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

$$\text{Var}[\bar{X}_1 - \bar{X}_2] = \text{Var}[\bar{X}_1] + \text{Var}[\bar{X}_2] + \text{C.O.V} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

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 C.O.V term in \( \text{Var}[\bar{X}_1 - \bar{X}_2] \) too hard!
2) Simpler way: "Make a new column"

<table>
<thead>
<tr>
<th>IQ before</th>
<th>IQ after</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_2 )</td>
</tr>
</tbody>
</table>

People 1 | | |

People 2 | | |

\[ \bar{d} \pm z^* \frac{sd}{\sqrt{n}} \]

I.e. 1 sample C.I., for \( \mu_1 - \mu_2 \) for paired data:

The Math is Trivial! Determining paired vs. not is not.

Paired vs. Not should be the first question you ask!
All of the CIs we have built are called "large-sample" CIs.

What if sample size is small? What breaks?

Even if \( \mu \) is normal, \( \frac{\bar{x} - \mu}{s/\sqrt{n}} \) is no longer normal!

I.e. \( \frac{\bar{x} - \mu}{s/\sqrt{n}} \) is no longer a good approximation.

So, then we do not know \( \bar{x} \) in self-evident fact.

(Recall that \( p \pm z^* \sqrt{\frac{p(1-p)}{n}} \) as the C.I. for \( \pi \) is already a large-\( n \) version of a more complicated expression.)

To see that \( \frac{\bar{x} - \mu}{s/\sqrt{n}} \) is not normal, ask yourself which of the following has the "wider" sampling distr?

\( \bar{x} \) is not normal, ask yourself

\( x \) or \( t \)?

\( x = \frac{\bar{x} - \mu}{s/\sqrt{n}} \) or \( t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \)

This one is "wider" because it has 2 sources of variability: \( \bar{x}, s \)

In fact,

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

\( t \sim \text{t-distribution with } \nu \text{ degrees of freedom} \)

\( f(t) = \frac{\Gamma\left(\frac{1}{2}(df+1)\right)}{\sqrt{\pi df} \Gamma\left(\frac{1}{2}df\right)} \left\{ df + \frac{t^2}{2} \right\}^{-\frac{1}{2}} \)

This is just FYI.

As far as you are concerned, the t-distr.

is just another Table

\( Table 6 \) not 4!
If \( df \to \infty \), then \( t \to z \).

\( df \sim n \) (see below)

So, as sample gets larger, \( t \to \text{Normal} \).

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

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

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

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

[Analogous to \( z = \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}} \) has a normal distr. with \( \mu=0, \sigma=1 \)].

If the pop. is not normal, we don't know the distr. of \( t \), for small samples. As a result of this, a small-sample C.I. requires the distr. of the population to be normal. (developed next)

This is a restriction that does not plague \( z \)-intervals.

But for \( t \)-intervals, pop. should be normal.

(or is assumed to be)

---

Or when \( \sigma \) is unknown

Let's do small sample C.I. for a single mean \( \mu_x \).

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

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

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

Either Table IV, or derive it from Table VI (6).
Example: Sample of 16, from a normal pop, yields $\bar{x} = 10, s = 2$

We are 95% confident that $\mu_x$ is in $10 \pm 2.13 \left( \frac{2}{\sqrt{16}} \right) \quad \text{df}=16-1=15$

I.e. $[8.9, 11.1]$

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

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

Remember that the C.I. is made so that some percentage of them would cover the pop. parameter.
In this case, 95% of the intervals with $t^* = 2.13$ would do the job. The one with $z^* = 1.96$ will not.

Q. How about small sample 2-sample C.I. for $\mu_x - \mu_1$?

Welch’s formula:

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

$\text{df} \geq \frac{(\frac{s_2^2}{n_1} + \frac{s_1^2}{n_2})^2}{\frac{1}{n_1 - 1} (s_1^2)^2 + \frac{1}{n_2 - 1} (s_2^2)^2}$

Try $t \approx 2$

And for paired data, the C.I. for $\mu_x - \mu_2$:

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

And for proportions? Not known!
Summary of CIs.

So far, we have

**Large-sample CIs for** $\mu$, $\bar{x}$, $\mu_1 - \mu_2$, $\bar{x}_1 - \bar{x}_2$

\[
\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \quad \bar{p} \pm z^* \sqrt{\frac{p(1-p)}{n}} \quad (\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}
\]

**Small-sample CIs for** $\mu$, $\bar{x}$, $\mu_1 - \mu_2$, $\bar{x}_1 - \bar{x}_2$

from Normal pop. $\implies t^*$, $df = n-1$, $df = \text{Welch}$

These come in the 2-sided and 1-sided variety, as well as paired and unpaired (i.e. independent samples).

\[
 t^* \quad df = n-1 \quad df = \text{Welch}
\]

This completes the list of "easy" CIs. But you can also compute a CI for $\sigma$, $\frac{\bar{x}}{\sigma x}$, ... and more, later. 

**Important:** Note that even though we call $z$-intervals "large sample CI," and we call $t$-intervals "small sample CI," the difference between them is really whether or not we know $\sigma_x$. So, you will see some books label their sections as "known $\sigma_x$" vs. "unknown $\sigma_x$"
In the above example, 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 \( \mu \). 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)?