**Hypothesis Testing and Significance**

Young Jin Lee

**2 types of methods for conducting a hypothesis test**

1. Via simulation (Unit 1)
2. Via theoretical methods that rely on the CLT (this Unit)

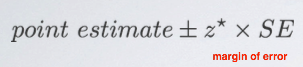
**Null hypothesis**

Often either a skeptical perspective or a claim to be tested.

**Alternative Hypothesis**

Represents an alternative claim under consideration and is often represented by a range of possible parameter values.

**Hypothesis testing via Confidence Interval**



If the null value falls in the confidence interval that is already constructed using a sample, you fail to reject the null hypothesis, and otherwise, reject it.

* A quick and dirty approach for hypothesis testing.
* Doesn't tell us the likelihood of certain outcomes under the null hypothesis; the p-value, based on which we can make decisions on the hypothesis.

**Example)**

You have constructed a 95% confidence interval for the average number of exclusive relationships college students have been to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

H0: μ = 3

H1: μ > 3

Since 3 falls in the interval, you fail to reject the null hypothesis that states college students have been in 3 exclusive relationships, on average.

* The interval says that any value within it could conceivably be the true population mean.

**Formal hypothesis testing using p-values**

**P-value**

**P(observed or more extreme outcome | H0 is true):** the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.

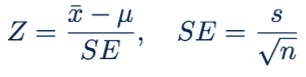
**Interpreting P-value**

1. p-value < alpha: it would be very unlikely to observe the data if the null hypothesis were true, and therefore we reject the null hypothesis.

1. p-value > alpha: it is indeed likely to observe the data even if the null hypothesis were true. And hence, we would not reject the null hypothesis.

Since p-value is high enough, you think the sample mean or an even higher sample mean (which is higher than what null hypothesis states) is likely to happen simply by chance. In other words, the difference between the null value and the observed sample mean is due to chance or sampling variability.

**Z score (test statistic)**



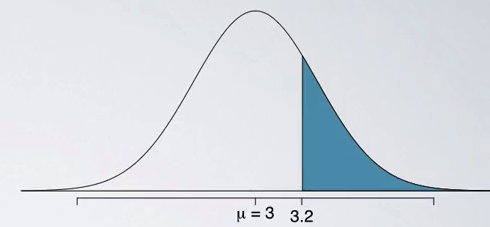
**Example)**

n = 50, x\_bar = 3.2, s = 1.74, SE = 0.246

P(x\_bar > 3.2 | H0: μ = 3)

X\_bar ~ N(μ = 3, SE = 0.246)

*Since we are assuming null hypothesis to be true, we can use that to construct the sampling distribution based on the central limit theorem.*



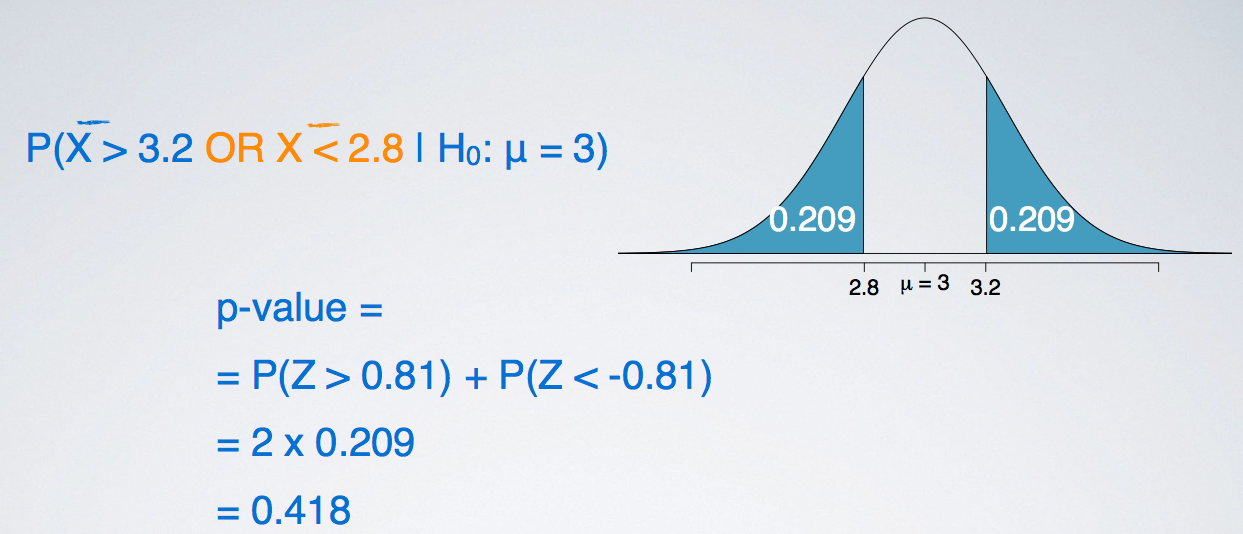
Calculate the z score (so called ‘test statistic’): z = (3.2 -3) / 0.246 = 0.81

