COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 16: A/B Testing

Spring 2025

March 30, 2026

Project Milestones and Deadlines

This week →

1. **Group Declaration**: Deadline **Wednesday, April 1**
   Please complete one of the two Google Forms to indicate your group preference: Group Declaration (if you know what group you want to be in) or Group Matching (if you do not have a group and would like us to form one for you)

~2.5 weeks away →

2. **Project Proposal**: Due **Friday, April 17** at 11:59pm
   Each group will select a final project notebook and dataset to work on for the final project and complete the Exploratory Data Analysis section. Based on your exploratory data analysis, you will state the hypothesis you are planning to test.

4 weeks away →

3. **Progress Report**: Due **Monday, April 27** at 11:59pm
   At this point, groups should be about ~60% done with the final project. For the progress report, groups should list out what analysis remains and how they plan on approaching it. Additionally, groups should share if they are running into any issues with their analysis that they may need assistance with or have questions about.

~ 5.5 weeks away →

4. **Final Project Report**: Due **Friday, May 8** at 11:59pm
   Groups will submit the completed reports along with a completed peer review.

Note: We will require all students to complete a peer review to share how work was distributed among team members. Any major discrepancies in the distribution of work will be factored into individual grades on this assignment.

Final Project Grading Breakdown:

- Group Declaration - 1%
- Project Proposal - 9%
- Progress Report - 25%
- Final Report - 65%

Mid-semester Feedback

- Thank you everyone who gave feedback!

- To receive extra credit, you were asked to schedule an email to send Friday

    - I received far fewer emails than responses to the feedback form…

    - If you received an “incomplete” on Courseworks for this extra credit, please check you sent the email to claim credit

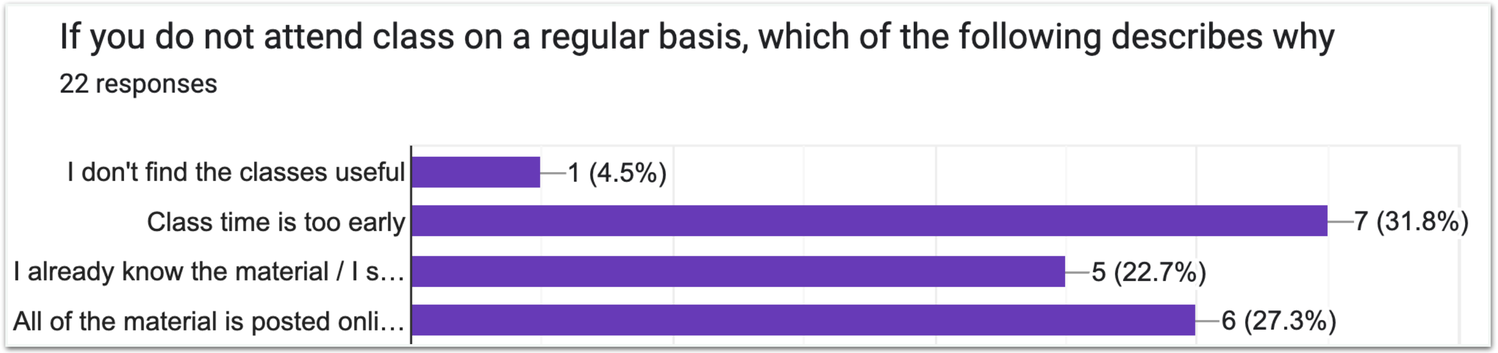
        - Please forward any relevant info to me so I can give you credit!

Mid-Semester Feedback

Common suggestions: (currently trying to implement with varying success)

- More interaction
- Slow down on code demo
- Sync slides and code demo better to reduce repetition

I appreciated the honesty with why people skip class:



Mid-semester Feedback

Unexpected results:

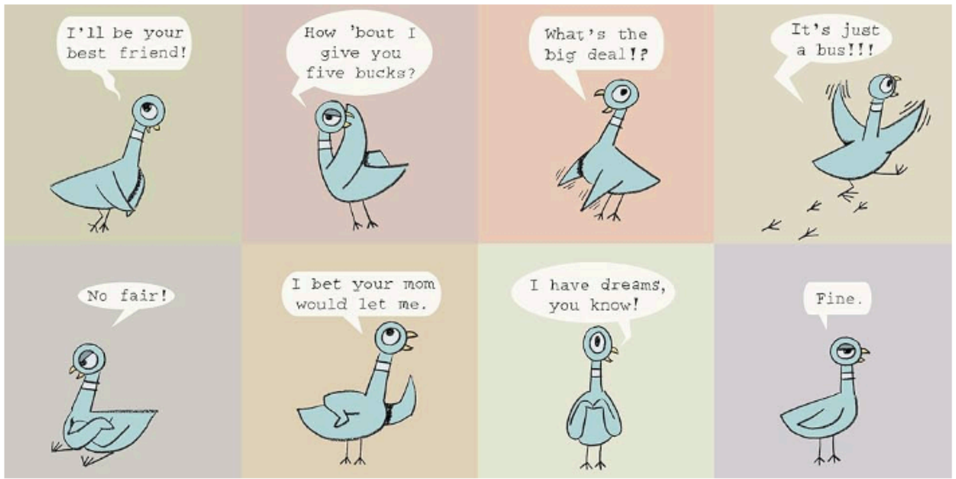
- I’m not the only one who loves pigeons

Hello Professor Lee,

Thank you for the extra credit. I attached a picture of Mr. Pigeon from Miraculous Ladybug.

Hi Prof. Lee,

Hi! I'm the bus driver! Listen, I've got to leave for a little while, so can you watch things for me until I get back? Thanks. Oh, and remember: Don't Let the Pigeon Drive the Bus!



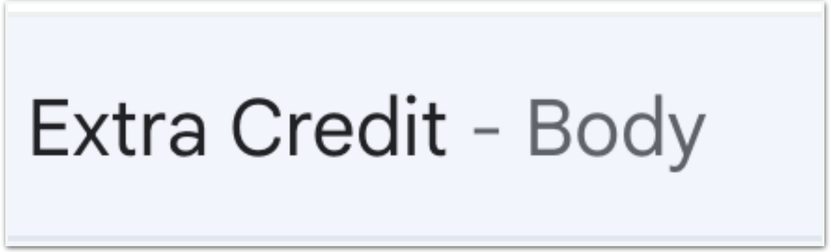
pigeon

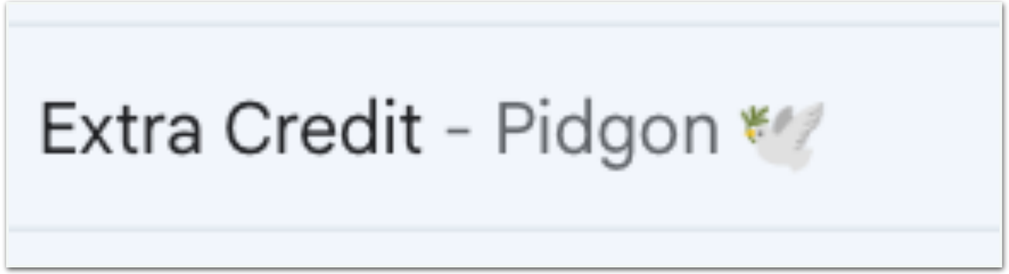
Hi Professor!

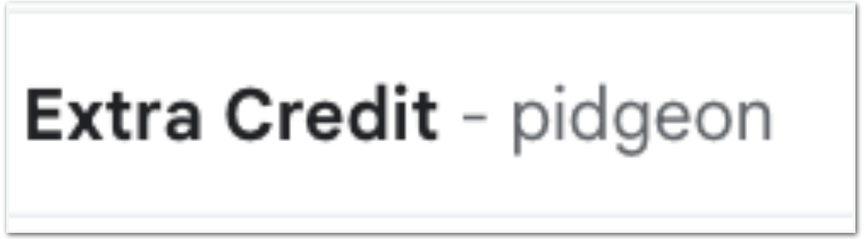
Mid-semester Feedback

Unexpected results:

- I’m not the only one who loves pigeons
- Pigeon is hard to spell







Decisions and Uncertainty

Incomplete Information

- With our hypothesis testing, we're trying to choose between two views of the world based on data in a sample
    - It's not always clear whether the data is consistent with one view or another
    - Random samples can sometimes end up in the extremes
        - This is unlikely but possible

Hypothesis Testing

- **Hypothesis Testing:** A statistical test in which we choose between two potential views
  - **Null Hypothesis:** Clearly defined model based on chance. Data is generated randomly and under clearly specified assumptions
    - This is the one we can simulate and test!
  - **Alternative Hypothesis**: The observed data differs from the null hypothesis in some way other than chance
    - Doesn’t say how or why the model isn’t good, just that it isn’t good

Test Statistic

Test statistic: The statistic we choose to simulate to decide between the two hypotheses

Questions before choosing the statistic:

