COMS BC1016
Introduction to Computational Thinking and Data Science

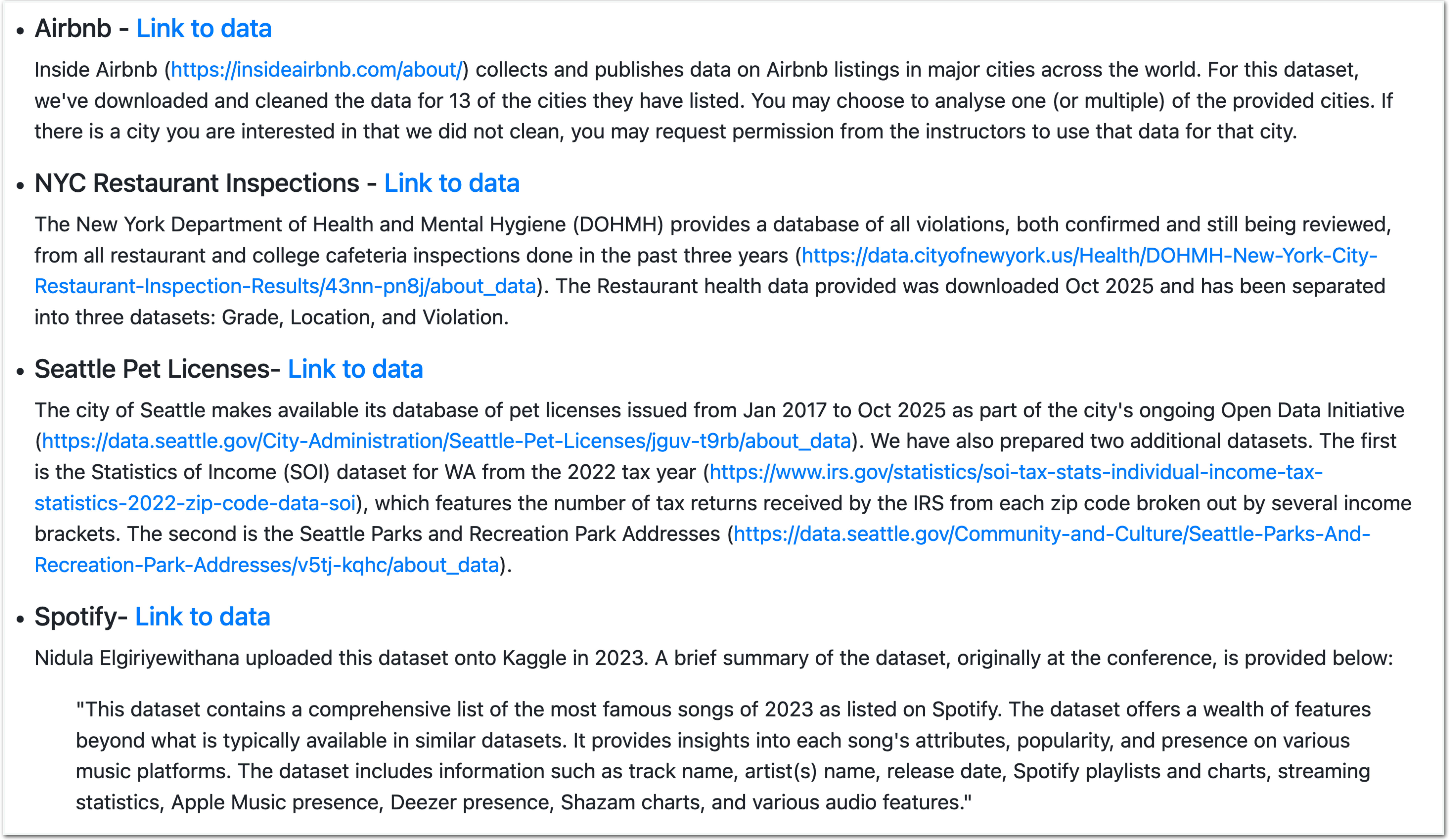
Lecture 14: AB Testing and Confidence Intervals

Nov 5, 2025

Reminders

- Welcome back from fall break! Unfortunately we're back to work…
  - HW 5 due today, HW 6 due next week
  - Labs again this week
- **Don't forget to sign up for final project groups by Friday, Nov 7!!**
  - If you want us to match you to a group: Link
  - If you want to be in a specific group: Link

Final Project Datasets



- Airbnb - Link to data

  Inside Airbnb (https://insideairbnb.com/about/) collects and publishes data on Airbnb listings in major cities across the world. For this dataset, we've downloaded and cleaned the data for 13 of the cities they have listed. You may choose to analyse one (or multiple) of the provided cities. If there is a city you are interested in that we did not clean, you may request permission from the instructors to use that data for that city.

- NYC Restaurant Inspections - Link to data

  The New York Department of Health and Mental Hygiene (DOHMH) provides a database of all violations, both confirmed and still being reviewed, from all restaurant and college cafeteria inspections done in the past three years (https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/about_data). The Restaurant health data provided was downloaded Oct 2025 and has been separated into three datasets: Grade, Location, and Violation.

- Seattle Pet Licenses- Link to data

  The city of Seattle makes available its database of pet licenses issued from Jan 2017 to Oct 2025 as part of the city's ongoing Open Data Initiative (https://data.seattle.gov/City-Administration/Seattle-Pet-Licenses/jguv-t9rb/about_data). We have also prepared two additional datasets. The first is the Statistics of Income (SOI) dataset for WA from the 2022 tax year (https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2022-zip-code-data-soi), which features the number of tax returns received by the IRS from each zip code broken out by several income brackets. The second is the Seattle Parks and Recreation Park Addresses (https://data.seattle.gov/Community-and-Culture/Seattle-Parks-And-Recreation-Park-Addresses/v5tj-kqhc/about_data).

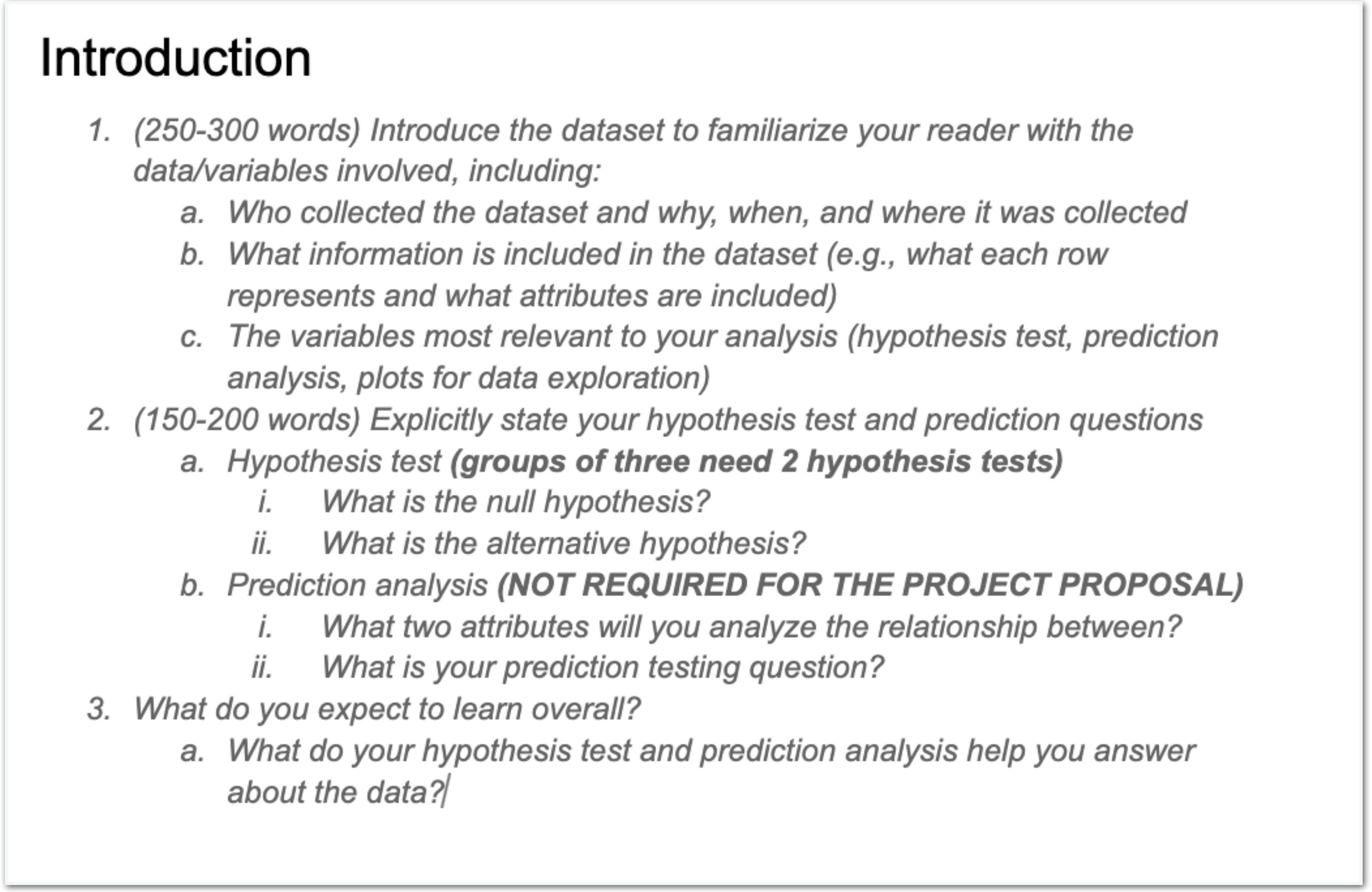
- Spotify- Link to data

  Nidula Elgiriyewithana uploaded this dataset onto Kaggle in 2023. A brief summary of the dataset, originally at the conference, is provided below:

  > "This dataset contains a comprehensive list of the most famous songs of 2023 as listed on Spotify. The dataset offers a wealth of features beyond what is typically available in similar datasets. It provides insights into each song's attributes, popularity, and presence on various music platforms. The dataset includes information such as track name, artist(s) name, release date, Spotify playlists and charts, streaming statistics, Apple Music presence, Deezer presence, Shazam charts, and various audio features."

Final Project Proposal

- Template is on 1017 Courseworks
    - Descriptions of each section are italicized gray and *should be deleted before submitting*
- For the proposal, you need to complete the introduction section *except for the prediction analysis*
    - Due next week **Friday, Nov 14**



Introduction

1. *(250-300 words) Introduce the dataset to familiarize your reader with the data/variables involved, including:*
    1. *Who collected the dataset and why, when, and where it was collected*
    2. *What information is included in the dataset (e.g., what each row represents and what attributes are included)*
    3. *The variables most relevant to your analysis (hypothesis test, prediction analysis, plots for data exploration)*
2. *(150-200 words) Explicitly state your hypothesis test and prediction questions*
    1. *Hypothesis test **(groups of three need 2 hypothesis tests)***
        1. *What is the null hypothesis?*
        2. *What is the alternative hypothesis?*
    2. *Prediction analysis **(NOT REQUIRED FOR THE PROJECT PROPOSAL)***
        1. *What two attributes will you analyze the relationship between?*
        2. *What is your prediction testing question?*
3. *What do you expect to learn overall?*
    1. *What do your hypothesis test and prediction analysis help you answer about the data?*

Lecture Outline

- Review of last lecture (p-values)

- AB Testing

- Confidence intervals

P-Value Review

Definition of the P-Value

P-value: Observed significance level

5% - statistically significant

1% highly statistically significant

The P-value is the chance under the null hypothesis that the test statistic is equal to the value that was observed in the data or is even further in the direction of the alternative

