EEP/IAS 118
Andrew Dustan
Section Handout 5

1. More practice with omitted variables bias

I've heard the following statement many times in my life:
(1) "Red cars cost more to insure than cars in other colors (even for the same amount of coverage)."

Similarly, I've received the following advice:
(2) "When you buy a car, make sure it's not red. Otherwise, the insurance company will charge you more for an equal amount of coverage."

I assert that while **(1)** is probably true, **(2)** is entirely false. When was the last time you got an insurance quote that asked you the color of your car? Let's use this myth to explore omitted variables bias further (and use dummy variables again).

Statement **(1)** is about predicting the price of insurance based only on a car's color (red or not):

$$\widehat{price} = \hat{\beta}_0 + \hat{\beta}_1 \, red$$

red = 1 when the car is red, 0 when it's not. Statement **(1)** says that if I run this regression, then $\hat{\beta}_1$ is

_____.

Statement **(2)** is about a ceteris paribus relationship between "redness" and insurance price in the population equation:

$$price = \beta_0 + \beta_1 \, red + u$$

I think $\beta_1 = 0$. Why won't its estimate be zero in the regression?

1) Think of an x_2 variable that should be in the population equation (i.e. it affects insurance price):

   _____.

2) How are this x_2 and car redness correlated? That is, when the car is red, is x_2 usually <u>higher</u> or <u>lower</u>?

   CIRCLE ONE: HIGHER $[cor(red, x_2) > 0]$ LOWER $[cor(red, x_2) < 0]$

3) How does x_2 affect insurance price? That is, what is the sign of β_2?

   CIRCLE ONE: POSITIVELY $[\beta_2 > 0]$ NEGATIVELY $[\beta_2 < 0]$

4) Put your answers from (2) and (3) together. Even if car color doesn't matter for insurance price, do we expect a randomly selected red car to cost more or less to insure than a randomly selected non-red one?

   CIRCLE ONE: MORE [positive bias, $\hat{\beta}_1$ too big] LESS [negative bias, $\hat{\beta}_1$ too small]

2. Estimating the mean of a population and its confidence interval

Why are we spending so much time on this?

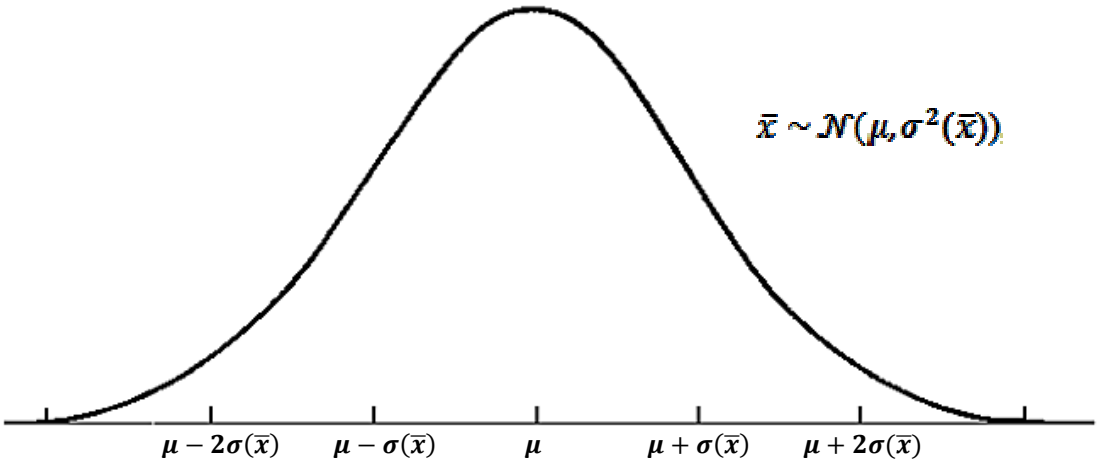
1) It's good to know how to estimate the mean of a population.
2) It reminds us that our estimates are guesses that are wrong, and we need to know how wrong we're likely to be.
3) Eventually, we'll apply these methods to talk about how wrong our $\widehat{\beta}$'s are likely to be.

