Hypothesis Testing II

Prof. Wells

STA 209, 4/3/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Use P-values to quantify the strength of evidence against the null hypothesis
- Investigate significance level as means of making decisions
- Discuss decision errors and statistical power

Section 1

Hypothesis Testing Review

Framework for Hypothesis Testing

Hypothesis Testing represents a type of scientific experiment, and so should follow the general scientific method.

1. Present research question
2. Identify hypotheses
3. Obtain data
4. Calculate relevant statistics
5. Compute likelihood of observing statistic under original hypothesis
6. Determine statistical significance and make conclusion on research question

Review: Coin-Flipping Hypotheses

A coin is to be flipped 8 times and the proportion of times it showed heads is recorded.

- We wish to test the claim that the coin is fair

Review: Coin-Flipping Hypotheses

A coin is to be flipped 8 times and the proportion of times it showed heads is recorded.

- We wish to test the claim that the coin is fair
- The **null hypothesis** H_0 is the claim we are testing. It often represents a skeptical perspective or that there is no relationship among several variables.
  - H_0: The probability of heads is 50%, or $p = 0.5$.
- The **alternative hypothesis** H_a is contrary to the null hypothesis. It is often the theory we would like to prove.
  - H_a: The probability of heads is greater than 50%, or $p > 0.5$.

Review: Coin-Flipping Hypotheses

A coin is to be flipped 8 times and the proportion of times it showed heads is recorded.

- We wish to test the claim that the coin is fair
- The **null hypothesis** H_0 is the claim we are testing. It often represents a skeptical perspective or that there is no relationship among several variables.
  - H_0: The probability of heads is 50%, or $p = 0.5$.
- The **alternative hypothesis** H_a is contrary to the null hypothesis. It is often the theory we would like to prove.
  - H_a: The probability of heads is greater than 50%, or $p > 0.5$.
- The alternate hypothesis in a **two-sided hypothesis test** proposes that the population parameter is not equal null value. (i.e. $p \neq .5$)
- The alternate hypothesis in a **one-sided hypothesis test** proposes that the population parameter is less than (or greater than) the null value (i.e. one of $p > .5$ or $p < .5$)

Review: Hypotheses

Approximating the Null Distribution

- The distribution of the statistic of interest, *if the null hypothesis were true*, is called the **Null Distribution**
- We can use R to approximate the null distribution by running 5000 experiments of 8 coin flips:

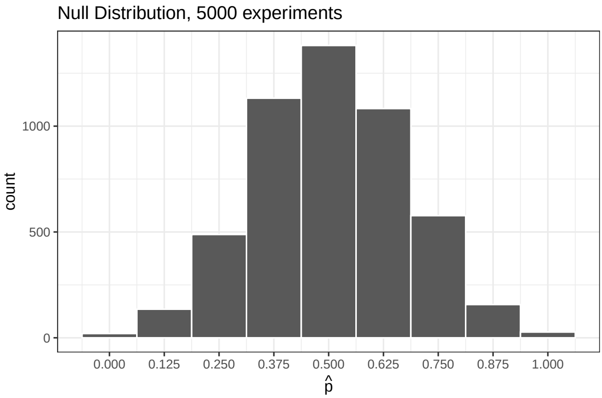
```
coin %>% rep_sample_n(size = 8, replace = T, reps = 5000) %>%
  summarize(n_heads = sum(face == "Heads")) %>% mutate(p_hat = n_heads/8)
```

```
## # A tibble: 5,000 x 3
##    replicate n_heads p_hat
##        <int>   <int> <dbl>
##  1         1       5 0.625
##  2         2       5 0.625
##  3         3       4 0.5
##  4         4       4 0.5
##  5         5       3 0.375
##  6         6       3 0.375
##  7         7       3 0.375
##  8         8       2 0.25
##  9         9       3 0.375
## 10        10       2 0.25
## # ... with 4,990 more rows
## # i Use `print(n = ...)` to see more rows
```

Visualizing the Null Distribution

- We can use a histogram to visualize the Null Distribution of the sample proportion $\hat{p}$

```
null_stats %>% ggplot(aes(x = p_hat))+geom_histogram(bins = 9, color = "white")
```



Section 2

Strength of Evidence

P-Values

- The **p-value** of a sample is the probability of observing a sample statistic at least as favorable to the alternative hypothesis as the current statistic, if H_0 were true.

P-Values

- The **p-value** of a sample is the probability of observing a sample statistic at least as favorable to the alternative hypothesis as the current statistic, if H_0 were true.
- To distinguish between sample statistics generally and the particular one obtained from the sample, we call the latter the **test statistic**

P-Values

- The **p-value** of a sample is the probability of observing a sample statistic at least as favorable to the alternative hypothesis as the current statistic, if H_0 were true.
- To distinguish between sample statistics generally and the particular one obtained from the sample, we call the latter the **test statistic**
  - In the prior experiment, we flipped a coin 8 times and obtained heads 100% of the time. The test statistic is $\hat{p} = 1.0$.