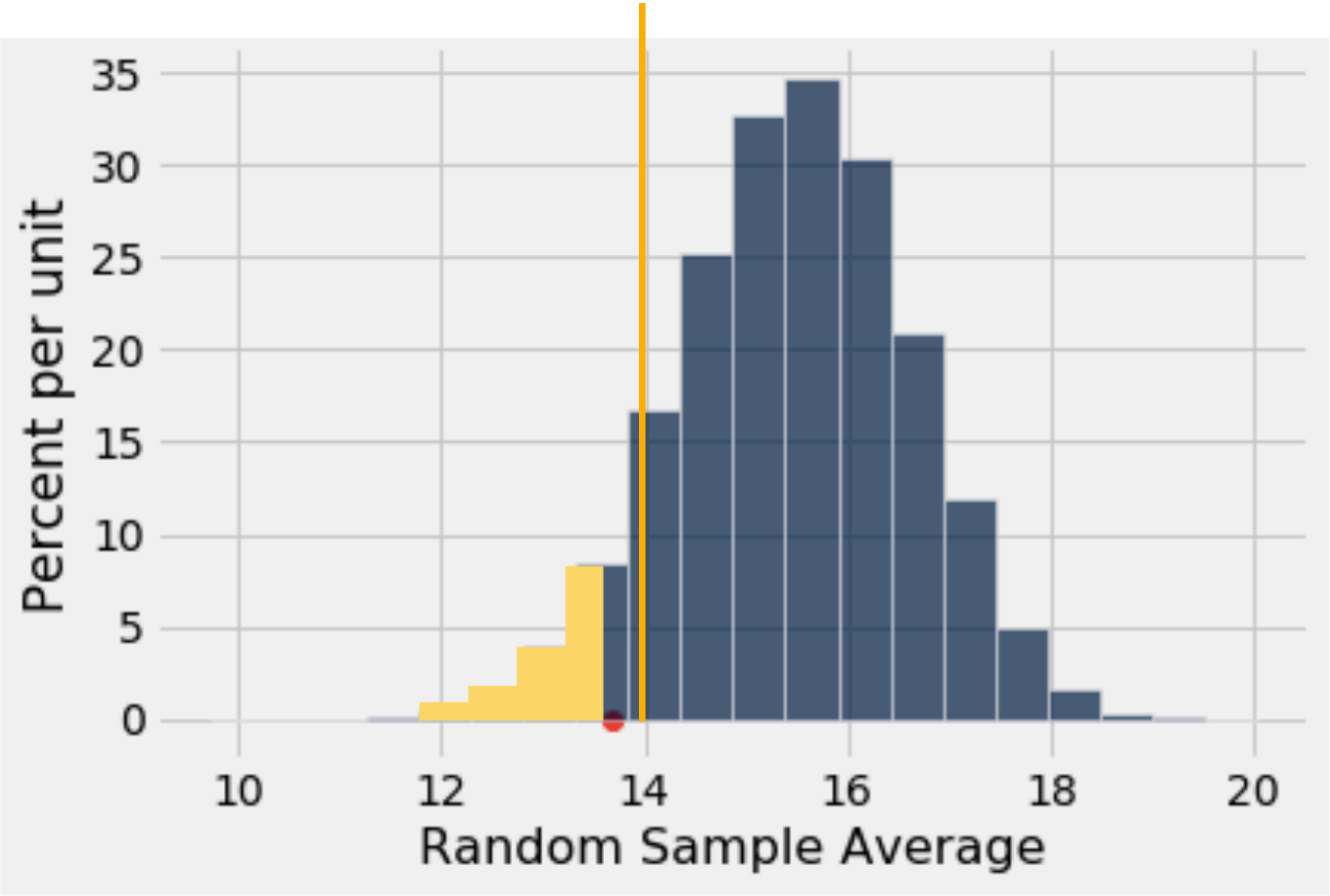
The P-Value as an Area

P-value is the area of the tail of the empirical distribution

Red dot: observed statistic

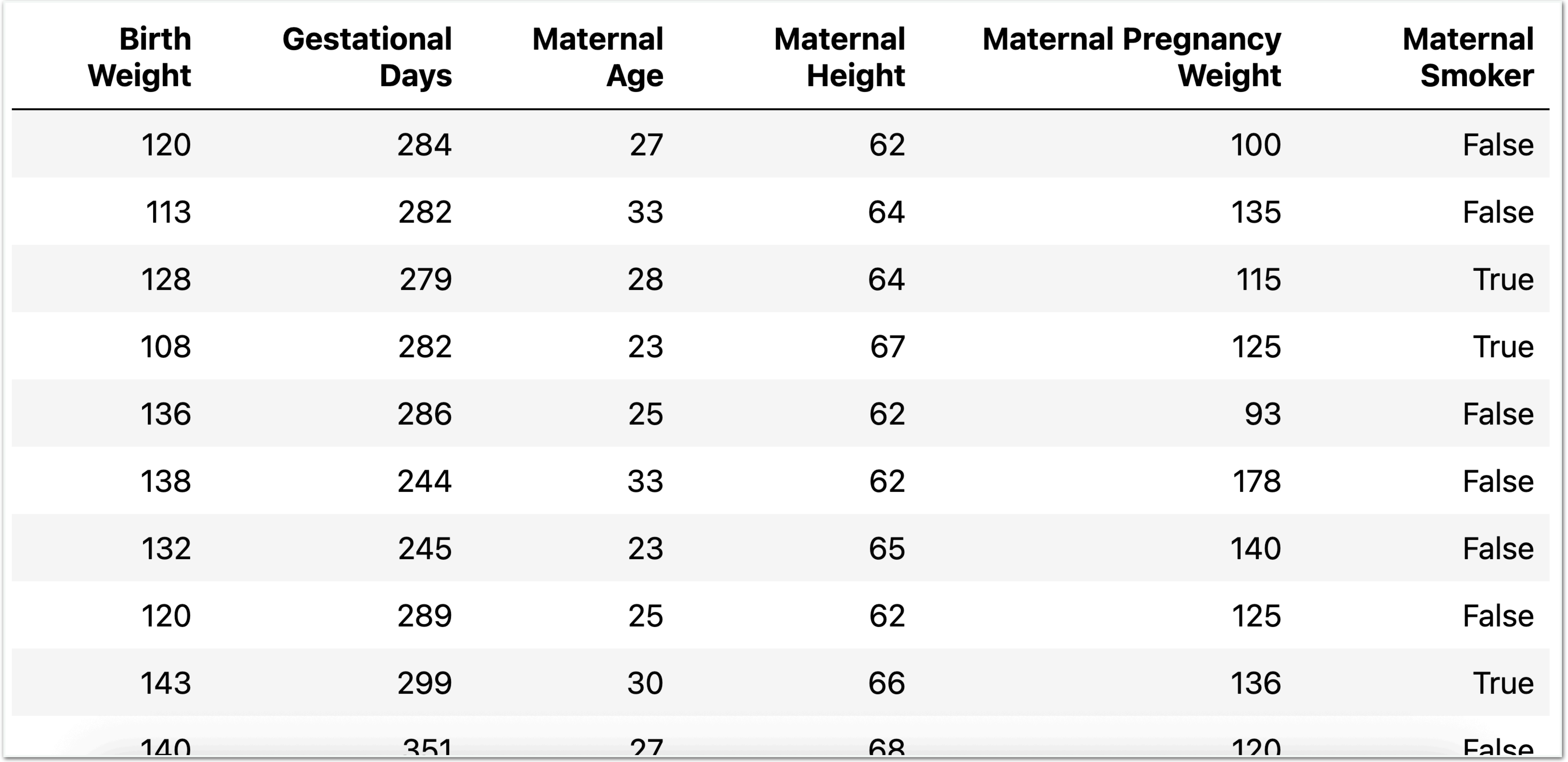
Yellow area: tail probability (p-value)

If your threshold (1% / 5%) is beyond the observed value in the direction of the alternative, you can reject the null hypothesis



A/B Testing

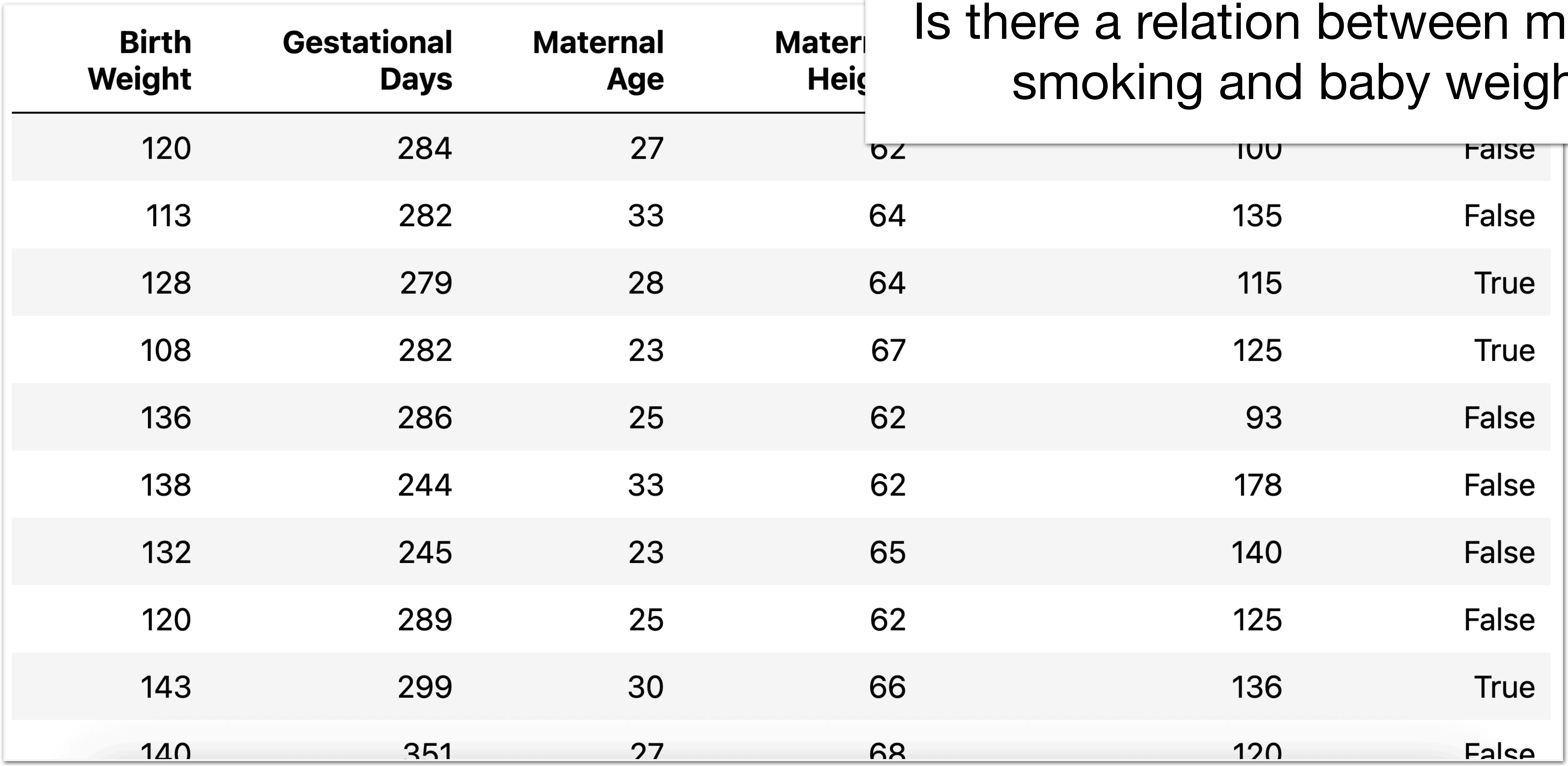
Scenario: Baby weights and Maternal Smoking



Birth Weight	Gestational Days	Maternal Age	Maternal Height	Maternal Pregnancy Weight	Maternal Smoker
120	284	27	62	100	False
113	282	33	64	135	False
128	279	28	64	115	True
108	282	23	67	125	True
136	286	25	62	93	False
138	244	33	62	178	False
132	245	23	65	140	False
120	289	25	62	125	False
143	299	30	66	136	True
140	351	27	68	120	False

Scenario: Baby weights and Maternal Smoking

Is there a relation between maternal smoking and baby weight?

