Hypothesis Testing

Ryan Miller

Outline

- Video #1
  - Null Models
- Video #2
  - p-values
- Video #3
  - An Example
- Video #4
  - Decision Errors
- Video #5
  - p-value Misconceptions

Introduction

- Last week, we focused on the *sample average* (proportion) as a *random variable*
  - Central Limit theorem gave us a Normal model for the *sampling distribution* of the sample average, which we could use to find a **confidence interval estimate**

Introduction

- Last week, we focused on the *sample average* (proportion) as a *random variable*
  - Central Limit theorem gave us a Normal model for the *sampling distribution* of the sample average, which we could use to find a **confidence interval estimate**
- Confidence intervals can be used to *statistically assess* the variability inherent to sample data
  - However, this week we'll learn about *complimentary approach*, known as **hypothesis testing**, that also uses the Central Limit theorem to evaluate the statistical variability in an observed outcome

Infants Choosing Toys (revisited)

- On the first day of class, we discussed an experiment published in *Nature* investigating whether infants have preference towards friendly behavior

Infants Choosing Toys (revisited)

- On the first day of class, we discussed an experiment published in *Nature* investigating whether infants have preference towards friendly behavior
- 16 infants repeatedly watched demonstrations of two scenarios
  - A "helper" toy assisting the main character
  - A "hinderer" toy blocking the main character
- When given the choice, 14 of 16 infants chose the "helper" toy

Infants Choosing Toys (revisited)

Statisticians analyze the results of an experiment by systematically trying to rule out possible explanations:

- Could the majority have chosen the "helper" toy due to a *confounding variable* like the toy's color or shape?

Infants Choosing Toys (revisited)

Statisticians analyze the results of an experiment by systematically trying to rule out possible explanations:

- Could the majority have chosen the "helper" toy due to a *confounding variable* like the toy's color or shape?
  - No, recall that these were *randomly assigned*

Infants Choosing Toys (revisited)

Statisticians analyze the results of an experiment by systematically trying to rule out possible explanations:

- Could the majority have chosen the "helper" toy due to a *confounding variable* like the toy's color or shape?
  - No, recall that these were *randomly assigned*
- Could the majority have chosen the "helper" toy due to some type of *bias*?

Infants Choosing Toys (revisited)

Statisticians analyze the results of an experiment by systematically trying to rule out possible explanations:

- Could the majority have chosen the "helper" toy due to a *confounding variable* like the toy's color or shape?
  - No, recall that these were *randomly assigned*
- Could the majority have chosen the "helper" toy due to some type of *bias*?
  - Probably not, measurement of this outcome is pretty clear-cut

Ideally, statisticians are left to decide between two explanations: random chance/variability or a real relationship

Infants Choosing Toys (revisited)

At this point, the remaining step is to try and rule out random chance as a viable explanation. To do so, statisticians apply the following logic:

1) Identify a suitable **null model** for the outcome of interest
2) Calculate the probability of seeing the outcome that occurred in the sample data if the null model were true
3) If this probability is sufficiently small, rule out random chance as an explanation

Infants Choosing Toys (revisited)

- In this example, a suitable null model would be $p = 0.5$, which implies that each baby's choice is just a coin-flip
  - Under this null model, we can investigate what outcomes could occur by random chance alone

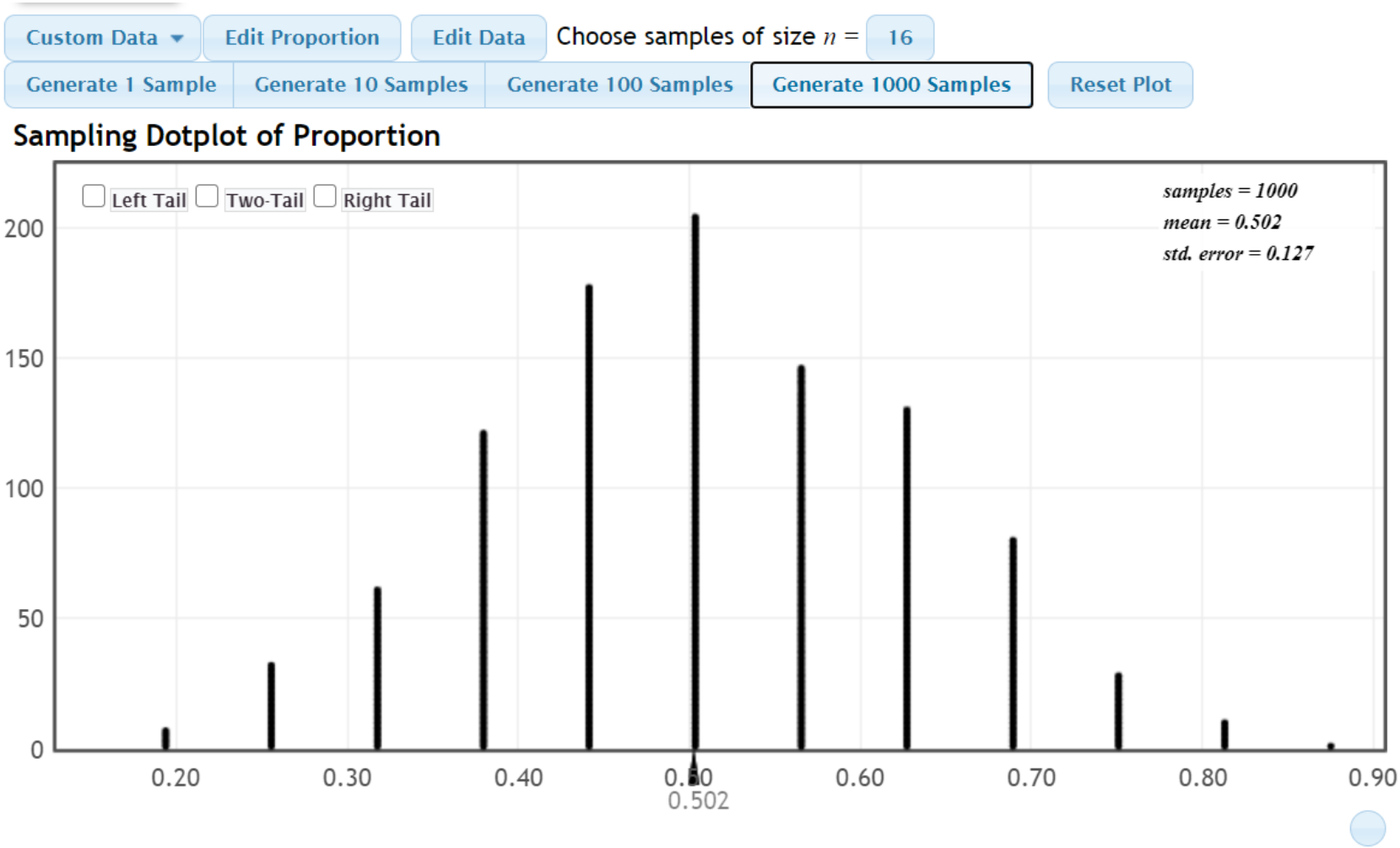
Infants Choosing Toys (revisited)