- What values of the statistic will make us lean towards the null hypothesis?
- What values will make us lean towards the alternative?
  - Preferable the answer should be just a “high” or just a “low” value
  - Try to avoid “both high and low”

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis
  - Draw the histogram of the simulated values
  - The empirical distribution of the statistic under the null hypothesis
- It is a prediction about the statistic made by the null hypothesis
  - It shows all the likely values of the statistic and how likely they are (if the null hypothesis is true)
- The probabilities are approximate based on our random samples

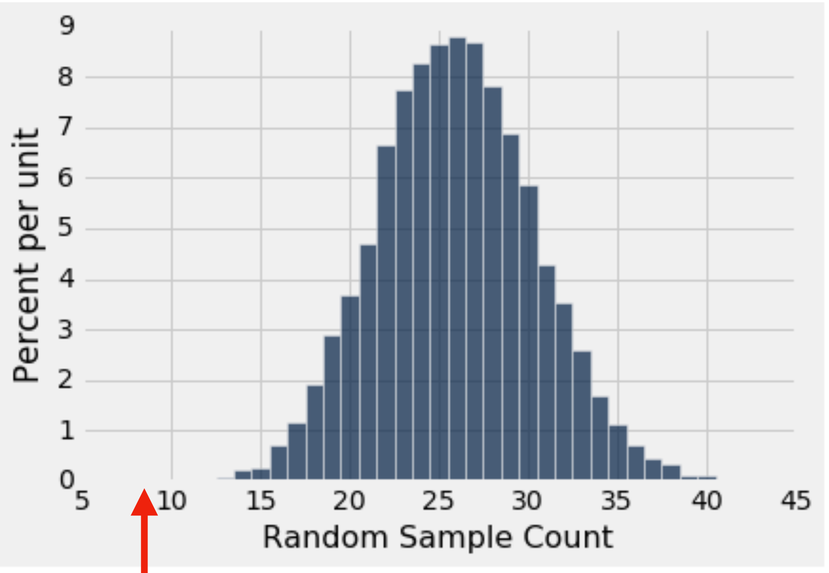
Conclusion of the Test

Resolve choice between null and alternative hypotheses:

- Compare the observed test statistic and its empirical distribution under the null hypothesis
    - If the observed value is *not* consistent with the empirical distribution, the test favors the alternative
- Whether a value is consistent with a distribution:
    - For some tests, a visualization may be sufficient
    - If not, there are conventions about “consistency” with the null

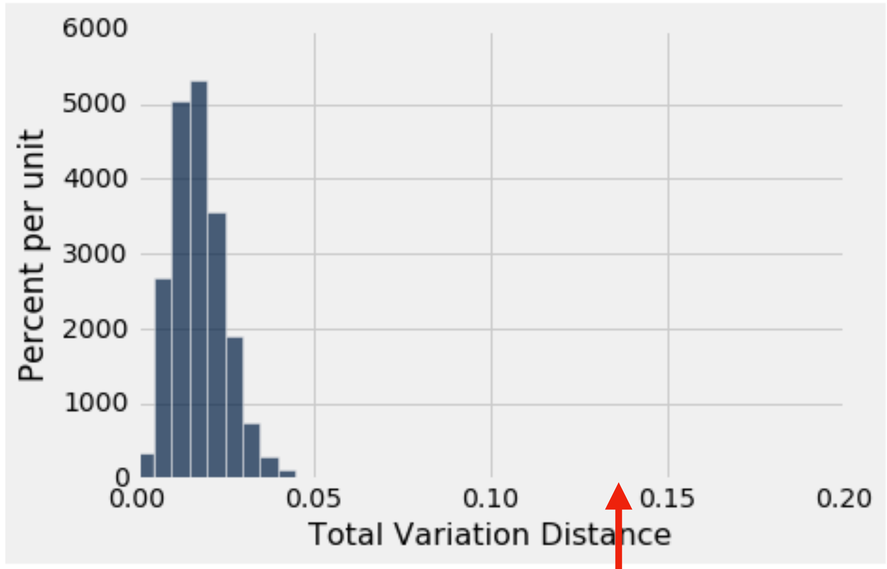
Our Examples So Far

Swain v Alabama



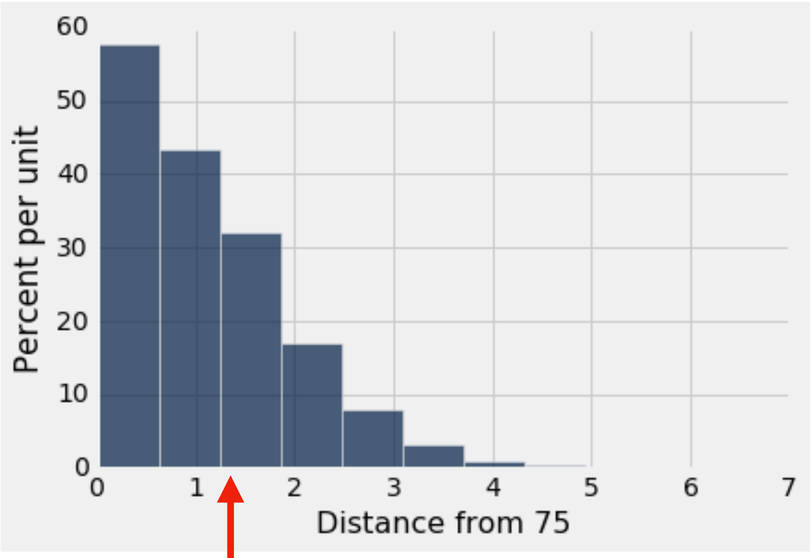
Observed Number (8)

Alameda Jury



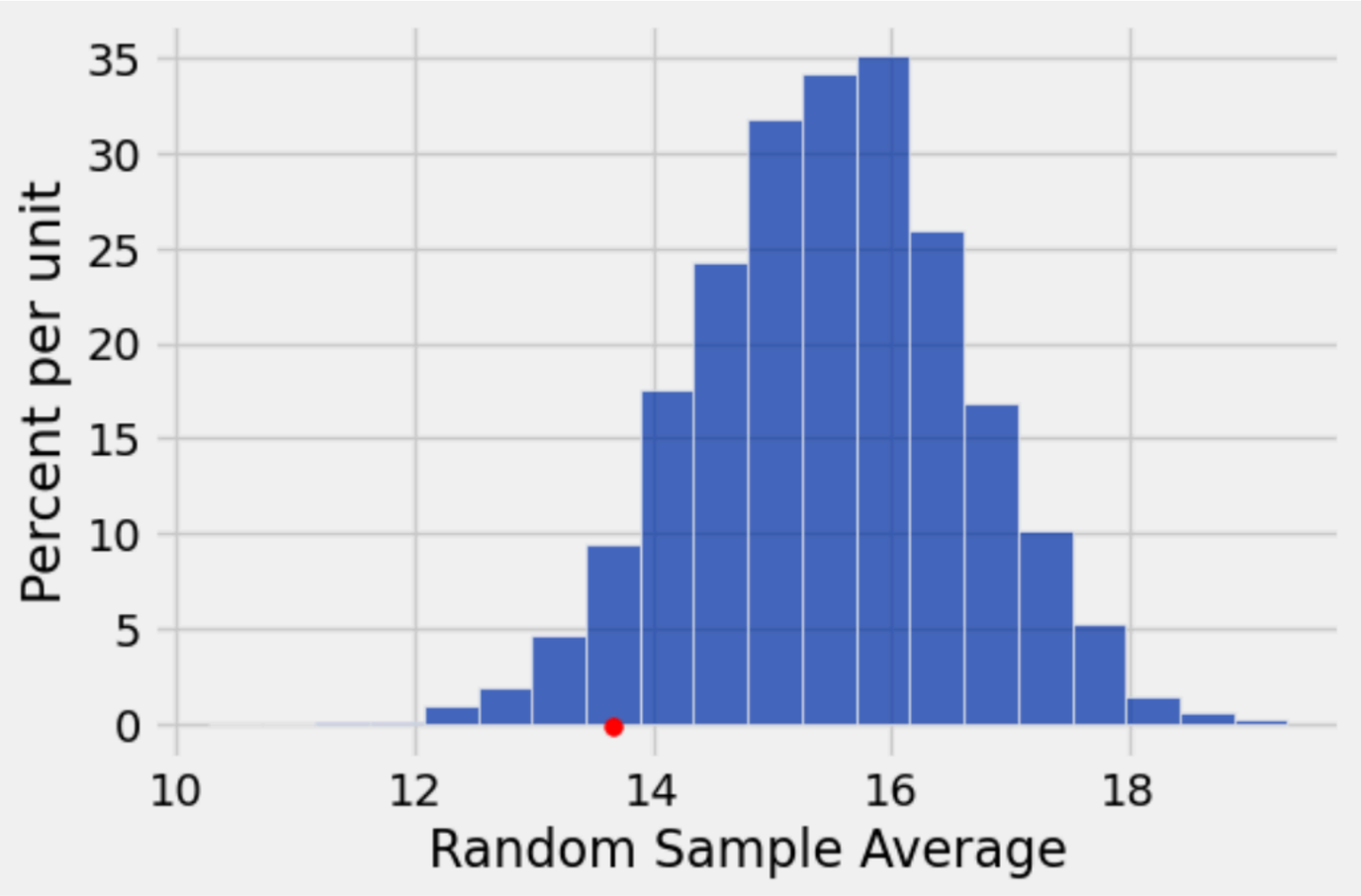
Observed TVD (0.14)

