Go over the examples in last lecture.

Consider the 1-sample, 2-sided CI for \( \mu \): \( \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \)

we derived it from \( z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \).

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 \( \frac{\bar{x} - \mu}{s_x/\sqrt{n}} \).

In fact, upon a little thinking you can see that it cannot have a normal dist.

To see that \( \frac{\bar{x} - \mu}{s_x/\sqrt{n}} \) is not normal, ask yourself:

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

\[ \frac{\bar{x} - \mu}{s_x/\sqrt{n}} \quad \text{or} \quad \frac{\bar{x} - \mu}{s_x/\sqrt{n}} \]

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:

\( z \sim \text{Normal}(0,1) \)

\( t \sim t\text{-distribution with } d\text{f } \) degrees of freedom.

\[ f(t) = \frac{\Gamma\left(\frac{1}{2}(d\text{f}+1)\right)}{\sqrt{\pi(d\text{f})} \Gamma\left(\frac{1}{2}d\text{f}\right)} \frac{1}{\sqrt{1+\frac{t^2}{d\text{f}}} \frac{d\text{f}+1}{d\text{f}}} \]

This is just FRI.

As far as you are concerned, the t-distrv. is just another Table VI, not IV!
If \( df \to \infty \), then \( t \to z \).

Table VI (6) gives right areas.

Then (Student's \( t \)) any size, small or large.

For a sample of size \( n \), from a normal pop.,

\[
t = \frac{\bar{x} - \mu_x}{S_x/\sqrt{n}}\]

has a \( t \)-dist. with \( df = n-1 \)

\[
\text{As } n \to \infty, \quad df \to \infty, \quad \therefore t \to z
\]

[Analogue to \( z = \frac{\bar{x} - \mu}{S_x/\sqrt{n}} \) 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 affect 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"

\[
\frac{\bar{x} - \mu_x}{S_x/\sqrt{n}} \Rightarrow \cdots \Rightarrow "< \mu_x < \cdots"
\]

\[
\therefore \text{C.I. for } \mu_x : \bar{x} \pm t^{*} \frac{S_x}{\sqrt{n}} \text{ with } df = n-1.
\]

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

This interval is also known as the "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 $\mu_x$ is in $10 \pm 2.13 \left( \frac{2}{\sqrt{16}} \right)$

I.e. $[8.9, 11.1]$

$\text{df} = 16 - 1 = 15$

Note that this is wider than the $z$-interval, Table IV.

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

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 $\mu_x$ less than 95% if the time.

sometimes called $z$-interval.

The $\pm \ldots$ formulas for $t$-intervals are the same as those for $z$-intervals, because they are both derived from "self-evident facts:"

$p(-z^* < z < z^*) = \text{conf. level}$

$p(-t^* < t < t^*) = \text{conf. level}$

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

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

Note $s_1^2, s_2^2$, not $s_1^2, s_2^2$.

But what about the df = ?

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

hard to show!

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

And don't forget $t^*$ still depends on 1-sided or 2-sided C.I.
Note that the basic difference between the z-interval and the t-interval is in whether or not we know $\sigma_x$ or not, respectively. So, the z-interval often appears under the header "Known $\sigma_x$", and the t-interval is under the header "Unknown $\sigma_x$". But these two 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 $\sigma_x$; so, we can use $\bar{x} \pm z^* S_x / \sqrt{n}$. When the sample is small, the $S_x$ is not a good approximation of $\sigma_x$, and so, we use $\bar{x} \pm t^* S_x / \sqrt{n}$. 
Recall that we required the 2 samples (in a 2-sample problem) to be independent. It happened when we wrote

\[ V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) + o \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \]

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 \( o \) term in \( V(\bar{x}_1 - \bar{x}_2) \). Too hard!
2) Simpler way: "Make a new column".

\[
\begin{array}{c|c|c}
\text{IQ before} & \text{IQ after} & \text{C.I. for } \mu_1 - \mu_2 \\
\hline
x_1 & x_2 & d = x_1 - x_2 \\
\hline
\text{person 1} & & \\
\text{person 2} & & \\
\end{array}
\]

\[ \bar{d} \pm t^* \frac{s_d}{\sqrt{n}} \], \( df = n-1 \)

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.
Consider the fish example again. The data:

<table>
<thead>
<tr>
<th>Type</th>
<th>n</th>
<th>$\bar{x}$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>56</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>

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

Important to define (the pop. parameters) clearly.

