COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 15: Statistical Significance

Spring 2025

Reminders/Announcements

- **Labs start again this week!**
  - Lab schedule has been updated on the main course website
- **Office hours updates:**
  - Madeline’s Wednesday office hours are being moved to 2pm-3:30pm
  - Nami’s Tuesday office hours are being moved online

Final Project Website

Webpage with details is now up: https://www.eysalee.com/courses/s26/bc1016_final.html

COMS BC1016 Final Project

Spring 2026
Barnard College

The final project for BC1016 provides an opportunity to bring together, apply, and communicate your knowledge of data science and statistics from this course. You will work in groups of 2 to choose one of the provided datasets to analyze and submit a writeup of your analysis and conclusions.

AI and Usage of Outside Resources

As a reminder, the syllabus states:

> For your final project, AI generated text is **not permitted** as part of your written descriptions in your final report. Your report must include your own original writing and reflections. Violations can result in a failing grade for the assignment and/or the course.

Please note this policy includes using generative AI to produce analysis. **You are not allowed to submit AI generated code**.

If you like to do anything more advanced than what we have covered in class, you must include a brief explanation of where you learned these concepts. For instance, if you have taken prior statistics classes and would like to perform a more complicated analysis or if you have experience with (and would like to use) Python libraries we have not covered in class.

Project Milestones and Deadlines

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. **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.

Project Milestones and Deadlines

1 weeks away →

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)

~3 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.

4.5 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.

~ 6 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%

Hypothesis Testing with Multiple Categories

Jury Selection in Alameda County

- In 2010, ACLU of Northern California reported that racial and ethnic groups were not properly represented in jury panels in Alameda County, CA

  - 11 felony trials over 2 years (2009 and 2010)

  - Collected demographic data on the 1453 panelists and compared to eligible jurors in the county

Comparing Distributions

- Mendel example we computed the distance of random samples from the model

$$\text{distance}(x, y) = |x - y|$$

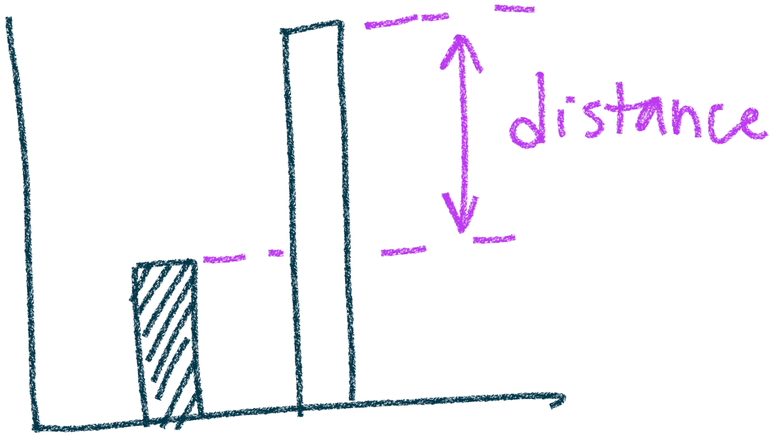
- For this, we can compute a generalized version of distance
  - **Total Variation Distance:** Measures the distance between two categorical distributions

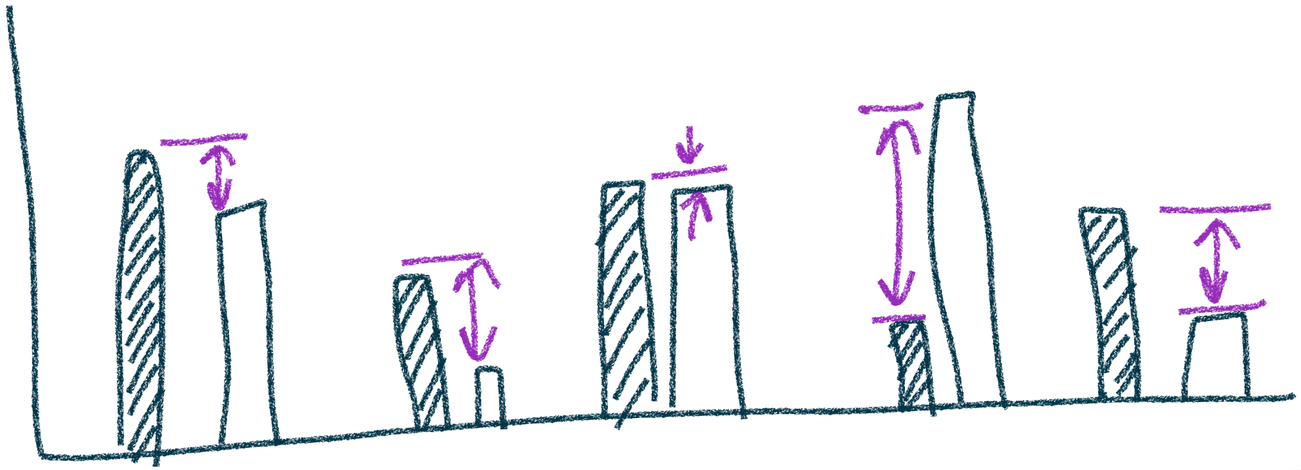
Ethnicity	% in Population
Asian	15
Black	18
Latino	12
White	54
Other	1