Pea Plants



Observed Distance (1.32)

A Less Clear Example



Definition of the P-Value

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

Definition of the P-Value

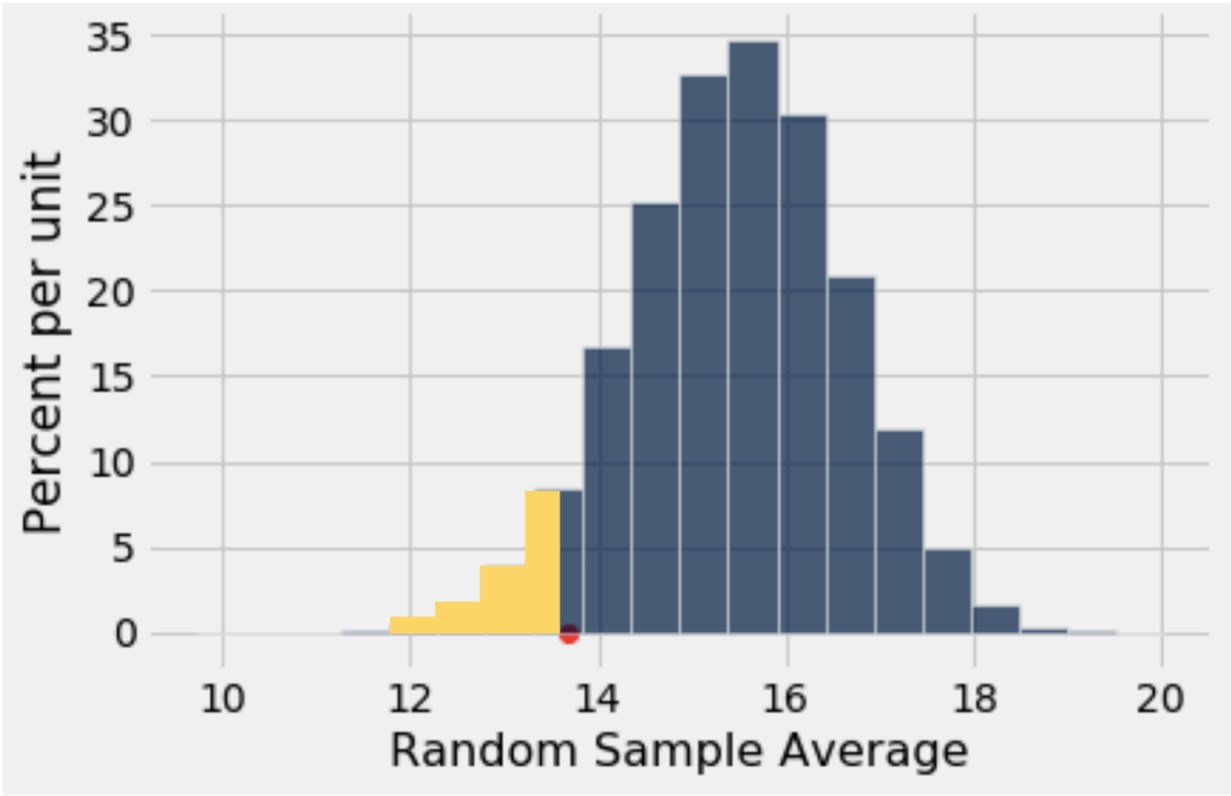
The P-value is the chance

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



Conventions about Inconsistency

- **“Inconsistent with the null”**: test statistic is in the tail of the empirical distribution under the null hypothesis

- **“In the tail”**

  - **< 5% (Area in the tail is less than 5%)**

    - The result is “statistically significant”

  - **< 1% (Area in the tail is less than 1%)**

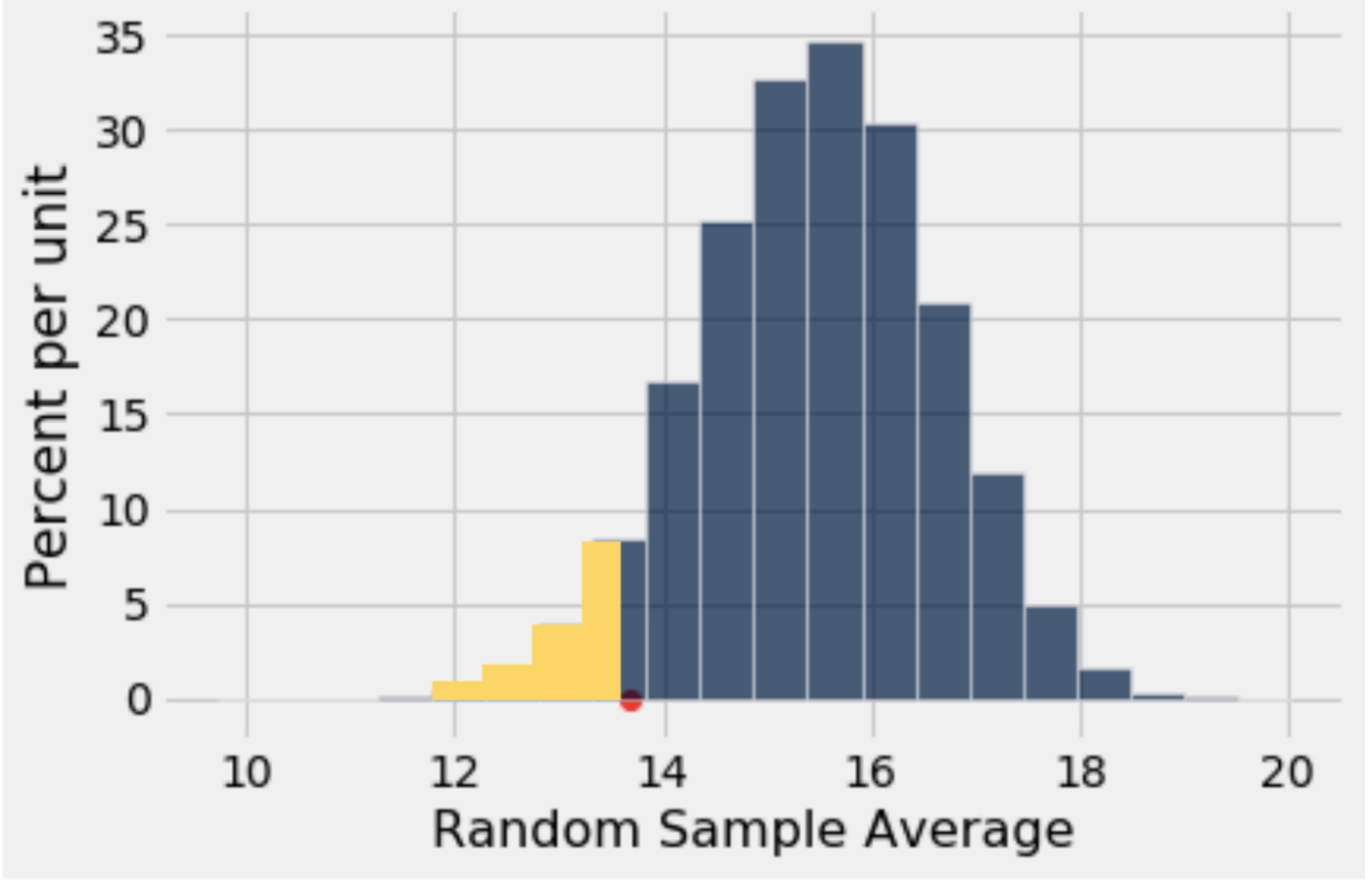
    - The result is “highly statistically significant”

Levels of Statistical Significance

Exam Notebook Demo (continued)