P-Values

- The **p-value** of a sample is the probability of observing a sample statistic at least as favorable to the alternative hypothesis as the current statistic, if H_0 were true.
- To distinguish between sample statistics generally and the particular one obtained from the sample, we call the latter the **test statistic**
  - In the prior experiment, we flipped a coin 8 times and obtained heads 100% of the time. The test statistic is $\hat{p} = 1.0$.
  - The p-value for this test statistic is

$$\text{Probability of at least 8 heads in 8 flips} = 0.5^8 = 0.0039$$

P-Values

- The **p-value** of a sample is the probability of observing a sample statistic at least as favorable to the alternative hypothesis as the current statistic, if H_0 were true.
- To distinguish between sample statistics generally and the particular one obtained from the sample, we call the latter the **test statistic**
  - In the prior experiment, we flipped a coin 8 times and obtained heads 100% of the time. The test statistic is $\hat{p} = 1.0$.
  - The p-value for this test statistic is

$$\text{Probability of at least 8 heads in 8 flips} = 0.5^8 = 0.0039$$

- The p-value quantifies the strength of evidence against the Null Hypothesis. Smaller p-values represent stronger evidence to reject H_0.

P-Values

- The **p-value** of a sample is the probability of observing a sample statistic at least as favorable to the alternative hypothesis as the current statistic, if H_0 were true.
- To distinguish between sample statistics generally and the particular one obtained from the sample, we call the latter the **test statistic**
  - In the prior experiment, we flipped a coin 8 times and obtained heads 100% of the time. The test statistic is $\hat{p} = 1.0$.
  - The p-value for this test statistic is

$$\text{Probability of at least 8 heads in 8 flips} = 0.5^8 = 0.0039$$

- The p-value quantifies the strength of evidence against the Null Hypothesis. Smaller p-values represent stronger evidence to reject H_0.
  - P-values very close to 0 represent statistics that were very unlikely to arise by chance, if the null hypothesis were true.

Calculating P-Values

- Method 1: We can approximate the null distribution using simulation, bootstrapping, and randomization.

Calculating P-Values

- Method 1: We can approximate the null distribution using simulation, bootstrapping, and randomization.
  - Then calculate the proportion of simulated statistics as extreme as the test statistic.

Calculating P-Values

- Method 1: We can approximate the null distribution using simulation, bootstrapping, and randomization.
  - Then calculate the proportion of simulated statistics as extreme as the test statistic.

```
null_stats %>% filter(p_hat >=1.0) %>%
  summarize(n = n()) %>%
  mutate(proportion = n/5000)
```

```
## # A tibble: 1 x 2
##       n proportion
##   <int>      <dbl>
## 1    27     0.0054
```

Calculating P-Values

- Method 1: We can approximate the null distribution using simulation, bootstrapping, and randomization.
    - Then calculate the proportion of simulated statistics as extreme as the test statistic.

```
null_stats %>% filter(p_hat >=1.0) %>%
  summarize(n = n()) %>%
  mutate(proportion = n/5000)
```

```
## # A tibble: 1 x 2
##       n proportion
##   <int>      <dbl>
## 1    27     0.0054
```

- Method 2: We use theory-based tools to create the theoretical null distribution.

Calculating P-Values

- Method 1: We can approximate the null distribution using simulation, bootstrapping, and randomization.
  - Then calculate the proportion of simulated statistics as extreme as the test statistic.

```
null_stats %>% filter(p_hat >=1.0) %>%
  summarize(n = n()) %>%
  mutate(proportion = n/5000)
```

```
## # A tibble: 1 x 2
##       n proportion
##   <int>      <dbl>
## 1    27     0.0054
```

- Method 2: We use theory-based tools to create the theoretical null distribution.
  - Then use the model to calculate the theoretical probability of observing a sample statistic as extreme as the test statistic.

Calculating P-Values

- Method 1: We can approximate the null distribution using simulation, bootstrapping, and randomization.
    - Then calculate the proportion of simulated statistics as extreme as the test statistic.

```
null_stats %>% filter(p_hat >=1.0) %>%
  summarize(n = n()) %>%
  mutate(proportion = n/5000)
```

```
## # A tibble: 1 x 2
##       n proportion
##   <int>      <dbl>
## 1    27     0.0054
```

- Method 2: We use theory-based tools to create the theoretical null distribution.
    - Then use the model to calculate the theoretical probability of observing a sample statistic as extreme as the test statistic.
    - Assuming that coin flips heads with probability 0.5 and that each flip is independent of the others, then the probability of 8 consecutive heads is

```
0.5^8
```

```
## [1] 0.00390625
```

P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?

P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?
  - No. The null distribution just depends on the null hypothesis. It describes the distribution of the statistic if the null hypothesis were true.

P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?
  - No. The null distribution just depends on the null hypothesis. It describes the distribution of the statistic if the null hypothesis were true.
- Does the specific alternative hypothesis play any role in calculating the p-value?