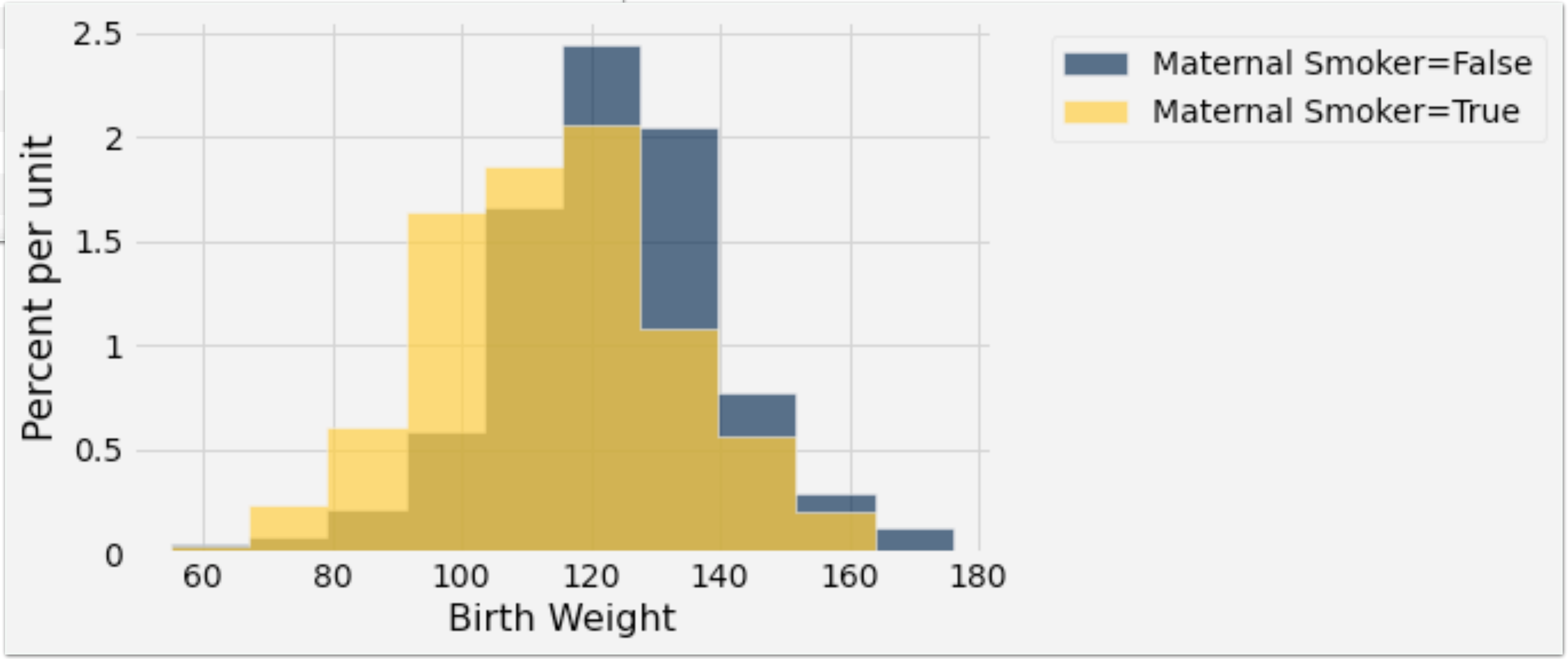


Birth Weight	Gestational Days	Maternal Age	Matern Heig		
120	284	27	62	100	False
113	282	33	64	135	False
128	279	28	64	115	True
108	282	23	67	125	True
136	286	25	62	93	False
138	244	33	62	178	False
132	245	23	65	140	False
120	289	25	62	125	False
143	299	30	66	136	True
140	351	27	68	120	False

Scenario: Baby weights and Maternal Smoking

Birth Weight	Gestational Days	Maternal Age	Maternal Height	Maternal Pregnancy Weight	Maternal Smoker
120	284	27	62	100	False
113	282	33	64	135	False
128	279	28	64	115	True
108	282	23	67	125	True
136	286				
138	244				
132	245				
120	289				
143	299				
140	351				

Is there a relation between maternal smoking and baby weight?



A/B Testing

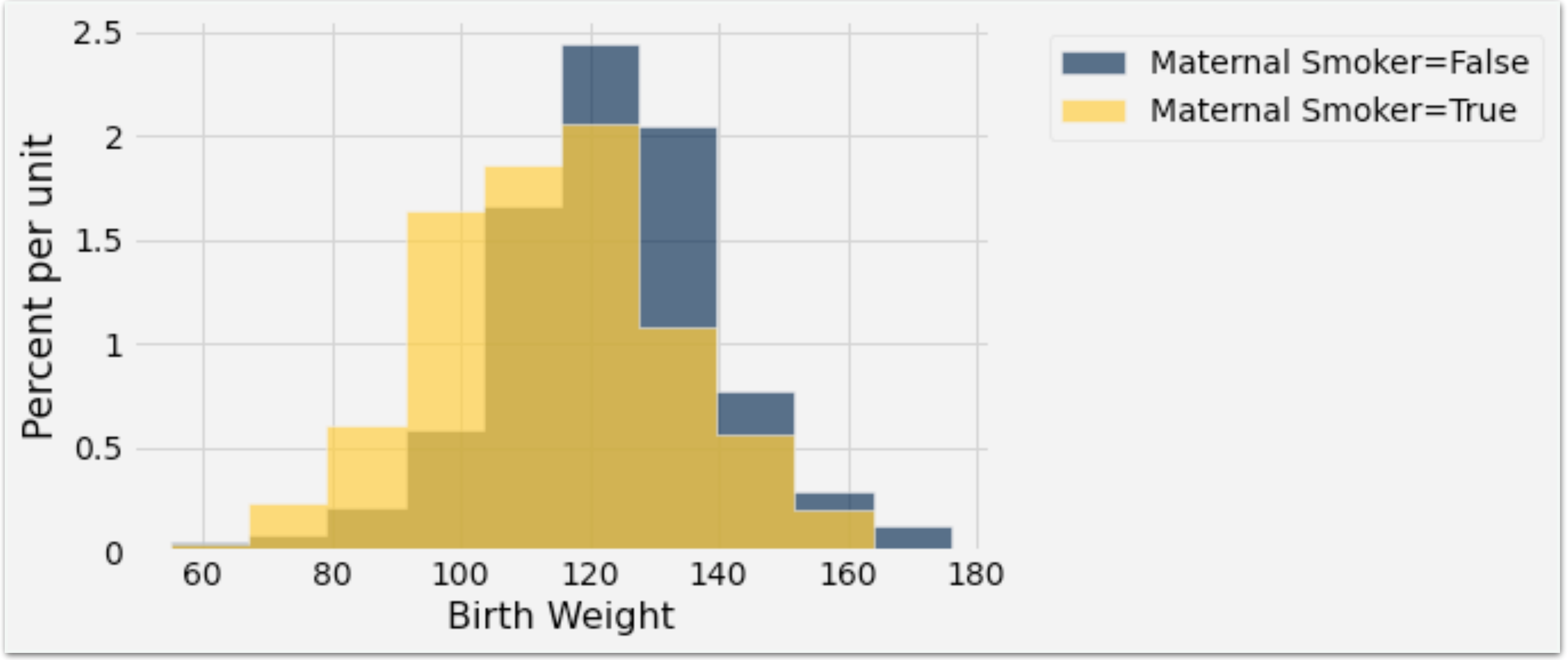
- Used when we want to compare two random samples with one another (from Group A and Group B)
    - Examples:
        - Outcomes in a medical trial (treatment / control group)
        - Outcomes of two different versions of a website
- Underlying question:
    - Do the two sets of values come from the same underlying distribution?

A/B Testing

- Testing whether Group A and Group B have the same underlying distribution or not
    - Null Hypothesis: The distributions of [test statistic] from both groups are the same
        - Any differences we observe are due to chance
    - Alternative Hypothesis: The distributions are different
- If the distributions look different, it supports the alternative hypothesis

A/B Testing Example: Birth Weight

- Going back to our example:
  - Group A: Mothers who smoked during pregnancy
  - Group B: Mothers who didn't smoke during pregnancy



Question: Can the difference in birth weight be due to chance alone?

A/B Testing Example: Birth Weight

Question: Can the difference in birth weight be due to chance alone?

- Null Hypothesis:

A/B Testing Example: Birth Weight

Question: Can the difference in birth weight be due to chance alone?

- Null Hypothesis: In the population, the distribution of birth weights of babies from both groups are the same.
    - That is, the difference we observe in the sample is due to chance

A/B Testing Example: Birth Weight

Question: Can the difference in birth weight be due to chance alone?

- Null Hypothesis: In the population, the distribution of birth weights of babies from both groups are the same.
    - That is, the difference we observe in the sample is due to chance
- Alternative:

A/B Testing Example: Birth Weight

Question: Can the difference in birth weight be due to chance alone?

- Null Hypothesis: In the population, the distribution of birth weights of babies from both groups are the same.
    - That is, the difference we observe in the sample is due to chance
- Alternative: Babies of mothers who smoke weigh less, on average, than babies of non-smokers

A/B Testing Example: Birth Weight

Question: Can the difference in birth weight be due to chance alone?

- Null Hypothesis: In the population, the distribution of birth weights of babies from both groups are the same.
    - That is, the difference we observe in the sample is due to chance
- Alternative: Babies of mothers who smoke weigh less, on average, than babies of non-smokers
- Test statistic:

A/B Testing Example: Birth Weight

Question: Can the difference in birth weight be due to chance alone?

- Null Hypothesis: In the population, the distribution of birth weights of babies from both groups are the same.
    - That is, the difference we observe in the sample is due to chance
- Alternative: Babies of mothers who smoke weigh less, on average, than babies of non-smokers
- Test statistic: Difference between average weights
    - Difference in averages = (Group B average) - (Group A average)

How to simulate differences between 2 groups?

Non-Smoker

120 oz

Non-Smoker

113 oz

Smoker

128 oz

…

Smoker

108 oz

Null Hypothesis: the distribution of birth weights of babies from both groups are the same.

Shuffling Labels Under the Null

Smoker

120 oz

Non-Smoker

113 oz

Non-Smoker

128 oz

…

Smoker

108 oz

Null Hypothesis: the distribution of birth weights of babies from both groups are the same.