The P-Value as an Area

- Area to the left of of our observed value



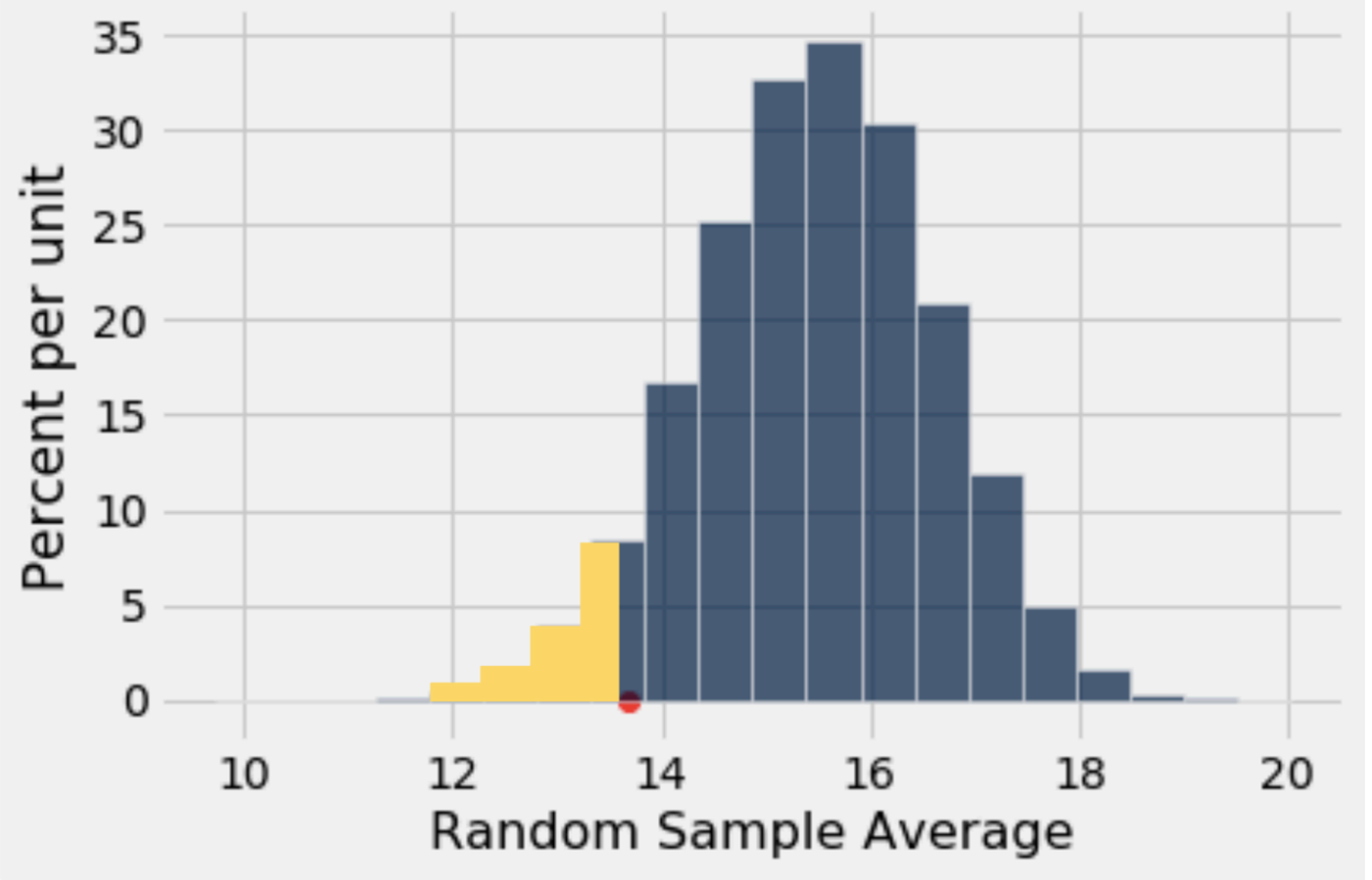
Discussion Questions

What does the red dot represent?

A. Our p-value

B. Our expected outcome

C. Our observed outcome



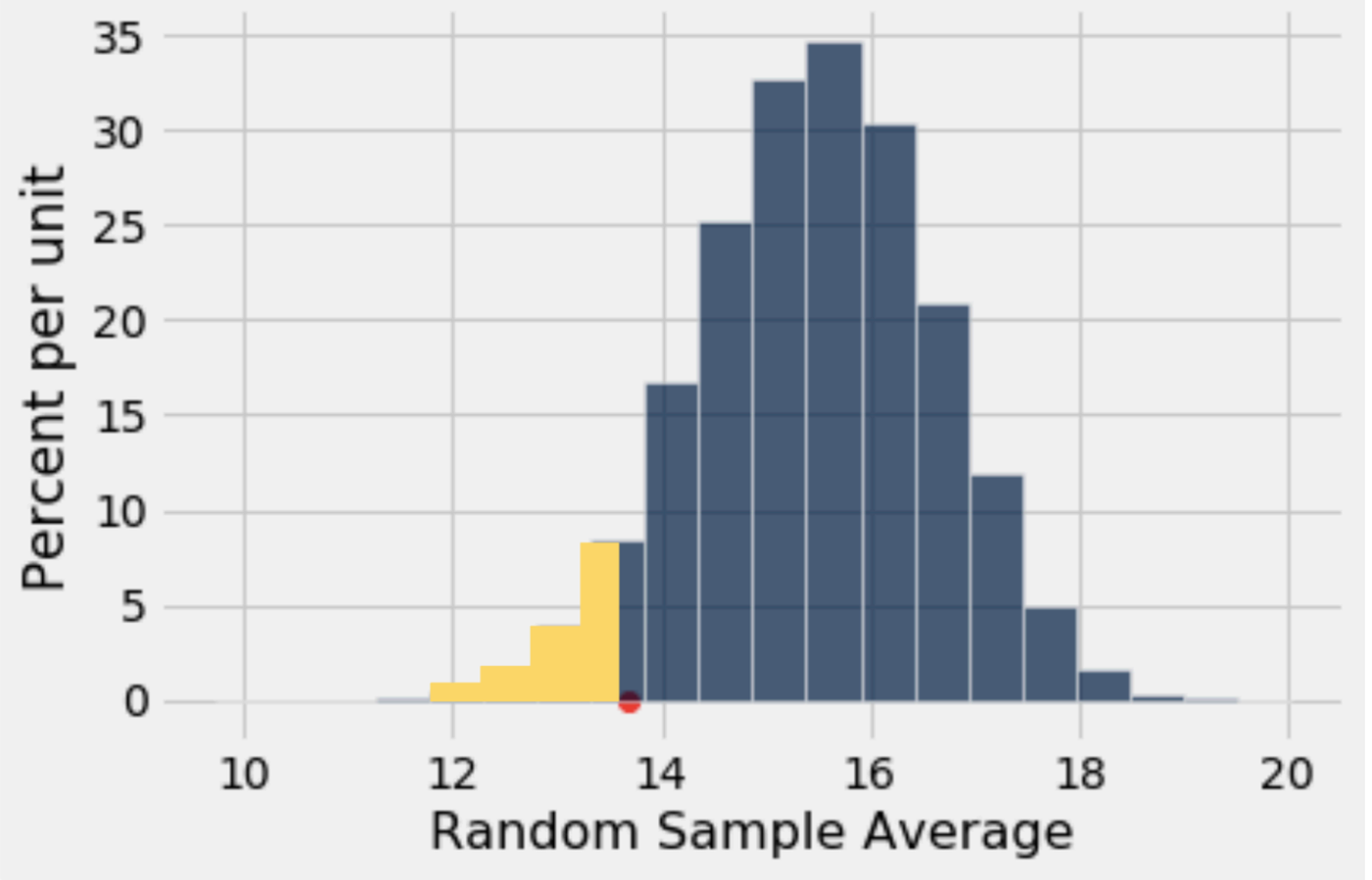
Discussion Questions

What does the red dot represent?