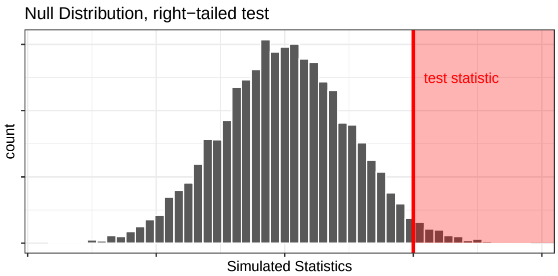
P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?
  - No. The null distribution just depends on the null hypothesis. It describes the distribution of the statistic if the null hypothesis were true.
- Does the specific alternative hypothesis play any role in calculating the p-value?
  - Yes! The **direction** of the alternative hypotheses determines which "tail(s)" of the null distribution correspond to *extreme* values.

P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?
  - No. The null distribution just depends on the null hypothesis. It describes the distribution of the statistic if the null hypothesis were true.
- Does the specific alternative hypothesis play any role in calculating the p-value?
  - Yes! The **direction** of the alternative hypotheses determines which "tail(s)" of the null distribution correspond to *extreme* values.

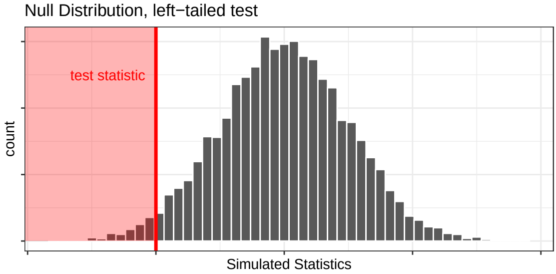
1. If H_a is of the form $\text{parameter} > \text{null value}$, then the p-value is the proportion of simulated statistics greater than or equal to the test statistic (i.e. the right tail)



P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?
  - No. The null distribution just depends on the null hypothesis. It describes the distribution of the statistic if the null hypothesis were true.
- Does the specific alternative hypothesis play any role in calculating the p-value?
  - Yes! The **direction** of the alternative hypotheses determines which "tail(s)" of the null distribution correspond to *extreme* values.

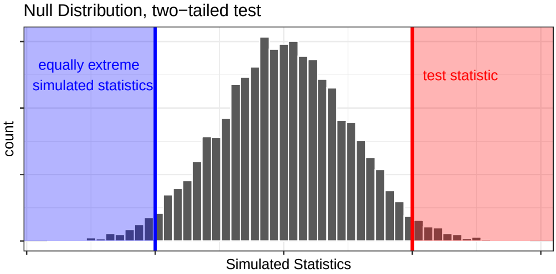
2. If H_a is of the form $\text{parameter} < \text{null value}$, then the p-value is the proportion of simulated statistics less than or equal to the test statistic (i.e. the left tail)



P-Values and the Alternative Hypothesis

- Does the specific alternative hypothesis play any role in making the null distribution?
  - No. The null distribution just depends on the null hypothesis. It describes the distribution of the statistic if the null hypothesis were true.
- Does the specific alternative hypothesis play any role in calculating the p-value?
  - Yes! The **direction** of the alternative hypotheses determines which "tail(s)" of the null distribution correspond to *extreme* values.

3. If H_a is of the form $\text{parameter} \neq \text{null value}$, then the p-value is twice the proportion of simulated statistics more extreme than the test statistic (i.e. both tails)



A Two-Tailed Example

- Suppose we want to determine whether a coin is fair, but don't have any prior expectation that it is biased towards heads or tails.

A Two-Tailed Example

- Suppose we want to determine whether a coin is fair, but don't have any prior expectation that it is biased towards heads or tails.
- Our hypotheses are:

$$H_0 : p = 0.5 \qquad H_a : p \neq 0.5$$

A Two-Tailed Example

- Suppose we want to determine whether a coin is fair, but don't have any prior expectation that it is biased towards heads or tails.
- Our hypotheses are:

$$H_0 : p = 0.5 \qquad H_a : p \neq 0.5$$

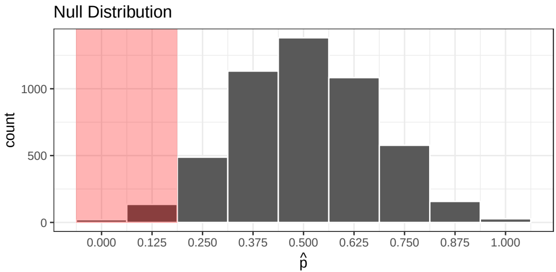
- We flip the coin 8 times and obtain 1 heads, for a proportion $\hat{p} = 0.125$.

A Two-Tailed Example

- Suppose we want to determine whether a coin is fair, but don't have any prior expectation that it is biased towards heads or tails.
- Our hypotheses are:

$$H_0 : p = 0.5 \qquad H_a : p \neq 0.5$$

- We flip the coin 8 times and obtain 1 heads, for a proportion $\hat{p} = 0.125$.
- Using the previous null-distribution, we shade values that are as extreme as our statistic:

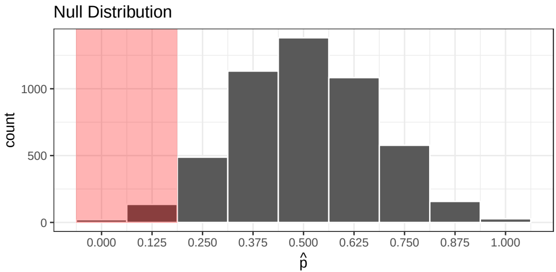


A Two-Tailed Example

- Suppose we want to determine whether a coin is fair, but don't have any prior expectation that it is biased towards heads or tails.
- Our hypotheses are:

$$H_0 : p = 0.5 \qquad H_a : p \neq 0.5$$

- We flip the coin 8 times and obtain 1 heads, for a proportion $\hat{p} = 0.125$.
- Using the previous null-distribution, we shade values that are as extreme as our statistic:



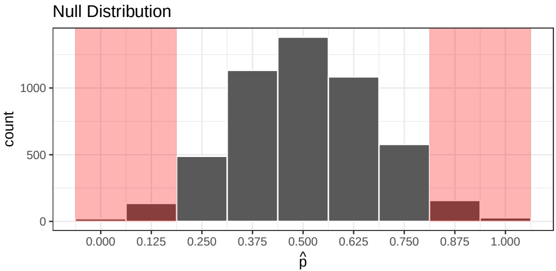
- We find the proportion of simulated statistics in the left tail is 0.034

A Two-Tailed Example

- Suppose we want to determine whether a coin is fair, but don't have any prior expectation that it is biased towards heads or tails.
- Our hypotheses are:

$$H_0 : p = 0.5 \qquad H_a : p \neq 0.5$$

- We flip the coin 8 times and obtain 1 heads, for a proportion $\hat{p} = 0.125$.
- Using the previous null-distribution, we shade values that are as extreme as our statistic:



- We double this to include the right-tail as well, and get a p-value of 0.068.

Section 3

Decision Rules

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.
  - Recall that smaller P-values (i.e. closer to 0) provide stronger evidence against H_0 and in favor of H_a.

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.
  - Recall that smaller P-values (i.e. closer to 0) provide stronger evidence against H_0 and in favor of H_a.
  - But what counts as "small"?

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.
  - Recall that smaller P-values (i.e. closer to 0) provide stronger evidence against H_0 and in favor of H_a.
  - But what counts as "small"?
- In general, the answer depends on the field of study, as well as the stakes of the investigation.

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.
  - Recall that smaller P-values (i.e. closer to 0) provide stronger evidence against H_0 and in favor of H_a.
  - But what counts as "small"?
- In general, the answer depends on the field of study, as well as the stakes of the investigation.
  - A P-value of 0.03 might provide compelling evidence when determining whether a coin is fair.

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.
  - Recall that smaller P-values (i.e. closer to 0) provide stronger evidence against H_0 and in favor of H_a.
  - But what counts as “small”?
- In general, the answer depends on the field of study, as well as the stakes of the investigation.
  - A P-value of 0.03 might provide compelling evidence when determining whether a coin is fair.
  - But the same p-value of 0.03 might not provide compelling evidence when determining whether a physics experiment gives evidence of the existence of the Higgs-boson particle (where a P-value less than 0.000001 might be required.)

Sufficient Evidence

- How do we decide what counts as *sufficient evidence* to reject the null hypothesis in favor of the alternative?
  - The p-value measures how unlikely a sample statitsic is, if the null hypothesis were true.
  - Recall that smaller P-values (i.e. closer to 0) provide stronger evidence against H_0 and in favor of H_a.
  - But what counts as "small"?
- In general, the answer depends on the field of study, as well as the stakes of the investigation.
  - A P-value of 0.03 might provide compelling evidence when determining whether a coin is fair.
  - But the same p-value of 0.03 might not provide compelling evidence when determining whether a physics experiment gives evidence of the existence of the Higgs-boson particle (where a P-value less than 0.000001 might be required.)
  - But if you are trying to determine whether pushing a crosswalk button more than once causes the stoplight to change faster, you might find a p-value of 0.25 compelling evidence.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
  - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
  - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.
- In the coin flip experiment, if we observe 1 out of 8 heads, our P-value is 0.068 using a 2-sided alternative hypothesis.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
  - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.
- In the coin flip experiment, if we observe 1 out of 8 heads, our P-value is 0.068 using a 2-sided alternative hypothesis.
  - The data is statistically significant at the $\alpha = 0.10$ significance level, but is **not** statistically significant at the $\alpha = 0.05$ level.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
  - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.
- In the coin flip experiment, if we observe 1 out of 8 heads, our P-value is 0.068 using a 2-sided alternative hypothesis.
  - The data is statistically significant at the $\alpha = 0.10$ significance level, but is **not** statistically significant at the $\alpha = 0.05$ level.
  - No matter whether we use a significance level of $\alpha = 0.10$ or $\alpha = 0.05$, the data provides identical strength of evidence: $P\text{-Value} = 0.068$.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
  - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.
- In the coin flip experiment, if we observe 1 out of 8 heads, our P-value is 0.068 using a 2-sided alternative hypothesis.
  - The data is statistically significant at the $\alpha = 0.10$ significance level, but is **not** statistically significant at the $\alpha = 0.05$ level.
  - No matter whether we use a significance level of $\alpha = 0.10$ or $\alpha = 0.05$, the data provides identical strength of evidence: $P\text{-Value} = 0.068$.
  - But whether we consider this "strong enough" depends on whether we are using $\alpha = 0.10$ or $\alpha = 0.05$.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
    - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.
- In the coin flip experiment, if we observe 1 out of 8 heads, our P-value is 0.068 using a 2-sided alternative hypothesis.
    - The data is statistically significant at the $\alpha = 0.10$ significance level, but is **not** statistically significant at the $\alpha = 0.05$ level.
    - No matter whether we use a significance level of $\alpha = 0.10$ or $\alpha = 0.05$, the data provides identical strength of evidence: $P\text{-Value} = 0.068$.
    - But whether we consider this “strong enough” depends on whether we are using $\alpha = 0.10$ or $\alpha = 0.05$.
- In general, we should **always** choose the value of α prior to conducting an experiment and observing data.