Simulating Under the Null

- If the null hypothesis is true, all rearrangement of labels are equally likely
- **Permutation Test**:
    - Shuffle all group labels
        - Keep the sizes of Group A and Group B same as before, but mix which weights fall into Group A and Group B
    - Find the difference between the average of two shuffled groups
    - Repeat

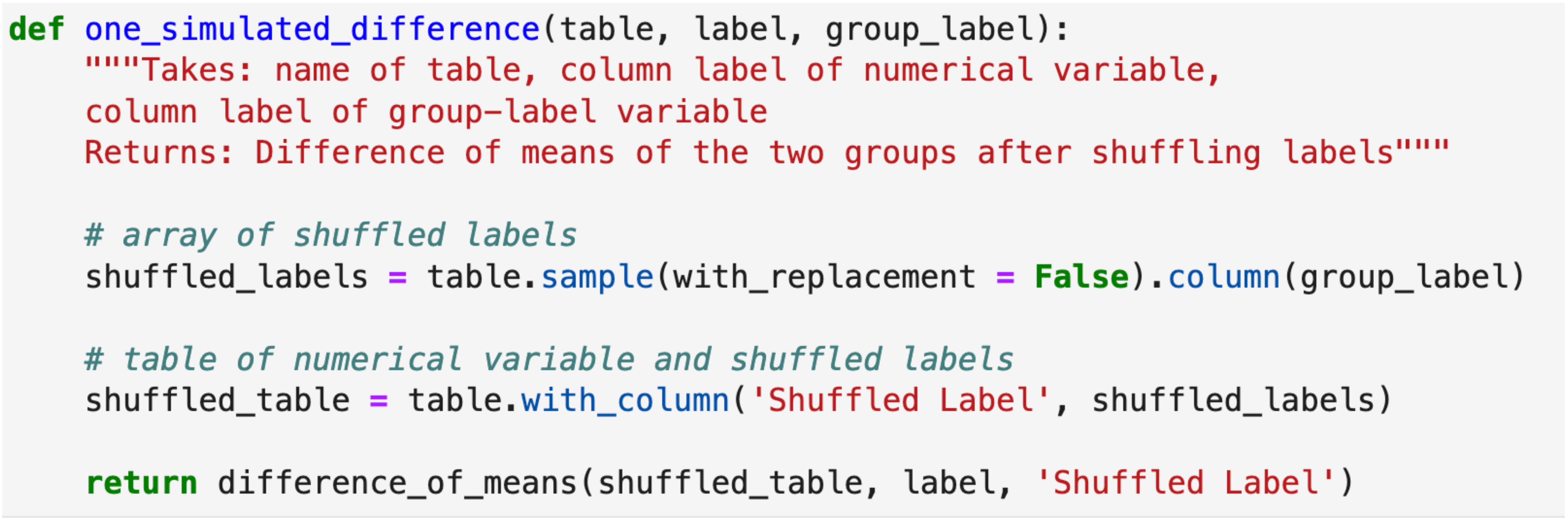
Shuffling with Random Permutation

- `tbl.sample()`
  - Table with same number of rows as original tbl, picked randomly with replacement
- `tbl.sample(n)`
  - Table of n rows picked randomly with replacement
- `tbl.sample(n, with_replacement = False)`
  - Table of n rows picked randomly without replacement
- `tbl.sample(with_replacement = False)`
  - All rows of tbl, in random order

Birth Weight Notebook Demo

A/B Testing Process

1. Write a function that calculates the test static for one simulation



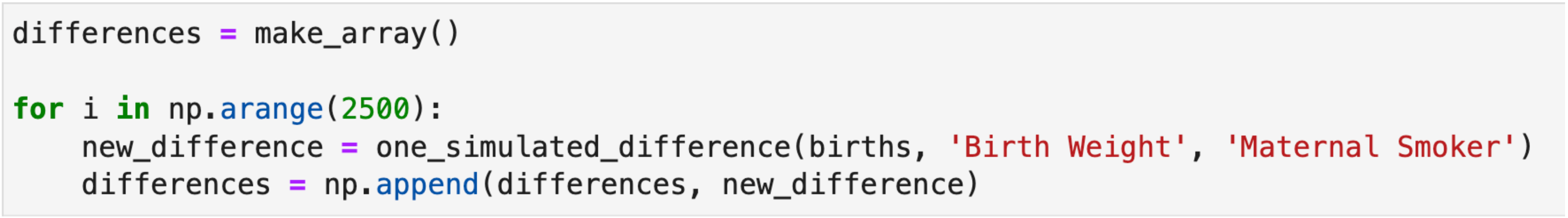
```python
def one_simulated_difference(table, label, group_label):
    """Takes: name of table, column label of numerical variable,
    column label of group-label variable
    Returns: Difference of means of the two groups after shuffling labels"""

    # array of shuffled labels
    shuffled_labels = table.sample(with_replacement = False).column(group_label)

    # table of numerical variable and shuffled labels
    shuffled_table = table.with_column('Shuffled Label', shuffled_labels)

    return difference_of_means(shuffled_table, label, 'Shuffled Label')
```

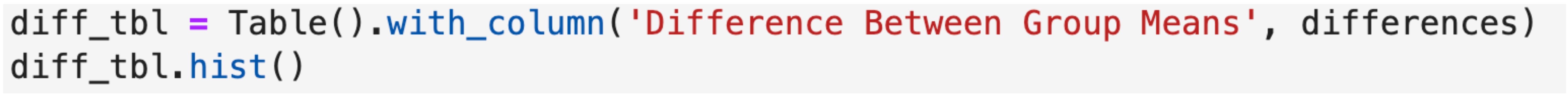
2. Repeat that process in a for loop many times



```python
differences = make_array()

for i in np.arange(2500):
    new_difference = one_simulated_difference(births, 'Birth Weight', 'Maternal Smoker')
    differences = np.append(differences, new_difference)
```

3. Plot the distribution and compare to our observed value

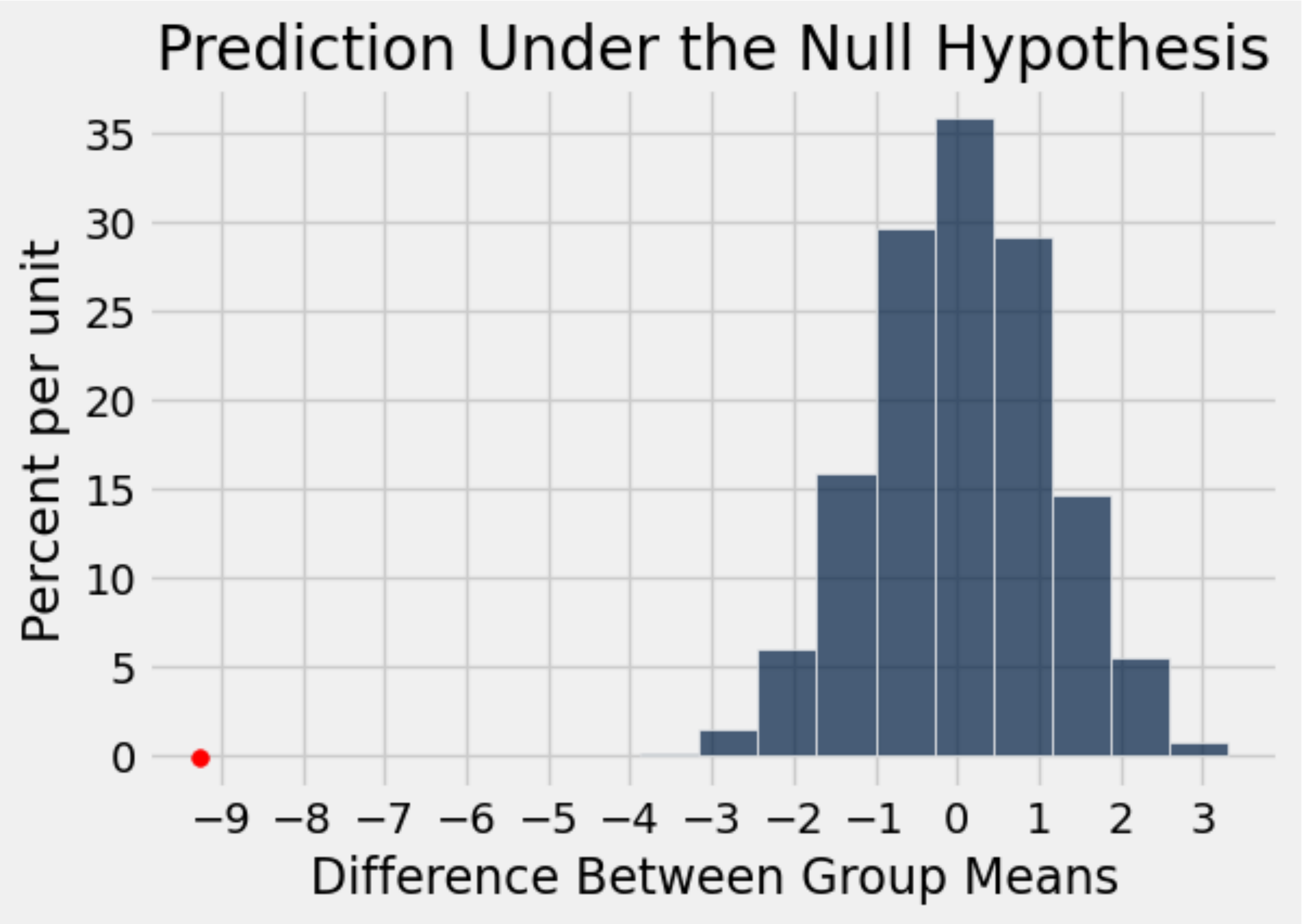


```python
diff_tbl = Table().with_column('Difference Between Group Means', differences)
diff_tbl.hist()
```

