Lecture 10
Chapter 6

* Confidence Intervals for μ_Y

So far the normal probability problems have required the knowledge of μ_Y, σ_Y, which are population values.

In reality we want to use $\bar{y}$ to estimate μ_Y, and would also like to know how good of an estimate $\bar{y}$ was.

With a few assumptions, we can use the fact
$\bar{Y} \sim N(\mu_Y, \sigma_Y/\sqrt{n})$ (if $n \geq 30$ or $Y \sim N(\mu_Y, \sigma_Y)$.)

We have to estimate both μ_Y and σ_Y, and more specifically, $\sigma_{\bar{Y}}$.

Definition: The standard error of $\bar{Y}$ (or the estimated standard deviation of $\bar{Y}$) is: $SE_{\bar{Y}} = s/\sqrt{n}$

It is a measure of how much error we expect to make when estimating μ_Y by $\bar{y}$.

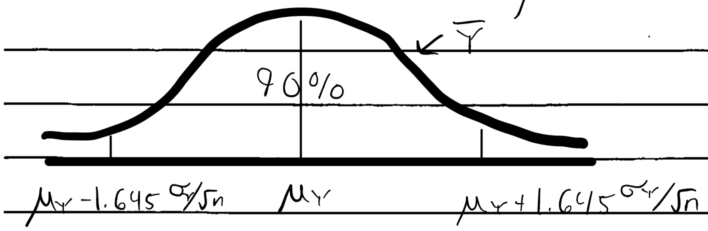
Facts: 1) As $n \to \infty$, $SE_{\bar{Y}} \to 0$ (if we sampled the population, we would have no error)
2) As $n \to \infty$, $\bar{y} \to \mu_Y$ and $s \to \sigma_Y$.

Now lets say we want a range of plausible $\bar{y}$'s we might observe if $\bar{Y} \sim N(\mu_Y, \sigma_Y/\sqrt{n})$. Say we want the middle 90%. In other words, we want the 5^{th} and 95^{th} percentiles of $\bar{Y}$.

First, the 5^{th} percentile for Z is: -1.645, and the 95^{th} percentile for Z is: $+1.645$,
Thus, the 5^{th} percentile for $\bar{Y}$ is: $\mu_Y + (-1.645)\sigma_Y/\sqrt{n} = \mu_Y - 1.645\frac{\sigma_Y}{\sqrt{n}}$
and the 95^{th} percentile for $\bar{Y}$ is: $\mu_Y + 1.645\,\sigma_Y/\sqrt{n}$

So the interval for the middle 90% is: $\mu_Y \pm 1.645\,\sigma_Y/\sqrt{n}$

I.e, 90% of the time we will draw a $\bar{y}$ that is in between $(\mu_Y - 1.645\ \sigma_Y/\sqrt{n}\ ,\ \mu_Y + 1.645\ \sigma_Y/\sqrt{n})$



Of course, the value of $\bar{y}$ we draw is random, but we know that 90% of the values of $\bar{y}$ we draw are within $(1.645\ \sigma_Y/\sqrt{n})$ of μ_Y.

So, if we took many random samples, and could calculate $\bar{y} \pm 1.645\ \sigma_Y/\sqrt{n}$ for them, ~90% of those intervals would contain μ_Y!

Problems: 1) σ_Y is unknown

2) $\bar{Y} \sim N(\mu_Y, \sigma_Y/\sqrt{n})$, so $(\bar{Y} - \mu_Y)/(\sigma_Y/\sqrt{n}) \sim Z$

(This is why we can use Z percentiles, i.e. 1.645)

Solutions: 1) Replace σ_Y with s (estimated std. dev).

BUT: We have added additional error to the distribution of $\bar{Y}$, because there is error associated with estimating σ_Y with s.

The result is that the distribution of $\bar{Y}$ when estimating σ_Y with s is <u>wider</u> than a normal distribution.

* The "t" distribution

Fortunately, the distribution of $\bar{Y}$ when using s instead of σ_Y is known, and called a "t" distribution.