A. Our p-value

B. Our expected outcome

C. Our observed outcome



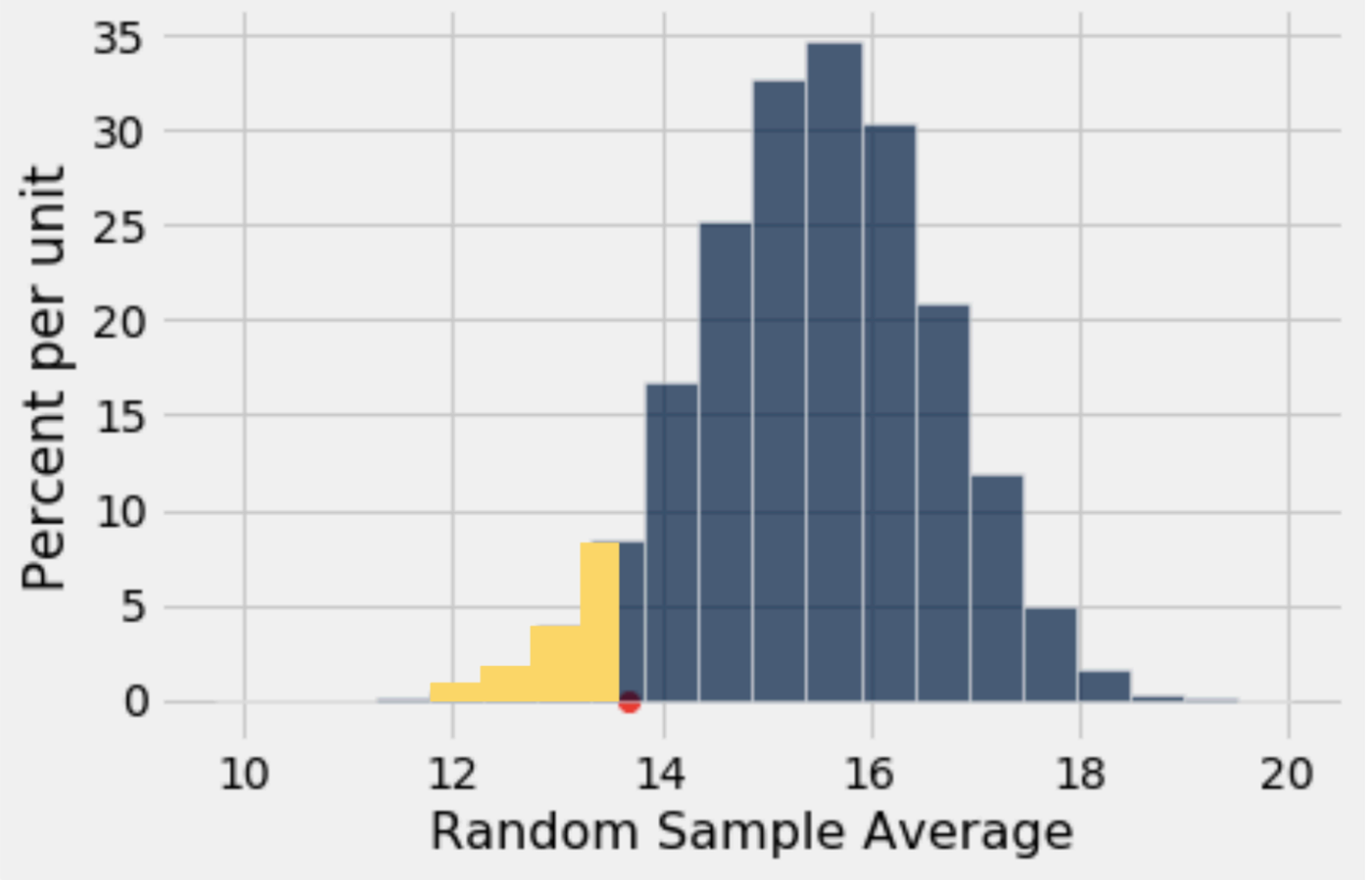
Discussion Questions

What do the yellow bars in this figure represent?

A. The tail of the distribution

B. Our level of statistical significance

C. The probability of getting our observed outcome



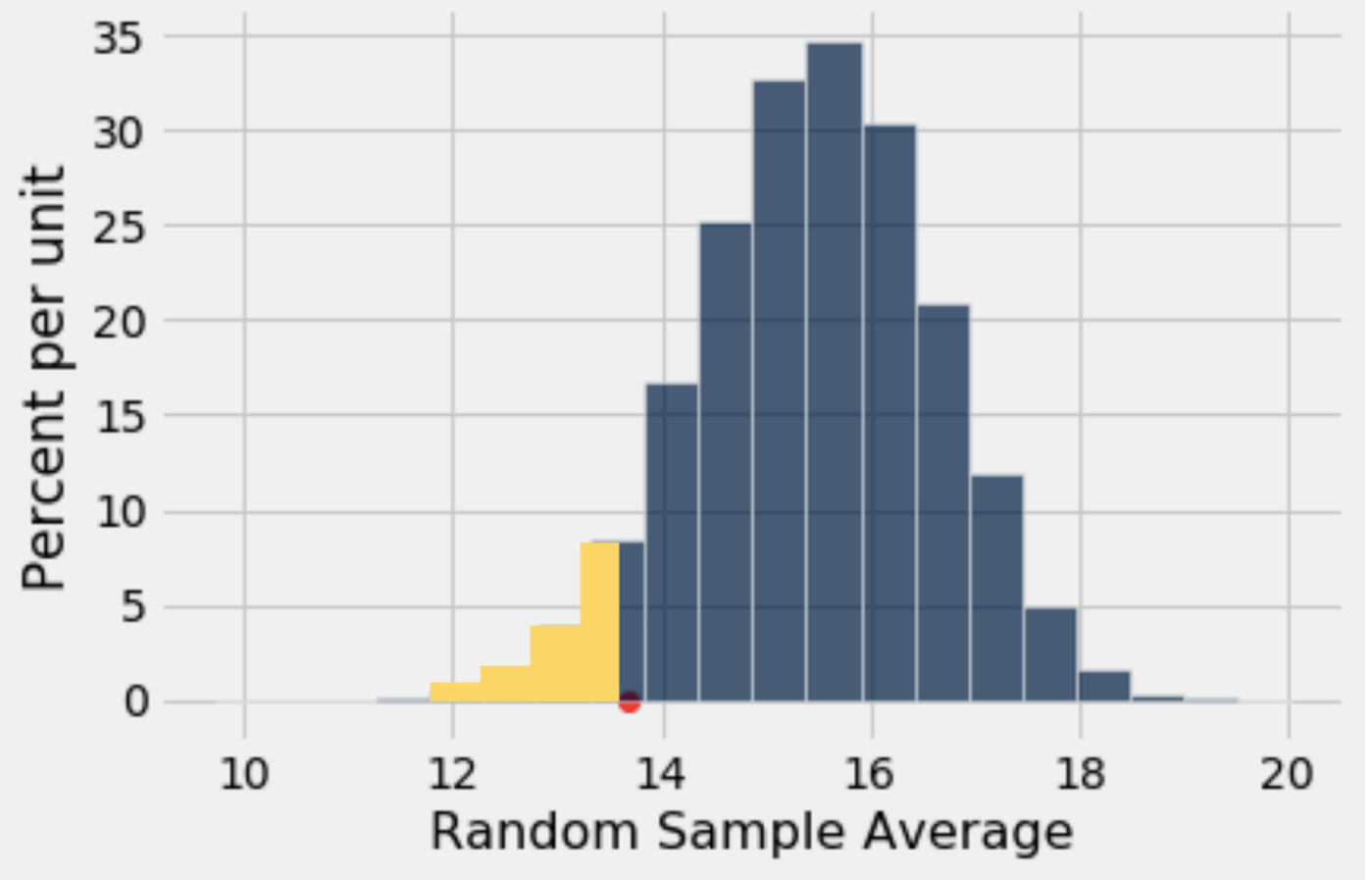
Discussion Questions

What do the yellow bars in this figure represent?