Birth Weight Conclusion

A p-value of 0.0 supports the alternative hypothesis

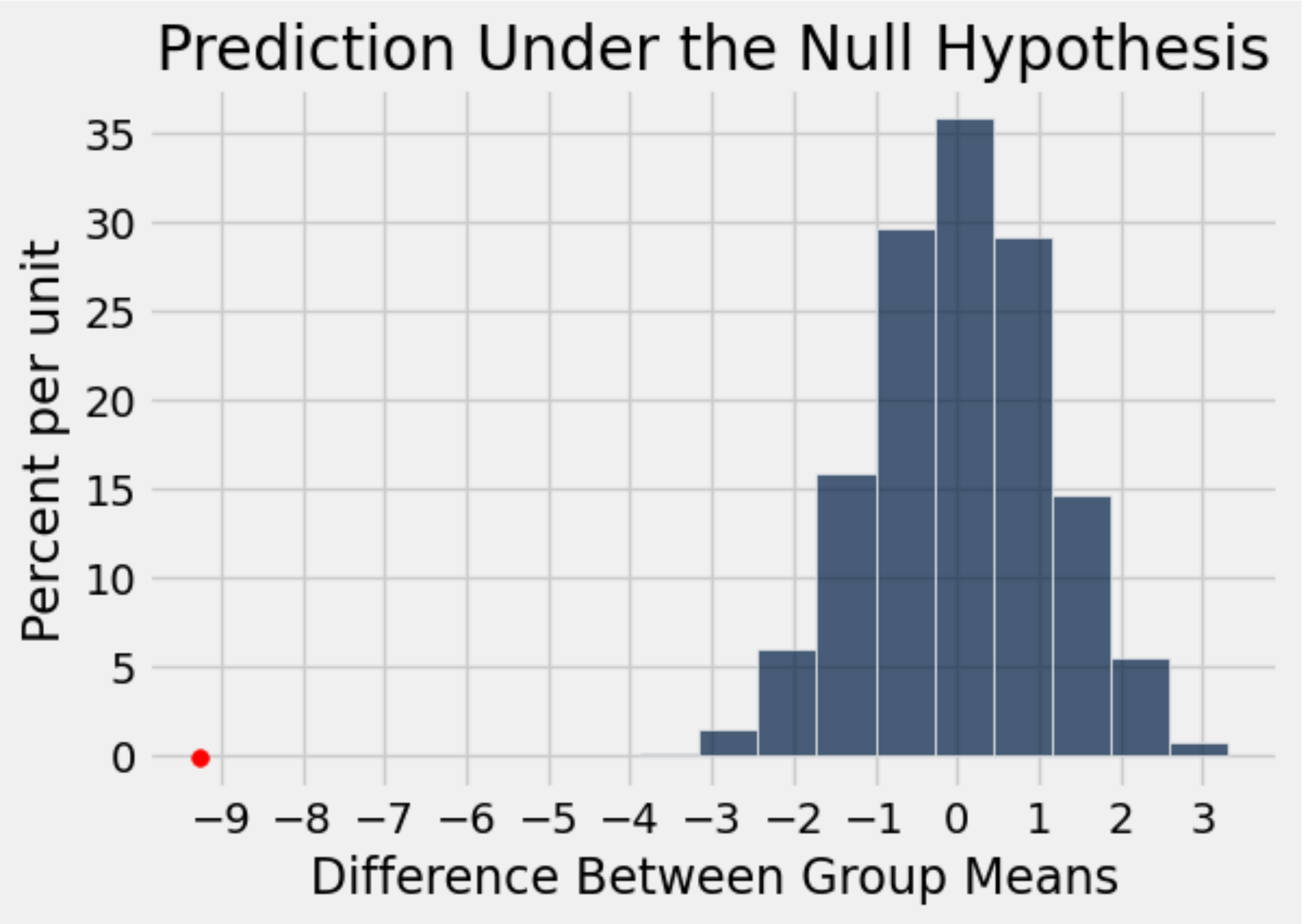
- Babies from smoking mothers weigh significantly less than babies from non-smoking mothers



Birth Weight Conclusion

A p-value of 0.0 supports the alternative hypothesis

- Babies from smoking mothers weigh significantly less than babies from non-smoking mothers



Question: Can we say that smoking causes lower birth rates? (Causation)

Observational Data vs Randomized Control Experiment

- Question: Can we say that smoking causes lower birth rates? (Causation)

- In data science, the gold standard for determining causation is a *randomized control experiment*

  - Group A: control group

  - Group B: treatment group

  - Participants are **randomly assigned to the groups**

- For observational data (e.g., our Maternal Smoking example) we can claim association but not causation

Confidence Intervals

Textbook Chapter on Confidence Intervals
Textbook Chapter on Using Confidence Intervals

Confidence Intervals

- A measure of uncertainty

- Lets us determine how different our sample is from the population (not just whether or hypothesis test is right / wrong)

- Based on the notion of percents, so helpful to understand what we mean by a percentile

Percentile

- The pth percentile of a collection is the smallest value in the collection that is *at least* as large as p% of all the values
- Example: Suppose you have an array of values [12, 17, 6, 9, 7].
    - What is the 80th percentile for this array?
        - Value at least as large as 4/5 of the five elements
- `percentile(p, values_array)`
    - Returns the smallest value in `values_array` that is at least as large as `p`% of the elements in the array

Estimation

Estimation

- How do you calculate the value of an unknown parameter?
- If you have the entire population: calculate it directly
- If you don’t have the entire population:
    - Take a random sample from the population
    - Use a **statistic** as an **estimate** of the parameter

Notebook Demo: SF Gov’t Salaries

Quantifying Uncertainty

- Our estimate depends on the sample we collected. How can we determine how accurate our estimate is?
    - In theory, we could collect a different sample and check how similar the statistic we calculated is
- What if we can’t go back and collect more samples?
    - How can we simulate having many random samples (to determine how accurately we can estimate the parameter) when we only have one sample?

Bootstrap Method

The Bootstrap Method

- Suppose we have a large random sample from the population
  - By the Law of Large Averages, it probably resembles the population from which it's drawn
  - We can replicate sampling from the population by *sampling from the sample*

The Bootstrap Method

- Re-sampling many times from the original random sample ≈ sampling from the population with high probability (Law of Averages)
    - To re-sample, we draw at random from the original sample with replacement to obtain the same sample size
    - The distribution we get from the bootstrap is the empirical distribution of the original sample
- This only works if the original sample is large enough (e.g., wouldn’t work for small samples)

