Introductory Applied Econometrics
EEP/IAS 118
Spring 2014

Andrew Crane-Droesch
Section #6
5 March 2014

WARM UP:

Match the terms in the table with the correct formula:

- Let X be a random variable with mean μ and variance σ^2, and we have a random sample $\{x_1, ..., x_n\}$ with average $\bar{x}$.

- Let Y be a binary random variable that can only take the values 0 or 1, its mean is p and we have a random sample $\{y_1, ..., y_n\}$ with average $\bar{y} = \hat{p}$.

	Term	Formula
1	Observed Variance of x	
2	Sample Variance of x	
3	Variance of $\bar{x}$	
4	Sample Variance of $\bar{x}$	
5	Sample Variance of y	
6	Sample Variance of $\bar{y}$ $(= \hat{p})$	

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \hat{p}(1-\hat{p}) \qquad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \frac{\sigma^2}{n} \qquad \frac{s^2}{n} \qquad \frac{\hat{p}(1-\hat{p})}{n}$$

1 Constructing and Interpreting a Confidence Interval

Suppose I took a random sample of 121 UCB students' heights in inches, and found that $\bar{x} = 65$ and $s^2 = 4$. Now, I'd like to construct a 95% confidence interval for the average height of UCB students.

Step 1. Determine the confidence level.

In the problem, it was given that we want to be 95% confident that our interval covers the true population parameter, so our confidence level is 0.95.

Step 2. Compute your estimates of $\bar{x}$ and s.

Again, we were given in this example that $\bar{x} = 65$ and $s^2 = 4$, so we don't need to do the calculations.

Step 3. Find c from the t-table.

The value of c will depend on both the sample size (n) and the confidence level (always use 2-Tailed for confidence intervals):

- If our confidence level is 80% with a sample size of 10: $c_{80} = 1.383$
- If our confidence level is 95%, with a sample size of 1000: $c_{95} = 1.960$

Step 4. Plug everything into the formula and **interpret**.

The formula for a W% confidence interval is:

$$CI_W = \left[\bar{x} - c_W \left(\frac{s}{\sqrt{n}}\right), \bar{x} + c_W \left(\frac{s}{\sqrt{n}}\right)\right]$$

Where c_W is found by looking at the t-table for $n-1$ degrees of freedom. So for our problem, we can plug everything in to get:

$$CI_{95\%} = \left[65 - 1.98 \left(\frac{4}{\sqrt{120}}\right), 65 + 1.98 \left(\frac{4}{\sqrt{120}}\right)\right]$$

The 95% confidence interval is _____. This interval has a 95% chance of covering the true average height of the UCB student population.

