Hypothesis Testing for Univariate Quantitative Data

Reminder

1. Identify population and parameter you are interested in.
    - Question: What is the average age at which BYU students find out Santa Claus isn't real? Specifically, is the average age at which BYU students find out Santa isn't real older than 8?
    - Parameter: The mean age at which all BYU students find out Santa Claus isn't real. We'll use the Greek letter μ to denote this value.
2. Collect data
    - A convenience sample of 1648 BYU students who are in my class and completed the student survey.
3. Posit a statistical model based on information in the sample
    - Explore the data.
    - Posited a normal population model.
4. Draw inference about the population using your model.

Types of Statistical Inference

3 ways of using sample to make inference about the population:

1. Point Estimation
2. Hypothesis Testing
3. Confidence Intervals (next set of lecture notes)

Point Estimation

Point estimation is a one number estimate of the population parameter.

Our model is

$$Y \sim N(\mu, \sigma)$$

with population parameters μ (a mean) and σ (a standard deviation).

Use sample statistics as estimates of population parameters:

- $\bar{y} \rightarrow \mu$
- $s \rightarrow \sigma$

Law of Large Numbers

How good of an estimate is $\bar{y}$ to μ?

- **Theorem: Law of Large Numbers**

  As the sample size (n) gets bigger, the probability of $\bar{y}$ being close to μ goes up.

Point Estimation

The Good

- It’s simple to wrap your head around

The Bad

- It’s always wrong (unless you sample the whole population)
- If the sample size is small (relative to the size of the population), it could be very wrong.
- If there is a lot of variability in the population, it could be very wrong.

Conclusion

- While useful, point estimation falls short so lets use a different method to draw conclusions about the population.

Hypothesis Testing

Intuition: Someone makes a claim about a parameter that you assume to be true until your data proves otherwise.

Example: A student claims that the average age BYU students learn about Santa is 8. I believe it may be older than that so I collect data to see if sample data is congruent with the student's claim.

Hypothesis Testing

Step 1 - Form null and alternative hypotheses:

- Null hypothesis: The claim about the parameter that we assume true. This is always a claim of *equality*. Denoted by H_0.
- Alternative hypothesis: The counter-claim that coincides with our beliefs about the parameter. Denoted by H_a.

Hypothesis Testing: Step 1

A student claims that the average age BYU students learn about Santa is 8. I believe it may be older than that so I collect data to see if sample data is congruent with the student's claim. Write out the null and alternative hypotheses for this research scenario.

$$H_0 : \mu = 8$$
$$H_a : \mu > 8$$

Note the following:

1. Hypotheses are written in terms of the parameter μ.
2. Null hypotheses are written with equality.
3. Alternative hypotheses can be either greater ($>$), less ($<$) or not equal to ($\neq$) - depending on the counter-claim.

Hypothesis Testing

Step 2 - gather the data and see if our sample data matches (or doesn't match) the null hypothesis.

$$H_0 : \mu = 8$$
$$H_a : \mu > 8$$

- Hypothetical: What if our sample has $\bar{y} = 8.1$? Is 8.1 far enough away from 8 to say $\mu >$ 8? What about $\bar{y} = 8.2$? How about $\bar{y} = 9$?
- What is "different enough" from the null hypothesis to make us think the null is wrong?
- We need (1) a measure of how different our sample statistic is from the hypothesized parameter and (2) is the observed difference reasonable to see when sampling from the population.
- Consider each of these in turn…

Standardizing Revisited

1. Measuring the difference between our sample and a hypothesized value:

Note: We'll use *standardized* differences because its a common scale for ALL problems.

Recall: to *standardize a value* is to calculate the number of standard deviations away from the mean a value is using the formula:

$$z = \frac{Y - \mu}{\sigma} = \frac{\text{value} - \text{mean}}{\text{std. dev.}}$$

Standardizing Revisited

1. Measuring the difference between our sample and a hypothesized value:

Note: We'll use *standardized* differences because its a common scale for ALL problems.

Recall: to *standardize a value* is to calculate the number of standard deviations away from the mean a value is using the formula:

$$z = \frac{Y - \mu}{\sigma} = \frac{\text{value} - \text{mean}}{\text{std. dev.}}$$

To *standardize a statistic* is to calculate the number of standard errors a statistic is away from a hypothesized value using the formula:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\text{sample mean} - \text{hyp. mean}}{\text{std. error}}$$

Key Points

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\text{sample mean} - \text{hyp. mean}}{\text{std. error}} = \frac{8.41 - 8}{2.65/\sqrt{1648}} = 6.326$$

1. Because we are standardizing $\bar{y}$ rather than Y, we use the letter t instead of z to highlight the difference.
2. s is the standard deviation of individuals (how much Y could change from individual to individual), but $s/\sqrt{n}$ is called the standard error of $\bar{y}$ and measures the standard deviation of $\bar{y}$ (how much $\bar{y}$ could change from sample to sample)
3. We interpret t similar to z by saying "our $\bar{y}$ is t standard errors away from the hypothesized value." Example, our sample mean of $\bar{y}$ = 8.41 is t = 6.326 standard errors away from the hypothesized mean.