Total Variation Distance

distribution A

distribution B





sum all these differences and divide by 2

+/- double counts distances

Computing Total Variation Distance (TVD)

- For each category, compute the difference in proportion between two distributions (under null hypothesis and empirical / observed)

- Take the absolute value of each difference

- Sum for all categories and then divide the sum by 2

```
def tvd(dist1, dist2):
    return sum(abs(dist1 - dist2))/2
```

Summary of Process of Applying TVD

- To assess whether a sample was drawn randomly from a known categorical distribution using TVD:
    - Sample at random from the population and compute the TVD from the random sample
    - Repeat many times
    - Compare the TVD empirical distribution of simulated to the actual TVD from the sample

Jury Selection Notebook Demo

Testing with Numerical Options

Example: Exam Scores

- A class has 12 sections, each led by a different TA
- After the midterm exam, students in Section 3 find that the average score in their section is lower than in other sections
  - Exam scores are numerical, not categorical
  - Students want to know if their grades are related to their section

Section	Midterm average
1	15.5938
2	15.125
3	13.6667
4	14.7667
5	17.4545
6	15.0312
7	16.625
8	16.3103
9	14.5667
10	15.2353
11	15.8077
12	15.7333

Example: Exam Scores

- A class has 12 sections, each led by a different TA
- After the midterm exam, students in Section 3 find that the average score in their section is lower than in other sections
  - Exam scores are numerical, not categorical
  - Students want to know if their grades are related to their section

What is a question we can hypothesis test?

Section	Midterm average
1	15.5938
2	15.125
3	13.6667
4	14.7667
5	17.4545
6	15.0312
7	16.625
8	16.3103
9	14.5667
10	15.2353
11	15.8077
12	15.7333

Example: Exam Scores

- Question:

Example: Exam Scores

- Question: Did the 27 students do lower by chance?

Example: Exam Scores

- Question: Did the 27 students do lower by chance?
- Potential Answers:
  - Null Hypothesis:

Example: Exam Scores

- Question: Did the 27 students do lower by chance?

- Potential Answers:

  - Null Hypothesis: The average score the students in Section 3 is like the average score of the same number of students picked at random from the class

Example: Exam Scores

- Question: Did the 27 students do lower by chance?

- Potential Answers:

  - Null Hypothesis: The average score the students in Section 3 is like the average score of the same number of students picked at random from the class

  - Alternative Hypothesis:

Example: Exam Scores

- Question: Did the 27 students do lower by chance?

- Potential Answers:

  - Null Hypothesis: The average score the students in Section 3 is like the average score of the same number of students picked at random from the class

  - Alternative Hypothesis: No, Section 3’s average is too low

Example: Exam Scores

- Question: Did the 27 students do lower by chance?
- Potential Answers:
  - Null Hypothesis: The average score the students in Section 3 is like the average score of the same number of students picked at random from the class
  - Alternative Hypothesis: No, Section 3’s average is too low
- Statistic to measure:

Example: Exam Scores

- Question: Did the 27 students do lower by chance?
- Potential Answers:
  - Null Hypothesis: The average score the students in Section 3 is like the average score of the same number of students picked at random from the class
  - Alternative Hypothesis: No, Section 3’s average is too low
- Statistic to measure: Average score per section (27 students)

Assessing a Model

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world
2. Simulate the statistic under the assumptions of the model
3. Compare the data to the model's predictions:
    a. Draw a histogram of the simulated values of the statistic
    b. Compute the observed statistic from the real sample

Null: Average score of the students in Section 3 is like the average score of the same number of students picked at random from the class

Alternative: Average score in Section 3 is too low

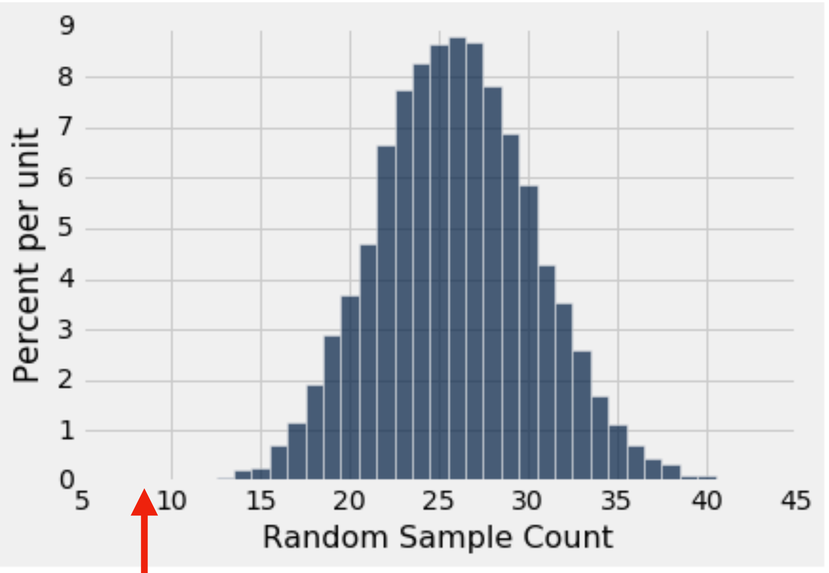
Statistic: average score per 27 students