- In this example, a suitable null model would be $p = 0.5$, which implies that each baby's choice is just a coin-flip
  - Under this null model, we can investigate what outcomes could occur by random chance alone
- One approach is to use simulation by generating many different repititions of 16 coin flips on StatKey

Infants Choosing Toys (revisited)

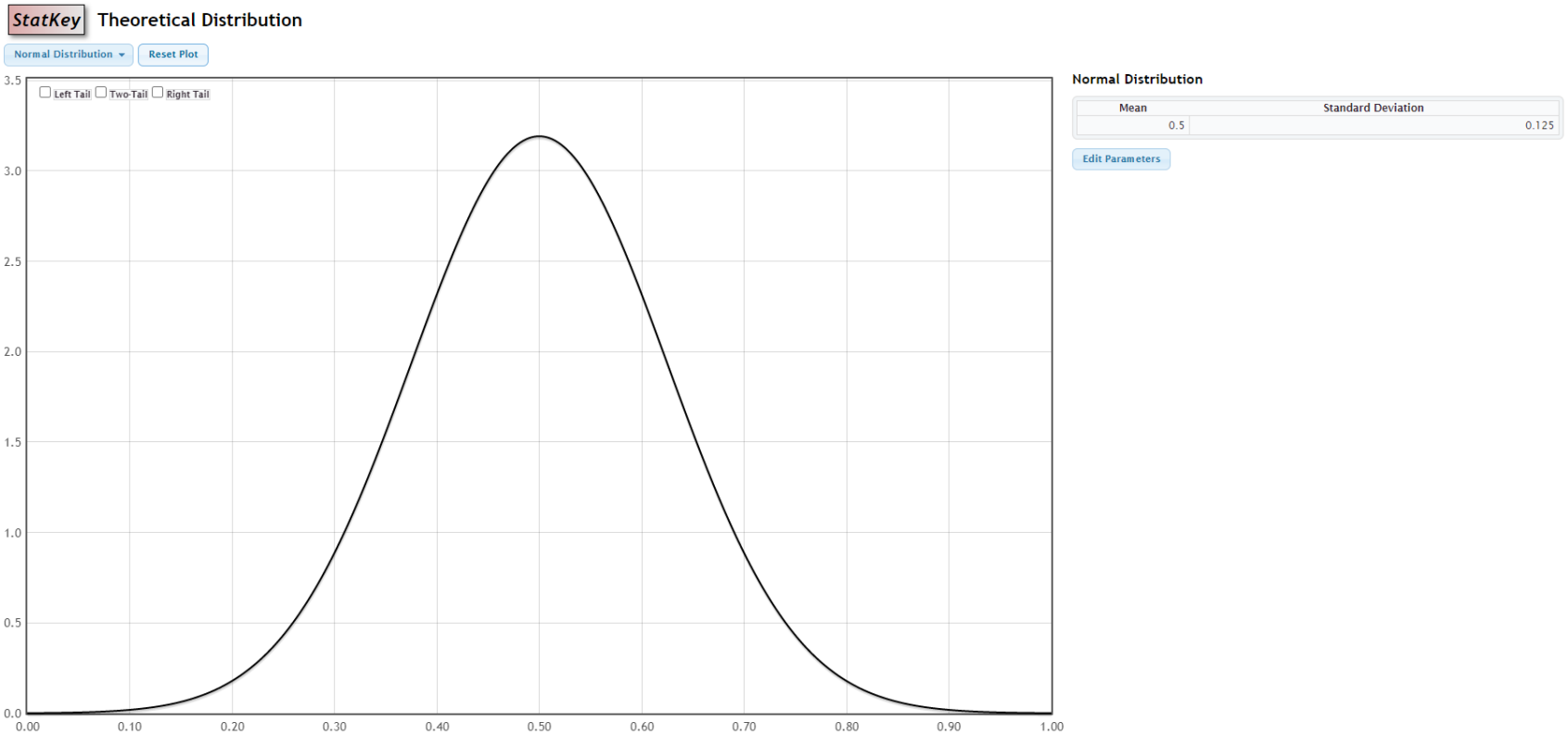
- In this example, a suitable null model would be $p = 0.5$, which implies that each baby's choice is just a coin-flip
  - Under this null model, we can investigate what outcomes could occur by random chance alone
- One approach is to use simulation by generating many different repititions of 16 coin flips on StatKey
- Another approach is to use Central Limit theorem to determine the distribution of sample proportions *under the null model*

The Simulation Approach



The CLT Approach

Note that according to CLT, $SE = \sqrt{.5 * .5/16} = 0.125$



The Null Distribution

- Both yield a distribution known as the **null distribution**, a display of outcomes that could have been observed in the study *if the null model were true*

The Null Distribution

- Both yield a distribution known as the **null distribution**, a display of outcomes that could have been observed in the study *if the null model were true*
- The null distribution is used to *assess the compatibility* of the actual sample data are with the outcomes we'd expect if the null model were true
  - In our example, we observed a sample proportion of $\hat{p} = 14/16 = 0.875$, which appears to be a very unlikely outcome under the null model