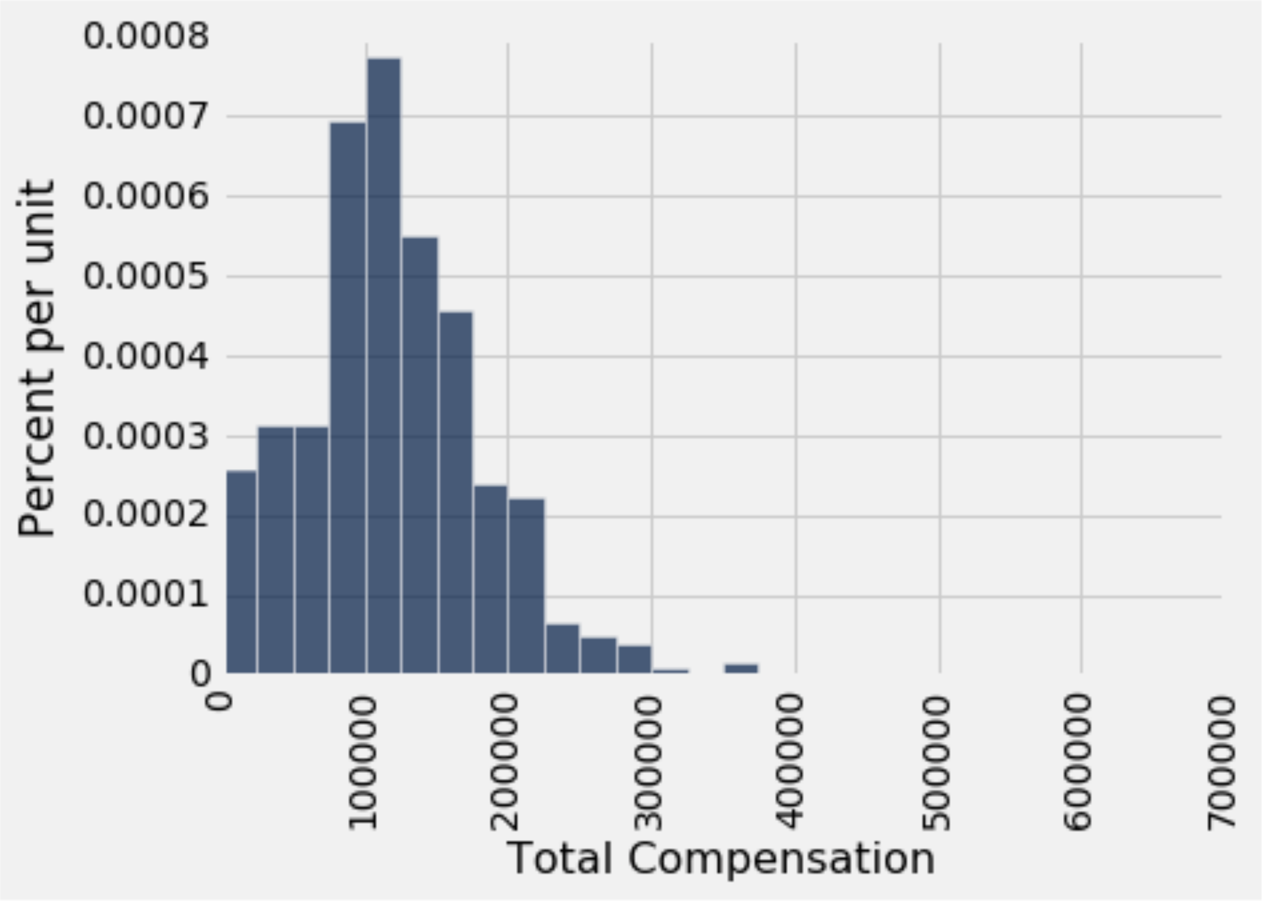
The Bootstrap

Population

?

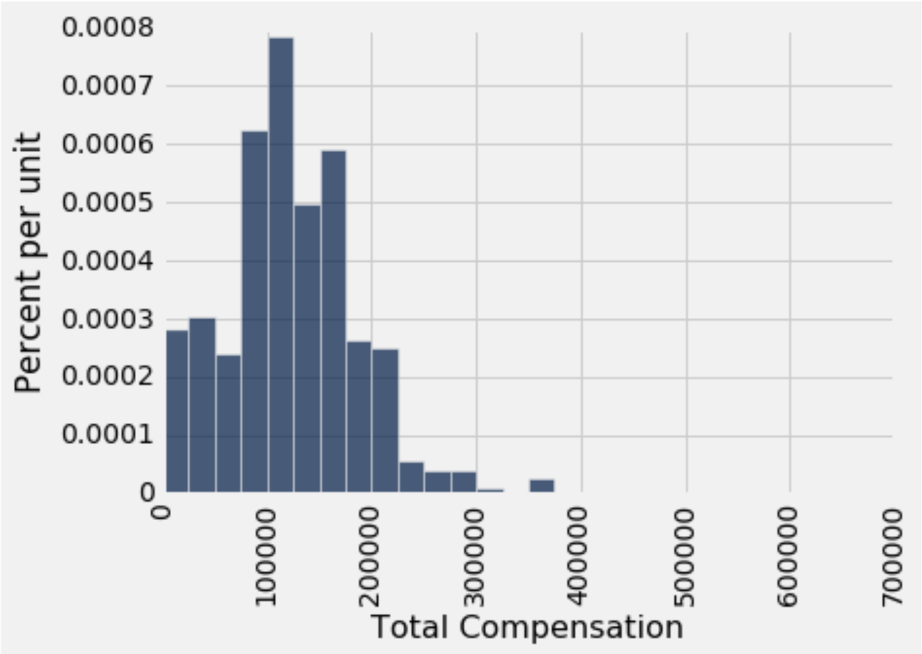
We don’t know the entire population and thus can’t calculate the parameter directly