Exam Notebook Demo

Statistical Significance

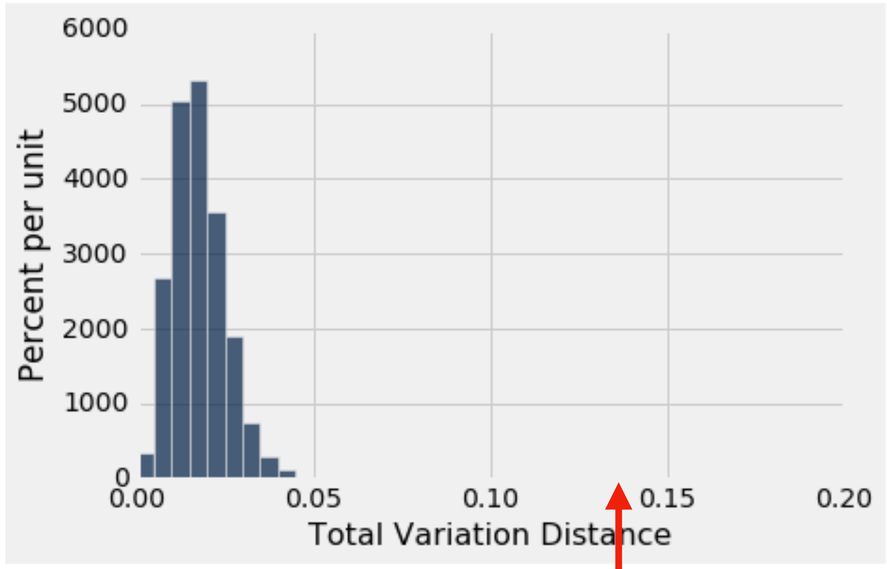
Our Examples So Far

Swain v Alabama



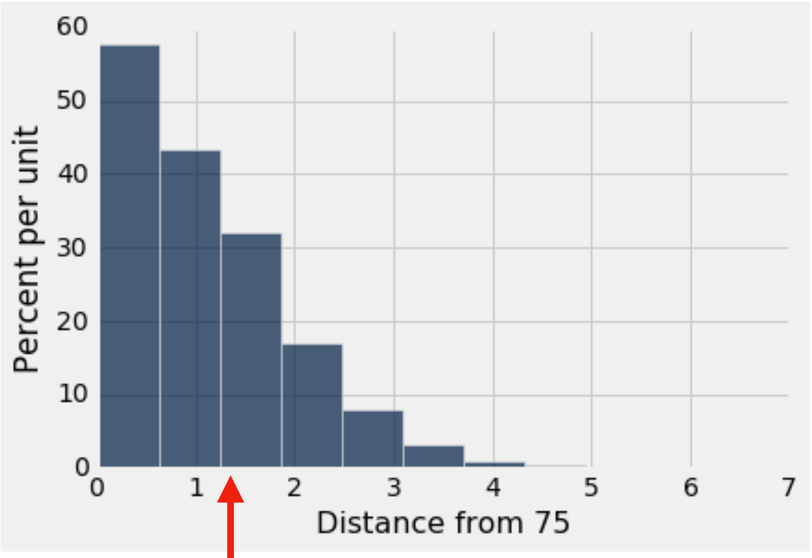
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

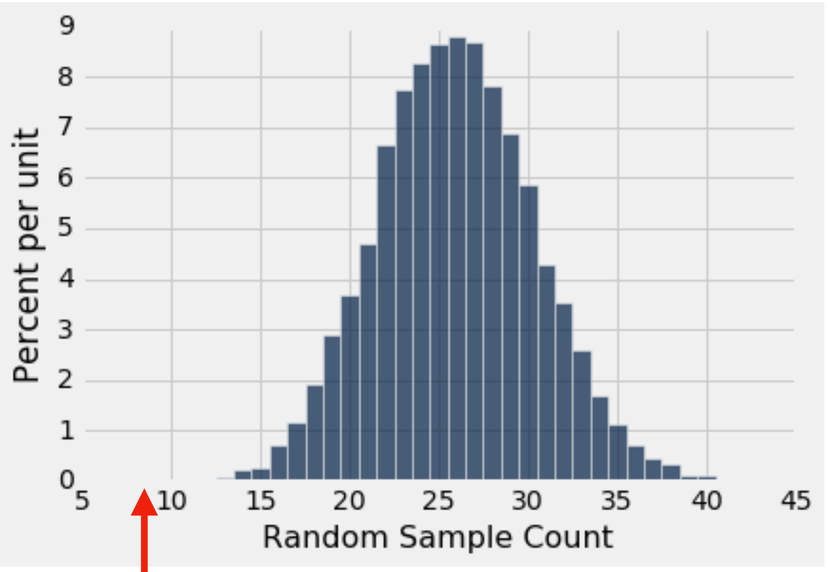
Pea Plants



Observed Distance (1.32)