P-value = P(z > 0.81) = 0.209

Even though we observed a sample mean slightly above 3, there is not enough evidence to reject the null hypothesis that sets the population average of number of exclusive relationships college students have been in to 3. If in fact college students have been in 3 exclusive relationships on average, there is a 21% chance that a random sample of 50 college students would yield a sample mean of 3.2 or higher. This is a pretty high probability, so we think that a sample mean of 3.2 or more exclusive relationships is likely to happen simply by chance. In other words, the difference between the null value of 3 relationships and the observed sample mean of 3.2 relationships is due to chance or sampling variability.

**Two-sided hypothesis tests**

Often, instead of looking for a divergence from the null hypothesis in a specific direction, so either a greater than or less than sign, we might be interested in divergence in any direction.



The definition of a p-value is the same regardless of doing a one or a two-sided test. However, the calculation becomes slightly different and ever so slightly more complicated since we need to consider at least as extreme as the observed outcome in both directions away from the mean.

If you're looking for the cutoff values for these two-sided hypothesis tests, all you need to do is to travel the same distance away from the mean.

**Hypothesis testing for a single mean:**

1. Set the hypothesis
2. Calculate the point estimate using a sample
3. Check conditions (same as CI condition)
4. Draw sampling distribution, calculate test statistic (z score), shade p-value
5. Make a decision based on [p-value vs alpha(significance level)] and interpret it in context of the research question:

* *See 2\_3\_HT\_examples.pdf*

**Inference for other estimators**

**Nearly normal sampling distributions**

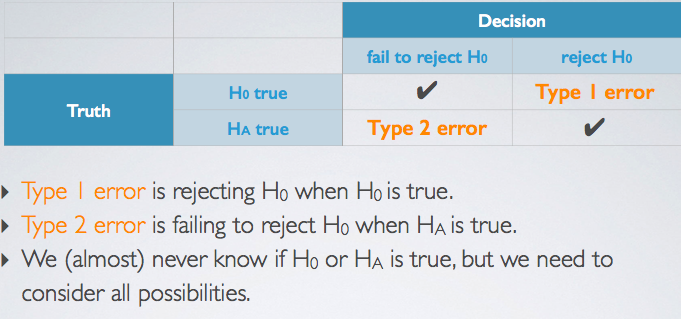
1. Sample means
2. Difference between sample means (difference between two groups/populations)
3. Sample proportion
4. Difference between sample proportions (difference between two groups/populations)

**Unbiased estimator**

An important assumption about point estimates is that they are unbiased, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates. That is, an unbiased estimate does not naturally over or underestimate the parameter, it provides a ‘good’ estimate. The sample mean is an example of an unbiased point estimate, as well as others we just listed above.