Sampling Distributions

2. Now that we have a measure of how different our sample is from the hypothesized value, is this difference "different enough" from H_0 for us to no longer believe H_0?

- Definition: A sampling distribution of a statistic is the possible values of that statistic you could observe when you sample from the population AND how often you will observe those values.
- Example: The sampling distribution of t is possible values of t that we could see when sampling from the population and how often we will see them (if we did repeated sampling).

Sampling Distribution of t

What the theorem says:

Theorem: Sampling Distribution of t

If the normal population model is appropriate and the null hypothesis $H_0 : \mu = \mu_0$ *is true, then*

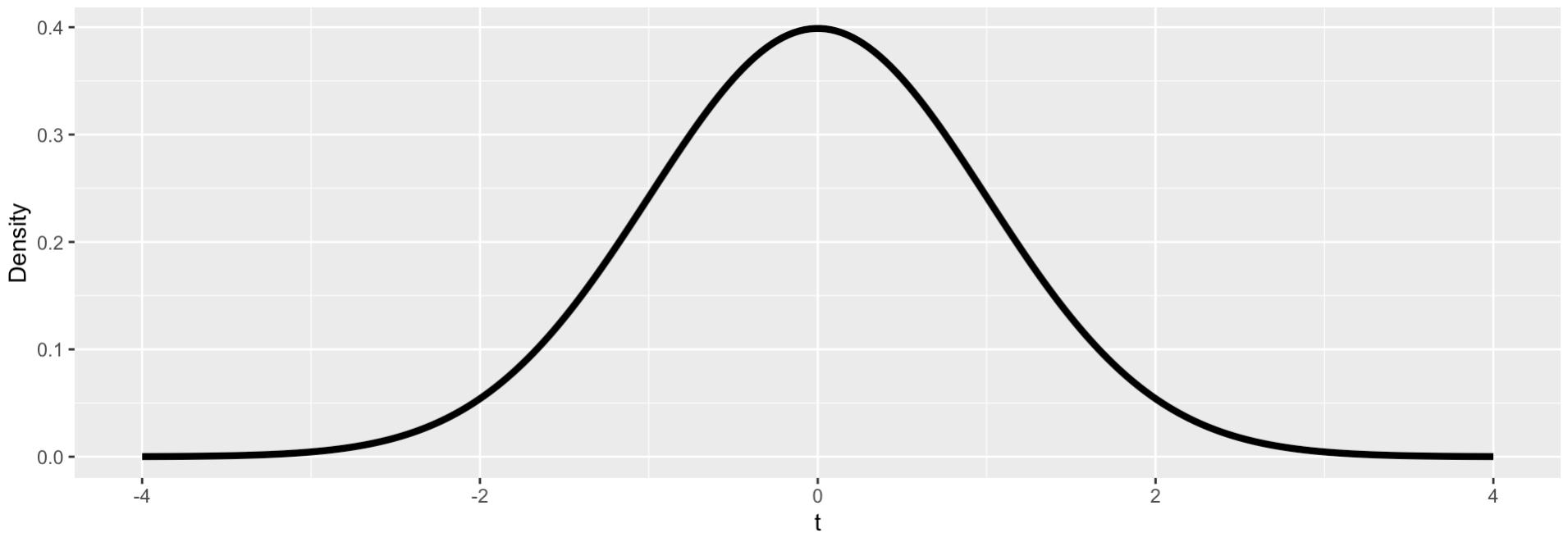
$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

is a standardized statistic and its sampling distribution is a t-distribution with center 0*, spread* 1 *and degrees of freedom* $n - 1$ *where* n *is the size of the sample.*

- Lets think about this theorem more intuitively!

Sampling Distribution of t

If the normal population model is appropriate and we have a claim (the null hypothesis) about the mean μ, then reasonable values of t that we *could* see if that null hypothesis is true follow this distribution:



Sampling Distributions

Sampling Distributions is, admittedly, a hard idea to wrap your head around the first time you see it. So, lets check out this page to illustrate this more clearly.