Facts about "t":

1) It is bell shaped and perfectly symmetric

2) Its percentiles are always larger than a normal; so $t_{\alpha/2}$ (the $(1-\alpha/2)100^{th}$ percentile) is always larger than $Z_{\alpha/2}$

3) The spread of t depends on n, specifically on degrees of freedom $= n-1$ (d.f. $= n-1$).
As n grows larger, the spread of t shrinks

4) As $n \to \infty$, $t \to Z$. $t_{\alpha/2}$ at d.f. $= \infty$ $= Z_{\alpha/2}$.

Now, our interval changes from

$\bar{y} \pm Z_{\alpha/2}\, \sigma_Y/\sqrt{n}$ or $\begin{pmatrix}\text{estimate}\\ \text{of } \mu_Y\end{pmatrix} \pm \begin{pmatrix}Z\\ \text{percentile}\end{pmatrix}\begin{pmatrix}\text{actual}\\ \text{std dev of } \bar{Y}\end{pmatrix}$

to

$\bar{y} \pm t_{\alpha/2}\, S/\sqrt{n}$ or $\begin{pmatrix}\text{estimate}\\ \text{of } \mu_Y\end{pmatrix} \pm \begin{pmatrix}t\\ \text{percentile}\end{pmatrix}\begin{pmatrix}\text{estimated}\\ \text{std dev of } \bar{Y}\end{pmatrix}$

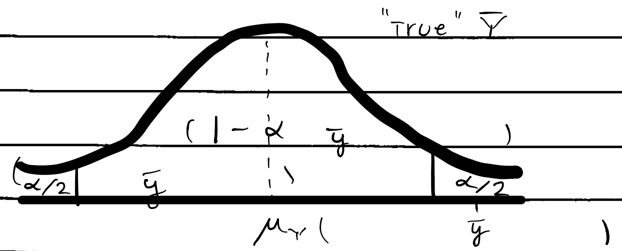
* A $(1-\alpha)100\%$ Confidence Interval for μ_Y

Note: CI = Confidence Interval

A $(1-\alpha)100\%$ CI for μ_Y is calculated as

$\bar{y} + t_{\alpha/2}\, S/\sqrt{n}$ at d.f. $= n-1$

For a specific sample mean $\bar{y}$, this interval will either contain μ_Y or not.
But, if we make many many CIs based on many samples, $\sim (1-\alpha)100\%$ of them will contain μ_Y. This is what is meant by "confidence"



Only when we sample a $\bar{y}$ "in the tails" do we "miss" μ_Y.

$(1-\alpha)100\%$ is the overall percentage in the middle, or the confidence level. Common values of α are: $\alpha = 0.10, 0.05, 0.01$

α	$(1-\alpha)100\%$	$\alpha/2$ (in the tails)
0.10	90%	.05
0.05	95%	.025
0.01	99%	.005

Finding $t_{\alpha/2}$

To find $t_{\alpha/2}$, we use the "t" table:

				Upper Tail Probability				
d.f.	0.20	0.10	0.05	0.04	0.03	0.025	....	0.005
1	$t_{.20}$		$t_{.05}$			$t_{.025}$		
2						⋮		
3								
⋮								
30								
40								
50								
⋮								
100								
⋮								
∞								

So you would go to row d.f. $= n-1$, column $\alpha/2$, and that would be $t_{\alpha/2}$.

Note: If a value is not in the table for d.f., round down to the next lowest.

* Assumptions for a CI for μ_Y

1) A random sample was taken (subjects indep)
2) Y must be distributed normal.
   i) $n \geq 30$ or ii) population is normal

We generally interpret CIs for μ as:
"We are $(1-\alpha)100\%$ confident that the true average of Y is between $(\bar{y} - t_{\alpha/2}\, s/\sqrt{n})$ and $(\bar{y} + t_{\alpha/2}\, s/\sqrt{n})$."
While stating the units of Y (interpreting "in terms of the problem").

