MSPP PPOL 501-02 & -06: Fall 2014
Course Notes # 11: Type I and Type II Errors; Power
Professor Carolyn Hill

I. TYPE I AND TYPE II ERRORS IN HYPOTHESIS TESTING

A. The Concepts of Type I and Type II errors

- As we've been testing hypotheses, we've been using the phrases "reject the null hypothesis" when the calculated statistic falls in the critical region. Or, "fail to reject the null hypothesis" when the calculated statistic does not fall in the critical region.

- Why don't we just say "accept the alternative hypothesis" or "accept the null hypothesis," respectively?

- We've touched on this a bit already, the framework here elaborates on those ideas:

  (1) we don't know what the true population mean μ is equal to (that's why we're using statistical inference to say something about this mean);

  (2) only one *true* population mean μ exists for the sampling distribution; when we run a hypothesis test, we don't know what this true mean is.

- Think about the various ways we can make the right and wrong decisions:

		"State of the World" – the Truth	
		H_0 is true	*H_0 is false*
Decision *after* Hypothesis Test	*Fail to Reject H_0*	Correct decision: probability = $(1-\alpha)$	Type II error: probability = β
	Reject H_0	Type I error: probability = α	Correct decision: probability = $(1-\beta)$ = "power of the test"

- Ideally, the probability of both Type I and Type II errors will be "small."

- However, there is a direct tradeoff between the two types of error *for a given sample size:* the smaller the significance level (α), the larger the probability of not rejecting a false null hypothesis (β).

- So, we have to assess the risks involved in committing either type of error and use that to balance the types of error probabilities.

B. Calculating Type II Error Probabilities and the Power of the Test[1]

- Key point: the probability of making a Type II error depends on: (1) the sample size, (2) the significance level, and (3) the true value of the parameter

- The following example shows how to compute a Type II error probability. We'll use a one-tail test. In practice, you can calculate power for two-tail hypothesis tests.

  Example: The program designers of a new approach for getting low-skilled clients into stable jobs claim that after completing the program, clients earn $35,000 per year, on average. An advocacy group is skeptical of this claim, and thinks that the program designers are overselling the effectiveness of their program. The advocacy group thinks that the true mean earnings for participants is actually less than $35,000 per year. The advocacy group plans to perform this test:

  H0: $\mu_{trainees}$ = *$35,000 per year*

  H1: $\mu_{trainees}$ < *$35,000 per year*

  They will use a 5% significance level, and a random sample of n=3,000 trainees. The standard deviation of trainee earnings in the sample is $10,000.

  Q1: At what sample mean will the null hypothesis be rejected, using alpha = 0.05 for a one-tail test? i.e., this is the usual hypothesis test that we are used to carrying out:

$$t_{calc}: \frac{\bar{X} - 35}{10/\sqrt{3{,}000}} = -1.645 \quad \Leftrightarrow \quad \bar{X} = 35 - 1.645\left(10/\sqrt{3{,}000}\right) = 34.7$$

  Decision criterion:
  If $\bar{X}$ <= $34.7 thousand, Reject Ho. If $\bar{X}$ > $34.7 thousand, Fail to Reject Ho.

  Q2: What is the probability of making a Type II error if the **<u>true mean earnings of trainees in the program</u>** is $34,850 per year?

  In this situation, the TRUE state of the world is that μ_{trainees} = $34.85 thousand, and the standard error is still $\left(10/\sqrt{3{,}000}\right) = 0.182574$.

  But remember we set up a decision rule in Q1 that we would reject Ho if $\bar{X}$ <= 34.7.

[1] This section draws from Weiss (2002) *Introductory Statistics, 6th ed*

We can think about a new, "true" curve of normally distributed $\overline{X}$ values about their true mean of \$34.85 thousand.

We know that a Type II error occurs if we do not reject Ho (i.e., if $\overline{X}$ >\$34.7 thousand). The probability of that happening is equal to the probability of our drawing a sample with a mean that is greater than 34.7, where that mean is drawn from a population where the *true mean is \$34.85 thousand:*

$$t_{calc}: \frac{34.7 - 34.85}{0.182574} = \quad -0.82158$$

Because we're calculating a t-statistic, we'd like to look up -0.82158 in the *t*-table. However, the *t*-table we have doesn't have this level of detail. Because we have a large enough sample size, to the point where the *t*-distribution approximates the Normal distribution, we can use the *Z*-table to look up this value. We see that -0.82 corresponds to an area to the left of Z of 0.2061, and an area to the right of Z of (1-0.2061) = 0.7939.

So if the true mean earnings of trainees in the program is \$34,850 per year, but we are testing the null Ho: $\mu_{trainees}$ = *\$35,000 per year* (against an alternative of $\mu_{trainees}$ < *\$35,000*), then the probability of making a Type II error is β **= 0.7939**.

This means that if we were testing a hypothesis that the true mean earnings of trainees in the program is \$35,000 per year, there would a probability of 0.7939, or about a 79% chance, that we would fail to reject this null hypothesis, if in fact the true mean earnings of trainees is \$34,850 per year.

- The probability of *not* making a Type II error is called **power**, or the **power of the test**:

  Power = (1-probability of Type II error) = ***1- β***

  ** The closer to 1 the power is, the better the hypothesis test is at detecting a false null hypothesis.

  ** The closer to zero the power is, the hypothesis test is not very good at detecting a false null.

  ** In the previous example, the power of the test **is 0.2061.** In other words, this test has relatively low power: it is not very good at detecting a false null.