Illustrating A Sampling Distribution

Key Points to Remember

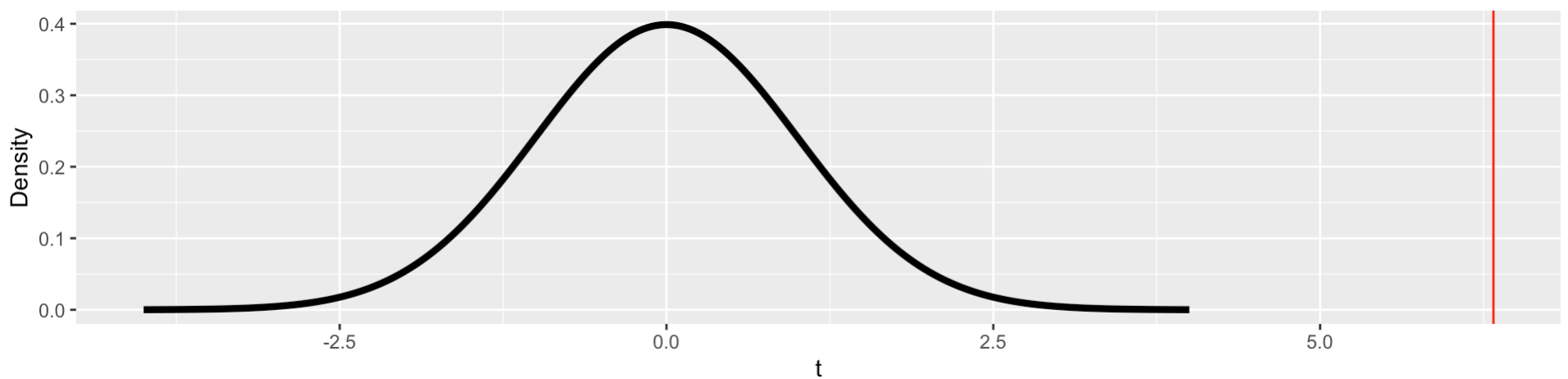
1. The sampling distribution of t tells us what values of t are compatible with H_0 (what values of t we could get if H_0 is true when we sample from the population).
2. The normal population model must be appropriate for this sampling distribution to be true.
3. How can we possibly know what values of t could happen if only ever take one sample?
    - Given a population model, we can do math OR computer simulation!

Hypothesis Testing

Revisiting measuring the difference between our sample and a hypothesize value:

1. Standardized test statistic: the number of standard errors away from the hypothesized value our data is

$$t = \frac{\text{value} - \text{mean}}{\text{std. error of value}} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{8.41 - 8}{2.65/\sqrt{1648}} = 6.326$$

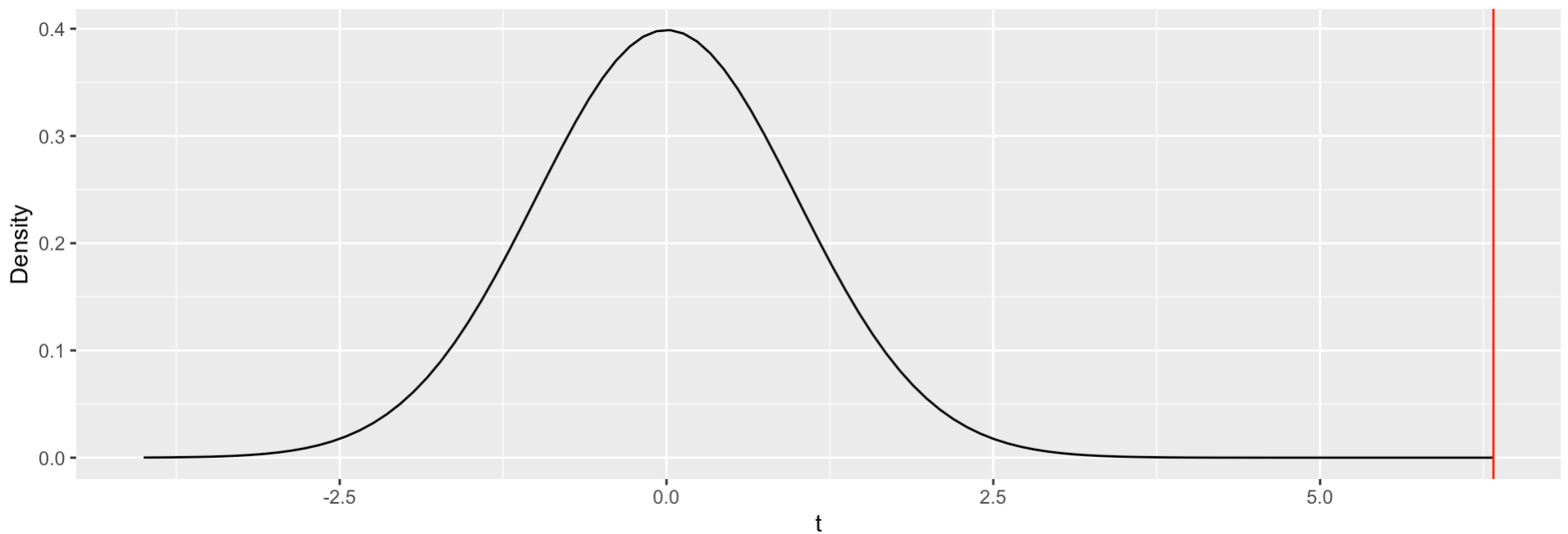


Hypothesis Testing

To measure if our data is consistent with the null hypothesis:

2. p-value: probability of observing our sample result or "more extreme" (as stated by H_a) if the null hypothesis is true.