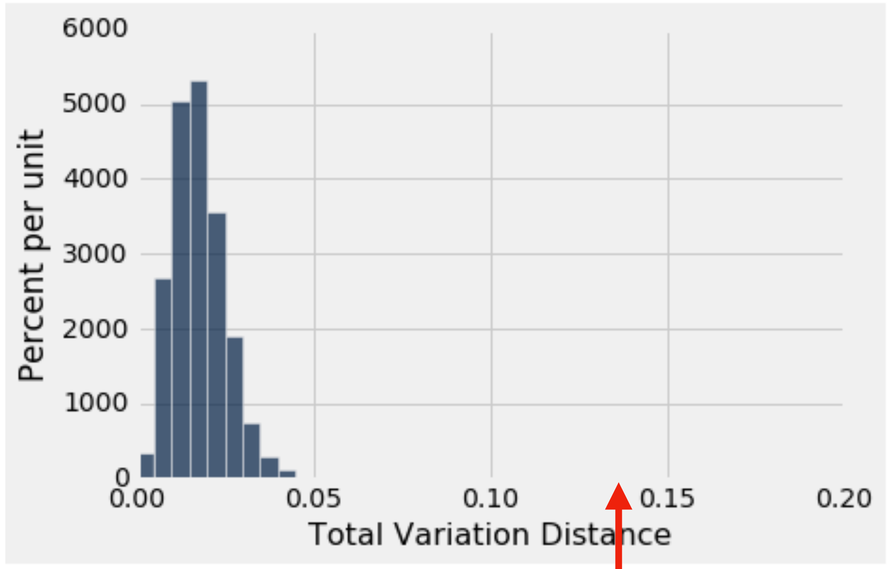
Our Examples So Far

Swain v Alabama



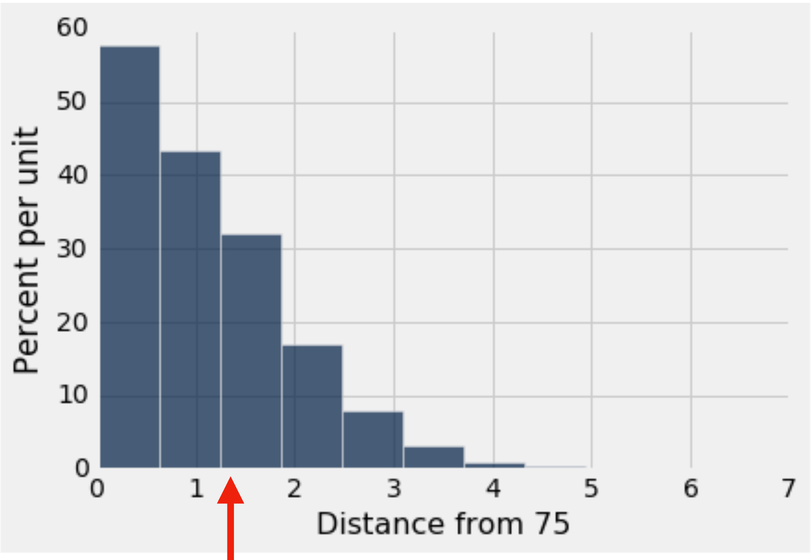
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Pea Plants

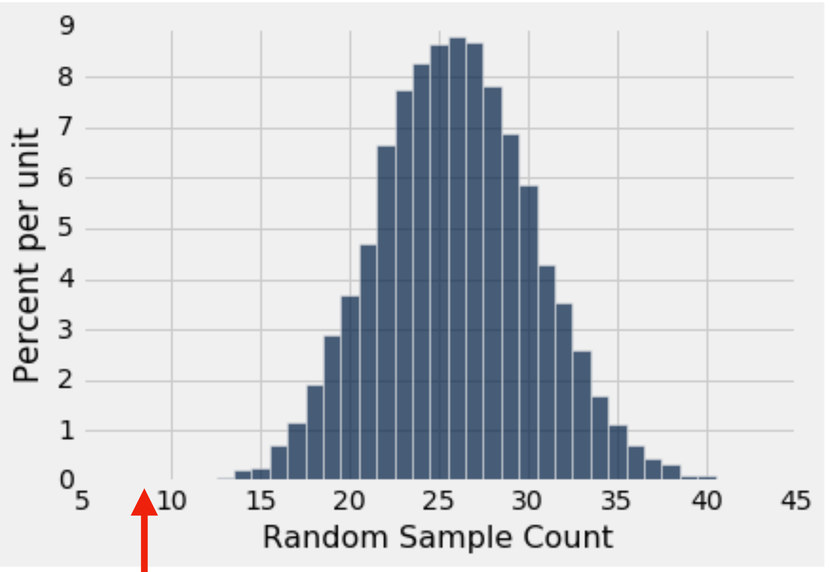


Observed Distance (1.32)

In a histogram, the area of each bar is the percent of individuals in the corresponding bin