The Null Distribution

- Both yield a distribution known as the **null distribution**, a display of outcomes that could have been observed in the study *if the null model were true*
- The null distribution is used to *assess the compatibility* of the actual sample data are with the outcomes we'd expect if the null model were true
  - In our example, we observed a sample proportion of $\hat{p} = 14/16 = 0.875$, which appears to be a very unlikely outcome under the null model
- In the next video, we'll introduce the p-value as a statistical tool to more precisely measure the degree of incompatibility between the observed data and the null model

p-values

- Probability theory allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*

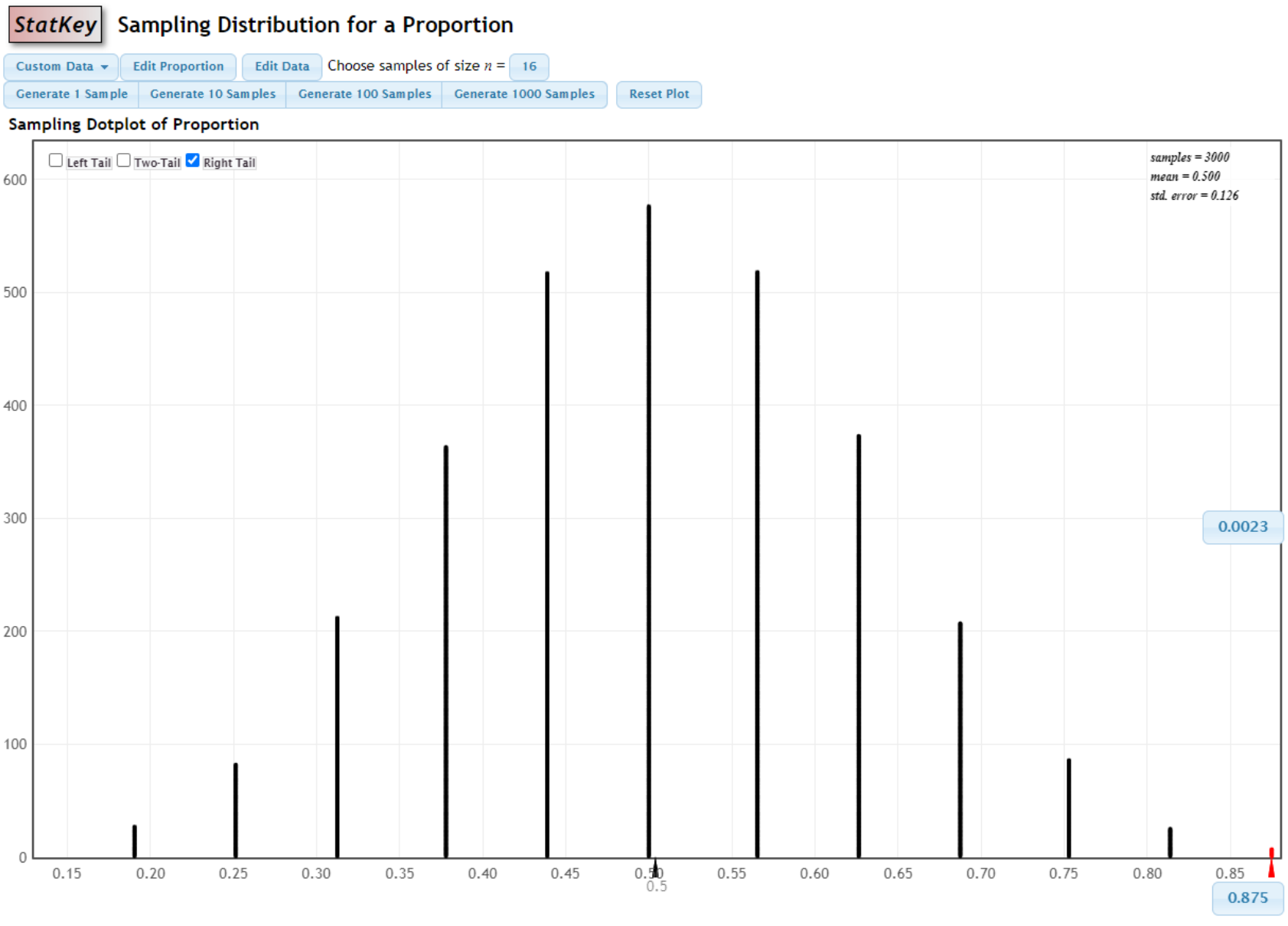
p-values

- Probability theory allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*
- The *smaller* the *p*-value, the *more incompatible* the sample data are with the null model, and thus the stronger the evidence is against random chance as a viable explanation