The p-value in our Santa example is the probability of observing $t > 6.326$ (because $H_a : \mu > 8$) assuming $\mu = 8$ (which is the null hypothesis).



In our example the p-value is 0.

Drawing Conclusions

1. If t is large then we reject H_0 and conclude that the data support H_a.
2. If p-value is small then we reject H_0 and conclude that the data support H_a.
3. How large is "large enough" or how small is "small enough" for us to reject H_0?
    - It depends! We'll revisit this in a minute but, for now, lets just say if the p-value is less than 0.05 we'll reject H_0.
4. In our Santa example, the p-value $= 0$ which is small. We conclude that our data is NOT consistent with the null hypothesis and that the mean is greater than 8.

Example: Chlorine in Swimming Pools

Lets put it all together using another example…

Chlorine is often used to eliminate bacteria and algae by disinfecting (killing) pool water while also oxidizes (chemically destroys) other materials such as dirt and chloramines. Adding the right amount of chlorine is sometimes a tough balancing act, but it is absolutely necessary to maintaining a healthy pool. On the one hand, not enough chlorine will not properly disinfect the water. On the other, too much chlorine can cause sickness and injuries, including rashes, coughing, nose or throat pain, eye irritation and bouts of asthma. A “safe” level of chlorine could be between 2ppm and 3ppm.

A pool technician takes regular measurements of the chlorine content in the water to ensure appropriate chlorine content. The *Chlorine* dataset (on the course analysis app) is a sample of chlorine content at 30 different locations in a public pool. The technician likes to keep the chlorine levels at about 2ppm and thinks the water is about that level. However, after swimming in the pool, you feel a little nauseous and think it might be lower. Carry out a test to see if you are correct.

Example: Chlorine in Swimming Pools

Step 0 - Open up the course analysis app

Step 1 - Write out the hypotheses.

- Let μ represent the mean chlorine content across the pool. Fill in the appropriate hypotheses:

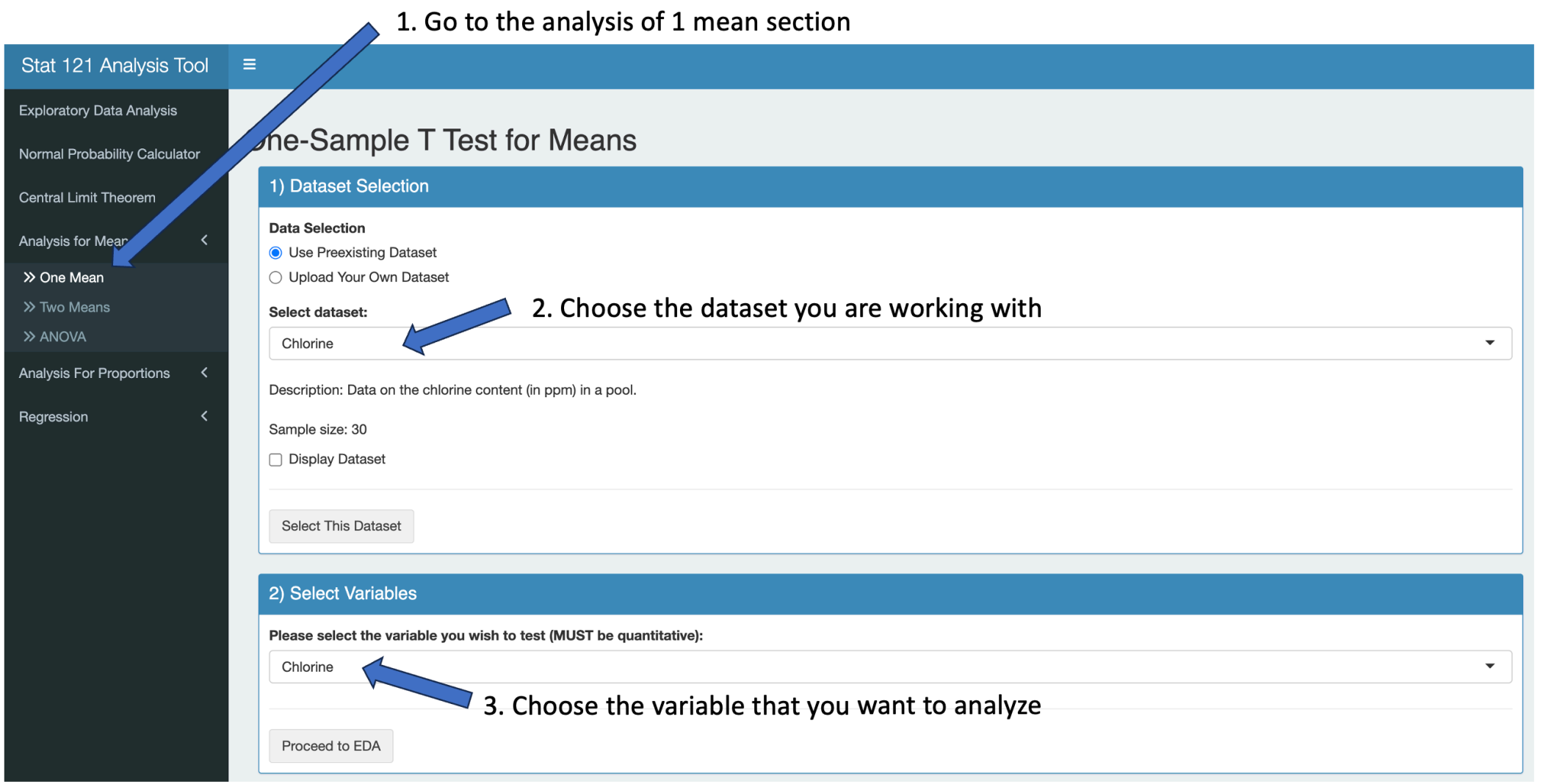
$$H_0 :$$
$$H_a :$$

Step 2 - Collect data and see if the data is consistent with H_0.

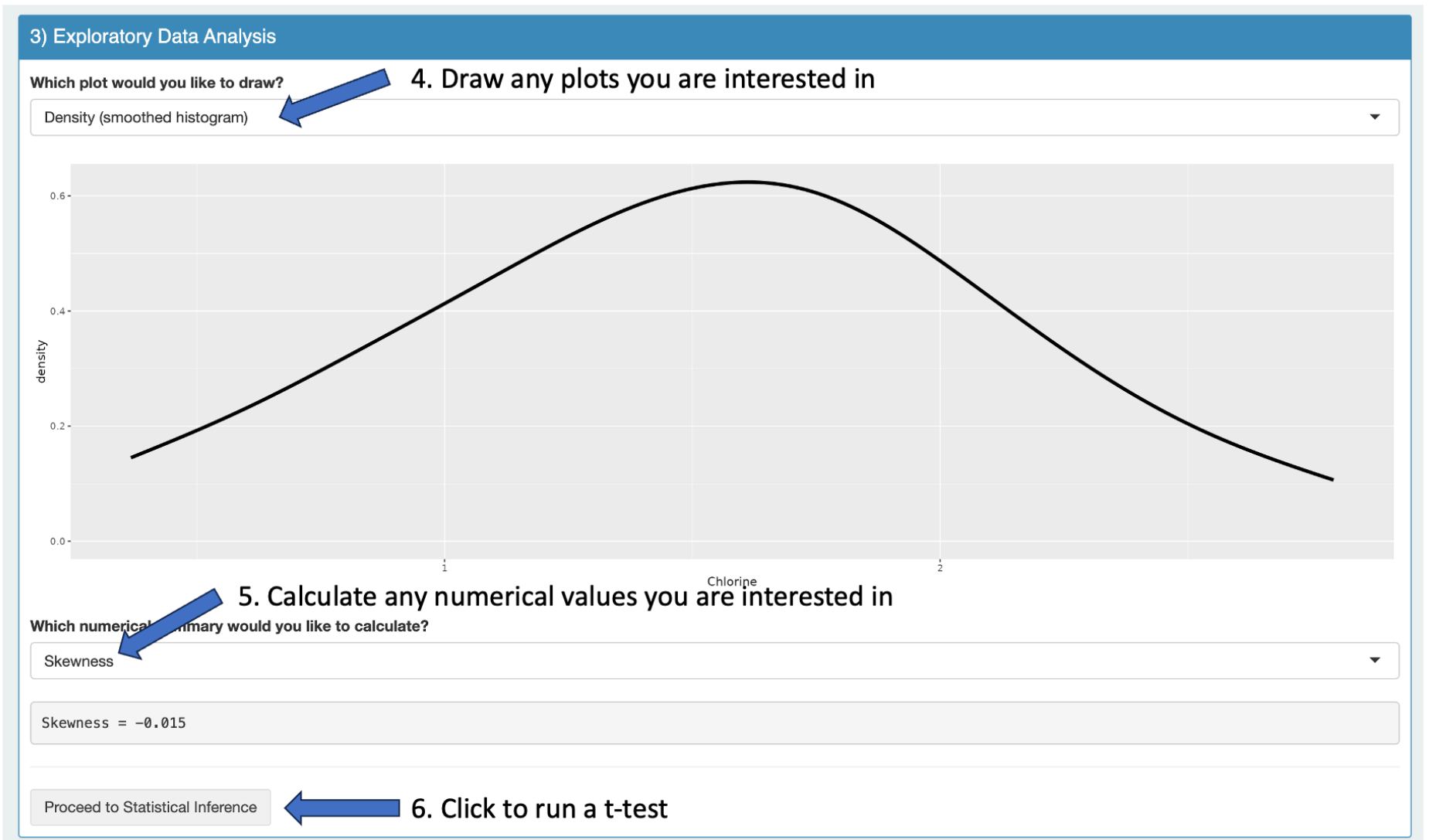
- Make sure to check if the t-distribution is appropriate in doing this.

Step 3 - Draw a conclusion.