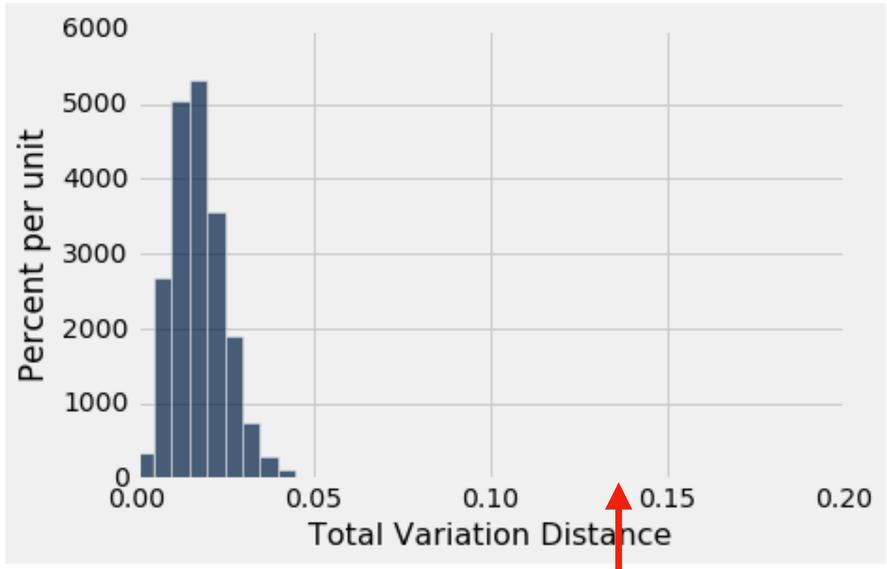
Our Examples So Far: Tail Area

Swain v Alabama



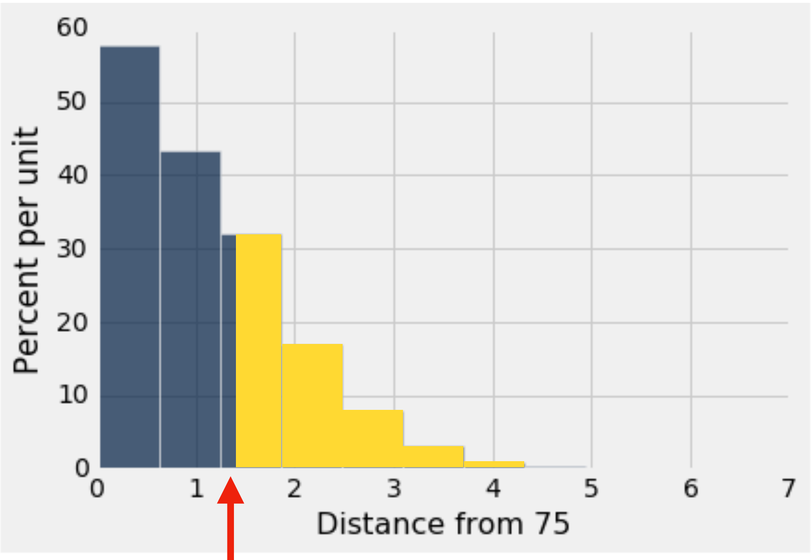
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Pea Plants

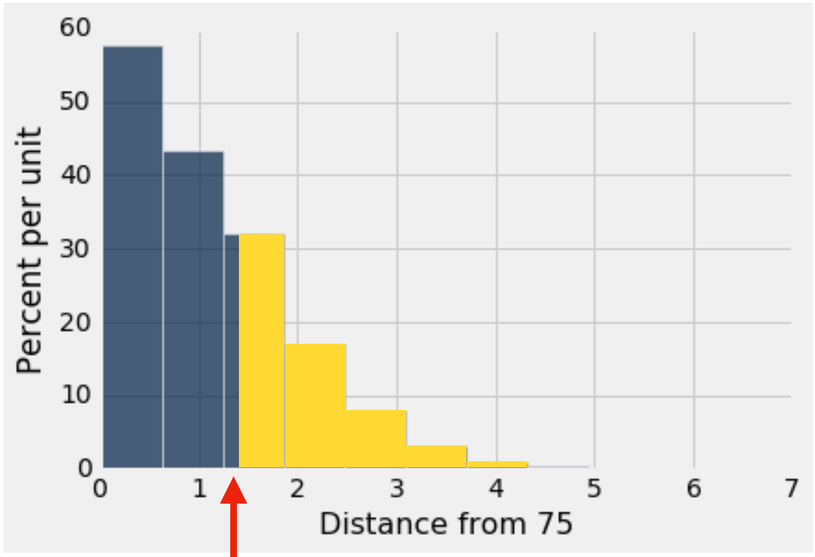


Observed Distance (1.32)

In a histogram, the area of each bar is the percent of individuals in the corresponding bin

Our Examples So Far: Tail Area

Pea Plants



Conventions look at to the the area starting from the observed statistic and looking in the direction that makes us lean toward the alternative hypothesis

Observed Distance (1.32)

In a histogram, the area of each bar is the percent of individuals in the corresponding bin

Determining Significance

- **P-Value**: Chance, based on the model in the null hypothesis, that the test statistic will be equal to the observed value in the sample or even further in the direction that supports the alternative

  - Also known as the “observed significance level” of a test

- If a p-value is small, the area beyond the observed statistic is small

  - This is far from what the null predicts, and suggests data supports 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%)**