A. The tail of the distribution

B. Our level of statistical significance

C. The probability of getting our observed outcome



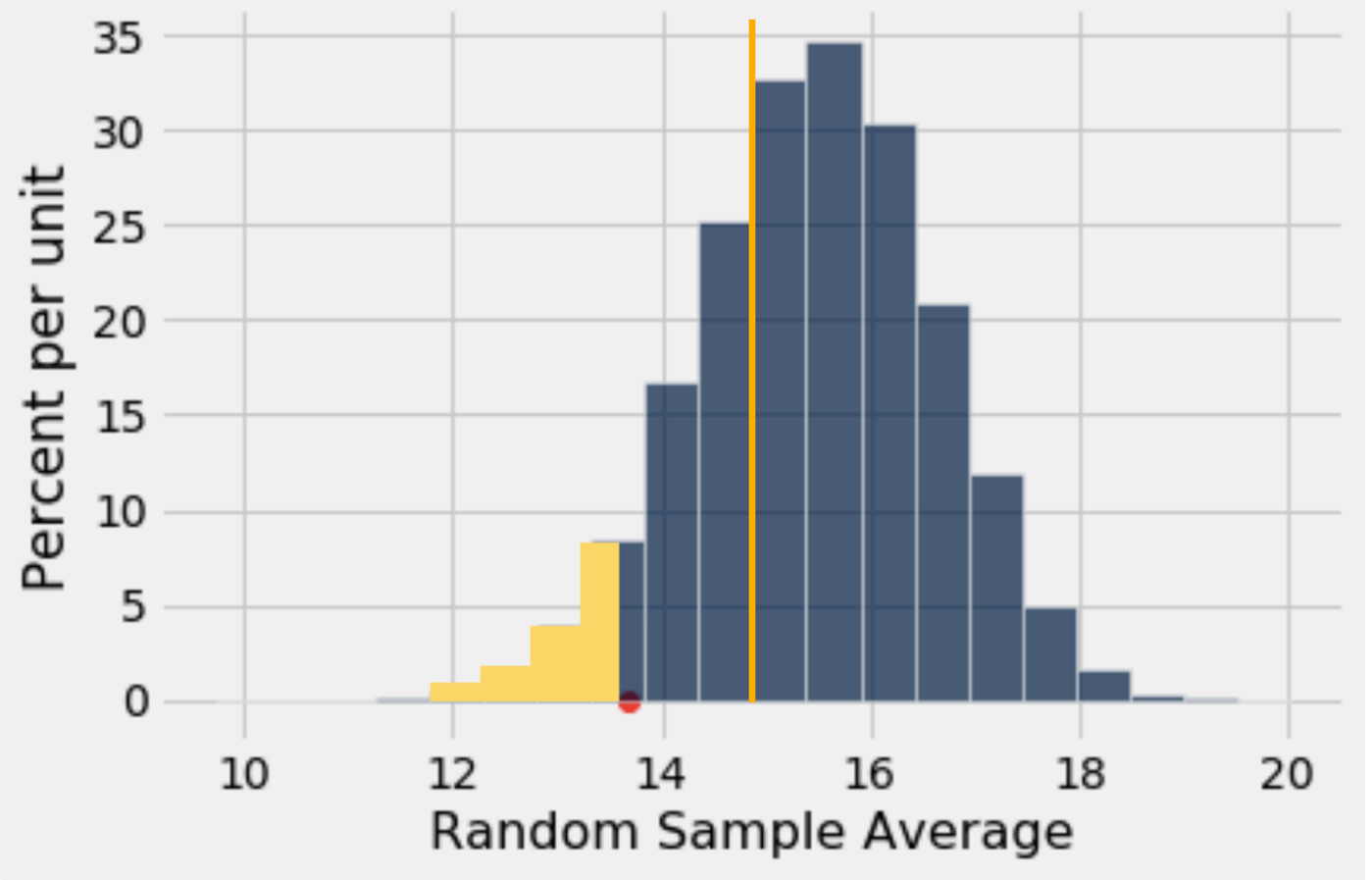
Discussion Questions

Imagine that the yellow vertical bar at ~15 represents our 5% threshold

Which of the following are true:

A. We can reject the null hypothesis, and our result is statistically significant at a 5% threshold

B. We cannot reject the null hypothesis



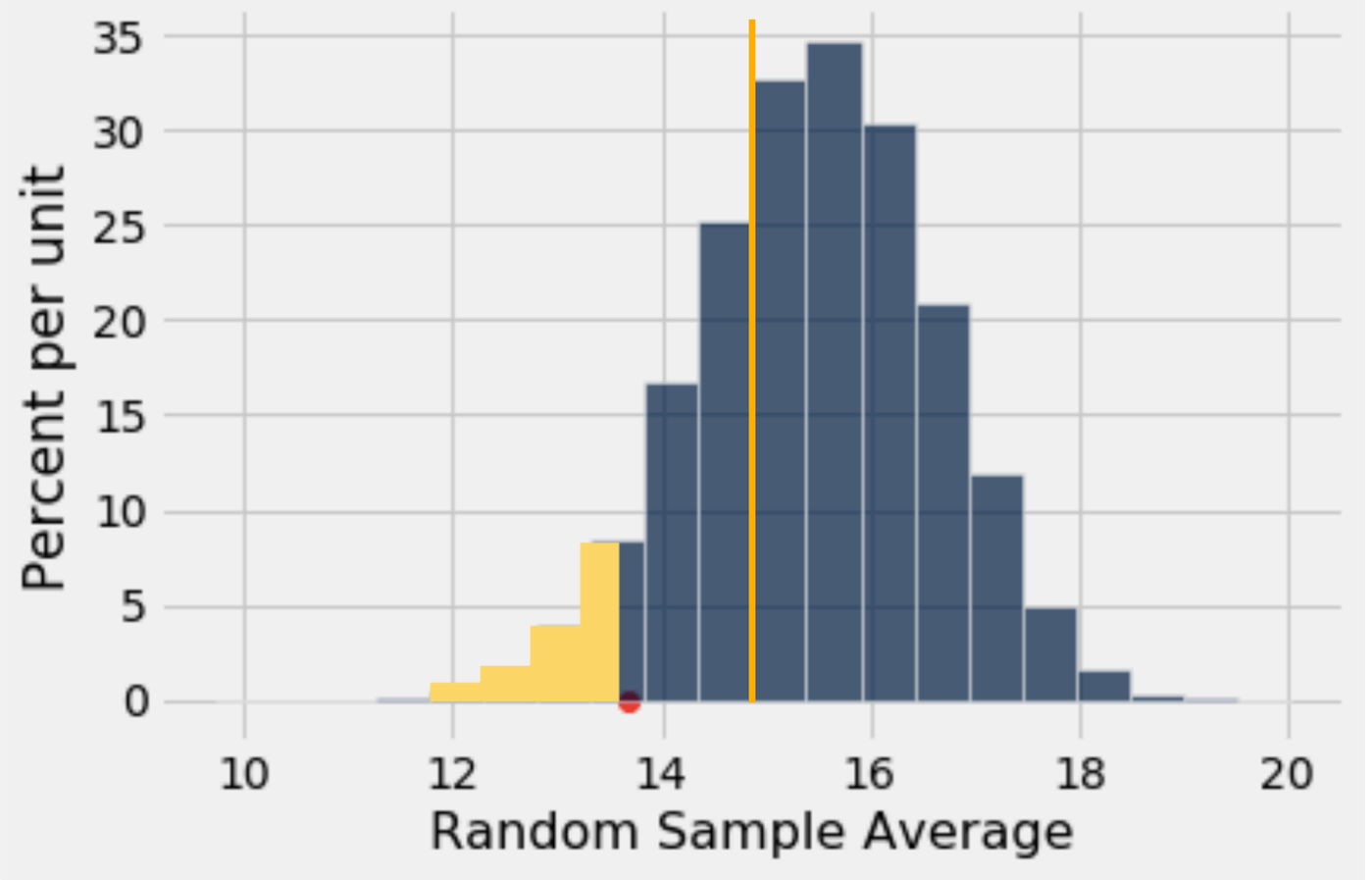
Discussion Questions

Imagine that the yellow vertical bar at ~15 represents our 5% threshold

Which of the following are true:

A. We can reject the null hypothesis, and our result is statistically significant at a 5% threshold

B. We cannot reject the null hypothesis



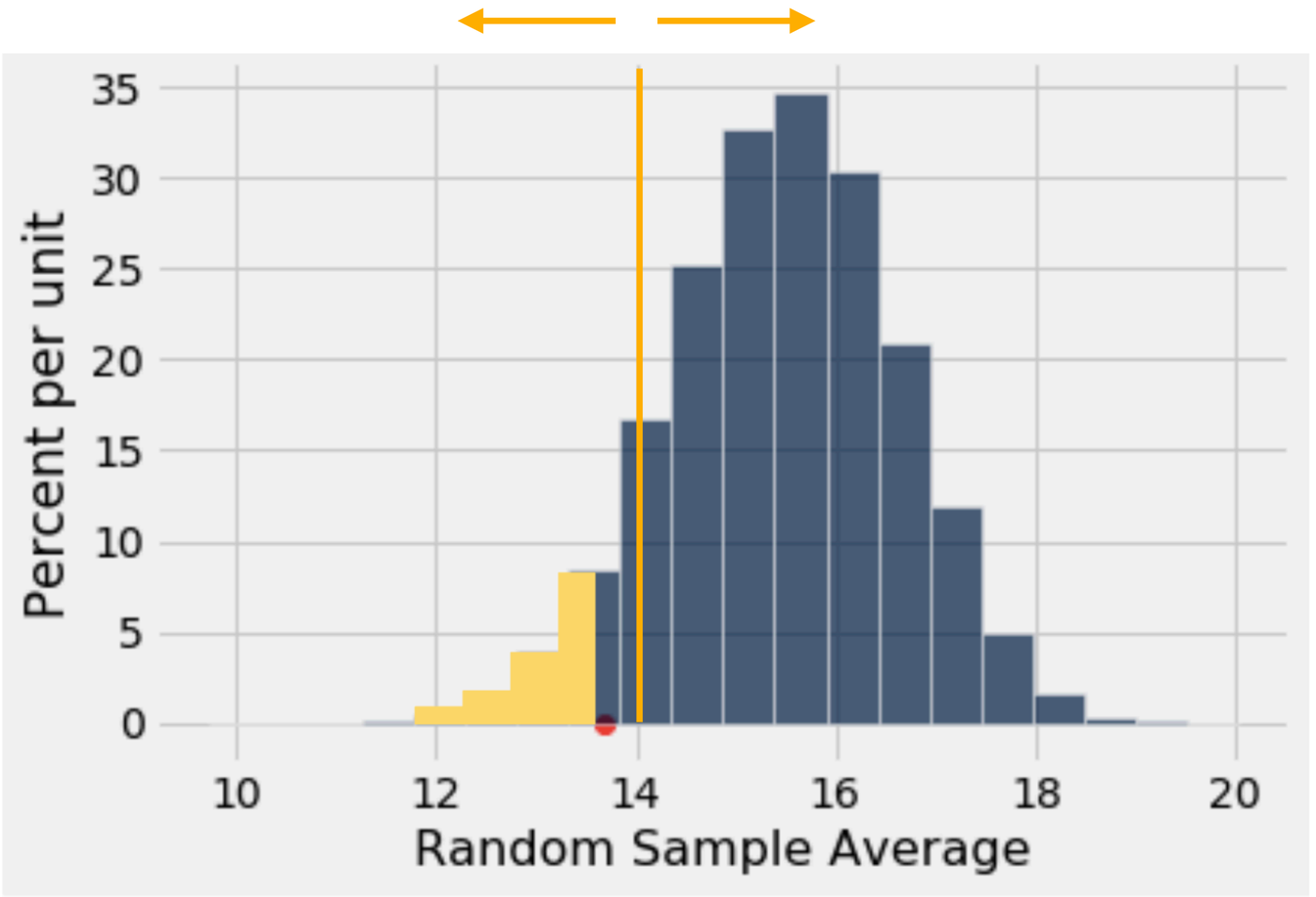
Discussion Questions

Imagine that the yellow vertical bar at ~14 represents our 5% threshold.