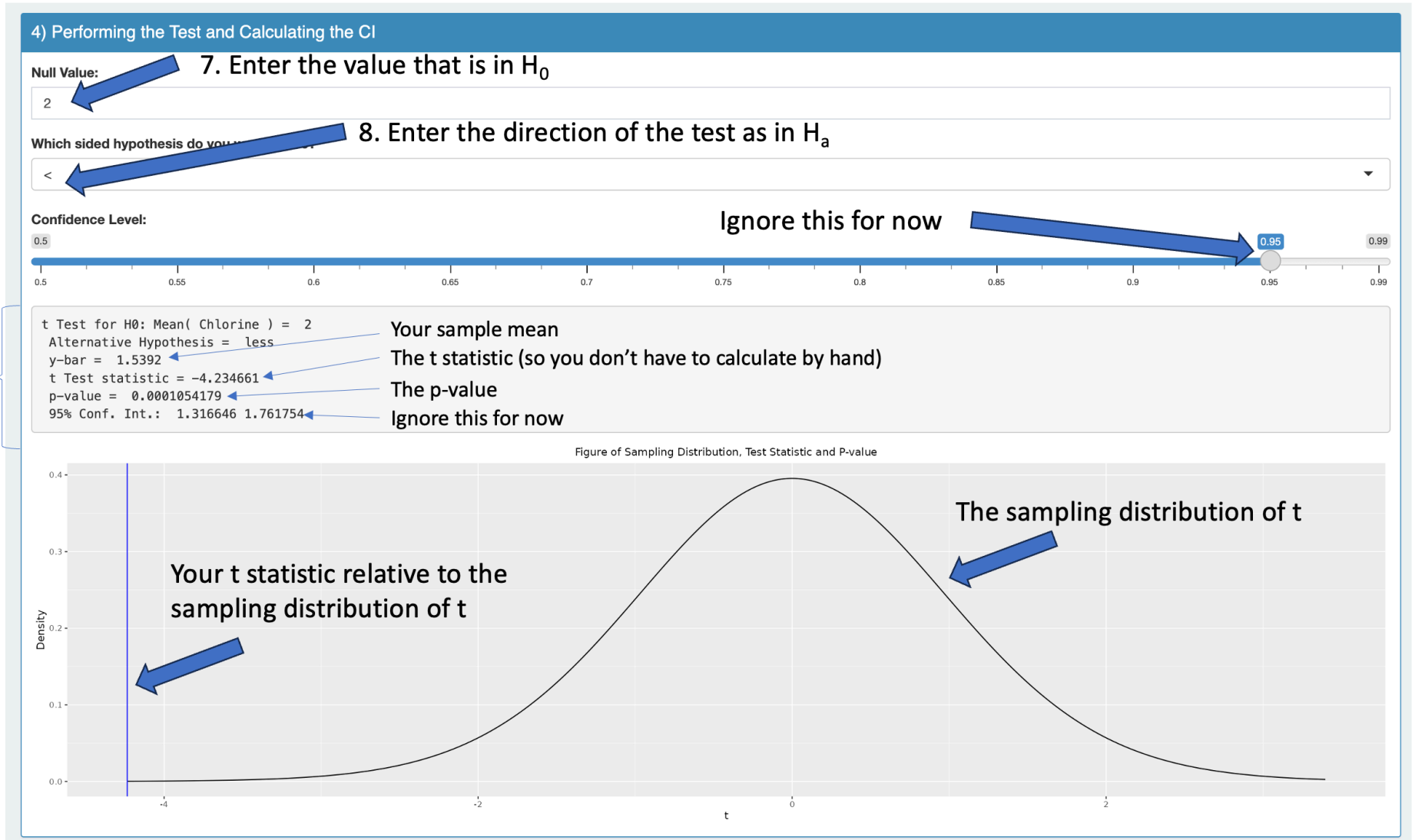
Using the Online Tool



Using the Online Tool



Using the Online Tool



Example: Chlorine in Swimming Pools

Step 0 - Open up the course analysis app

Step 1 - Write out the hypotheses.

- Let μ represent the mean chlorine content across the pool. Fill in the appropriate hypotheses:

$$H_0 : \mu = 2$$
$$H_a : \mu < 2$$

Example: Chlorine in Swimming Pools

Step 2 - Collect data and see if the data is consistent with H_0.

- t-distribution is appropriate.
- $t = -4.23$
- p-value $= 0.0001$

Step 3 - Draw a conclusion.

- Our data is not consistent with the null hypothesis so we conclude that the mean content is less than 2.

Nuances of Hypothesis Testing

1. What do we do if the normal population model doesn't apply to our problem (we need it to so that our results are trustworthy)?

- Key Issue: If the normal population model doesn't apply, then we don't know what values of t are "reasonable" under the null hypothesis and we can't know if our t is consistent with the null or not.
- Solution -

Central Limit Theorem

If the normal population model is not appropriate BUT you have a large sample size, then reasonable values of t under the null is is still approximately a t-distribution.

