Stuff you now know:

Dealing with ambiguity
Random variable
histograms
Comparative boxplots
quantiles
distributions
probability (e.g. from Poisson)
sample mean and variance
distr mean and variance
qqplots

scatterplots
correlation
regression (multiple, polynomial, ...)
ANOVA (R^2, $s_e \sim$ RMSE)
overfitting, collinearity, interaction

sampling distribution
1-sample Confidence Interval for ...
2-sample CI for ...
t-distribution

In the recording of today's lecture, I'll first go over the examples in the last lecture. So, please study them carefully.

Lecture 19 (Ch. 7-end)

Here are the CI's we have derived: (If σ_x = known (or large n) ↙ else ↘)

1-sample
- C.I. for μ_x: $\bar{x} \pm z^* \sigma_x/\sqrt{n}$ or $\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$, df = n − 1
- C.I for π_x: $p \pm z^* \sqrt{\frac{p(1-p)}{n}}$ No t^* (FYI: use bootstrap)

2-sample
- C.I for $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2 \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ or ... t^*, df = Welch
- C.I for $\pi_1 - \pi_2$: $p_1 - p_2 \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ No t^* (FYI ...)

Random vs. observed : probabilistic vs. confidence interpretation

The formulas/math are easy. It's their interpretation that's HARD!

1) Confident ... 2) prob ... 3) Corollary

(FYI: In addition to these 2-sided CI's, there is also UCB and LCB)

The formulas for these are derived the same way:

Start with $pr(-z^* < z < z^*) =$ Conf. level (solve for)

or $pr(-t^* < t < t^*) =$ "

substitute $z = \frac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}} \sim N(0,1)$ or $\frac{\bar{x} - \mu_x}{s_x/\sqrt{n}} \sim t$, df = n−1 → μ_x to get CI

$z = \frac{p - \pi_x}{\sqrt{\frac{\pi_x(1-\pi_x)}{n}}} \sim N(0,1)$ → π_x ---

$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$ or ...t, df = Welch → $\mu_1 - \mu_2$ ---

$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$ → $\pi_1 - \pi_2$ ---

Not-independent samples

We required the 2 samples (in a 2-sample problem) to be indep.

It happened when we wrote

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

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

E.g.1 : Does a certain pill increase IQ.

→ { Take 100 people, measure their IQ, w/o pill
  , 110 " , , " " with pill } versus

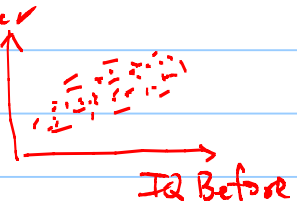
→ Take 100 people, measure their IQ, both before and after pill.

E.g.2: Is there a diff in the mean "speed" of App X and App Y?

→ { Ask 100 people to tell you how fast App X was,
  ask 120 " " " " App Y was. } versus

→ Time App X and App Y on each of 100 computers.

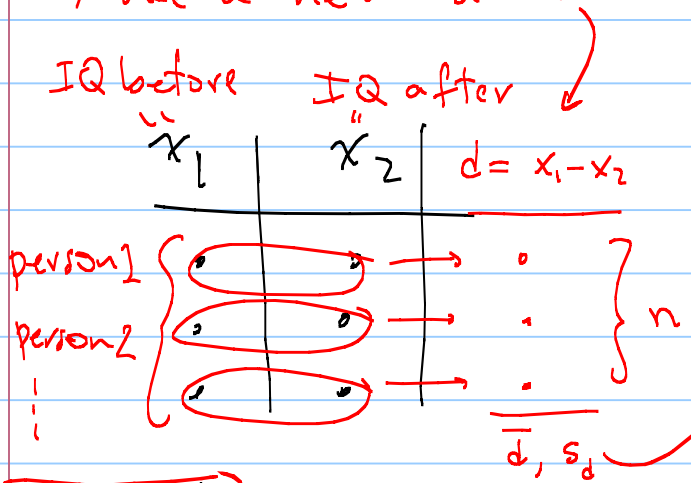
→ Such data are called "paired".



You can usually see/test this by looking at:

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

"Make a new column"



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

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

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

Benefit: paired CIs are often narrower (more precise). → a good thing!

The Math is trivial. "Addressing" pairedness is Complex! →

Example Here is how we did the fish example in last lecture:

Concentration of zinc in 2 types of fish.

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

Are the true/pop. means different?

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

} Important to define μ_1, μ_2 (the pop. parameters) clearly.

95% C.I. for $\mu_1 - \mu_2$: $(9.15 - 3.08) \pm 1.96\sqrt{\frac{(1.27)^2}{56} + \frac{(1.71)^2}{61}}$

$6.07 \pm 0.54 \Rightarrow [5.53, 6.61]$

There are many ways of collecting paired data, e.g.

1) Take one Type I and one Type II, from n lakes.
2) " " " " " " , with the same weight.
3) ---

In every example, a Type I fish is somehow matched/paired with a Type II fish.

Benefit: In every example, the CI for $\mu_1 - \mu_2$ will often be narrower (ie. more precise)

⇒ Once again: If you are analyzing existing data, the 1st question should be "Are the data paired?"

⇒ If you are collecting the data yourself, do your best to assure the data are paired. Create as much dependency as you can. The more factors you "control", the narrower your CI will be.

Jerry Wei (our TA), has found that a negative correlation between the 2 samples can lead to a wider CI (larger p-value, in ch 8) for the paired test! The width of a CI is related to something called power. A wider CI has a lower power - and that's not a good thing.

FYI

```
library(MASS)
set.seed(123)
 s  = matrix(c(1, -0.8, -0.8, 1), 2)
df = mvrnorm(n=100, mu=c(0, 0.3), Sigma=s, empirical=F)
t.test(df[,1], df[, 2], alternative="less",paired=FALSE)
t.test(df[,1], df[, 2], alternative="less", paired=TRUE)
```

https://stats.stackexchange.com/questions/38102/paired-versus-unpaired-t-test

The wiki page on "paired difference test" compares the power of the unpaired and paired tests, and indeed shows that a positive correlation is required for higher power. Ie, there *is* such a thing as a "bad pairing" which can lead to lower

FYI

It may be tempting to think that a paired design does not exist for proportions. But it does. 2 genders (boy/girl), 2 computer types (Mac/Dell)

Here is an example data from an unpaired design:

Boy	Girl
Mac	Mac
Dell	Mac
Dell	Dell
⋮	⋮

prop. of Boys with Mac = p_1

prop. of Girl with Mac = p_2

C.I. for $\pi_1 - \pi_2$: $(p_1 - p_2) \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Here is an example data from a paired design:

Husband	Wife
Mac	Mac
Dell	Mac
Dell	Dell
⋮	⋮

prop. of Husbands with Mac = p_1

prop. of wives with Mac = p_2

However, there is no simple formula for the C.I. of $\pi_1 - \pi_2$ when data are paired. (After all, we cannot make a new column of differences, because we cannot subtract "Mac" from "Mac", ie. the data are categorical.

hw_lect19_1 Use t-distr.

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 at least 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?