The basics:

Here is a table of the population quantities we want to know and the sample estimates we have for them:

POPULATION	SAMPLE
Mean (μ)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance of x [$var(x)$ or $\sigma^2(x)$]	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Variance of $\bar{x}$ [$var(\bar{x})$ or $\sigma^2(\bar{x})$] $= \frac{var(x)}{n}$	$\frac{s^2}{n}$

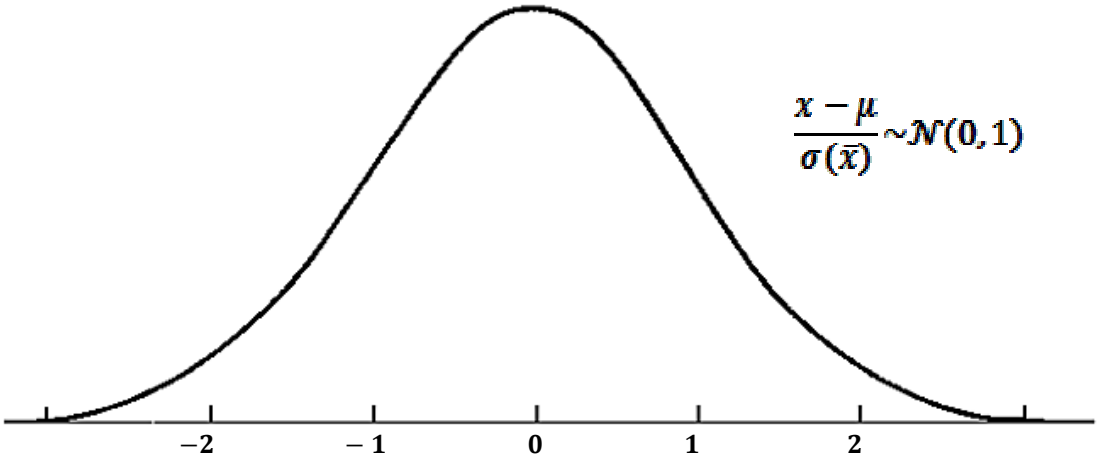
We know by the Central Limit Theorem that as our sample gets big, $\bar{x} \sim \mathcal{N}(\mu, \sigma^2(\bar{x}))$, or in words, that the sample mean is distributed normally, is centered at the true population mean μ, and has variance $\sigma^2(\bar{x})$, which goes down as we get a larger sample. If we knew μ and $\sigma^2(x)$ [and hence $\sigma^2(\bar{x})$] we could draw the distribution of $\bar{x}$:



For us to make use of our statistical tables (rather than having to use a computer), we need to transform this normal distribution into the **standard normal distribution**. We do this by using the following fact:

$$\frac{\bar{x} - \mu}{\sigma(\bar{x})} \sim \mathcal{N}(0, 1)$$

In words, to normalize the distribution of the mean, *subtract off the mean and divide by the standard deviation*. This gives us the following picture:



The reason we do this transformation is to get to this fun fact:

$$P\left(-1.96 < \frac{\bar{x}-\mu}{\sigma(\bar{x})} < 1.96\right) = 0.95$$

This lets us do algebra to get the formula we wanted for the confidence interval:

$$P\big(\bar{x} - 1.96\sigma(\bar{x}) < \mu < \bar{x} + 1.96\sigma(\bar{x})\big) = 0.95$$

In other words, if I take a sample and get its mean, I can say with 95% chance of being right that μ is inside this interval:

$$CI_{95} = [\bar{x} - 1.96\sigma(\bar{x}), \bar{x} + 1.96\sigma(\bar{x})]$$

In practice, we don't know $\sigma(\bar{x})$ so we replace it with the standard error of the sample mean, $s/\sqrt{n}$. Using the standard error changes the distribution of the sample mean a little bit (to a Student's t distribution which has a bit fatter tails) but since we are working with a large sample, it won't really affect what we are doing. Here is the distribution that we want:

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

I don't want to dwell on the difference between the t and standard normal distributions because they look almost the same when you get to a sample size of about 100. (In any case we were already assuming the sample size was large enough to apply the Central Limit Theorem.) When you do this you get **the formula we'll always use for a confidence interval**:

$$\boldsymbol{CI_{95} = \left[\bar{x} - c_{95}\left(\frac{s}{\sqrt{n}}\right), \bar{x} + c_{95}\left(\frac{s}{\sqrt{n}}\right)\right]}$$

We get c_{95} by looking at the t-table entry corresponding to the degrees of freedom we have and the percentage we use for the confidence interval. (When you look at the t-table, remember that 95% confidence corresponds to the 5% significance level, two-tailed test.) It's really close to 1.96 if the sample size is big.

PRACTICE:

You take a random sample of n=100 people's height in inches and find the following:

$\bar{x} = 65$

$s^2 = 4$ [note that we got this from multiplying the *sample variance* by n/(n-1), a very minor adjustment]

Form the 95% confidence interval for the mean of height in the population.

Steps:

1. Get *s*: _____$\sqrt{4} = 2$_____
2. Get $\sqrt{n}$: _____$\sqrt{100} = 10$_____

3. Get c_{95} from the t-table: _____1.98_____

4. Form the interval using the CI_{95} definition above: $\left[65 - 1.98\left(\frac{2}{10}\right), 65 + 1.98\left(\frac{2}{10}\right)\right]$

5. Simplify CI_{95} so that it's just an interval with two numbers in it: $[64.60, 65.39]$

6. Interpret this confidence interval in words: *Based on the information available to us from the sample, we are 95% confident that the mean height in this population is between 64.60 and 65.39*.

Bottom line: if you're confused about how it all works, that is fine. Learning these steps will ensure you can perform the calculations until you sort out the concepts.

MORE PRACTICE:
Use this Stata summary table of a sample of Michigan State University undergraduate GPAs to form a **90%** confidence interval (corresponds to 10% significance level two-tailed test) for undergraduate GPA in the MSU population:

colGPA

Percentiles		Smallest		
1%	2.3	2.2		
5%	2.5	2.3		
10%	2.6	2.4	**Obs**	**141**
25%	2.8	2.4	Sum of Wgt.	141
50%	3		**Mean**	**3.056738**
		Largest	Std. Dev.	.3723103
75%	3.3	3.8		
90%	3.6	3.9	**Variance**	**.138615**
95%	3.7	3.9	Skewness	.3246205
99%	3.9	4	Kurtosis	2.59999

1. Get *s*: _____$s^2 = 0.138615\left(\frac{141}{140}\right) = 0.1396 \rightarrow s = \sqrt{0.1396} = 0.3736$_____
2. Get $\sqrt{n}$: _____$\sqrt{141} = 11.874$_____

3. Get c_{90} from the t-table: ___1.66____

4. Form the interval using the CI_{90} definition above: $\left[3.057 - 1.66\left(\frac{0.3736}{11.874}\right), 3.057 + 1.66\left(\frac{0.3736}{11.874}\right)\right]$

5. Simplify CI_{90} so that it's just an interval with two numbers in it: $[3.005, 3.109]$

6. Interpret this confidence interval in words: *Based on the information available to us from the sample, we are 90% confident that the mean GPA in this population is between 3.005 and 3.109.*

3. Hypothesis Testing

Hypothesis testing is pretty similar to creating a confidence interval in many ways. We are still using the fact that $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, except now we'll actually calculate this quantity directly.