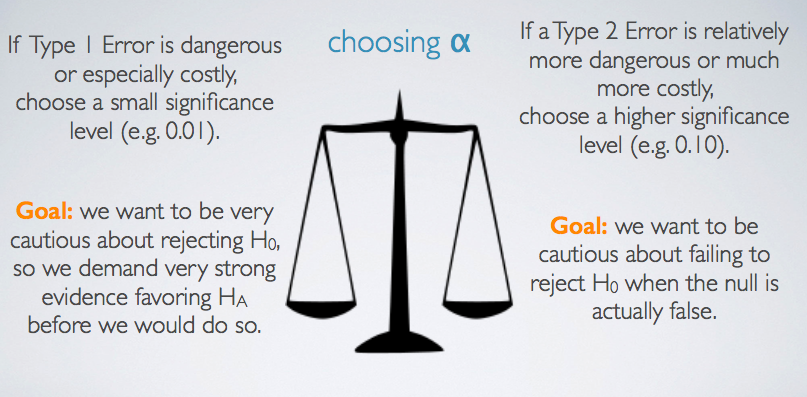
* The way to do inference with both confidence intervals and hypothesis testing is the same for all nearly normal point estimates.

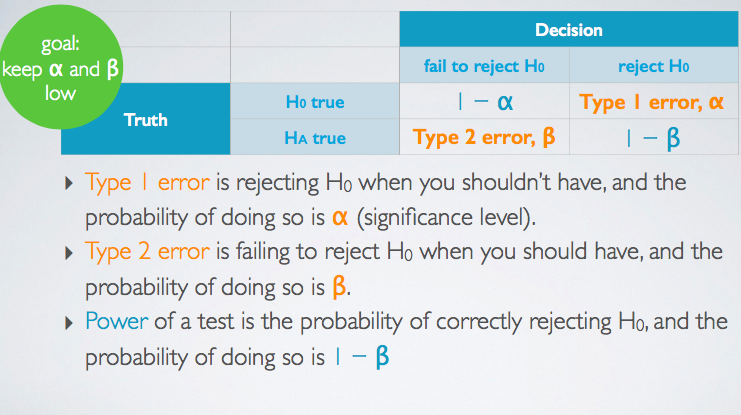
**Decision Errors**

****

The likelihood of making a type 1 error and likelihood of making a type 2 error are actually inversely proportional. So, it's actually not that easy to keep both of those error rates down. So sometimes we have to choose and we're going to talk about how do we choose between which one we are okay with being a little higher versus which one we really want to minimize as much as possible.

For example, Declaring the defendant guilty when they are actually innocent (type 1) is much worse than declaring the defendant innocent when they are actually guilty (type 2).

****

****

**Type 1 error**

We reject H0 when the p-value is less than 0.05 (alpha = 0.05). This means that, for those cases where H0 is actually true, we do not want to incorrectly reject it more than 5% of those times. In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

P(Type 1 error | H0 true) = alpha

This is why we prefer small values of alpha – increasing alpha increases the Type 1 error rate.

**Type 2 error**

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 error, i.e. we fail to reject the null hypothesis even when we should reject it?

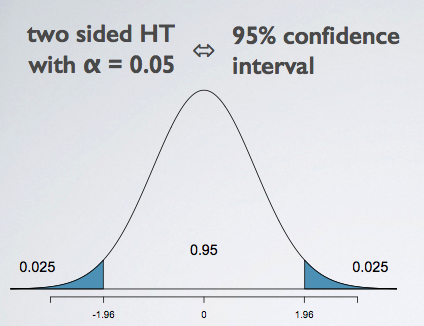
* The answer is not obvious.
* If the true population average is very close to the null value, it will be difficult to detect a difference (and reject H0).
* If the true population average is very different from the null value, it will be easier to detect a difference.
* Clearly, Beta depends on the effect size (the difference between point estimate and null value)

**Significance vs Confidence Level**

The interplay between significance levels used in hypothesis testing as well as confidence levels used in construction of confidence intervals.

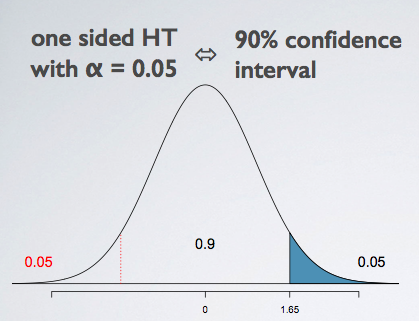
So far, we've used two inferential techniques, hypothesis tests and confidence intervals. And it makes sense that the results of these two techniques agree with each other. And they will, if we're using equivalent levels of significance and confidence.