Ex: The average systolic blood pressure (BP) for 10 women was found to be 119 mmHg, with a std dev of 2.1. Assume a random sample was taken.

a) Calculate the 95% confidence interval for the true average systolic BP for women.
$t_{\alpha/2}$: $\alpha = 0.05$, d.f. $= n-1 = 10-1 = 9$, at row 9 and column 0.025 $t_{.025} = 2.262$
CI: $119 \pm 2.262\,(2.1/\sqrt{10}) \Rightarrow 119 \pm 1.502$
$\Rightarrow (117.498,\ 120.502)$

b) Interpret your interval from b) in terms of the problem.
We are 95% confident that the true average systolic BP for women is between 117.498 and 120.502.

c) Would the interval widen or narrow if the sample size increased?
Since more data should mean less error, narrow.
OR Since $s/\sqrt{n}$ should decrease, narrow.

d) If we increased our confidence level to 99%, would our interval widen or narrow?
Widen, since we must cover more plausible values.
OR widen, since $t_{\alpha/2}$ will increase.

Note: We had to assume the population distribution of systolic BP was normal, since $n < 30$.

* Calculating n

Suppose we wanted to plan ahead, and we knew we wanted our CI to have a certain half-width, or Margin of Error.

Definition: The margin of error for a CI for μ_Y is
$MOE = t_{\alpha/2}\, s/\sqrt{n}$

Let e = desired value of MOE. Then, we want to plan how many people we need to sample to get e.

The formula is:

$$n = \frac{(Z_{\alpha/2})^2 s^2}{e^2}$$

note: $Z_{\alpha/2} = t_{\alpha/2}$ at d.f $= \infty$, ALWAYS ROUND UP

We use Z instead of t since t relies on n.

Note: This requires us to have an estimate s of σ_Y, and a desired Confidence Level (CL), or $(1-\alpha)100\%$.

Ex: A sample of 35 graduate students showed they drank an average of 3.5 cups of coffee per day, with a std. dev. of 4.1.

a) Find the 99% CI for the true average # of cups of coffee graduate students drink.

$t_{\alpha/2}$; $\alpha = 0.01$, $n-1 = 35-1 = 34$ (round down, use 30)

$t_{.005} = 2.750$

CI: $3.5 \pm (2.750)\, 4.1/\sqrt{35} \Rightarrow 3.5 \pm 1.906 \Rightarrow (1.594, 5.406)$

b) Does your interval support the claim that graduate students drink more than four cups of coffee on average? Explain

Since both bounds not strictly over 4, it does not support the claim. The interval represents plausible values of the true average based on our CI, and the plausible values are not all above 4.

c) What assumptions did you need?

A random sample.

d) How many subjects should we measure so that the margin of error for a 99% CI is 1 day?

$e = 1$, $s = 4.1$, $t_{\alpha/2}$ at d.f. $= \infty = 2.575$

$$n = \frac{(2.575)^2(4.1)^2}{1^2} = 111.4608 \approx 112 \text{ subjects}$$

* Comparing two group means

Instead of considering one mean, we often want to assess if two groups of subjects are significantly different. We often do this by comparing means.

Notation: Let group 1 be from population 1, with parameters μ_1, σ_1. The sample has statistics $\bar{y}_1, S_1$, with sample size n_1

Let group 2 be from population 2, with parameters μ_2, σ_2. The sample has statistics $\bar{y}_2, S_2$, with sample size n_2

If we make the following assumptions:

1) A random sample was taken from both populations
2) Populations 1 and 2 are independent
3) $\bar{Y}_1$ and $\bar{Y}_2$ are normally distributed, eith because
   i) $n_1 \geq 30$ and $n_2 \geq 30$ or
   ii) Populations 1 and 2 are normally distributed.

If these assumptions hold, then you could show

$$(\bar{Y}_1 - \bar{Y}_2) \sim N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right)$$

This uses linear combinations of random variables.