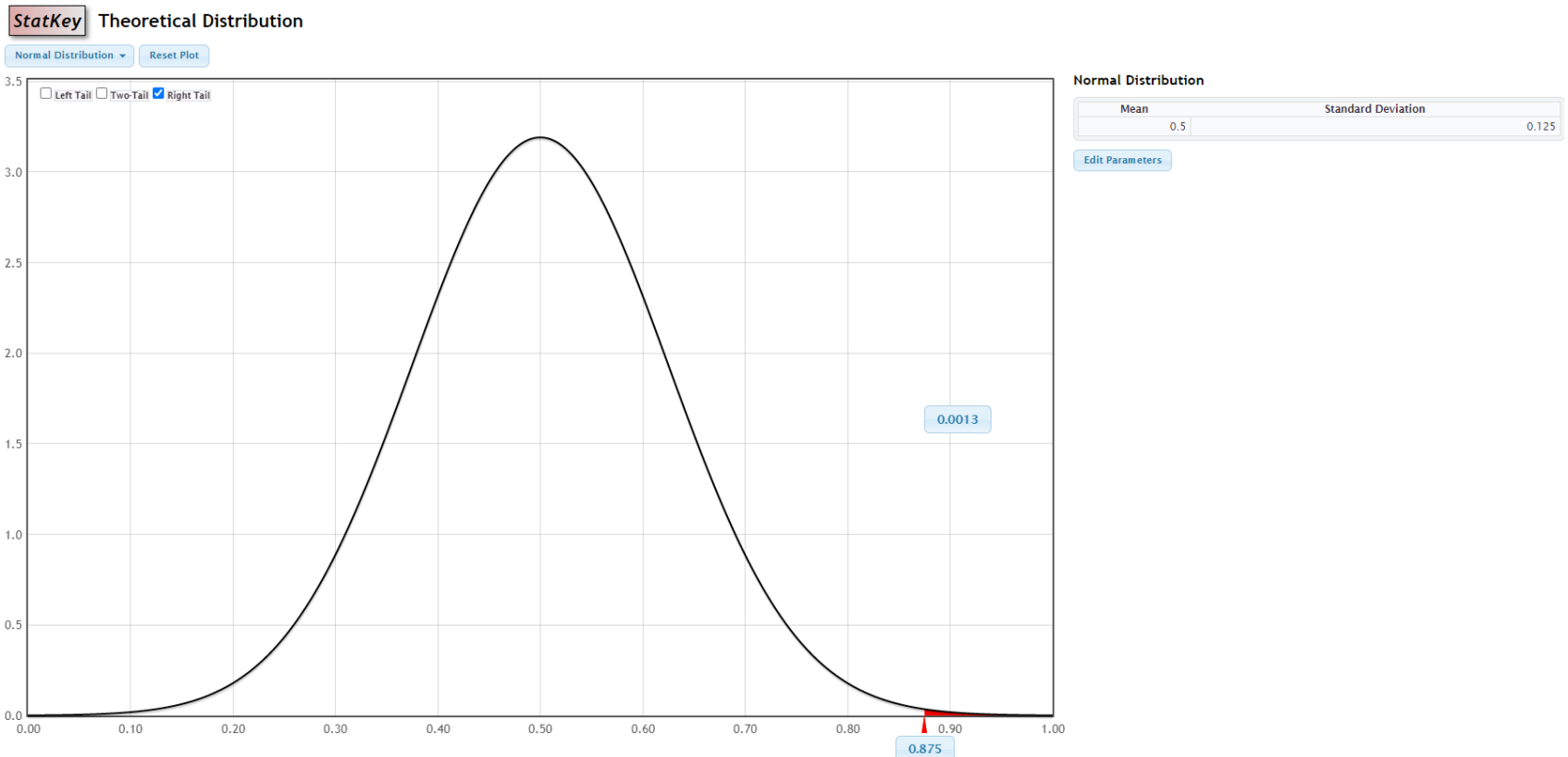
p-values

- Probability theory allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*
- The *smaller* the *p*-value, the *more incompatible* the sample data are with the null model, and thus the stronger the evidence is against random chance as a viable explanation
  - For example, a *p*-value of 0.01 indicates a 1/100 chance of seeing results as extreme as the sample data if the null model were true (high degree of incompatibility)
  - A *p*-value of 0.2 indicates a 1/5 chance of seeing results as extreme as the sample data (low degree of incompatibility)

The Simulated Null Distribution

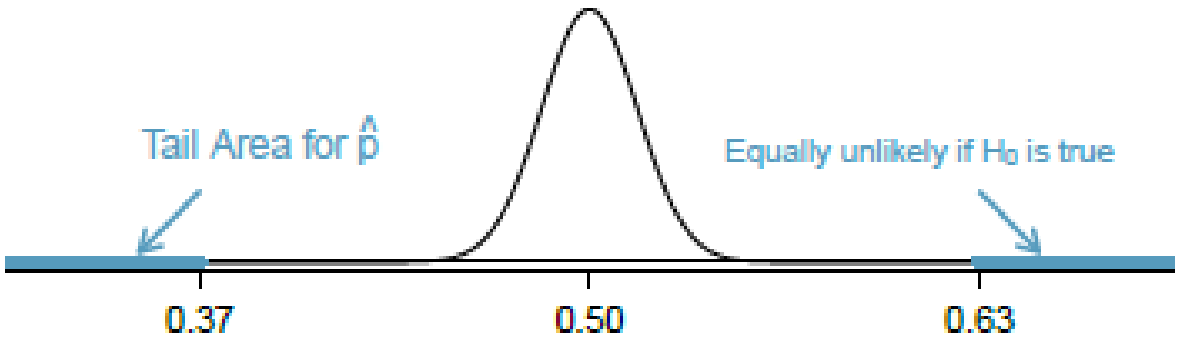


The CLT Null Distribution



Two-sided *p*-values

- The *p*-values (0.0023 and 0.0013) we calculated using the simulated/CLT null distributions aren't actually the ones that a researcher would report in an scientific journal
  - Instead, they are a special case known as a *one-sided* *p*-value, which are rarely used by statisticians
- Instead, *two-sided* *p*-values tend to be prefered:



- The practical implication is that we must *double* the one-sided tail area to find *all* areas of the null distribution that are *as unlikely as the observed outcome* in our sample

Alternative Hypotheses

There are many reason why statisticians prefer two-sided p-values, and one is the notion that any null model should be paired with a *complementary* alternative:

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5$$

Under this setup, an observed sample proportion that is either very large or very small would provide substantial evidence against the null model

p-values as a Measure of Evidence

Ronald Fisher, creator of the *p*-value, and described by his peers as "a genius who almost single-handedly created the foundations of modern statistical science", suggests the following guidelines:

p-value	Evidence against the null
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

p-values as a Measure of Evidence

Ronald Fisher, creator of the *p*-value, and described by his peers as "a genius who almost single-handedly created the foundations of modern statistical science", suggests the following guidelines:

p-value	Evidence against the null
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

- Many scientific fields use $\alpha = 0.05$ as a "significance threshold" for *rejecting* a null hypothesis
- Thus, *p*-values < 0.05 are described as "statistically significant"

Arguments Against "Statistical Significance"

- $p < 0.05$ is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- That is, a p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"

Arguments Against "Statistical Significance"

- $p < 0.05$ is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- That is, a p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"
- When reporting results you should always include the p-value itself, not just whether it met some arbitrary significance threshold
  - Imagine your weather app only telling you: "it's cold" or "it's not cold"
  - This is bad because "cold" is subjective, it's better to provide the temperature and let you decide for yourself

Closing Remarks

- The null distribution is an intermediate step in calculating the p-value, or the probability of observing an outcome at least as unusual that seen in the sample data if the null model were true