Practice

Use the Stata output below to construct a 90% confidence interval for Michigan State University undergraduate GPA from a random sample of the MSU student body:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      colGPA |       101       2.984    .3723103        2.2          4
```

1. Confidence level: 90% (Significance level is 100-90=10%)

2. $\bar{x}, s$: $\bar{x} = 2.984$ and $s = 0.3723$

3. Find c_{90}: $c_{90} = 1.662$

4. Compute & Interpret interval:

$$\left[2.984 - 1.662\frac{.3723}{\sqrt{101}},\quad 2.984 + 1.662\frac{.3723}{\sqrt{101}}\right]$$
$$\left[2.922,\quad 3.046\right]$$

Interpretation: We are 95% confident that this interval covers the true MSU average GPA.
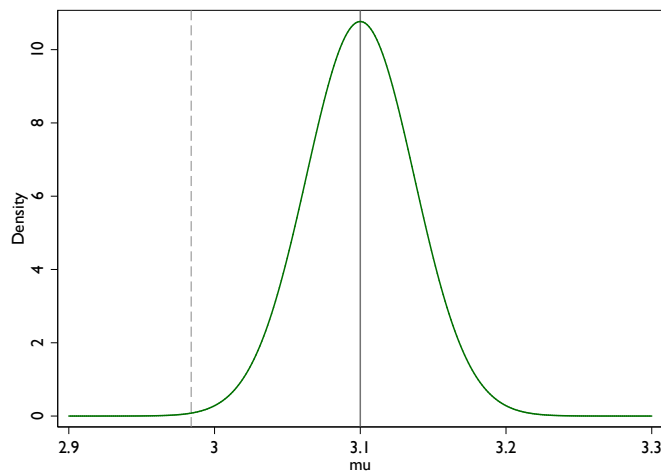
2   Hypothesis Testing

Hypothesis testing is intimately tied to confidence intervals. When we interpret our confidence interval, we specify that there's a 95% chance that the interval covers the true value. But then there's only a 5% chance that it doesn't—so after calculating a 95% confidence interval, you would be skeptical to hear that μ is equal to something above or below your interval. To see this, think about the above practice problem, except suppose (just for the moment) that we *know* the unknown true μ and $Var(\bar{x})$; so we know that the true MSU average GPA is 3.0 and $Var(\bar{x}) = 0.0015$. Below is the distribution of $\bar{x}$.

Is there a high probability that a random sample would yield an $\bar{x} = 2.984$? It certainly looks like it, given what we know about the true distribution of $\bar{x}$. Is there a high probability that a random sample would yield an $\bar{x} = 3.1$?
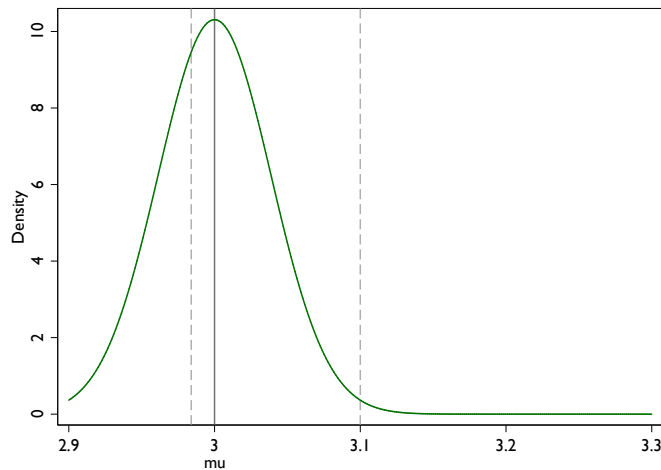
Hypothesis testing reverses this scenario, but uses the exact same intuition. Suppose the dean of MSU, Lou Anna Simon, firmly believes the true average GPA of her students is 3.1, and she's convinced our random sample isn't an accurate reflection of the caliber of MSU students. Suppose we decide to entertain her beliefs for the moment that the *true* mean GPA is 3.1. According to Lou Anna, the distribution of $\bar{x}$ looks like this:

Should we be skeptical of Lou Anna's claim?

How to test a hypothesis:

Now we can formalize what we did above into the 5 steps in hypothesis testing that we covered in lecture:

Step 1: Define hypotheses: H_0 and H_1

Here we write down the null and alternative hypotheses in precise mathematical language.

$$H_0 :$$
$$H_1 :$$

Since we don't have a good reason to think the average GPA should be higher or lower than 3.1, our alternative hypothesis is:

$$H_1 : \mu \neq 3.1$$

Step 2: Compute test statistic: $t = \frac{\bar{x} - \mu}{SE(\bar{x})} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

This step is pretty mechanical. To compute the t-statistic (often referred to as the t-stat), we need the sample mean $\bar{x}$, the sample variance s^2, and the sample size n:

$$
\begin{aligned}
\bar{x} &= 2.984 \\
s^2 &= 0.13861 \\
n &= 101 \\
\Rightarrow t &= \frac{2.984 - 3.1}{\sqrt{\frac{0.13861}{101}}} = -3.13
\end{aligned}
$$

You can think of the t-stat as: the sample mean converted to standard units, assuming the null hypothesis is true. Equivalently, if the null hypothesis is true, then we drew the sample mean from a normal distribution with mean 3.1 and using this distribution, we can re-write $\bar{x}$ in standard units.

Step 3: Choose significance level (α) and find the critical value: c_α from table

The significance level is:

Which is the same as the probability of Type 1 Error:

Since Type I error is obviously bad, we want the probability that we make this type of error to be small. We can set this probability to be low by choosing a low significance level. Once we've picked a significance level, we can find the critical value from a statistical table. Let's choose a 5% significance level for our example, so that the probability of Type I Error is just 5%. Check the t-table for the critical value when our significance level for a 2-tailed test is $\alpha = 0.05$ and we have 100 degrees of freedom:

$$c_\alpha = c_{.05} = 1.987$$

Step 4: Reject the null hypothesis or fail to reject it: Reject the null if $|t| > |c_\alpha|$

If $|t| > |c_\alpha|$, then we reject the null hypothesis. We established above that *if the null hypothesis were true* and the true mean is 3.1, then the probability of observing a sample mean with a t-stat of $|t| > |c_{.05}|$ is equal to the significance level, or 5%. So if we do observe a sample mean with a t-stat of $|t| > |c_{.05}|$ then we say we reject the null hypothesis. Just as before, we can think of a picture to help with the intuition here:

If $|t| < |c_\alpha|$, then we fail to reject the null hypothesis. Note that we never really "accept" the alternative or null hypothesis, we simply do or do not find evidence against the null. In our example:

$$|t| = |-3.13| = 3.13 > 1.987 = c_{.05} \quad \Rightarrow \text{Reject the null}$$

Step 5: Interpret

If we *reject* the null hypothesis: There is statistical evidence at the 5% level that the average GPA at MSU is different from 3.1. It's unlikely that Lou Anna's assertion is correct.

If we *fail to reject* the null hypothesis: There is no statistical evidence at the 5% level that the average GPA at MSU is different from 3.1. It's possible that Lou Anna's assertion is correct.

3 Two Variations: Proportions and Difference Between Two Means

There are two variations on hypothesis testing that we've covered so far:

1. When $\bar{x}$ is actually a proportion, i.e. the average of a binary variable. Hypothesis tests for binary variables vary from the steps outlined above in two ways:

   (a) Under the null hypothesis, we know the true proportion, p. Recall the formula given for the mean and variance of a binary variable—knowing the mean implies that you know the variance. Thus, when we use the null hypothesis to compute the test statistic, we don't use the sample variance. Instead we use the variance specified by the null hypothesis.

   $$\mathbb{E}[x] = p \quad Var(x) = \frac{p(1-p)}{n}$$

   (b) Remember that if we know the true mean *and* the true variance, we use a normal distribution instead of a t-distribution. So to find the critical value, we use the normal table instead of the t-table.

2. When we're interested in how two sample means differ. Hypothesis tests for a difference in means vary from the steps outlined above in one way:

   (a) To compute the test statistic, we use the standard deviation of the *difference.*

Example 1. Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. Call X the binary voting variable. An NGO suspects that the dictator is lying and contracts you, a skilled econometrician, to investigate this claim. You have a small budget, so you can only collect a random sample of 200 voters in the country. From your sample of 200, you find that 115 supported the dictator, so the sample proportion $\hat{p} = 0.575$.

Step 1.

$$
\begin{aligned}
H_0: & \quad p = 0.65 \\
H_1: & \quad p < 0.65
\end{aligned}
$$

(The reason why we're using a 1-tailed test here is because we suspect the proportion supporting the dictator is below 65%.)

Step 2. Remember that the variance of X under the null hypothesis is: $Var(X) = p(1-p)$

Then for large samples, we know that:

$$Var(\hat{p}) = \frac{p(1-p)}{n} = \frac{.65(1-.65)}{200} = 0.0011375 \quad \Rightarrow \quad SD(\hat{p}) = 0.033727$$

This means that our test statistic, $z = \frac{\hat{p}-p}{SD(\hat{p})} \sim N(0,1)$:

$$z = \frac{\hat{p}-p}{SD(\hat{p})} = \frac{.575 - .65}{0.033727} = -2.224$$

Step 3. Here we'll choose a 5% significance level again, so $\alpha = .05 \quad \Rightarrow c_{.05} = -2.57$

Step 4. Reject null hypothesis if $|z| > |c_{.05}|$, so we: Reject   Fail to reject

Step 5. Interpret: We fail to reject the null hypothesis (at the 5% significance level) that the proportion of people who support the dictator is 65%. In this sample, there is no evidence that the dictator is fixing the vote results.

Example 2. Now let's consider an example from actual data for a poverty alleviation program in Mexico. In 1997, 24,059 households in rural Mexico were randomly allocated between treatment and control groups for a conditional cash transfer program called Oportunidades to keep kids in school. When analyzing the results of a randomized experiment, the first step is to verify that the control group is, on average, very much like the treatment group in terms of characteristics that we observe and have data for. For example, data was collected on household assets. Your data reveals that while 14.47% of the 14,846 treatment households have a refrigerator, and 16.53% of the 9,213 control households have one. In order to confirm that about the same proportion of households in each group have a refrigerator, we need to perform a hypothesis test.

Call the sample proportion of households with a refrigerator in the treatment group $\hat{p}_t$, the true treatment proportion p_t, the sample proportion of households with a refrigerator in the control group $\hat{p}_c$, and the true control proportion p_c. Also, call the whole sample proportion of households (in either treatment or control) with a refrigerator $\hat{p}$.

Step 1. Define the hypotheses

$$H_0: \quad p_t - p_c = D = 0$$
$$H_1: \quad p_t - p_c = D \neq 0$$

Step 2. Compute a test statistic (t-stat) How do we compute this test statistic? We know that the null hypothesis specifies $\mathbb{E}[p_t - p_c] = 0$, so what's left is the standard deviation. Whenever we're testing a difference of means, remember the formula: $Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y})$.

So applying the formula, we have that:

$$Var(\hat{p}_t - \hat{p}_c) = Var(\hat{p}_t) + Var(\hat{p}_c)$$
$$Var(\hat{p}_t) = \frac{\hat{p}(1-\hat{p})}{n_t}$$
$$Var(\hat{p}_c) = \frac{\hat{p}(1-\hat{p})}{n_c}$$
$$\text{Which means } SD(\hat{p}_t - \hat{p}_c) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_t} + \frac{\hat{p}(1-\hat{p})}{n_c}}$$

The trickiest part here is keeping track of what your null hypothesis is! Now we're ready to calculate our z-statistic:

$$
\begin{aligned}
\hat{D} &= \hat{p}_t - \hat{p}_c = -.0206 \\
\hat{p} &= \frac{14846}{24059}(.1447) + \frac{9213}{24059}(.1653) = .1526 \\
SD(\hat{D}) &= \sqrt{\frac{.1526(1-.1526)}{14846} + \frac{.1526(1-.1526)}{9213}} = .00477 \\
\Rightarrow \quad z &= \frac{-.0206 - 0}{.00477} = -4.32
\end{aligned}
$$

Step 3. By the null hypotheses we chose, we're doing a two-sided test. Let's choose the 5% significance level as this is the most common test that economists evaluate. Check the normal table to find that $c =$

Step 4. Reject   Fail to reject

Step 5. Interpret:At the 1% significance level, there is statistical evidence that the proportion of households with a refrigerator in the control group is not the same as the proportion of households with a refrigerator in the treatment group. What does this mean for the study? *Probably not much. In randomized experiments such as this one, many household characteristics are checked for "balance" across treatment and control. Statistically, we expect that some of our hypothesis tests will reject the null simply because a 5% significance level indicates that 5% of the time we will reject the null even though it's true.*