But again, we can't know σ_1 and σ_2, so we estimate them with s_1 and s_2, so that $(\bar{Y}_1 - \bar{Y}_2)$ is t distributed. The d.f. are much more complicated, but the details behind this CI are exactly the same as the one for μ. The above will give us a CI for $\mu_1 - \mu_2$.

The d.f. (which you would be given) are:

$$r = (s_1^2/n_1 + s_2^2/n_2)^2 \Big/ \left[\frac{(s_1^4/n_1^2)}{n_1 - 1} + \frac{(s_2^4/n_2^2)}{n_2 - 1}\right] \quad \text{rounded down.}$$

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}, \quad d.f. = \nu$$

Note: You may choose whichever group you like is 1 or 2, the interpretation will be similar.

Note: This CI is for a difference in two averages, so gives plausible values for the difference in two averages.

The three interpretations are:

1) Both bounds are below zero. This gives evidence that $\mu_1 < \mu_2$ (group 1 has a smaller average than group 2)
2) Both bounds are above zero. This gives evidence that $\mu_1 > \mu_2$ (group 1 has a larger average than group 2)
3) The bounds cover zero. This impiles that 0 is a plausible value, so we say there is no significant difference between μ_1 and μ_2

Ex: A researcher is studying the difference in average BMI between men and women. A small random sample shows:

	M	W
n	35	49
s	5.6	3.2
$\bar{y}$	21.2	23.4

Assume $\nu = 49.79$

Let group 1 = Men, group 2 = Women } arbitrary

a) Find the 95% confidence interval for the true difference in average BMI for men vs. women.

$t_{\alpha/2}$: $\alpha = 0.05$, d.f. $= 49$ (use 40 (we round down)), $t_{.025} = 2.021$

CI:

$$(21.2 - 23.4) \pm 2.021 \sqrt{\frac{5.6^2}{35} + \frac{3.2^2}{49}}$$

$\Rightarrow \quad -2.2 \pm 2.021(1.051) \quad \Rightarrow \quad (-4.324, -0.076)$

b) Interpret your CI in terms of the problem, being as specific as you can.

We are 95% confident that mens average BMI is less than womens by between 0.076 and 4.324.

c) What is the largest difference in average BMI we should expect between men and women?

This is the largest difference implied by the CI, which is 4.324.

* Hypothesis Tests for $\mu_1 - \mu_2$ (Chap 7)

Notice that we can use sample data to see if CIs support a particular claim or not. We can also perform hypothesis tests, which quantify how likely our data is, if a claim were true.

There are always 4 steps to hypothesis tests (HT), where the details differ depending on what test we are doing.

Step 1) State the null and alternative hypotheses.
The null hypothesis is the statement which we mathematically assume is true. It always contains an $=$, $\leq$, or $\geq$.
The alternative hypothesis is what must be true if the null is not, and so is the opposite of the null.
The null is denoted H_0, the alternative is denoted H_A.

For testing statements to do with $\mu_1 - \mu_2$, we have three pairs:

two sided $\{$ $H_0: \mu_1 - \mu_2 = 0$ $(\mu_1 = \mu_2)$ vs. $H_A: \mu_1 - \mu_2 \neq 0$ $(\mu_1 \neq \mu_2)$

one sided $\begin{cases} H_0: \mu_1 - \mu_2 \leq 0 \ (\mu_1 \leq \mu_2) \text{ vs. } H_A: \mu_1 - \mu_2 > 0 \ (\mu_1 > \mu_2) \\ H_0: \mu_1 - \mu_2 \geq 0 \ (\mu_1 \geq \mu_2) \text{ vs. } H_A: \mu_1 - \mu_2 < 0 \ (\mu_1 < \mu_2) \end{cases}$

A claim we are interested in can be in H_0 or H_A, but the $\leq$, $\geq$ or $=$ <u>must</u> be in H_0.

Step 2) Calculate a test-statistic
A test-statistic measures how much our sample data differs from H_0. Since H_0 has the $=$ sign, we use the value of 0 in our test-statistic.
Our test-statistic is:

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

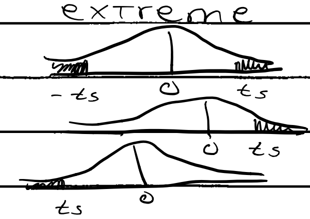
= the # of estimated std. dev our sample difference is from 0.