**Two-tailed tests**



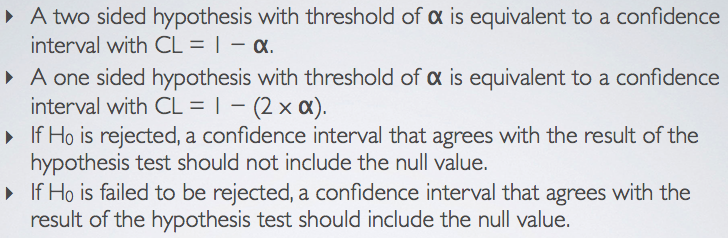
If we add up the 95% in the middle with the alpha on the two tails, we get to the one, which is the entire area under the curve. So indeed, for two-sided hypothesis tests, the significance level and the confidence level are complements of each other.

**One-tailed tests**



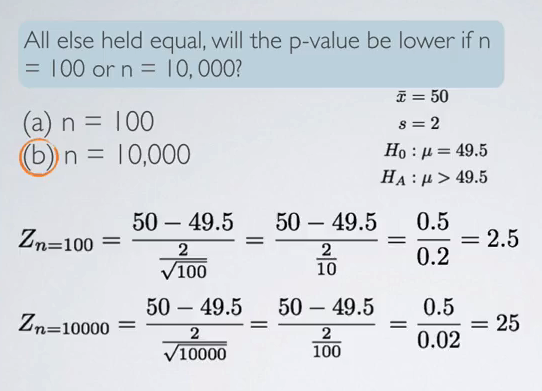
While the lower tail in this figure is of no interest to us within the framework of the hypothesis test. Since confidences always have to be symmetric, and the confidence level is always about the middle whatever percent of the distribution. we can’t have a confidence interval that only goes a certain amount of distance in one direction but more in the other direction, we are actually going to need to think about the 5% in the lower tail, even though we are not going to use it for any sort of decision making within the hypothesis test.

**Agreement of CI and HT**



**Statistical vs Practical Significance**

Increase in sample size -> decrease in standard error -> increase test statistic -> smaller p value



A z-score of 25 basically means almost no p-value. Or in other words, highly statistically significant finding. However, is it practically significant? When we're thinking about practical significance, we focus on the effect size. And remember, we define the effect size as the difference between your point estimate, and then your null value. So that would be in the calculation of the test statistic as the numerator. In both instances, we have the same exact effect size. And it's a small effect size to be fair as well. And even though we have a small effect size, which may not be practically significant. We are able to find the statistically significant result simply by inflating our sample size. And remember that the sample size is something the researcher has control over, because after all, you get to decide how many observations you want to sample. Sure, there's going to be a bound based on how many, how much resources you have but at the end of the day that's the human controlled part of a study. So, when you see highly statistically significant results make sure that you have a critical eye and make sure that you also inquire whether the effect size is reported and what the sample size is as well. And not only should you inquire this stuff but if you are reporting these highly statistically significant results it's always a good idea to let your readers know of your effect size and your sample size so that it, the discussion is made clear if a statistically significant finding is also practically significant or not.

So, to summarize real differences between the point estimate and the null value are easier to detect with large samples. However very large samples will result in statistical significance even for tiny differences between the sample mean and the null value or our effect size, even when the difference is not practically significant. So, in order to make sure that your findings don't suffer from this problem of being statistically significant, but not practically significant, oftentimes what we do is we do some a priori analysis before you actually do the data collection to figure out, based on characteristics of the variable you are studying, how many observations to collect. So, it is highly recommended that researchers, either they should do this themselves, or consult with statisticians, but to figure out how many observations to sample before they actually go into doing that. Because the last thing you want to do is having to find out is, we have already put in the researches to collect some data, and you either don't have enough or you have too many observations. This brings to mind a quote from a famous statistician, R.A. Fisher. To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination. He may be able to say what the experiment died of.