Nuances of Hypothesis Testing

1. What do we do if the normal population model doesn't apply to our problem (we need it to so that our results are trustworthy)?

Even if the normal model doesn't perfectly apply, using the sampling distribution of t is still OK if we have a large sample size according to the central limit theorem.

- What is a "large" sample size?
  - It depends! If you have strong skewness or big outliers then you need a bigger sample size.
  - $n > 30$ is usually a pretty good "rule-of-thumb" and we'll use this for this class.
  - See the Central Limit Theorem part of the course analysis app

Nuances of Hypothesis Testing

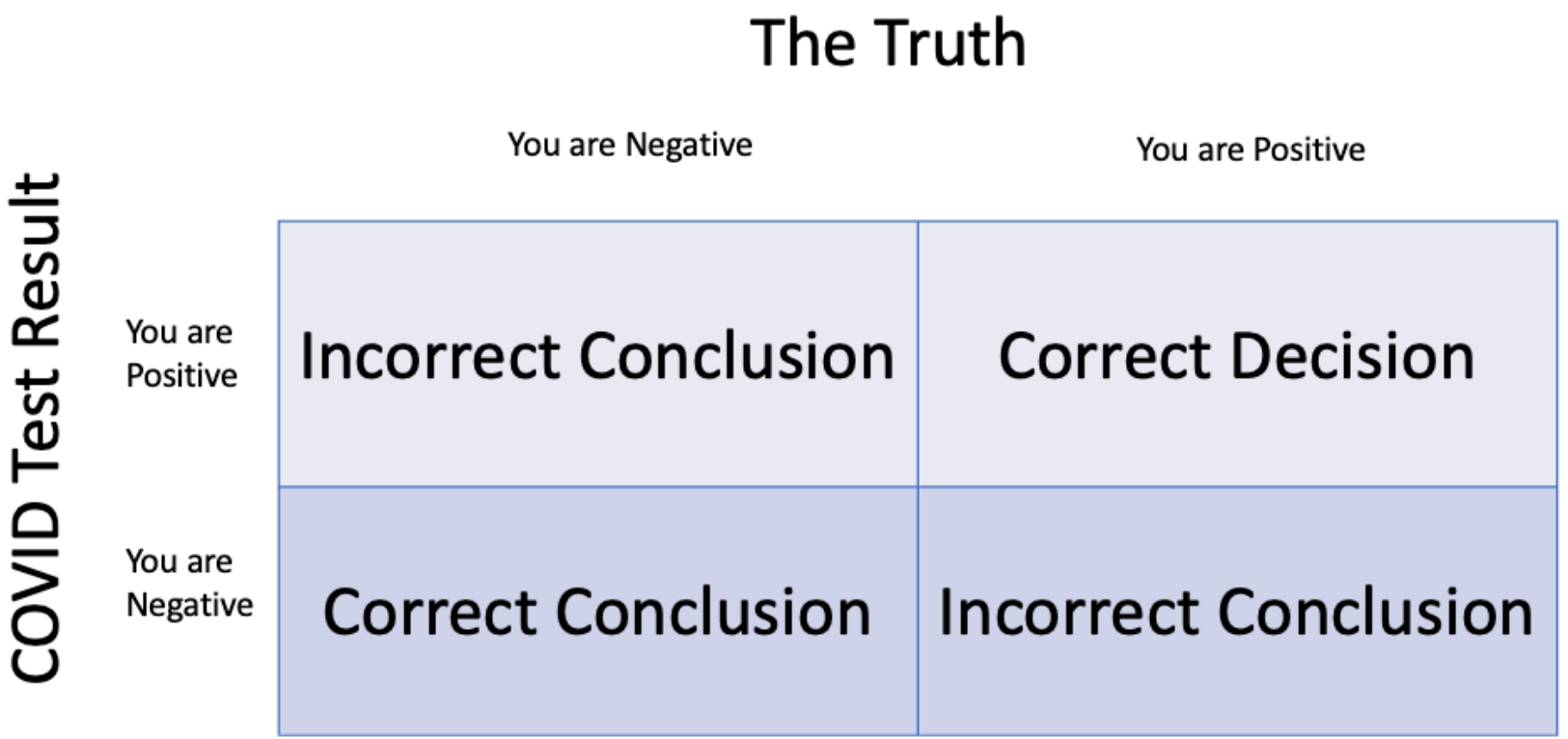
2. What is "small" for a p-value? In other words, what cutoff should I use for my p-value to make conclusions?

- Key Issue: what we define as "small" dictates whether we feel our data is consistent with the null hypothesis or not.
- Solution: Choose a cutoff value for the p-value (called a significance level) between 0 and 1 that takes into account consequences of making a mistake in your conclusions (errors in hypothesis testing).

