Last time we built the CI for $\mu_x$: 
$$x \pm z \frac{s_x}{\sqrt{n}}$$
and the CI for $\pi_x$: 
$$p \pm z \sqrt{\frac{p(1-p)}{n}}$$

Let's take care of the business of $s_x$.

Consider the (1-sample, 2-sided) CI for $\mu_x$: 
$$x \pm z \frac{s_x}{\sqrt{n}}$$
We derived it from 
$$z = \frac{x - \mu_x}{s_x/\sqrt{n}} \sim N(0,1).$$
In practice, however, the CI is computed as 
$$x \pm z \frac{s_x}{\sqrt{n}}$$
So, it's natural to ask what is the dist. of 
$$\frac{x - \mu_x}{s_x/\sqrt{n}}.$$
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?

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

This one has a wider distr. because it has 2 sources of variability, $x, s_x$.

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

- $z \sim N(0,1)$
- $t \sim t$-distribution with $df$ degrees of freedom

The $t$-distribution is just another Table of normal.

Table VI (6) gives right areas.

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

This is just FYI:

As far as you are concerned, the $t$-dist. is just another Table. 

Table VI (6) not 4!
(Student's t) for any size, small or large.

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

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

has a t-dist. with \( df = n-1 \)  \[\text{As } n \to \infty, \quad df \to \infty, \quad \therefore t \to z\]

Analogous to \( z = \frac{\bar{x} - \mu_x}{s/\sqrt{n}} \) 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, every thing 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.:

\[
\text{prob}(-t^{*} < t < t^{*}) = \text{conf. level}
\]

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

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

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

We are 70% confident that \( \mu \) is in

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

Note: This is wider than the z-interval;

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

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 \( \mu \) less than 95% of the time.

Table VI Tail areas for t curves

<table>
<thead>
<tr>
<th>t</th>
<th>df 1</th>
<th>df 2</th>
<th>df 3</th>
<th>df 4</th>
<th>df 5</th>
<th>df 6</th>
<th>df 7</th>
<th>df 8</th>
<th>df 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>0.1</td>
<td>0.468</td>
<td>0.465</td>
<td>0.463</td>
<td>0.462</td>
<td>0.462</td>
<td>0.462</td>
<td>0.462</td>
<td>0.461</td>
<td>0.461</td>
</tr>
<tr>
<td>0.2</td>
<td>0.437</td>
<td>0.430</td>
<td>0.427</td>
<td>0.426</td>
<td>0.425</td>
<td>0.424</td>
<td>0.424</td>
<td>0.423</td>
<td>0.423</td>
</tr>
<tr>
<td>0.3</td>
<td>0.407</td>
<td>0.396</td>
<td>0.392</td>
<td>0.390</td>
<td>0.388</td>
<td>0.387</td>
<td>0.386</td>
<td>0.386</td>
<td>0.386</td>
</tr>
<tr>
<td>0.4</td>
<td>0.379</td>
<td>0.364</td>
<td>0.358</td>
<td>0.355</td>
<td>0.353</td>
<td>0.352</td>
<td>0.351</td>
<td>0.350</td>
<td>0.349</td>
</tr>
<tr>
<td>0.5</td>
<td>0.352</td>
<td>0.333</td>
<td>0.326</td>
<td>0.322</td>
<td>0.319</td>
<td>0.317</td>
<td>0.316</td>
<td>0.315</td>
<td>0.315</td>
</tr>
<tr>
<td>0.6</td>
<td>0.328</td>
<td>0.305</td>
<td>0.295</td>
<td>0.290</td>
<td>0.287</td>
<td>0.285</td>
<td>0.284</td>
<td>0.283</td>
<td>0.282</td>
</tr>
<tr>
<td>0.7</td>
<td>0.285</td>
<td>0.254</td>
<td>0.241</td>
<td>0.234</td>
<td>0.230</td>
<td>0.227</td>
<td>0.225</td>
<td>0.223</td>
<td>0.222</td>
</tr>
<tr>
<td>0.8</td>
<td>0.267</td>
<td>0.232</td>
<td>0.217</td>
<td>0.210</td>
<td>0.205</td>
<td>0.201</td>
<td>0.199</td>
<td>0.197</td>
<td>0.196</td>
</tr>
</tbody>
</table>

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 \( \mu \) or a single \( \pi \). 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 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 \( \mu_1 - \mu_2 \) or for \( \pi_1 - \pi_2 \)

Dropping the \( x \) for simplicity.

