
y1 = c(-0.27, -0.14, 1.61, 0.09, 0.00, 2.07, 0.56, -1.67, -0.51, -0.54)
y2 = c(0.20 , 0.54, -0.33, 1.93, -0.32, 1.77, 0.75, 0.17, -0.29, 0.84)

d) Provide one interpretation of the observed CI, AND state the conclusion in English, i.e., the "corollary." .

e) Which one is narrower?