Closing Remarks

- The null distribution is an intermediate step in calculating the *p*-value, or the probability of observing an outcome at least as unusual that seen in the sample data if the null model were true
  - We will almost always report *two-sided* *p*-values, which involve extreme outcomes on both ends of the null distribution
  - The two-sided *p*-value is found by multiplying the relevant one-sided tail-area by 2

Closing Remarks

- The null distribution is an intermediate step in calculating the *p*-value, or the probability of observing an outcome at least as unusual that seen in the sample data if the null model were true
  - We will almost always report *two-sided* *p*-values, which involve extreme outcomes on both ends of the null distribution
  - The two-sided *p*-value is found by multiplying the relevant one-sided tail-area by 2
- You should think of the *p*-value as a measure of incompatibility between the null model and the observed data
  - A smaller *p*-value suggests lower compatibility (ie: it's likely that the null model is wrong)

How to Conduct a Hypothesis Test

The key steps in any hypothesis test are as follows:

1) State the null and alternative hypotheses
2) Find the null distribution (the distribution of possible outcomes that could occur if the null hypothesis were true)
3) Using the null distribution, locate the estimate observed in the sample data to find the p-value
4) Use the p-value to make a conclusion

We'll now go through a full example of this process

Example

According to Wikipedia, babies born 15-weeks prematurely have a 70% survival rate. A recent study of babies born at Johns Hopkins University found that 31 of 39 (79.5%) babies born 15-weeks early survived. Does this study provide statistically compelling evidence that Wikipedia's claim is wrong?

Step 1 - State the Hypotheses

In order to evaluate whether the sample data are incompatible with Wikipedia's claim, we begin by assuming that Wikipedia's claim is true:

$$H_0 : p = 0.7$$

$$H_A : p \neq 0.7$$

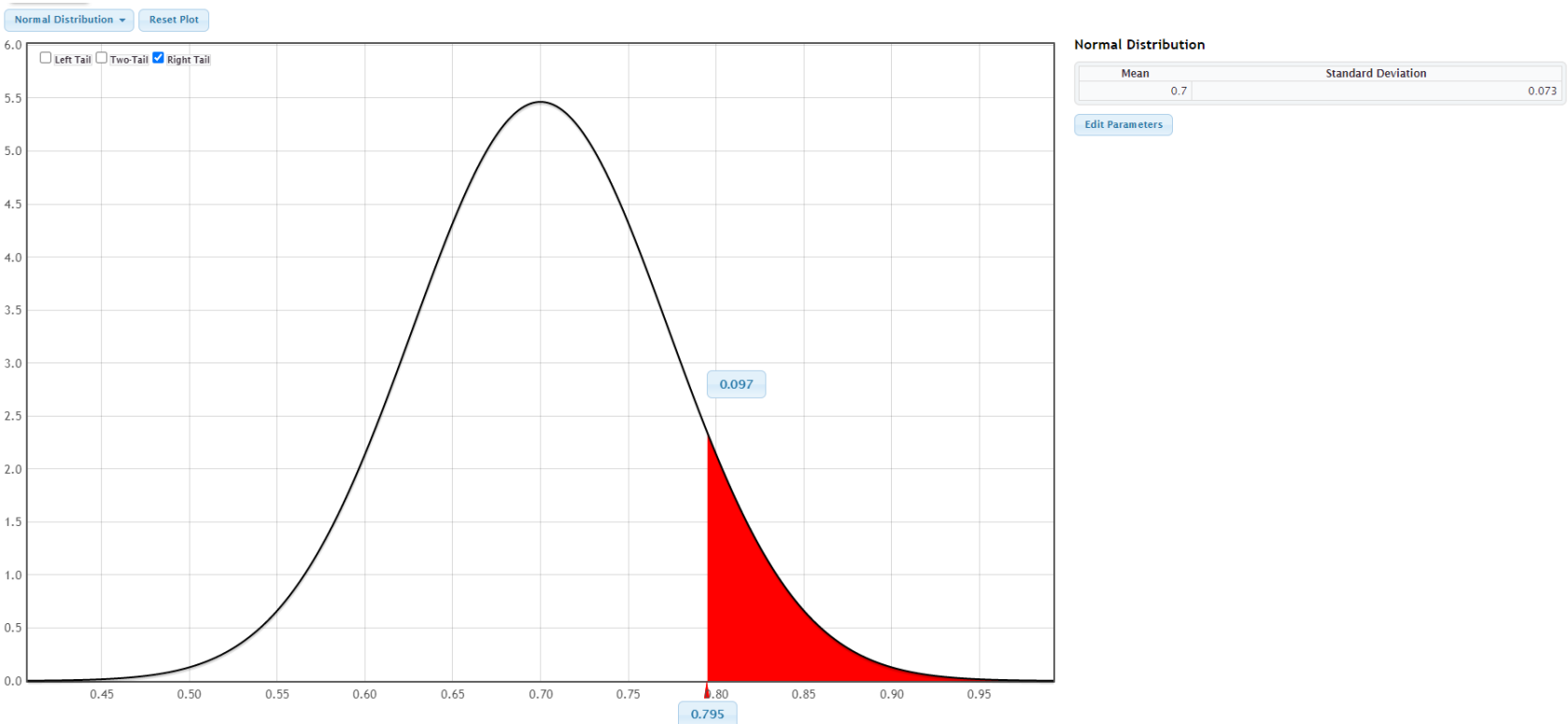
Step 2 - Find the Null Distribution

Central Limit theorem suggests we might use a Normal model to tell us which sample proportions we might expect to see if the $H_0 : p = 0.7$ is true:

$$\hat{p} \sim N\left(0.7, \sqrt{\frac{0.7(1-0.7)}{39}}\right)$$

In words, the expected value of our null model is 0.7 and the standard error is $\sqrt{\frac{0.7(1-0.7)}{39}} = 0.073$