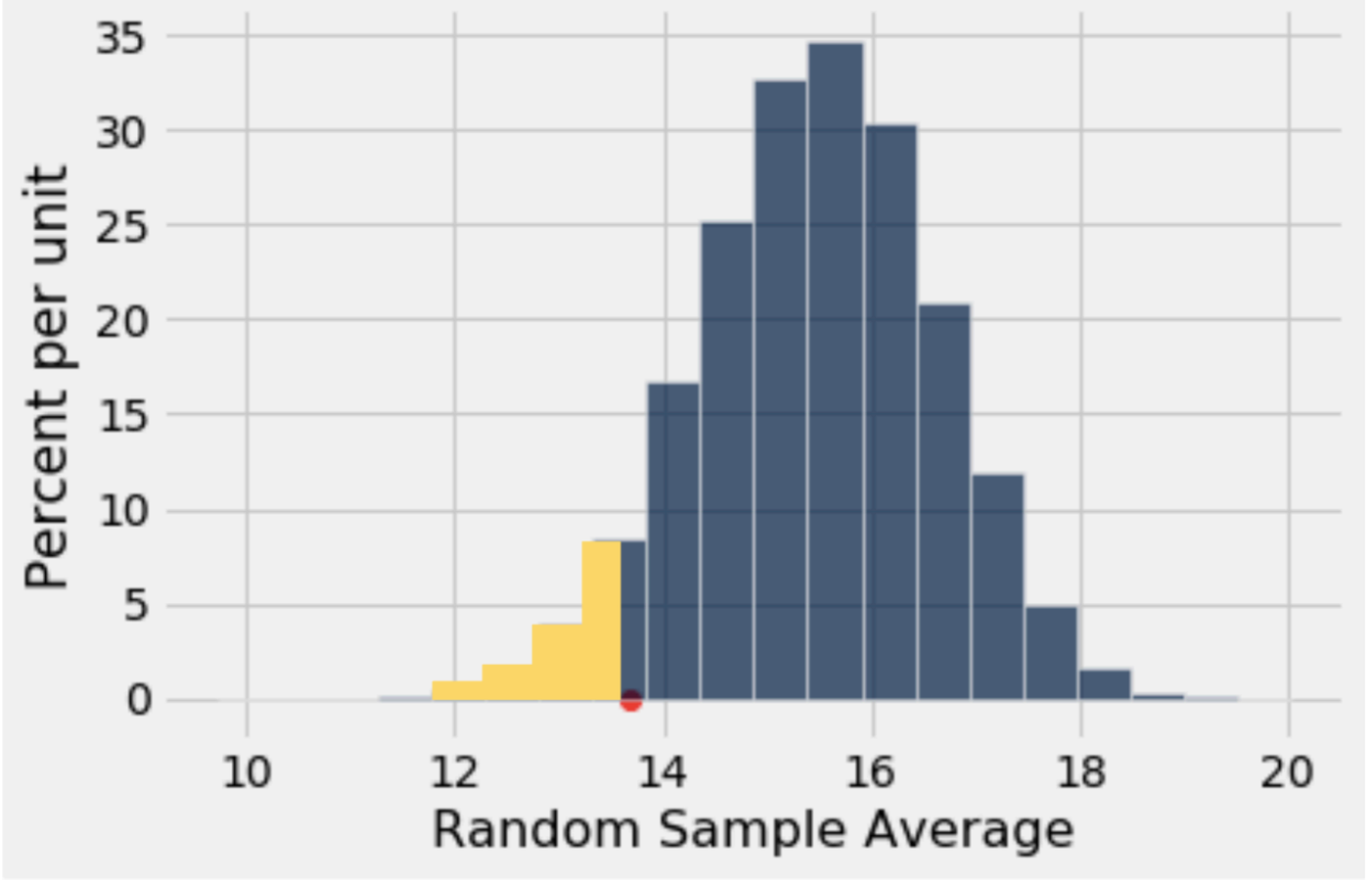
    - The result is “highly statistically significant”

Levels of Statistical Significance

Exam Notebook Demo (continued)

