Stats 9.3: Another statistic/parameter pair

It is not unusual to have two lists of data, say a list X and a list Y, and our interest is in the way they are related (somehow).

Examples:
a) X is the list of IQ's at ENC, and Y is the list of IQ's at UMass, and we are interested in if the average IQs are the same or not.
b) X is the list of scores on Exam 1 in Stats, and X is the list of scores on Exam 2, and we are interested in if, on the average, students scored better on the second exam than on the first.

You might notice that (a) consists of samples from two different populations, and those populations are independent, so the samples are independent. On the other hand, (b) consists of samples taken from the same population, so the samples are dependent. In fact, (b) would be most interesting if we had what we called paired data. That is, we got both the first exam and the second exam from a list of students, and measured how far apart those pairs of scores are.

Our interest today is in the case where we have independent populations, and we are interested in how the means of those populations compare to one another.

<u>Example 1</u>: I used to collect all of the exam scores for the course EMES. I took random samples of size 18 from two different years, and looked at the first exam in each year. I wanted to see if the two groups did the same, or if one group did better than the other. Here's the data:

X: 70 60 45 43 64 64 73 67 60 61 79 67 71 53 80 75 57 68
Y: 75 77 29 19 52 75 75 55 44 38 72 38 51 52 77 69 31 68

To make the comparison, I'm going to look at the means of the two groups to make my decision. Let's call the mean of the first group μ_1, and we will call the mean of the second group μ_2 (don't get excited, Pokemon fans!). We want to measure how far apart those means are, so our parameter of interest is

$$\mu_1 - \mu_2$$

Here's a question for you, and the answer is on the top of the next page.
If we are trying to find out about the parameter $\mu_1-\mu_2$, what do you think is the best statistic to use?

That is, what is the unbiased estimator for $\mu_1-\mu_2$?

Answer to the question on the previous page:

The unbiased estimator for μ_1-μ_2 is $\overline{X}$-$\overline{Y}$. (We could call it $\overline{X}_1$-$\overline{X}_2$.)

If we're going to do either do a hypothesis test or a confidence interval for μ_1-μ_2 then we need to know the pdf for $\overline{X}$-$\overline{Y}$. To find that, we need a little bit of theory (but not anything you will need to repeat to me):

Start the theory!

An important property of normal distributions:

If you add or subtract independent normals, you get another normal.
The means add or subtract.
The variances add.

We know that, under the right conditions,

$\overline{X}$ is ND(μ_1, $\sigma_1/\sqrt{n_1}$)
and $\overline{Y}$ is ND(μ_2, $\sigma_2/\sqrt{n_2}$)

So, under these same conditions,

$\overline{X}$-$\overline{Y}$ is ND, with $\mu = \mu_1$-μ_2, and $\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(For that last thing, we take the standard deviations, square them to get the variances, add those variances together, then take the square root to get back to the standard deviation.)

End the theory!

OKAY, if you missed out on what I was saying right above, join in again here

Two things about $\overline{X}$-$\overline{Y}$ that will probably sound familiar.

A: $\overline{X}$-$\overline{Y}$ has that distribution above, under the right conditions. Those conditions are the same we ran into before

1) Both populations are normal
or 2) Both samples are large (each n≥30)

If we don't meet either condition, we will have to assume that the populations are normal, and write on our page: Assume pops ND

B: $\overline{X}$-$\overline{Y}$ can turn into either a z-score or a t-score, depending on what?

(Answer on the top of the next page.)

Answer to the question on the previous page:
B: $\overline{X}$-$\overline{Y}$ can turn into either a z-score or a t-score, depending on whether we know the σ's or not. That is:
- If we know the σ's, then we use z.
- If we don't know the σ's, then we use the s's and we use a t. In this case, we will calculate the degrees of freedom df as follows:

df = the smaller of n_1-1 and n_2-1 .

Now we put it all together.

Confidence interval for μ_1-μ_2

$$(\overline{X}-\overline{Y}) \pm z \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

if the σ's are known

$$(\overline{X}-\overline{Y}) \pm t \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

if the σ's are not known.

Hypothesis testing for μ_1-μ_2

The statistics we will use will be either of these

$$z_{\overline{X}_1-\overline{X}_2}=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

if the σ's are known

$$t_{\overline{X}_1-\overline{X}_2}=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

if the σ's are not known.

On to the next page to try some!

Example 1, continued: I used to collect all of the exam scores for the course EMES. I took random samples of size 18 from two different years, and looked at the first exam in each year. I wanted to see if the two groups did the same, or if one group did better than the other. Here's the data:

X: 70 60 45 43 64 64 73 67 60 61 79 67 71 53 80 75 57 68
Y: 75 77 29 19 52 75 75 55 44 38 72 38 51 52 77 69 31 68

Find a 95% confidence interval for the difference of the means of the two groups. For your convenience, $\overline{X}$ = 64.3, $\overline{Y}$ = 55.4, s_1 = 10.3, s_2 = 18.9.

Solution: 95% CI for μ_1-μ_2
Notice that the σ's are not known, so we will have to use the t-version.

$$(\overline{X}-\overline{Y}) \pm t\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

We have 18-1 = 17 degrees of freedom. On the t-chart we look up for 17 degrees of freedom, in the two tail with α=0.05 column, finding t = 2.110.

Plugging in we get (64.3-55.4) ± 2.110 $\sqrt{\frac{10.3^2}{18}+\frac{18.9^2}{18}}$ = 5.1

which is 8.9 ± 2.110(5.1) ^^^ ASSUME POPS ND

which is 8.9 ± 10.8 $\Rightarrow$ [-1.9 , 19.7]

A follow-up question: Can we be sure that class 1 did better than class 2?

Answer: No, we cannot be sure. Since the CI contains both positive numbers (meaning μ_1 is bigger) and negative numbers (meaning μ_2 is bigger). we cannot tell.

Example 2. An experiment was conducted to test the effects of alcohol. The errors were recorded in a test of visual and motor skills for a treatment group of people who drank ethanol and another group given a placebo. The results are shown below. Test the claim that there is a difference between the treatment and control groups.

Treatment group	$\overline{X}_1 = 4.20$	$n_1 = 22$	$s_1 = 2.20$
Control group	$\overline{X}_2 = 1.71$	$n_2 = 22$	$s_2 = 0.72$

Solution: Hyp Test for μ_1-μ_2
Notice that the σ's are unknown, so we use the t-version.
Since this will be a 2-tail test, if we choose α=0.05, our t-value (from the chart) is 2.080.

H_o: μ_1-μ_2
H_1: $\mu_1 \neq \mu_2$

$\alpha = 0.05$

DR: If $t_{\overline{X}-\overline{Y}} > 2.080$ or $t_{\overline{X}-\overline{Y}} < -2.080$ then reject H_o.

$$t_{\overline{X}_1-\overline{X}_2} = \frac{(\overline{X}_1 - \overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{(4.20-1.71)-(0)}{\sqrt{\frac{2.20^2}{22}+\frac{0.72^2}{22}}} = \frac{2.49}{0.49} = 5.08$$

^^^ ASSUME POPS ND

Since 5.08 is well above 2.080, we reject H_o

The means are not the same.

<u>Homework for today</u>

Section 9.3, # 5-8,13,14,15,16
This time I want you to email some answers to me.
For #5-8, send me your answers. (Each is either “matched pair” or “indep.”
For #14, send me your final confidence interval.
For #16, send me your two conclusions to the hypothesis test. So your first answer should be either “Rej H_o” or “Fail to Rej H_o”, and your second answer should be translating that first answer into English.
Don’t forget about hints and videos on the website.