Let's go back to the MSU grade-point average example. Say an administrator there insists that the average university-wide GPA is exactly 3.0. Without access to the entire population's GPA *we can't say if he's right, but we can say if we're pretty sure he's wrong*. Following the steps from class:

STEP 1. Define hypotheses:

$$H_0: \mu = 3.0$$
$$H_1: \mu \neq 3.0$$

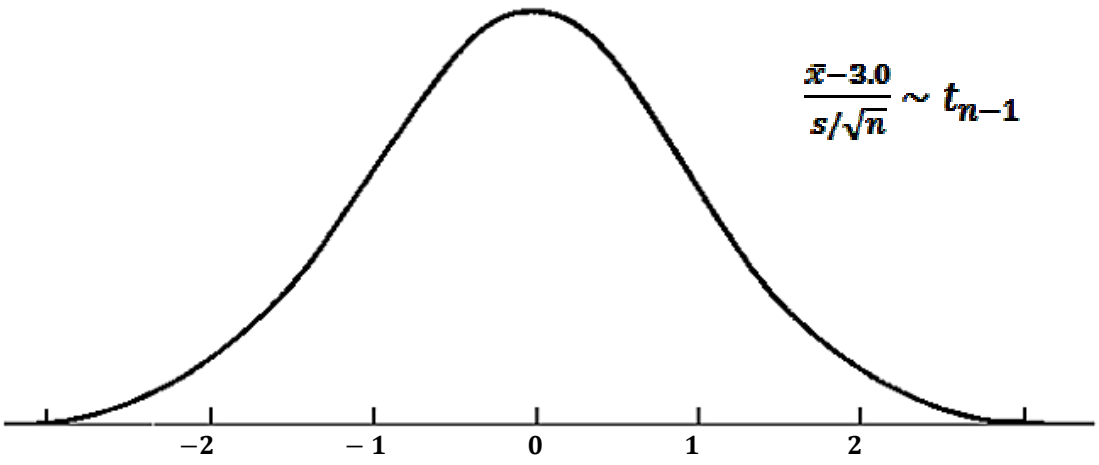
Here we're doing a two-tailed test, unlike the one-tailed test from class. Two-tailed tests are the most common.

Under the null hypothesis, μ is 3.0, so we assume this is true for now. If true, then $\frac{\bar{x}-3.0}{s/\sqrt{n}} \sim t_{n-1}$. Let's actually compute this "test statistic", t_{n-1}.

STEP 2. Compute the test statistic:

1. Get *s*: _____
2. Get $\sqrt{n}$: _____
3. Get n-1: _____
4. Compute t_{n-1} (this will just be a number): _____

Now it's time to think about what t_{n-1} actually means. If μ really is 3.0, then the distribution of t_{n-1} looks like this:



How likely is it that I'd get my $\bar{x}$ if the true mean of GPA is 3.0? To find out, compare the t-statistic you got to the values on the axis of this graph. That'll show you how improbable your result was. If it's too improbable then we can confidently reject the notion that average GPA really is 3.0.

STEP 3. Get the significance level of the test:

Here we'll choose the 5% significance level, meaning we'll wrongly reject a **true** H_0 5% of the time. We go to the t-table and find out what the corresponding critical value (c) is. It's about 1.98 for n-1 = 141-1 = 140.

STEP 4. Reject the null hypothesis or fail to reject it:

Our critical value, c, is 1.98. Our t-statistic was _____.

If our $|t_{n-1}| > 1.98$ then we reject the null hypothesis because our $\bar{x}$ was so far away from the null hypothesis of 3.0. If $|t_{n-1}| < 1.98$ then we can't reject the null hypothesis because $\bar{x}$ is close enough to the null hypothesis of 3.0 that we can't say it's wrong with enough confidence.

Did we reject H_0? YES NO

STEP 5. Interpret:
Either:
There is statistical evidence at the 5% significance level that the average GPA at Michigan State is different from 3.0. We have good reason to believe that the administrator was incorrect.

Or:
There is no statistical evidence at the 5% significance level that the average GPA at Michigan State is different from 3.0. We would not be confident in saying that the administrator is incorrect.