Significance Levels

- The decision threshold is called the **significance level** (usually denoted as α).
- If the P-Value is less than the prescribed significance level α, we say the data is **statistically significant** (or *statistically discernible*) at the level α.
  - In this case, the sample provides compelling evidence to reject H_0 in favor of H_a.
- In the coin flip experiment, if we observe 1 out of 8 heads, our P-value is 0.068 using a 2-sided alternative hypothesis.
  - The data is statistically significant at the $\alpha = 0.10$ significance level, but is **not** statistically significant at the $\alpha = 0.05$ level.
  - No matter whether we use a significance level of $\alpha = 0.10$ or $\alpha = 0.05$, the data provides identical strength of evidence: $P\text{-Value} = 0.068$.
  - But whether we consider this "strong enough" depends on whether we are using $\alpha = 0.10$ or $\alpha = 0.05$.
- In general, we should **always** choose the value of α prior to conducting an experiment and observing data.
  - Otherwise, we are liable to choose a significance level that conforms to whichever decision we would prefer to make.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
    - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.
- A **Type 2 Error** occurs when we fail to reject H_0 when it is in fact false.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.
- A **Type 2 Error** occurs when we fail to reject H_0 when it is in fact false.
  - The coin was indeed biased. But we withheld judgment since unlikely events do happen from time to time.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.
- A **Type 2 Error** occurs when we fail to reject H_0 when it is in fact false.
  - The coin was indeed biased. But we withheld judgment since unlikely events do happen from time to time.
- In general, we will never know we made an error at all (but we can still quantify the probability that we made a particular error)

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

  - In general, computing power can be difficult, and requires we investigate the distribution of a sample statistic under the alternative hypothesis.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

  - In general, computing power can be difficult, and requires we investigate the distribution of a sample statistic under the alternative hypothesis.

With great power comes...

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

  - In general, computing power can be difficult, and requires we investigate the distribution of a sample statistic under the alternative hypothesis.

With great power comes...greater chance of Type I error.

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are informal statements of the Null and Alternate Hypotheses?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to assess whether the person has COVID?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to assess whether the person has COVID?
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to assess whether the person has COVID?
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to assess whether the person has COVID?
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?
5. What significance level are you willing to use for this COVID test? *Remember, decreasing significance level also decreases the power of the test.*

DNA Tests

DNA testing allows researchers to compare the DNA profile of a suspect to the profile of DNA at the crime scene. Suppose that the perpetrator's DNA profile will **always** match profile of the DNA found at the crime scene. However, there is a small chance that profile of an innocent person matches the crime scene DNA profile, as well.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA profile matches the crime scene DNA profile.

1. What are informal statements of the Null and Alternate Hypotheses?

DNA Tests

DNA testing allows researchers to compare the DNA profile of a suspect to the profile of DNA at the crime scene. Suppose that the perpetrator's DNA profile will **always** match profile of the DNA found at the crime scene. However, there is a small chance that profile of an innocent person matches the crime scene DNA profile, as well.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA profile matches the crime scene DNA profile.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to determine whether the person has committed the crime?

DNA Tests

DNA testing allows researchers to compare the DNA profile of a suspect to the profile of DNA at the crime scene. Suppose that the perpetrator's DNA profile will **always** match profile of the DNA found at the crime scene. However, there is a small chance that profile of an innocent person matches the crime scene DNA profile, as well.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA profile matches the crime scene DNA profile.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to determine whether the person has committed the crime?
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?

DNA Tests

DNA testing allows researchers to compare the DNA profile of a suspect to the profile of DNA at the crime scene. Suppose that the perpetrator's DNA profile will **always** match profile of the DNA found at the crime scene. However, there is a small chance that profile of an innocent person matches the crime scene DNA profile, as well.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA profile matches the crime scene DNA profile.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to determine whether the person has committed the crime?
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?

DNA Tests

DNA testing allows researchers to compare the DNA profile of a suspect to the profile of DNA at the crime scene. Suppose that the perpetrator's DNA profile will **always** match profile of the DNA found at the crime scene. However, there is a small chance that profile of an innocent person matches the crime scene DNA profile, as well.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA profile matches the crime scene DNA profile.

1. What are informal statements of the Null and Alternate Hypotheses?
2. What 'statistic' is being used to determine whether the person has committed the crime?
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?
5. What significance level are you willing to use for this DNA test? *Remember, decreasing significance level also decreases the power of the test.*

Section 4

(Mis)Intepreting P-Values

The Problem with P-Values

- In the early days of statistical theory, P-values were introduced as an *informal* measure to indicate whether a phenomenon warrants further investigation.