The P-Value as an Area

- Area to the left of of our observed value



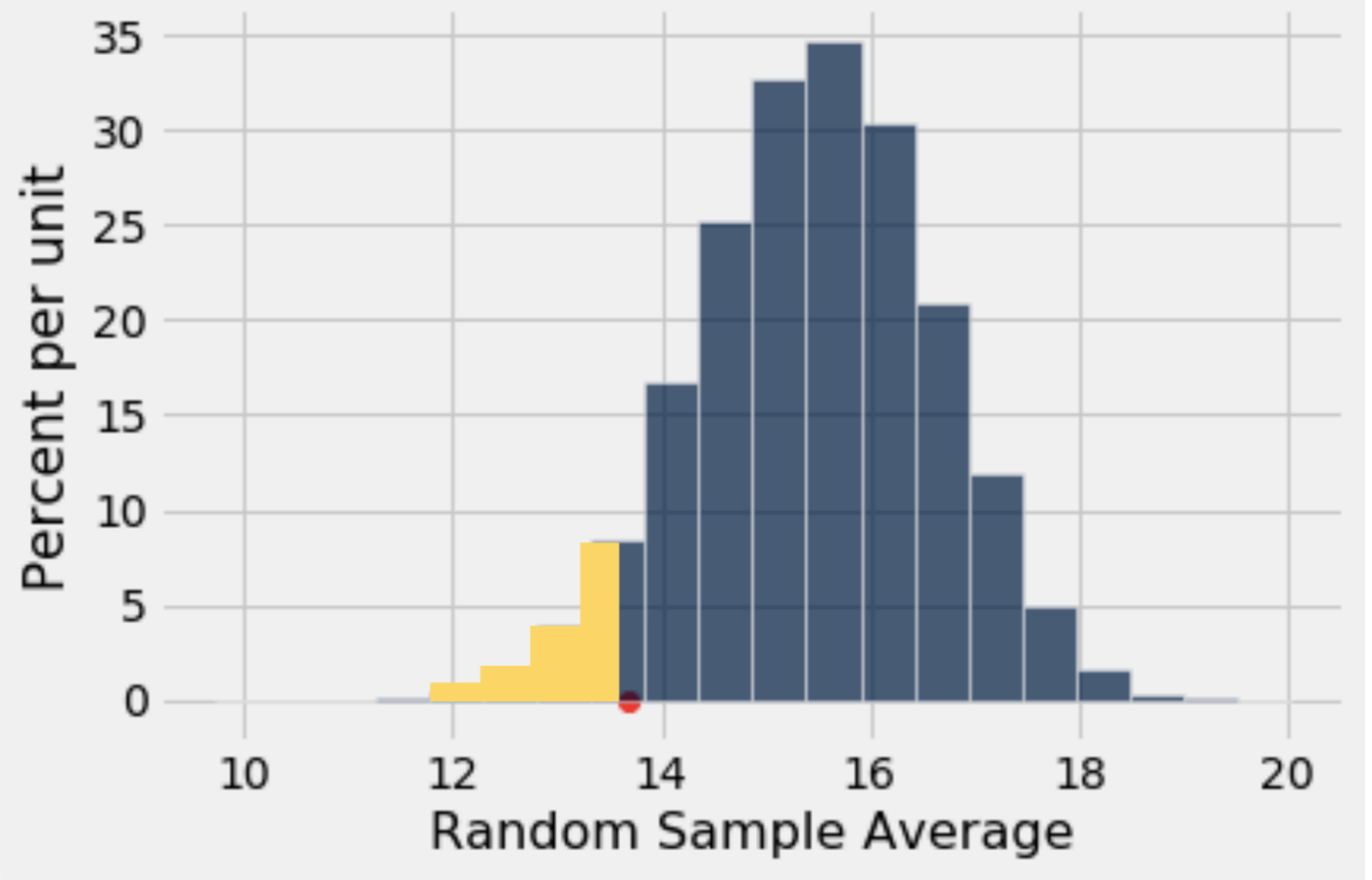
Discussion Questions

What does the red dot represent?