Step 3 - Locate the Sample Estimate to find the *p*-value



Remember we need to double this area to get a two-sided *p*-value of 0.194

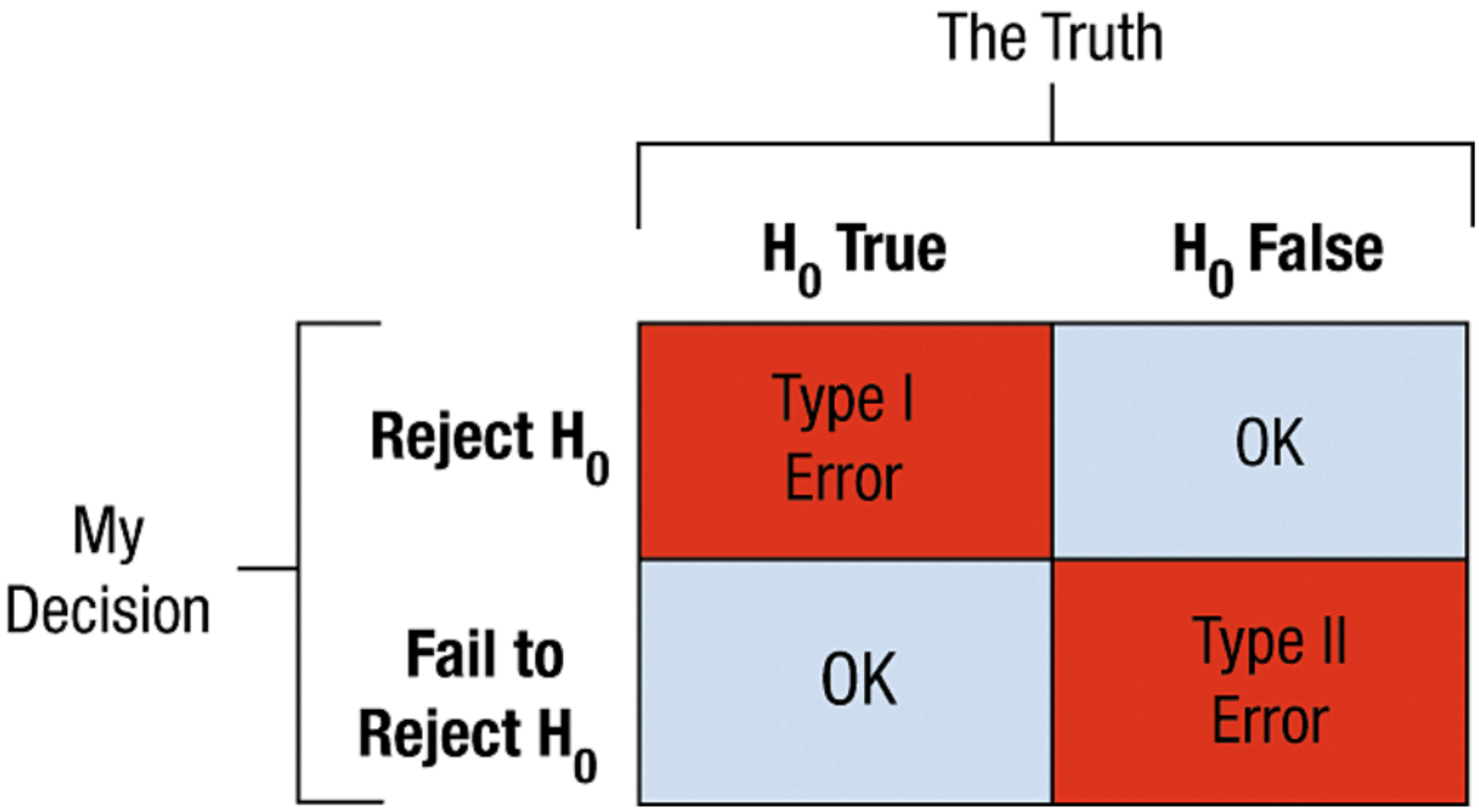
Step 4 - Make a Conclusion

Based upon the p-value of 0.194, there is a roughly 1 in 5 chance of seeing a sample proportion like this one (31 of 39) if Wikipedia's claim of 70% survival is correct. Therefore, we conclude these data do not provide sufficient evidence to disprove Wikipedia's claim.

Decision Thresholds

- The final step in a hypothesis test is to use the *p*-value to make a decision
  - Many scientific fields use $\alpha = 0.05$ as a "significance threshold" for *rejecting* a null hypothesis
- More generally, we could let α denote a *decision threshold*
  - If *p*-value $\leq \alpha$ we'd *reject* H_0 in favor of the alternative
  - If *p*-value $> \alpha$ we'd decide there *isn't enough evidence* to reject H_0

Decision Errors



Example #1

- Consider a jury trial for Person A
  - H_0 : Person A is not guilty vs. H_A : Person A is guilty
- In words, what would a Type I and Type II error represent?

Example #1 (solution)

- A Type I error would mean that Person A is not guilty (H_0 is true), but the jury decides they are guilty (reject H_0)
- A Type II error would mean that Person A is guilty (H_0 is false), but the jury decides they are not guilty (not enough evidence to reject H_0)

Example #2

- Consider a clinical trial evaluating a new medication for disease B
  - H_0 : The medication doesn't cure disease B vs. H_A : The medication cures disease B
- In words, what would a Type I and Type II error represent?

Example #2 (solution)

- A Type I error would mean the new medication is not effective (H_0 is true), but the study concludes it cures disease B (reject H_0)
- A Type II error would mean the new medication cures disease B (H_0 is false), but the study concludes it is ineffective (not enough evidence to reject H_0)

Error Rates

- By design, using a *decision threshold* of α means the probability of making a Type I error (when H_0 is true) is α
  - If $\alpha = 0.05$, we'd expect a Type I to occur in 5% of tests where the null hypothesis is true

Error Rates

- By design, using a *decision threshold* of α means the probability of making a Type I error (when H_0 is true) is α
  - If $\alpha = 0.05$, we'd expect a Type I to occur in 5% of tests where the null hypothesis is true
- If we wanted to reduce the rate of Type I errors, we might consider a more stringent threshold of $\alpha = 0.01$
  - This comes at the expense of making more Type II errors (we've made it harder to reject H_0, which includes scenarios when H_0 is false)

Error Rates

- By design, using a *decision threshold* of α means the probability of making a Type I error (when H_0 is true) is α
  - If $\alpha = 0.05$, we'd expect a Type I to occur in 5% of tests where the null hypothesis is true
- If we wanted to reduce the rate of Type I errors, we might consider a more stringent threshold of $\alpha = 0.01$
  - This comes at the expense of making more Type II errors (we've made it harder to reject H_0, which includes scenarios when H_0 is false)