Errors in Hypothesis Testing

Example: You take a COVID test with

$$H_0 : \text{You are negative}$$
$$H_a : \text{You are positive}$$



Errors in Hypothesis Testing

My Conclusion	The Truth: H_0 is correct	The Truth: H_0 is not correct
H_0 is rejected	Type 1 Error	Correct Decision
H_0 is not rejected	Correct Conclusion	Type 2 Error

- Type 1 Error = rejecting H_0 when you shouldn't
  - What would a Type 1 error be for COVID example?
  - What would it be for Santa example?
- Type 2 Error = failing to reject the H_0 when you should reject H_0
  - What would a Type 2 error be for COVID example?
  - What would it be for Santa example?

Nuances of Hypothesis Testing

2. What is "small" for a p-value? In other words, what cutoff should I use for my p-value to make conclusions?

Significance level (denoted by α) - your p-value "cutoff" and the probability of a Type 1 Error if null is true (a bad thing).

- The choice of cutoff (signficance level) depends on the problem!
- If Type 1 error is a big deal then set α small.
- If Type 1 error is not a big deal then set α bigger.
- Generally, set α so that it is a good balance between making a mistake (Type 1 Error) and being correct.
    - $\alpha = 0.05$ is usually a pretty good balance

Nuances of Hypothesis Testing

3. Drawing conclusions.

We assume the null hypothesis is true in order to use the t-distribution theorem. Therefore:

1. We either "reject" or "fail to reject" the null hypothesis as our conclusion.
2. We can never "accept" the null hypothesis.

Nuances of Hypothesis Testing

4. Statistical vs. practical significance.

Two key definitions:

1. Statistical significance - you reject H_0.
2. Practical significance - the difference between your observed result and the hypothesized result is big enough to matter in real life.

Statistical vs. Practical Significance

Example: Researchers are studying a new weight-loss program. Using a large sample they carry out a test of $H_a : \mu > 0$ where μ is the mean weight loss (in pounds) and conclude that their results were statistically significant at the 0.05 α level. However, the mean weight loss of participants in the study was $\bar{y} = 0.25$ pounds. Most people would say that the results are not practically significant because a six month weight-loss program should yield a mean weight loss much greater than the one observed in this study.

- Lesson: Don't just look at p-value $< \alpha$ to make scientific conclusions. Look at $\bar{y}$ relative to H_0 as well.

Nuances of Hypothesis Testing

5. The effect of sample size

All else being equal, a bigger sample size means:

- t gets bigger (further from zero)
- p-value gets smaller
- easier to reject H_0 (find statistical significance but not necessarily practical significance)

Practice with Hypothesis Testing

I claim that BYU students think BYU was founded in 1900. You think that BYU students know more than that and they know it was founded before 1900. Carry out a hypothesis test to settle the debate.

1. What are the null and alternative hypothesis?
2. What does the sampling distribution of t tell us under the null hypothesis?
3. Is the t-distribution appropriate for this problem? Why or why not?
4. Assuming the t distribution in OK, what is the t test statistic? How do we interpret it?
5. What is the p-value? How do we interpret it?

Practice Continued

6. Am I right or are you (use $\alpha = 0.1$)? Meaning, do we reject H_0 or fail to reject it?
7. What is a Type 1 error in this case?
8. What is a Type 2 error in this case?
9. Should α be lower or higher in this case?
10. Is the result practically significant? Why or why not?

Practice (Answers)

I claim that BYU students think BYU was founded in 1900. You think that BYU students know more than that and they know it was founded before 1900. Carry out a hypothesis test to settle the debate.

1. What are the null and alternative hypothesis? $H_0 : \mu = 1900$ $H_a : \mu < 1900$
2. What does the sampling distribution of t tell us under the null hypothesis? The possible values of t that we could get when we sample IF the null hypothesis is true.
3. Is the t-distribution appropriate for this problem? Why or why not? The t distribution is OK because we have a large sample size AND there is not strong skewness.
4. Assuming the t distribution in OK, what is the t test statistic? How do we interpret it? $t = -21.872$. Our sample mean of 1884.515 is -21.872 standard errors away from the hypothesized value of 1900.
5. What is the p-value? How do we interpret it? p-value = 0; IF the null hypothesis is true, the probability of us seeing $\bar{y} = 1884.515$ or less is 0.

Practice (Answers)

6. Am I right or are you (use $\alpha = 0.1$)? Meaning, do we reject H_0 or fail to reject it? We reject the null and conclude that students know BYU was founded before 1900.
7. What is a Type 1 error in this case? Saying that students know BYU was founded before 1900 when they think it was founded about 1900.
8. What is a Type 2 error in this case? Saying that students think BYU was founded around 1900 when, in fact, they know it was founded before 1900.
9. Should α be lower or higher in this case? It depends!
10. Is the result practically significant? Why or why not? Maybe.

Key Terminology

- Point Estimation
- Hypothesis testing
- Law of large numbers
- Steps in hypothesis testing
- Type 1 Error
- Type 2 Error
- Errors in hypothesis testing
- Test statistics
- p-value
- Significance level
- Central limit theorem
- Statistical significance
- Practical significance
- t-distribution
- Standard error