The Problem with P-Values

- In the early days of statistical theory, *P*-values were introduced as an *informal* measure to indicate whether a phenomenon warrants further investigation.
- Later, and primarily to create simple statistical manuals for untrained practitioners, this informal measure became the unassailable rule:

"*If* $p\text{-}value < 0.05$, *reject* H_0; *If* $p\text{-}value > 0.05$, *do not reject* H_0"

The Problem with P-Values

- In the early days of statistical theory, P-values were introduced as an *informal* measure to indicate whether a phenomenon warrants further investigation.
- Later, and primarily to create simple statistical manuals for untrained practitioners, this informal measure became the unassailable rule:

  "If p-$value < 0.05$, reject H_0; If p-$value > 0.05$, do not reject H_0"

- As a result, many academic journals used this threshold to determine whether or not a claim is true, and therefore, publication-worthy

The Problem with P-Values

- In the early days of statistical theory, P-values were introduced as an *informal* measure to indicate whether a phenomenon warrants further investigation.
- Later, and primarily to create simple statistical manuals for untrained practitioners, this informal measure became the unassailable rule:

  "*If p*-$value < 0.05$, *reject* H_0; *If p*-$value > 0.05$, *do not reject* H_0"

- As a result, many academic journals used this threshold to determine whether or not a claim is true, and therefore, publication-worthy
  - Non-technical reports (i.e. news media, pop-literature, word-of-mouth) further propagate this rule

The Problem with P-Values

- In the early days of statistical theory, P-values were introduced as an *informal* measure to indicate whether a phenomenon warrants further investigation.
- Later, and primarily to create simple statistical manuals for untrained practitioners, this informal measure became the unassailable rule:

"*If p-value* < 0.05*, reject* H_0*; If p-value* > 0.05*, do not reject* H_0"

- As a result, many academic journals used this threshold to determine whether or not a claim is true, and therefore, publication-worthy
  - Non-technical reports (i.e. news media, pop-literature, word-of-mouth) further propagate this rule
- This editorial bias also leads to the practice of "data dredging" or "p-hacking":
  - Researchers prioritize the search for phenomenon with small p-values, at the expense of otherwise noteworthy or important outcomes, and often eschewing other statistical and scientific reasoning.

The Problem with P-Values

- In the early days of statistical theory, P-values were introduced as an *informal* measure to indicate whether a phenomenon warrants further investigation.
- Later, and primarily to create simple statistical manuals for untrained practitioners, this informal measure became the unassailable rule:

  "*If p-value* < 0.05*, reject* H_0*; If p-value* > 0.05*, do not reject* H_0"

- As a result, many academic journals used this threshold to determine whether or not a claim is true, and therefore, publication-worthy
  - Non-technical reports (i.e. news media, pop-literature, word-of-mouth) further propagate this rule
- This editorial bias also leads to the practice of "data dredging" or "p-hacking":
  - Researchers prioritize the search for phenomenon with small p-values, at the expense of otherwise noteworthy or important outcomes, and often eschewing other statistical and scientific reasoning.
- This may be one cause of the *Reproducibility Crisis* currently faced in the fields of Psychology and Medicine (and to some extent, other natural and social sciences)

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about p-values:

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about *p*-values:

1. P-Values indicate how incompatible the data are with a specific statistical model.

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about *p*-values:

1. P-Values indicate how incompatible the data are with a specific statistical model.
2. P-values do not measure the probability that the null hypothesis is true; or that the data were produced by random chance alone.

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about *p*-values:

1. P-Values indicate how incompatible the data are with a specific statistical model.
2. P-values do not measure the probability that the null hypothesis is true; or that the data were produced by random chance alone.
3. Scientific, business, or policy decisions should not be based only on whether a p-value is less than a specific threshold.

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about *p*-values:

1. P-Values indicate how incompatible the data are with a specific statistical model.
2. P-values do not measure the probability that the null hypothesis is true; or that the data were produced by random chance alone.
3. Scientific, business, or policy decisions should not be based only on whether a p-value is less than a specific threshold.
4. Proper statistical inference requires full reporting and transparency

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about *p*-values:

1. P-Values indicate how incompatible the data are with a specific statistical model.
2. P-values do not measure the probability that the null hypothesis is true; or that the data were produced by random chance alone.
3. Scientific, business, or policy decisions should not be based only on whether a p-value is less than a specific threshold.
4. Proper statistical inference requires full reporting and transparency
5. A p-value does not measure the size of an effect, or the importance of a result.

Guidelines for the Responsible Use of P-Values

In 2016, the American Statistical Association put forth 6 guidelines to address misconceptions about *p*-values:

1. P-Values indicate how incompatible the data are with a specific statistical model.
2. P-values do not measure the probability that the null hypothesis is true; or that the data were produced by random chance alone.
3. Scientific, business, or policy decisions should not be based only on whether a p-value is less than a specific threshold.
4. Proper statistical inference requires full reporting and transparency
5. A p-value does not measure the size of an effect, or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
    - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small p-value suggests greater incompatibility)

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
  - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small *p*-value suggests greater incompatibility)
- But p-values are **NOT** the probability that the null hypothesis is true.