Error Rates and Study Replication

- The *decision threshold* of $\alpha = 0.05$ is very widely used because it is thought to balance the rates of Type I and Type II errors
- While we'd expect a Type I error in 5% of studies, if others are repeating the same research the chance of two independent studies both resulting in a Type I error is very small
  - 0.05*0.05 $= 0.0025$ (or $1/400$)

Error Rates and Multiple Tests

- It is important to draw a distinction between testing the same hypothesis in multiple different studies and testing multiple hypotheses in the same study
  - In the later scenario, a *decision threshold* of $\alpha = 0.05$ can be problematic

Error Rates and Multiple Tests

- It is important to draw a distinction between testing the same hypothesis in multiple different studies and testing multiple hypotheses in the same study
  - In the later scenario, a *decision threshold* of $\alpha = 0.05$ can be problematic
- As an example, consider a genetic association study testing differences in the expression levels of 7129 genes across two patients with two different types of leukemia
  - This single study involves 7129 different hypothesis tests
  - If all of the tests used $\alpha = 0.05$, and none of the genes were related to the type of leukemia, we'd expect to see 356 "statistically significant" genes

Error Rates and Multiple Tests

- It is important to draw a distinction between testing the same hypothesis in multiple different studies and testing multiple hypotheses in the same study
  - In the later scenario, a *decision threshold* of $\alpha = 0.05$ can be problematic
- As an example, consider a genetic association study testing differences in the expression levels of 7129 genes across two patients with two different types of leukemia
  - This single study involves 7129 different hypothesis tests
  - If all of the tests used $\alpha = 0.05$, and none of the genes were related to the type of leukemia, we'd expect to see 356 "statistically significant" genes
- As you'd expect, it is wise to use a more stringent significance threshold in this type of study (one involving many different related hypotheses)

The Bonferroni Adjustment

- A simple fix is to divide the desired Type I error rate, α, by the number of hypothesis tests, h, to get an *adjusted significance threshold* ($\alpha^* = \alpha/h$)

The Bonferroni Adjustment

- A simple fix is to divide the desired Type I error rate, α, by the number of hypothesis tests, h, to get an *adjusted significance threshold* ($\alpha^* = \alpha/h$)
- This procedure is known as the "Bonferroni Adjustment" and it will limit the *entire study's* Type I error rate to α (known as the *family-wise error rate*)
  - For the Leukemia example (involving 7129 different genes), we might use an adjusted significance threshold of $\alpha^* = 0.05/7139 = 0.00007$ if we wanted to limit the probability of making at least one Type I error to 5%

Closing Remarks

- Hypothesis testing is a decision making tool, but it isn't perfect
  - Type I errors occur when the null hypothesis is *true*, but the data say to *reject it*
  - Type II errors occur when the null hypothesis is *false*, but the data *do not provide enough evidence to reject it*

Closing Remarks

- Hypothesis testing is a decision making tool, but it isn't perfect
  - Type I errors occur when the null hypothesis is *true*, but the data say to *reject it*
  - Type II errors occur when the null hypothesis is *false*, but the data *do not provide enough evidence to reject it*
- The Type I error rate is controlled by the *significance threshold*, α
  - There is a trade-off between using more/less stringent values of α (lowering α will reduce the chances of making a Type I error but increase the likelihood of making a Type II error)

Closing Remarks

- Hypothesis testing is a decision making tool, but it isn't perfect
  - Type I errors occur when the null hypothesis is *true*, but the data say to *reject it*
  - Type II errors occur when the null hypothesis is *false*, but the data *do not provide enough evidence to reject it*
- The Type I error rate is controlled by the *significance threshold*, α
  - There is a trade-off between using more/less stringent values of α (lowering α will reduce the chances of making a Type I error but increase the likelihood of making a Type II error)
- Performing a large number of hypothesis tests within a single study can be problematic

Hypothesis Testing Misconceptions

- We've now introduced the general framework for hypothesis testing, which is based upon using the p-value as a measure of evidence against a null hypothesis
- Unfortunately, p-values are frequently misunderstood and are often used incorrectly
  - The misuse of p-values has become such a problem that *Basic and Applied Social Psychology* has banned their use (source)

Hypothesis Testing Misconceptions

- We've now introduced the general framework for hypothesis testing, which is based upon using the p-value as a measure of evidence against a null hypothesis
- Unfortunately, p-values are frequently misunderstood and are often used incorrectly
  - The misuse of p-values has become such a problem that *Basic and Applied Social Psychology* has banned their use (source)
- However, it is my belief that p-values, if used properly, are an important statistical tool
  - But in order to be used properly, you need to be aware of the mistakes that others are making

Mistake #1 - "Proving" the Null Model

- Let's consider a silly example where the NBA's Steph Curry and myself compete by each shooting 5 three-point shots
  - I make 2 of 5, and Steph makes 5 of 5

Mistake #1 - "Proving" the Null Model

- Let's consider a silly example where the NBA's Steph Curry and myself compete by each shooting 5 three-point shots
  - I make 2 of 5, and Steph makes 5 of 5
- We might use a hypothesis test to evaluate the null hypothesis that we're both equally good three-point shooters (ie: $H_0 : p_{\text{Miller}} = p_{\text{Curry}}$)
  - The p-value for this scenario is 0.17
  - Does that mean we are equally good 3-pt shooters?

Mistake #1 - "Proving" the Null Model

- The answer is a resounding "no", Steph Curry and I are *not equally good* three-point shooters!