Q3: What is the probability of making a Type II error if the **true mean earnings of trainees in the program** is $34,250 per year?

In this situation, the TRUE state of the world is that μ_{trainees} = $34.25 thousand, and the standard error is still $\left(10/\sqrt{3{,}000}\right) = 0.182574$.

But remember we set up a decision rule in Q1 that we would reject Ho if $\overline{X}$ <= 34.7.

We can think about a new, "true" curve of normally distributed $\overline{X}$ values about their true mean of $34.25 thousand.

We know that a Type II error occurs if we do not reject Ho (i.e., if $\overline{X}$ >$34.7 thousand). The probability of that happening is equal to the probability of our drawing a sample with a mean that is greater than 34.7, where that mean is drawn from a population where the *true mean is $34.25 thousand:*

$$t_{calc}: \frac{34.7 - 34.25}{0.182574} = \quad 2.46475$$

Remember that because the sample size is large enough, we can look up 2.46 in the *Z*-table, and see that this corresponds to an area up to Z of 0.9931, and thus an area beyond Z of 0.0069.

So if the true mean earnings of all trainees in the program is $34,250 per year, the probability of making a Type II error when our null hypothesis is Ho: $\mu_{trainees}$ = *$35,000 is:* **β = 0.0069**. The ***power of the test*** is equal to 0.9931.

If we are testing a hypothesis that the true mean earnings of trainees in the program is $35,000 per year, there would be less than a 1% chance that we would fail to reject this null hypothesis, if in fact the true mean earnings of trainees is $34,250 per year.

- The hard thing is that the true value of μ is not known: otherwise, we wouldn't be doing hypothesis tests! Often, researchers will construct a table of powers for various possible values of μ for a given sample size (in this case, n=3,000) e.g.:

IF the True Mean: μ=	**Prob. of Type II error: β**	**Power: *1- β***
35.25		
35.10		
35.00		
34.90		
34.80		
34.70		
34.60		
34.60		
34.50		
34.40		
34.30		
34.20		

- Power is often a concern in evaluation research, when detecting the effects of a program may be crucial. The relevant point here is that they may be concerned about the "effect size" of a program – how large of an effect can their sample design pick up, given a particular sample size?

- And speaking of sample size….

- Instead of a sample of *n=3,000,* what if instead the group sampled *n*=101 trainees? We can repeat the questions above, now using this smaller sample size:

  *Q1**: At what sample mean will the null hypothesis be rejected, using alpha = 0.05 for a one-tail test?

  $$t_{calc}: \frac{\overline{X} - 35}{\left(10 \big/ \sqrt{101}\right)} = -1.660 \quad \Leftrightarrow \quad \overline{X} = 35 - 1.660\left(10 \big/ \sqrt{101}\right) = 33.348$$

  Decision criterion:
  If $\overline{X}$ <= 33.348, Reject Ho. If $\overline{X}$ > 33.348, Fail to reject Ho.

We could now construct a table now using n=201:

IF the True Mean: μ=	**Prob. of Type II error: β**	**Power: *1-* β**
35.25		
35.10		
35.00		
34.90		
34.80		
34.70		
34.60		
34.50		
34.40		
34.30		
34.20		

••

- Let's take another example: Suppose we're interested in the average family income in the U.S.

  First, suppose we draw a random sample of n=3,000 where the sample mean income is $33,000 and the standard deviation is *s* =$16,000.

- We can construct an interval around this sample mean: $\overline{X} \pm 1.96\,\sigma_{\overline{X}}$

  $$= 33 \pm 1.96\ (\frac{16}{\sqrt{3000}}) = 33 \pm 0.57255$$

  [$32,427, $33,573]: We can say with 95% confidence that the population mean is between $32,427 and $33,573 per year.

- Now, suppose that the true population mean family income is μ_{true} = $32,750 and population standard deviation is $16,000.

  If I were to conduct a hypothesis test and use a null hypothesis of μ_{Ho} =$33,000, what is the probability that I would correctly reject that null hypothesis (i.e., 1-β)? What is the probability that I would incorrectly fail to reject that null hypothesis (i.e., β)?

- To calculate these probabilities, we want to map the lower and upper bounds of the confidence interval calculated above onto the distribution of the true mean.

  -- Calculate how many standard errors $32,427 is away from the true mean of $32,750:

  $$\frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}} = \frac{32.427 - 32.750}{16/\sqrt{3{,}000}} = -1.10418 = t \text{ lower}$$

  -- Calculate how many standard errors $33,573 is away from the true mean of $32,750:

  $$\frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}} = \frac{33{,}573 - 32{,}750}{16/\sqrt{3{,}000}} = 2.815816 = t \text{ upper}$$

  -- Look up these *t*-scores in the Z-table (because at this sample size, the *t* approximates the *Z)* to find the area under the curve: The probability of rejecting the null in the left-hand tail of the true curve (centered on μ_{true}=$32,750) is approximately equal to 0.1357 (i.e., the area to the left of Z= -1.10); and the probability of rejecting the null in the right-hand tail of the true curve is approximately equal to 0.0024 (i.e., the area to the right of Z= 2.82)

  -- Thus, the probability of correctly rejecting the null hypothesis of μ_{HO}=$33,000 if the true population mean is $\mu_{true=}$$32,750 = 0.1357 + 0.0024 = **0.1381.** The probability of a **Type II** error under these conditions is (1-0.1381) = **0.8619.**