Do we expect the 1% threshold to lie:

A. To the left of 14

B. To the right of 14



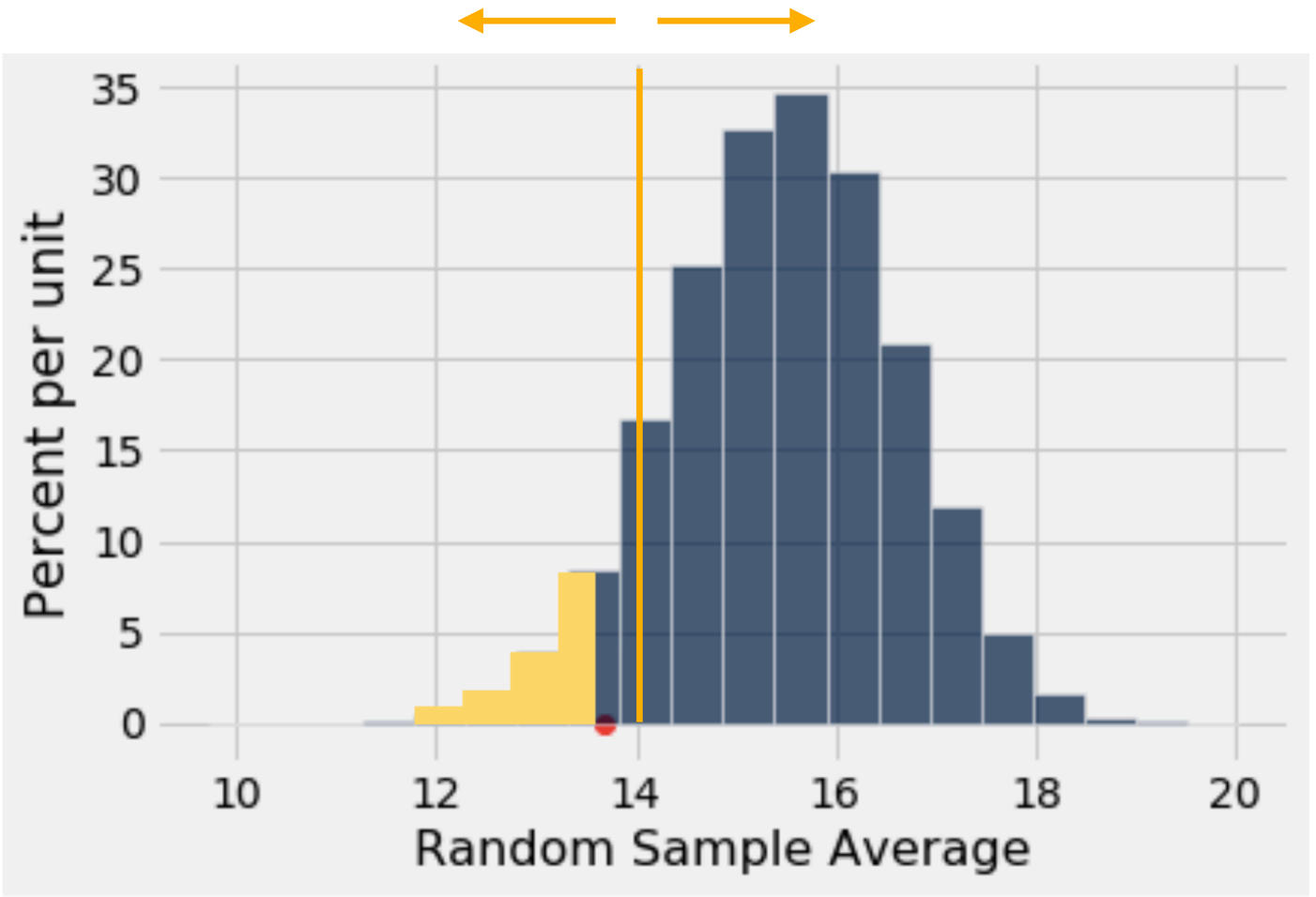
Discussion Questions

Imagine that the yellow vertical bar at ~14 represents our 5% threshold.

Do we expect the 1% threshold to lie:

A. To the left of 14

B. To the right of 14



Hypothesis Testing Review

Two Categories (e.g. percent of flowers that are purple)

- Test Statistic (1): `observed_proportion`
- Test Statistic (2): `abs(observed_proportion - null_proportion)`
- Simulate with: `sample_proportions(n, null_dist)`

Multiple Categories (e.g. ethnicity distribution of jury panel)

- Test Statistic: `tvd(observed_distribution, null_distribution)`
- Simulate with: `sample_proportions(n, null_distribution)`

Numerical Data (e.g. scores in a lab section)

- Test Statistic: `observed_mean`
- Simulate with: `population_data.sample(n, with_replacement=False)`

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

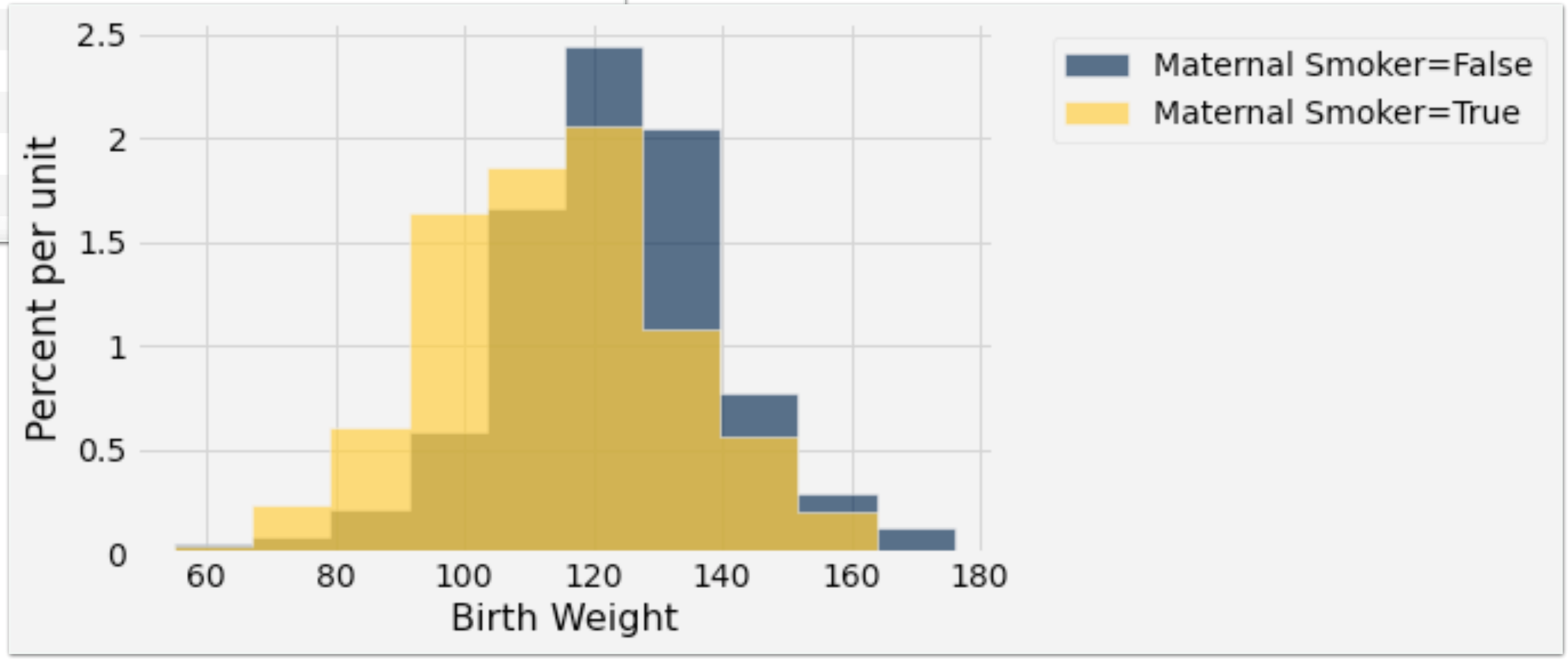
Is there a relation between maternal smoking and baby weight?