The larger the $|t_s|$, the more our data differs from the null. This has d.f. ν

Step 3) Calculate the p-value
The definition of a p-value is "The probability of observing our data (our test-statistic) or more extreme, if in reality the null were true"
"or more extreme" means different than H_0.
If H_0 were true, we know $(\bar{Y}_1 - \bar{Y}_2)$ has a "t" distribution, and we assume the equality $(\mu_1 - \mu_2 = 0)$
We can use this to calculate the p-value.

(extreme)

H_A	p-value	Tailed	"extreme" area
$\mu_1 - \mu_2 \neq 0$	$2\Pr\{t > \lvert t_s \rvert\}$	Two	
$\mu_1 - \mu_2 > 0$	$\Pr\{t > t_s\}$	one	
$\mu_1 - \mu_2 < 0$	$\Pr\{t < t_s\}$	one	



The smaller the p-value, the less likely it is to observe our data or more extreme, if H_0 was true $(\mu_1 - \mu_2 = 0)$.

Step 4) State your conclusion, and interpret.
How small of a p-value is small enough to say "It is too unlikely, we reject H_0"?
We use the cutoffs $\alpha = .10, .05$ or $.01$
If p-value $< \alpha$, reject H_0 (our data supports H_A)
If p-value $\geq \alpha$, fail to reject H_0 (our data supports H_0).
Note: We often call $\alpha(100)\%$ the level of significance.
Note: Since a t-distribution is symmetric, $\Pr\{t < -a\} = \Pr\{t > +a\}$.

* Calculating p-values with t-table.

With our t-table, we can only calculate ranges of p-values at particular d.f. The will look like

i) p-value $< a$ ii) p-value $> b$, or
iii) $a <$ p-value $< b$

We find a, b with the columns of the t-table.

Ex: Say $H_0: \mu_1 \leq \mu_2$, $H_A: \mu_1 > \mu_2$, $t_s = 3.381$, $\nu = 27$

Our p-value is $\Pr\{t > 3.381\}$ at d.f. $= 27$.
At row 27, 2.771 is in column 0.005 $\Rightarrow \Pr\{t > 2.771\} = 0.005$
3.690 is in column 0.0005 $\Rightarrow \Pr\{t > 3.690\} = 0.0005$
Since t_s is between 2.771 and 3.690, our p-value is between 0.005 and 0.0005. I.e, $0.0005 <$ p-value < 0.005

Note: If t_s is larger than any value in the row, $\Pr\{t > t_s\} < 0.0005$

Note: If t_s is smaller than any value in the row, $\Pr\{t < t_s\} > 0.20$

If your p-value is two sided, multiply the range by 2.
If you have a negative t_s, use $\Pr\{t < -a\} = \Pr\{t > +a\}$

Ex: A study was trying to determine if the average weight gain for two breeds of cows differed.

Breed	I	II	
mean	18.3	13.9	(lbs)
std. dev.	17.8	19.1	$r = 71$
n	33	51	

a) State H_0, H_A

$H_0: \mu_1 - \mu_2 = 0$ vs $H_A: \mu_1 - \mu_2 \neq 0$ (claim)

b) Calculate t_s

$$t_s = \frac{(18.3 - 13.9) - 0}{\sqrt{17.8^2/33 + 19.1^2/51}} = 1.075$$

c) Calculate the p-value.

At d.f. = 70 (71 rounded down),

$Pr\{t > .847\} = 0.20$, $Pr\{t > 1.294\} = 0.10$

Thus, $2(0.10) < \text{p-value} < (.20)2 \Rightarrow .20 < \text{p-value} < .40$

d) State your decision.

Since p-value > any $\alpha = .10, .05, .01$, we fail to reject H_0 and conclude there is no sig. diff in average weight gain of the two breeds.