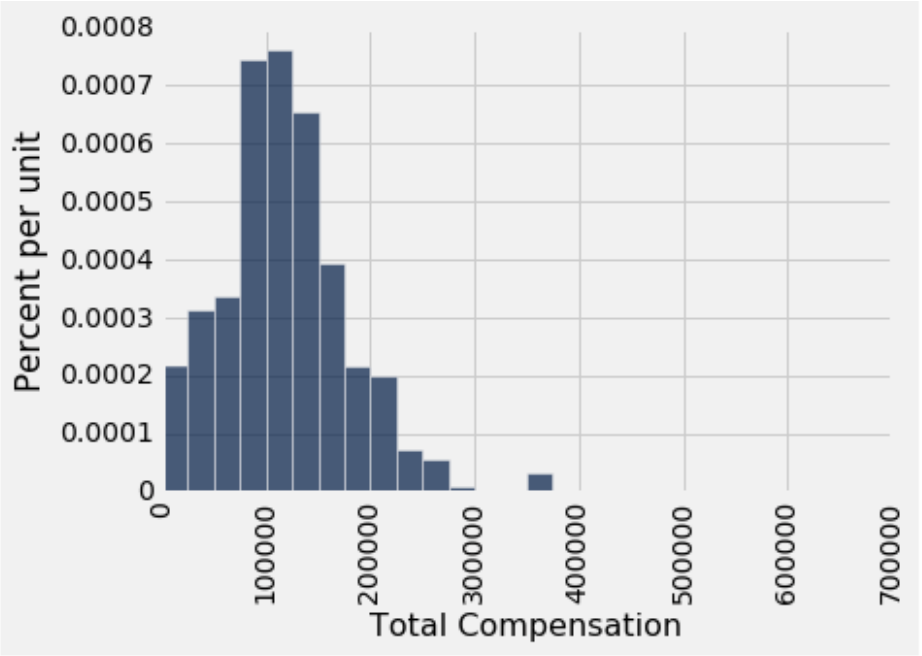
Sample

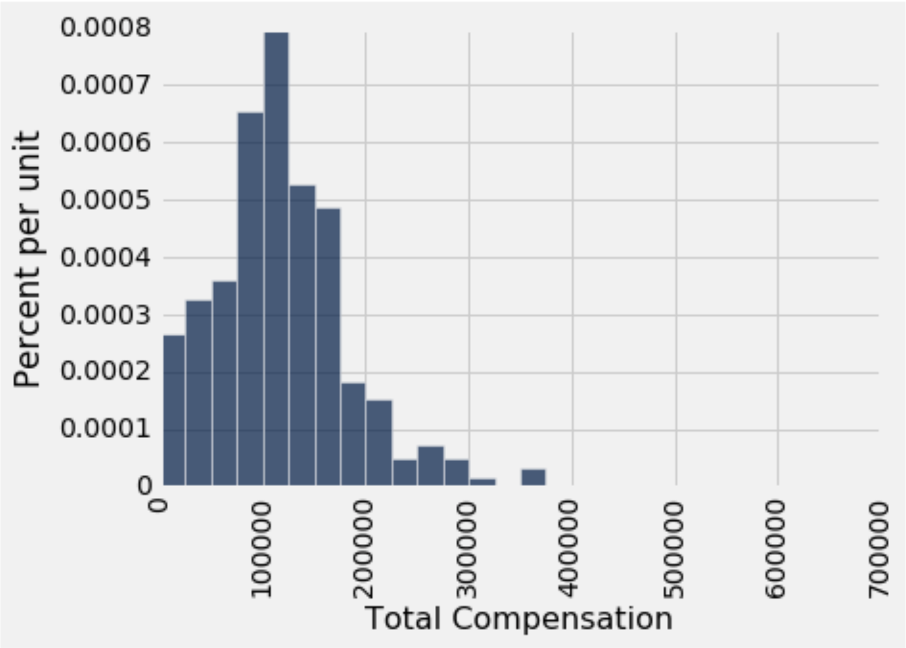


However, we can take a single sample…

Resamples







…and generate lots of resamples

Interpreting Confidence Interval

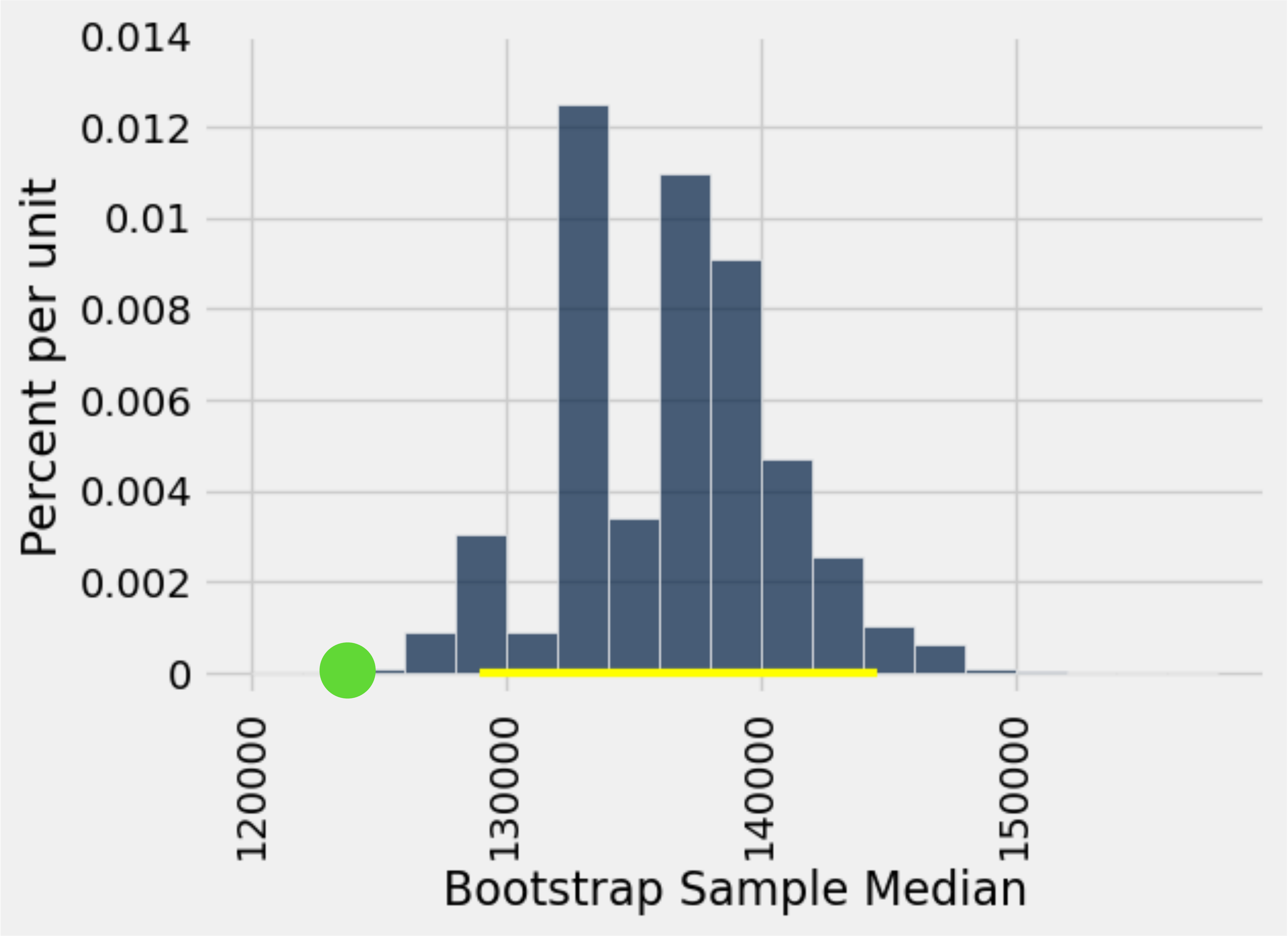
- The confidence interval helps us state how confident we are that our sample statistic estimates / contains the true population parameter

Relationship between Confidence Interval and P-value

- If our threshold (significance level) for p-value is p%
    - We construct a 100-p% confidence interval

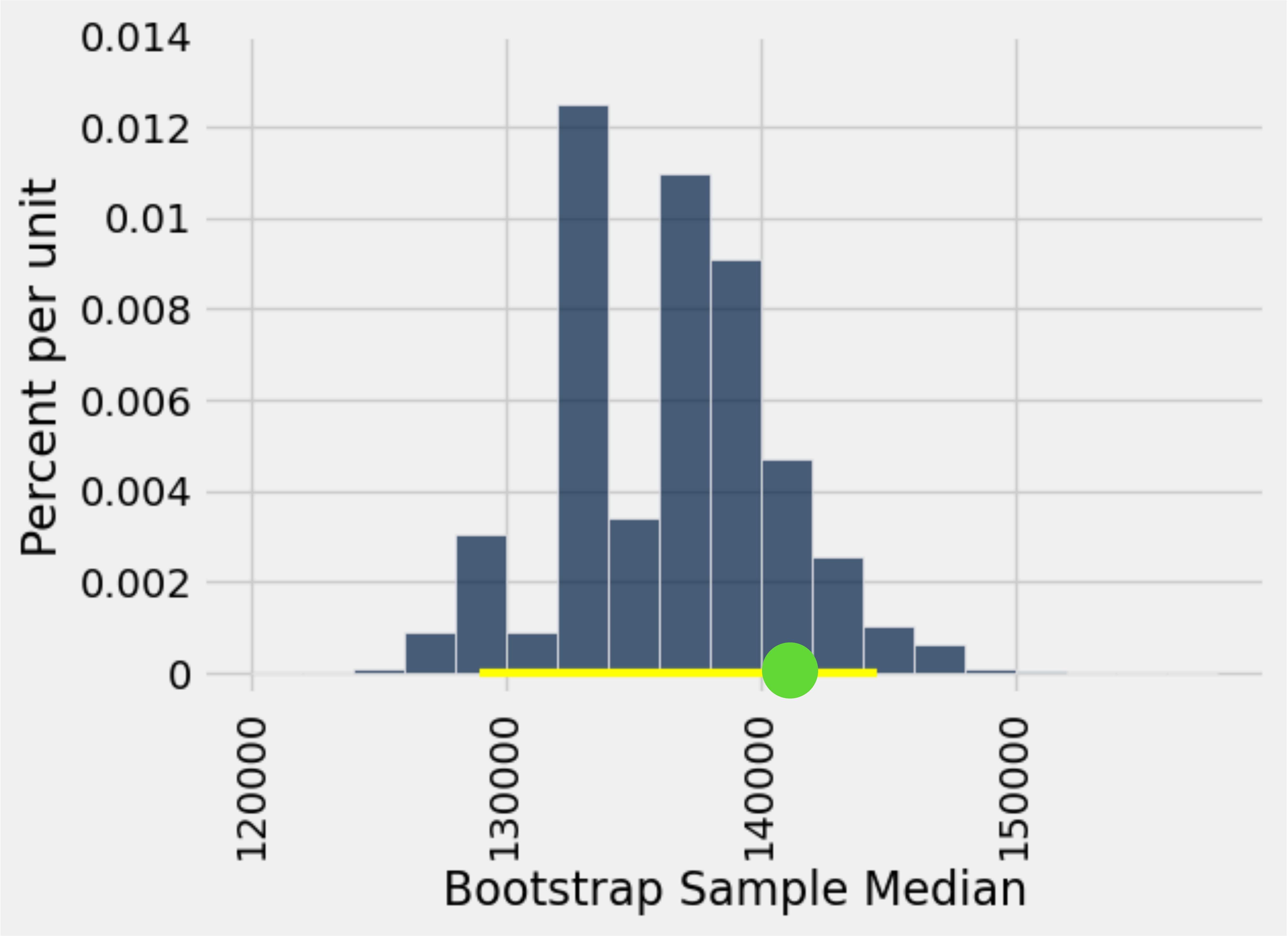
- Example:
    - Null hypothesis - Population average = x
    - Alternative Hypothesis - Population average ≠ x
    - If x is not in our 100-p% interval, then we reject the null

Rejecting the null



Our x is outside the 95% confidence interval

Cannot reject null



Our x is inside the 95% confidence interval

Relationship between Confidence Interval and P-value

- The smaller the p-value threshold, the wider the confidence interval
    - Another way to consider: if you don’t need to be as confident in the result, you can more safely say that the values fall into a narrower range

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for % heads is 65%

Let's say your 95% confidence interval is [45, 60]

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for % heads is 65%

Let’s say your 95% confidence interval is [45, 60]

Questions:

1. For a 5% p-value cutoff, can we reject the null?

2. For a 10% p-value cutoff, can we reject the null?

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for % heads is 65%

Let's say your 95% confidence interval is [45, 60]

Questions:

1. For a 5% p-value cutoff, can we reject the null?
    - Yes - 65% is outside our confidence interval
2. For a 10% p-value cutoff, can we reject the null?

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for % heads is 65%

Let's say your 95% confidence interval is [45, 60]

Questions:

1. For a 5% p-value cutoff, can we reject the null?
   - Yes - 65% is outside our confidence interval
2. For a 10% p-value cutoff, can we reject the null?
   - Yes - we expect the confidence interval to be even narrower, so 65% would still be outside the confidence interval

Next time

- Normal Distributions