These are called 2-sample C.I.

\[ \Rightarrow \text{2-sample problems involving 2 means are easy to recognize.} \]

Examples involving 2 props are more tricky. Here is a correct one:

Is the prop. of Mac users among boys different \( = \pi_1 \) from \( = \pi_2 \) of Mac \( = \pi \) girls? \( = \pi_2 \)

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

\[ \Rightarrow \text{Here is an incorrect example:} \]

Is the proportion of people who use Macs different \( = \pi \) from \( = \pi \) other computers? \( = \pi \)

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

So, it's like the lab example, above, there is only 1 indep. prop.
To build a CI for $\mu_1 - \mu_2$ or $\pi_1 - \pi_2$, whose sampling distribution do we need?

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

The $E$ and $V$ of the sampling distributions are

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

$CLT$: $\bar{x}_1 - \bar{x}_2 \sim N(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

What's the quantity that has a $Z(t)$ dist?

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim t\text{-dist. with } df = n - 1$$

$df = n - 1$ (SE Welch)

Self-evident fact:

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

Solve for $\mu_1 - \mu_2$

$CI$ for $\mu_1 - \mu_2$:

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

Interpretation: Same as before.

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

$$df \approx \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\sigma_1^2\right)^2 + \frac{1}{n_2 - 1} \left(\sigma_2^2\right)^2}$$
That was CI for $M_1 - M_2$. The analog for $\pi_1 - \pi_2$ is:

$$C.I. \text{ for } \pi_1 - \pi_2 : \quad (\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$t$-based CI for props does not exist.

↓ ↓ ↓ Go over all these examples ↓ ↓ ↓

Example: Here is another data set:

\[
\begin{align*}
\text{Pop. 1} & \quad \text{Pop. 2} \\
\text{Winter quarter} & \quad \text{Spring quarter} \\
\text{Lab is good} & \quad 10 (0.152) \quad 17 (0.262) \\
\text{Bad} & \quad 56 (0.848) = P_1 \quad 48 (0.738) = P_2 \\
\hline
& \quad \frac{66}{65}
\end{align*}
\]

Does data provide sufficient evidence to claim that the proportion of “bads” in the 2 populations are different? \(90\% \text{ Conf. Int.}\)

\[\pi_1 = \text{prop. of students in pop. who don’t like Lab, in Winter} \]
\[\pi_2 = \text{‘‘‘‘‘‘ ‘‘‘‘, in Spring} \]

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

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

\[
(0.848 - 0.738) \pm 1.645 \sqrt{\frac{0.848(1-0.848)}{66} + \frac{0.738(1-0.738)}{65}} = (-0.005, 0.225)
\]

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

Cov. Syl. \[2\] \[0.11 \pm 0.115\]

Correct Conclus. \[\text{Cannot conclude that } \pi_1 \text{ and } \pi_2 \text{ are different.} \]

\[\text{Sure, you are thinking that it is possible that they are equal.} \]
\[\text{But the data provide no evidence for it!} \]
\[\text{The data provide no evidence that they are different either!} \]
\[\text{Basically, we cannot conclude anything about } \pi_1 - \pi_2: \]

\[\text{Incorrect Concl.} \quad \pi_1 \text{ and } \pi_2 \text{ are same. Very big error!} \]
Example: 82 students have picked up their test, but 30 have not, even 1 week after the test was returned.

Call these 2 groups "Attendees" and "Non-attendees".

<table>
<thead>
<tr>
<th>Group</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>

Sample population.

\[ \mu_1 = \text{mean of test 1 for non-attend students who have ever taken 390.} \]
\[ \mu_2 = \text{mean of test 1 for attend students.} \]

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

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

\[
(\bar{x}_2 - \bar{x}_1) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

95% CI

\[
(13.25 - 11.8) \pm 1.96 \sqrt{\frac{(3.32)^2}{30} + \frac{(3.04)^2}{82}} = 1.45 \pm 1.96 (1.693)
\]

\[
1.45 \pm 1.36 = (0.09, 2.81) \Rightarrow [\frac{\mu_2 - \mu_1}{0.09}] \approx 2.81
\]

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.

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>Type</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?

$\mu_1 = \text{pop. mean zinc in Type I}$ \quad \{ Important to define $\mu_1, \mu_2$ \}

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

$df \text{ Welch} = \ldots = 110.32 \Rightarrow \text{not important}$

\[ t^* \approx 2.0 \quad \text{Table VI} \Rightarrow t^* \approx 2.0 \]

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

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

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.

Because the C.I. is entirely to the right of $0$, 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 $\mu_1 - \mu_2$. 

\[ \boxed{\text{LCB}} \]
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 \( \pi_1 \) denote the true proportion of defective bridges in the USA, and \( \pi_2 \) .... 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 \( \pi_1 \) is larger than \( \pi_2 \)?