Lecture 22 (Ch. 7)

Paired data

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

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

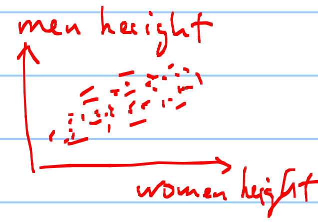
But there exist problems where the 2 samples are not independent.

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 usualy see/test this by looking at:

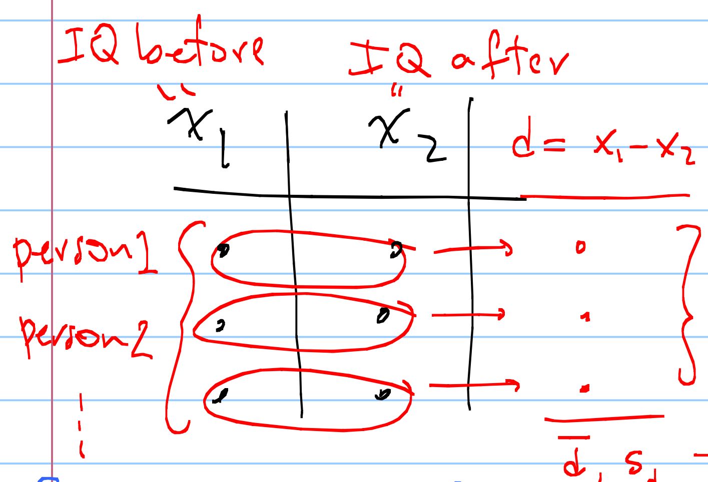


E.g. 2: IQ before and after some pill.

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"



I.e. 1 sample C.I.

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

$\bar{d} \pm z^* \frac{s_d}{\sqrt{n}}$

$\uparrow$ Depends on 1-sided or 2-sided.

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

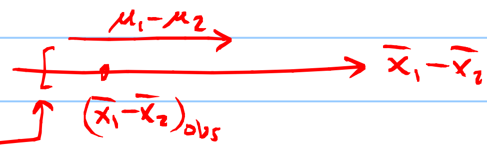
μ_1 = pop. mean zinc in Type I
μ_2 = " " II

Important to define (the pop. parameters) clearly.

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

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

$$6.07 - 0.455 = 5.53$$

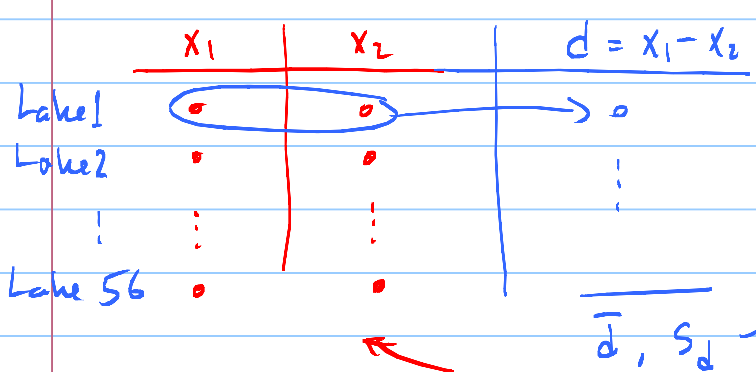


Conclusion: We are 95% confident that $\mu_1 > \mu_2 + 5.53$

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:

$$\bar{d} - 1.645\frac{s_d}{\sqrt{56}}$$

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!

$\boxed{\text{Q1:}}$ We want to see if $\mu_1 < \mu_2$. Then the appropriate quantity to compute is

a) upper conf. bound for $\mu_1 - \mu_2$, only if data are paired

b) upper " " " $\mu_1 - \mu_2$.

c) lower " " " $\mu_2 - \mu_1$.

We want to know if $\mu_1 \overset{?}{<} \mu_2$, ie. $\mu_1 - \mu_2 \overset{?}{<} 0$

So, we have to see how large it can get

ie. upper conf. bound for $\mu_1 - \mu_2$.

But $\mu_1 - \mu_2 \overset{?}{<} 0$ is equivalent to $\mu_2 - \mu_1 \overset{?}{>} 0$

ie. lower conf. bound for $\mu_2 - \mu_1$

The notion of pairedness is completely unrelated to the question you are trying to answer. The notion of pairedness effects <u>how</u> the CI is computed.

Because of the multiple answers to this ~~qz~~, it will not be graded. Only participation point will be given.

Unknown σ_x

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 (see below).

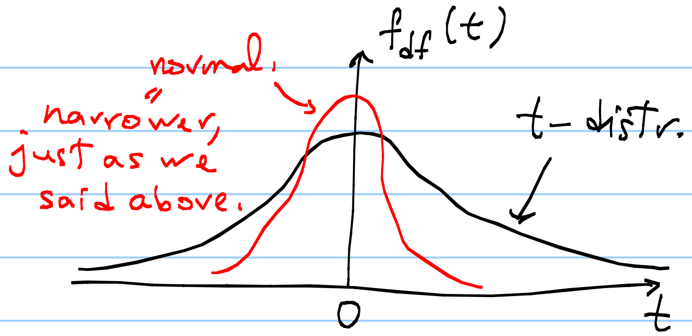
param. of t-distr.

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

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 $t \to z$.

$df \sim n$ (see below)

So, as sample gets larger, $t \to$ Normal.

Thm (Student's t)

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

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

[Analogous to $z = \frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}}$ 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"

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

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

Either Table IV, or derive it from Table VI (6).

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.

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 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 $\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$.

hw_lect 22-1

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% LARGE-sample CI for the difference between the two true means. Provide one interpretation of the observed CI. You may use R to do simple claculations, but use the CI formulas derived in class.

b) 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 LARGE-sample CI, and interpret it.

c) Which one is narrower?

hw_lect22_2

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 μ. 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)?