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
    - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small *p*-value suggests greater incompatibility)
- But p-values are **NOT** the probability that the null hypothesis is true.
- Consider the following *hypothetical* example:
    - Pro basketball player Stephen Curry and I each take five 3-point shots. Stephen Curry makes all 5, while I make 2.

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
  - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small *p*-value suggests greater incompatibility)
- But p-values are **NOT** the probability that the null hypothesis is true.
- Consider the following *hypothetical* example:
  - Pro basketball player Stephen Curry and I each take five 3-point shots. Stephen Curry makes all 5, while I make 2.
  - The null hypothesis that we have the same shot probability, while the alternative is that Stephen Curry has higher probability.

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
  - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small *p*-value suggests greater incompatibility)
- But p-values are **NOT** the probability that the null hypothesis is true.
- Consider the following *hypothetical* example:
  - Pro basketball player Stephen Curry and I each take five 3-point shots. Stephen Curry makes all 5, while I make 2.
  - The null hypothesis that we have the same shot probability, while the alternative is that Stephen Curry has higher probability.
  - The p-value for this experiment (i.e. probability of a result as extreme or more) is 0.17.

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
  - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small p-value suggests greater incompatibility)
- But p-values are **NOT** the probability that the null hypothesis is true.
- Consider the following *hypothetical* example:
  - Pro basketball player Stephen Curry and I each take five 3-point shots. Stephen Curry makes all 5, while I make 2.
  - The null hypothesis that we have the same shot probability, while the alternative is that Stephen Curry has higher probability.
  - The p-value for this experiment (i.e. probability of a result as extreme or more) is 0.17.
  - Is it reasonable to conclude that there is a 17% chance that Stephen Curry and I are equally good shooters?

P-Values and Probability

- By definition, a P-value is the probability of observing data as extreme as the data collected **if the null hypothesis were true**
  - P-values are one measure indicating how incompatible the data is with a specified hypothesis (Small *p*-value suggests greater incompatibility)
- But p-values are **NOT** the probability that the null hypothesis is true.
- Consider the following *hypothetical* example:
  - Pro basketball player Stephen Curry and I each take five 3-point shots. Stephen Curry makes all 5, while I make 2.
  - The null hypothesis that we have the same shot probability, while the alternative is that Stephen Curry has higher probability.
  - The p-value for this experiment (i.e. probability of a result as extreme or more) is 0.17.
  - Is it reasonable to conclude that there is a 17% chance that Stephen Curry and I are equally good shooters?
  - No. We would also need to take into account our prior beliefs about the likelihood of this hypothesis.

Effect Size and Practical Significance

- A small p-value indicates a result that is unlikely to occur due to chance, if the null hypothesis were true.

Effect Size and Practical Significance

- A small *p*-value indicates a result that is unlikely to occur due to chance, if the null hypothesis were true.
  - But the size of the p-value gives **NO** indication about the actual size of the effect measured;
  - Moreover, the p-value gives no indication about whether the observed difference is of *practical* importance.

Effect Size and Practical Significance

- A small *p*-value indicates a result that is unlikely to occur due to chance, if the null hypothesis were true.
  - But the size of the p-value gives **NO** indication about the actual size of the effect measured;
  - Moreover, the p-value gives no indication about whether the observed difference is of *practical* importance.
  - A large sample is able to detect extremely minuscule differences between populations, producing very small *p*-values.

Effect Size and Practical Significance

- A small *p*-value indicates a result that is unlikely to occur due to chance, if the null hypothesis were true.
  - But the size of the p-value gives **NO** indication about the actual size of the effect measured;
  - Moreover, the p-value gives no indication about whether the observed difference is of *practical* importance.
  - A large sample is able to detect extremely minuscule differences between populations, producing very small *p*-values.
- The **effect size** is the difference between the true value of the parameter and the null value.

Effect Size and Practical Significance

- A small *p*-value indicates a result that is unlikely to occur due to chance, if the null hypothesis were true.
  - But the size of the p-value gives **NO** indication about the actual size of the effect measured;
  - Moreover, the p-value gives no indication about whether the observed difference is of *practical* importance.
  - A large sample is able to detect extremely minuscule differences between populations, producing very small *p*-values.
- The **effect size** is the difference between the true value of the parameter and the null value.
  - Effect size determines whether a result is *practically significant* (i.e. is noteworthy or worth changing behavior over).