Birth Weight	Gestational Days	Maternal Age	Mater 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

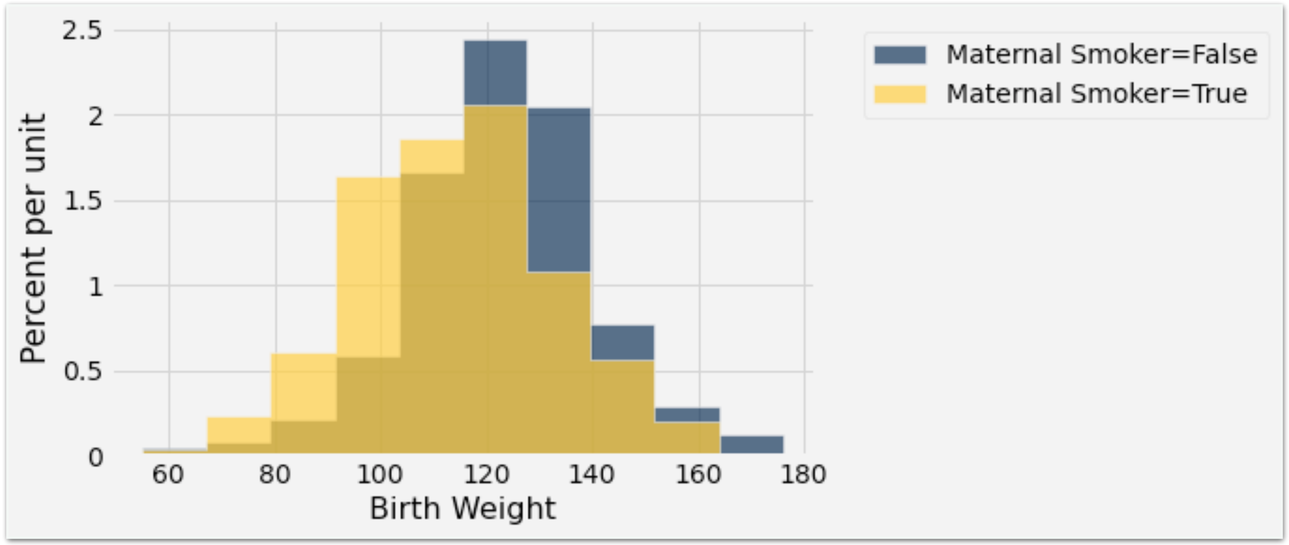
- Used when we want to compare two random samples with one another (from Group A and Group B)
  - Examples:
    - Outcomes in a medical trial (treatment / control group)
    - Outcomes of two different versions of a website
- Underlying question:
  - Do the two sets of values come from the same underlying 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	Non-Smoker	Smoker	...	Smoker
120 oz	113 oz	128 oz		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

```
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

```
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)
```

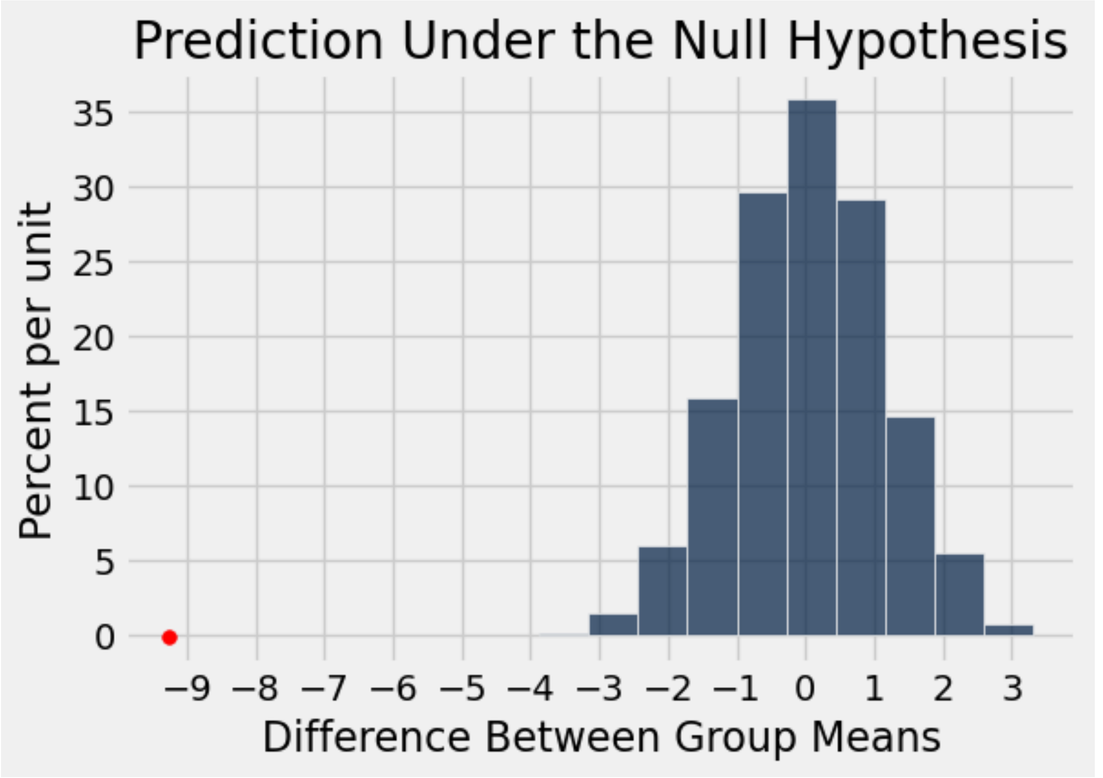
3. Plot the distribution and compare to our observed value

```
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

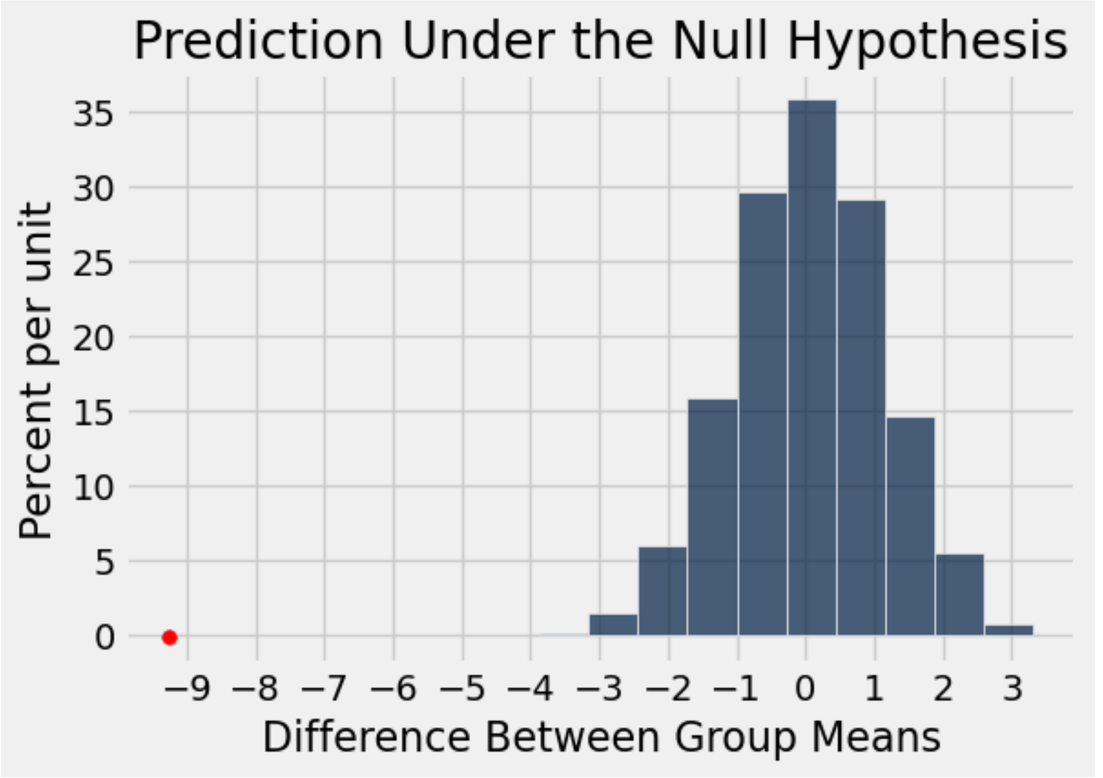
- Babies from smoking mothers weigh significantly less than babies from non-smoking mothers



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 weights? (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

Next time

- Bootstrapping
- Confidence Intervals