The appropriate “interval” is a lower confidence bound for $\mu_1 - \mu_2$:

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

$$= 6.07 - 0.455 = 5.6$$

Conclusion: We are 95% confident that $\mu_1 > \mu_2 + 5.6$

Corollary: Yes, there is evidence that $\mu_2 > \mu_1$. [not with 95% confidence.]

Q1: Which of the following designs will lead to paired-data for answering the same question asked above?

A) Collect 56 Type I from one lake and 56 Type II from another lake.

B) Collect 56 + 61 Type I from one lake and 56 + 61 Type II from another lake.

B) Collect 1 Type I from one lake and 1 Type II from another lake, and repeat 56 times in the same 2 lakes.

C) Collect 1 Type I and 1 Type II from each of 56 lakes.
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:

\[
\begin{align*}
\text{Lake 1} & \quad x_1 \quad \circ \quad d = x_1 - x_2 \\
\text{Lake 2} & \quad \circ \quad \circ \\
\text{Lake 56} & \quad \circ \quad \circ \\
\end{align*}
\]

95\% paired C.I.:

\[
\text{I} = \bar{d} \pm t^* \frac{s_d}{\sqrt{56}}
\]

\(df = n-1\)

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 5.6. 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:

\[ z\text{-based CI's for } (\text{If } \sigma_x = \text{known. If not, then } n \text{ large}) \]

\[
\begin{align*}
X & \pm z^* \frac{\sigma_x}{\sqrt{n}} \\
\frac{X_1 - X_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} & \pm z^* \left( \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2} \right) \\
\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} & \pm z^* \left( \frac{p_1}{n_1} + \frac{p_2}{n_2} \right)
\end{align*}
\]

\[ t\text{-based CI's for } (\text{If } \sigma_x = \text{unknown. Must have pop = normal}) \]

\[
\begin{align*}
\bar{X} & \pm t^* \frac{s}{\sqrt{n}} \\
\frac{X_1 - X_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} & \pm t^* \left( \frac{s_1}{n_1} + \frac{s_2}{n_2} \right) \\
\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} & \pm t^* \left( \frac{p_1}{n_1} + \frac{p_2}{n_2} \right)
\end{align*}
\]

\[ \text{df = Welch} \]

\[ \text{use bootstrap (see lab)} \]

These come in the \underline{2-Sided} and \underline{1-Sided} variety

Don't forget that we also saw C.I. for \( \sigma_x, \frac{\sigma_1}{\sigma_2}, \)...

And on top of all that, you need to decide \underline{paired vs. unpaired}.

\[ \sqrt{\frac{\text{num}}{\text{den}}} > 7.29 \]

\[ \text{Let this be the first question you ask yourself!} \]
For the data you collected, consider one of the continuous variables (call it $y$), and one of the categorical variables (call it $x$). Let $\mu_1$ denote the true mean of $y$ when $x = \text{first level of } x$, and $\mu_2$ denote the true mean of $y$ when $x = \text{2nd level of } x$.

a) Compute a $t$-based, 2-sided, 95% CI for $\mu_1 - \mu_2$.

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

Consider the following data on $x_1$ and $x_2$ which was collected in a paired design:

$x_1 = c(-0.27, -0.14, 1.61, 0.09, 0.00, 2.07, 0.56, -1.67, -0.51, -0.54)$

$x_2 = c(-0.32, 0.20, 1.93, 0.54, 0.75, 1.77, 0.84, -0.29, -0.33, 0.17)$

a) Compute a 2-sided, 95% CI for the difference between the two true means. You may use R to do simple calculations, but use the CI formulas derived in class. BTW, you can "test" that $x_1$ and $x_2$ 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 $x_2$ have been randomly shuffled. Compute an appropriate 95% 2-sided CI.

$y_1 = c(-0.27, -0.14, 1.61, 0.09, 0.00, 2.07, 0.56, -1.67, -0.51, -0.54)$

$y_2 = c(0.20, 0.54, -0.33, 1.93, -0.32, 1.77, 0.75, 0.17, -0.29, 0.84)$

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

e) Which one is narrower?