Mistake #1 - "Proving" the Null Model

- The answer is a resounding "no", Steph Curry and I are *not equally good* three-point shooters!
- The p-value measures the strength of evidence against the null hypothesis
  - In a sample involving only 5 shots, there isn't enough data to provide sufficient evidence against the null hypothesis
  - A lack of evidence does not mean the null hypothesis is likely to be correct

Mistake #1 - A Non-hypothetical Example

- It might seem professionals would easily avoid the mistake highlighted in that silly Steph Curry example, but unfortunately it happens quite often

Mistake #1 - A Non-hypothetical Example

- It might seem professionals would easily avoid the mistake highlighted in that silly Steph Curry example, but unfortunately it happens quite often
- In 2006, the Woman's Health Initiative evaluated the relationship between low-fat diets and reduced risk of breast cancer risk and found a p-value of 0.07

Mistake #1 - A Non-hypothetical Example

- It might seem professionals would easily avoid the mistake highlighted in that silly Steph Curry example, but unfortunately it happens quite often
- In 2006, the Woman's Health Initiative evaluated the relationship between low-fat diets and reduced risk of breast cancer risk and found a p-value of 0.07
  - The NY Times ran the headline: "Study Finds Lowfat Diets Won't Stop Cancer or Heart Disease"
  - The article described the study's results as: "The death knell for the belief that reducing the percentage of fat in the diet is important for health"
- In reality, these results simply indicates insufficient evidence linking dietary fat and breast cancer, it's very possible there is a small benefit but we cannot rule out random chance

Comments - "Proving" the Null Hypothesis

- Hypothesis testing is not designed to "prove" a null hypothesis, so you should never use it to try and do so
  - The null hypothesis is intended to be a "straw man" that researchers want to "knock down"

Comments - "Proving" the Null Hypothesis

- Hypothesis testing is not designed to "prove" a null hypothesis, so you should never use it to try and do so
  - The null hypothesis is intended to be a "straw man" that researchers want to "knock down"
- The closest thing to "proving" a null hypothesis is a *very narrow confidence interval around the null value*
  - This interval estimate would suggest the only plausible values for the parameter of interest are extremely close to what the null hypothesis suggests

Comments - Confidence Intervals vs. Hypothesis Tests

- Confidence intervals and hypothesis tests are two complementary tools for evaluating the variability in sample data
  - A confidence interval provides a range of plausible estimates for a population characteristic
  - A hypothesis test considers a null model for the population characteristic and measures how compatible the sample data are with this model

Comments - Confidence Intervals vs. Hypothesis Tests

- Confidence intervals and hypothesis tests are two complementary tools for evaluating the variability in sample data
  - A confidence interval provides a range of plausible estimates for a population characteristic
  - A hypothesis test considers a null model for the population characteristic and measures how compatible the sample data are with this model
- Consider $H_0 : p = 0.5$, and suppose our sample produces a 95% CI estimate for p of (0.53, 0.63)
  - This interval says that it is *not plausible* that $p = 0.5$, so we expect the hypothesis test to have a $p - value < 0.05$ (based upon the 95% confidence level)

Comments - Confidence Intervals vs. Hypothesis Tests

- Confidence intervals and hypothesis tests are two complementary tools for evaluating the variability in sample data
  - A confidence interval provides a range of plausible estimates for a population characteristic
  - A hypothesis test considers a null model for the population characteristic and measures how compatible the sample data are with this model
- Consider $H_0 : p = 0.5$, and suppose our sample produces a 95% CI estimate for p of (0.53, 0.63)
  - This interval says that it is *not plausible* that $p = 0.5$, so we expect the hypothesis test to have a $p - value < 0.05$ (based upon the 95% confidence level)
- Again consider $H_0 : p = 0.5$, but now suppose a different sample leads to a sample proportion of $\hat{p} = 0.53$ and a p-value of 0.11, we'd expect the 95% confidence interval estimate from this sample to suggest that 0.5 *is* a plausible value (ie: the 95% CI would contain 0.5)

Mistake #2 - Clinical vs. Statistical Significance

- Confidence intervals and hypothesis tests lead to similar conclusions, but provide complementary information
- In the 1980s, *AstraZeneca* developed *Prilosec*, a very successful medication for healing erosive esophagitis (heart burn)
  - In the 2001, just before the company's patent on *Prilosec* was about to expire, *AstraZeneca* developed a new drug, *Nexium*

Mistake #2 - Clinical vs. Statistical Significance

- Confidence intervals and hypothesis tests lead to similar conclusions, but provide complementary information
- In the 1980s, *AstraZeneca* developed *Prilosec*, a very successful medication for healing erosive esophagitis (heart burn)
  - In the 2001, just before the company's patent on *Prilosec* was about to expire, *AstraZeneca* developed a new drug, *Nexium*
- To get *Nexium* approved by the FDA, *AstraZeneca* conducted a large randomized experiment comparing it to *Prilosec*
  - The experiment resulted in a p-value < 0.001, well below significance threshold of $\alpha = 0.05$ used by the FDA
- After its approval, *AstraZeneva* spent millions of dollars marketing *Nexium* and it soon became one of the top selling drugs in the world, leading to billions in profits

Mistake #2 - Clinical vs. Statistical Significance

- While p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)

Mistake #2 - Clinical vs. Statistical Significance

- While p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)
- Further, the active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting too far into the chemistry (not my area of expertise), Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers

Mistake #2 - Clinical vs. Statistical Significance

- While p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)
- Further, the active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting too far into the chemistry (not my area of expertise), Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers
- Critics of the pharmaceutical industry argue the results of the Nexium study were not **clinically significant**, meaning the differences in the two drugs aren't substantial enough to be influencing clinical practices

Mistake #2 - Clinical vs. Statistical Significance

- A very small p-value does not mean an observed relationship is large, meaningful, or important
  - The p-value is a tool for evaluating how plausible it is for an observed relationship to be explained by random chance

Mistake #2 - Clinical vs. Statistical Significance

- A very small p-value does not mean an observed relationship is large, meaningful, or important
  - The p-value is a tool for evaluating how plausible it is for an observed relationship to be explained by random chance
- With enough data, it is possible to show small/inconsequential relationships are unlikely to occur by chance alone
  - This doesn't mean those relationships have any real-world significance
  - Reporting confidence intervals along side hypothesis test results is one way to address this shortcoming

Closing Remarks

- A large or non-significant p-value does not mean that the null hypothesis is likely true
  - Instead, a large p-value only means there is insufficient evidence in the sample

Closing Remarks

- A large or non-significant p-value does not mean that the null hypothesis is likely true
  - Instead, a large p-value only means there is insufficient evidence in the sample
- A small or significant p-value does not mean the observed relationship is important or meaningful
  - Instead, a small p-value only means the sample data are unlikely to have occurred by random chance alone if the null model were true