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online. . .
  - Are less likely to divorce ($p-\mathrm{value} < 0.002$)
  - Are more likely to have high marital satisfaction ($p-\mathrm{value} < 0.001$)

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online...
    - Are less likely to divorce ($p-\text{value} < 0.002$)
    - Are more likely to have high marital satisfaction ($p-\text{value} < 0.001$)
- BUT! The estimated *effect sizes* are tiny (and perhaps not practically significant)

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online. . .
  - Are less likely to divorce ($p-\text{value} < 0.002$)
  - Are more likely to have high marital satisfaction ($p-\text{value} < 0.001$)
- BUT! The estimated *effect sizes* are tiny (and perhaps not practically significant)
  - Divorce rates are 5.96% for those who met online, versus 7.07% for those who met in-person (Effect Size $= 1.11$)

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online...
  - Are less likely to divorce ($p-\text{value} < 0.002$)
  - Are more likely to have high marital satisfaction ($p-\text{value} < 0.001$)
- BUT! The estimated *effect sizes* are tiny (and perhaps not practically significant)
  - Divorce rates are 5.96% for those who met online, versus 7.07% for those who met in-person (Effect Size = 1.11)
  - On a 7 point scale, happiness values were 5.64 for those who met online, versus 5.48 for those who met in-person (Effect Size = 0.16)

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online...
  - Are less likely to divorce ($p-\text{value} < 0.002$)
  - Are more likely to have high marital satisfaction ($p-\text{value} < 0.001$)
- BUT! The estimated *effect sizes* are tiny (and perhaps not practically significant)
  - Divorce rates are 5.96% for those who met online, versus 7.07% for those who met in-person (Effect Size $= 1.11$)
  - On a 7 point scale, happiness values were 5.64 for those who met online, versus 5.48 for those who met in-person (Effect Size $= 0.16$)
- Does this provide compelling evidence that those seeking spouses should do so online?

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online. . .
    - Are less likely to divorce ($p-\text{value} < 0.002$)
    - Are more likely to have high marital satisfaction ($p-\text{value} < 0.001$)
- BUT! The estimated *effect sizes* are tiny (and perhaps not practically significant)
    - Divorce rates are 5.96% for those who met online, versus 7.07% for those who met in-person (Effect Size $= 1.11$)
    - On a 7 point scale, happiness values were 5.64 for those who met online, versus 5.48 for those who met in-person (Effect Size $= 0.16$)
- Does this provide compelling evidence that those seeking spouses should do so online?
    - Are the estimated effect sizes meaningful?

Effect Size and Practical Significance

- A recent *Nature* study of 19,000 people found that those who met their spouses online...
  - Are less likely to divorce ($p-\text{value} < 0.002$)
  - Are more likely to have high marital satisfaction ($p-\text{value} < 0.001$)
- BUT! The estimated *effect sizes* are tiny (and perhaps not practically significant)
  - Divorce rates are 5.96% for those who met online, versus 7.07% for those who met in-person (Effect Size $= 1.11$)
  - On a 7 point scale, happiness values were 5.64 for those who met online, versus 5.48 for those who met in-person (Effect Size $= 0.16$)
- Does this provide compelling evidence that those seeking spouses should do so online?
  - Are the estimated effect sizes meaningful?
  - Can we deduce causal relationships from this investigation? (This is unrelated to significance and effect size)

Conclusion

- Despite their issues, p-values are still quite popular and widely used (although are perhaps over-used)

Conclusion

- Despite their issues, p-values are still quite popular and widely used (although are perhaps over-used)
- The following quote, attributed to George Cobb of Mount Holyoke College, summarizes this as:

Conclusion

- Despite their issues, p-values are still quite popular and widely used (although are perhaps over-used)
- The following quote, attributed to George Cobb of Mount Holyoke College, summarizes this as:

Q: Why do so many caolleges and grad schools teach $p = 0.05$*?*
A: Because that's what the scientific community and journal editors use.
Q: Why do the scientific community and journal editors still use $p = 0.05$*?*
A: Because that's what they were taught in college or grad school.

Conclusion

- Despite their issues, p-values are still quite popular and widely used (although are perhaps over-used)
- The following quote, attributed to George Cobb of Mount Holyoke College, summarizes this as:

  Q: Why do so many caolleges and grad schools teach $p = 0.05$?
  A: Because that's what the scientific community and journal editors use.
  Q: Why do the scientific community and journal editors still use $p = 0.05$?
  A: Because that's what they were taught in college or grad school.

- Understanding p-values and interpreting p-values in context is an important goal for STA 209

Conclusion

- Despite their issues, p-values are still quite popular and widely used (although are perhaps over-used)
- The following quote, attributed to George Cobb of Mount Holyoke College, summarizes this as:

  Q: Why do so many caolleges and grad schools teach $p = 0.05$?
  A: Because that's what the scientific community and journal editors use.
  Q: Why do the scientific community and journal editors still use $p = 0.05$?
  A: Because that's what they were taught in college or grad school.

- Understanding p-values and interpreting p-values in context is an important goal for STA 209
- Determining an appropriate significance level that balances the rate of Type I and Type II error, for your specific research question, is also an important goal for STA 209.

Conclusion

- Despite their issues, p-values are still quite popular and widely used (although are perhaps over-used)
- The following quote, attributed to George Cobb of Mount Holyoke College, summarizes this as:

  Q: Why do so many caolleges and grad schools teach $p = 0.05$?
  A: Because that's what the scientific community and journal editors use.
  Q: Why do the scientific community and journal editors still use $p = 0.05$?
  A: Because that's what they were taught in college or grad school.

- Understanding p-values and interpreting p-values in context is an important goal for STA 209
- Determining an appropriate significance level that balances the rate of Type I and Type II error, for your specific research question, is also an important goal for STA 209.
- Determining whether a given number is less than 0.05 is not an important goal for STA 209