A. Our p-value

B. Our expected outcome

C. Our observed outcome



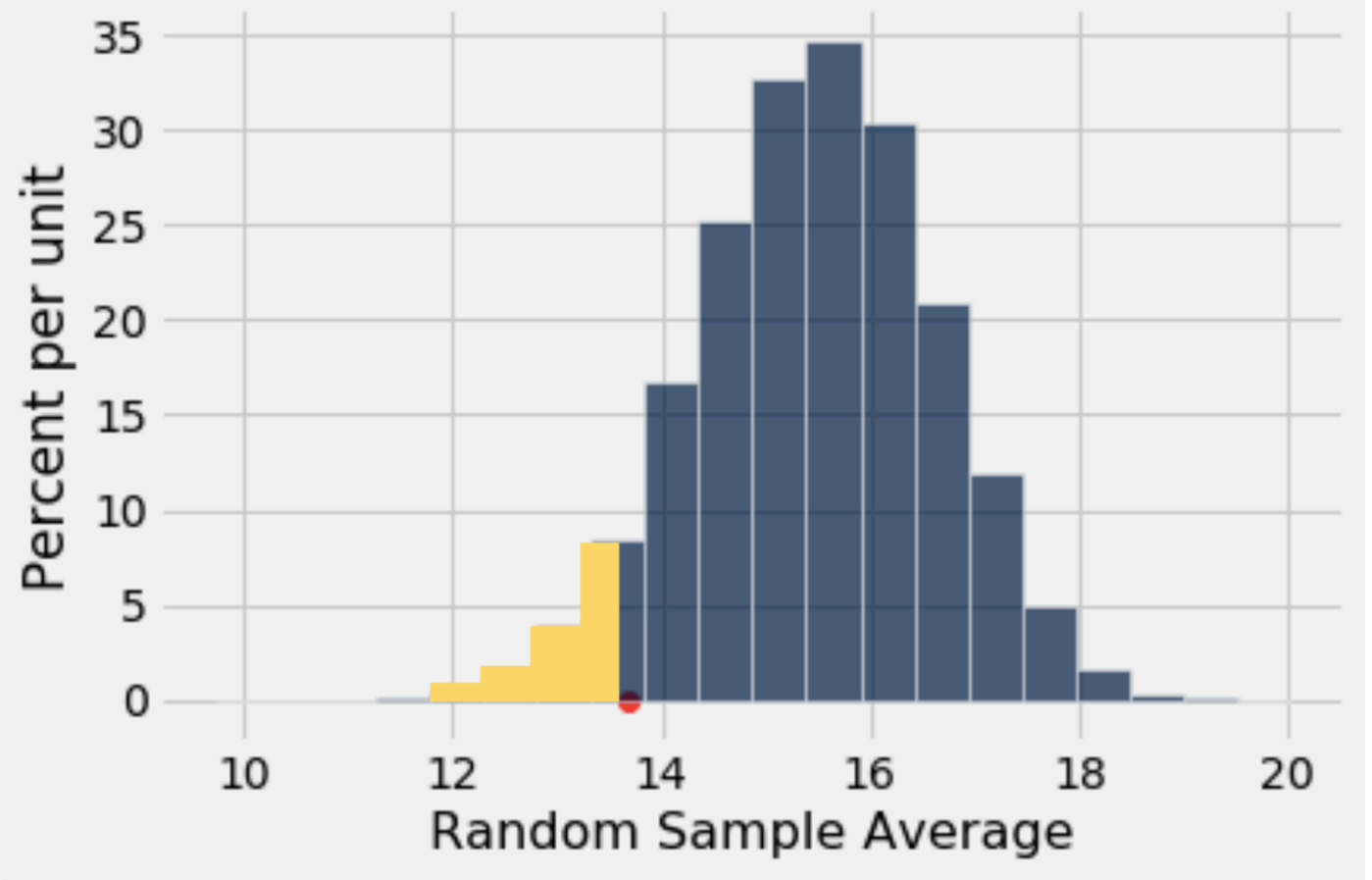
Discussion Questions

What does the red dot represent?

A. Our p-value

B. Our expected outcome

C. Our observed outcome



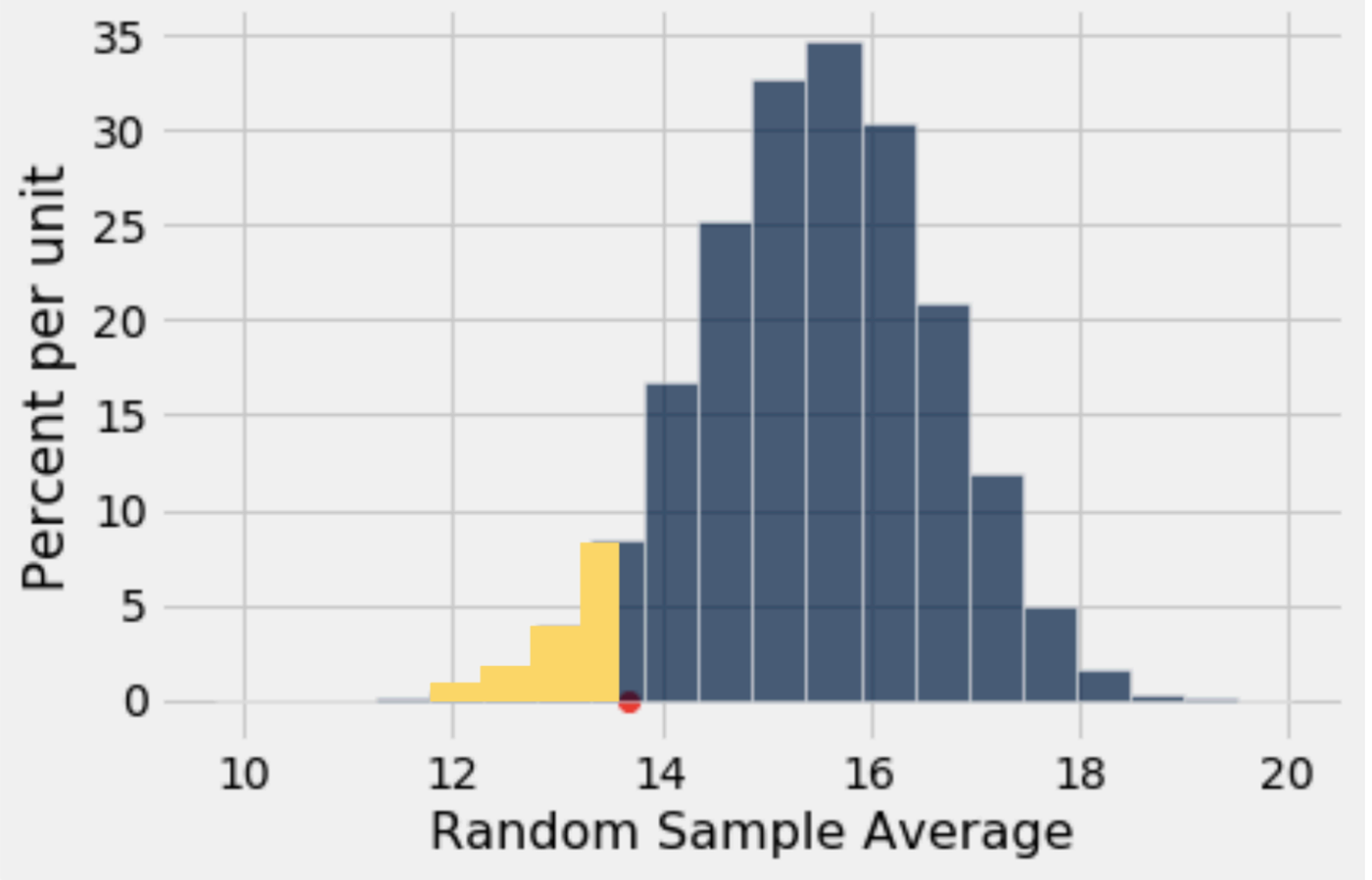
Discussion Questions

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



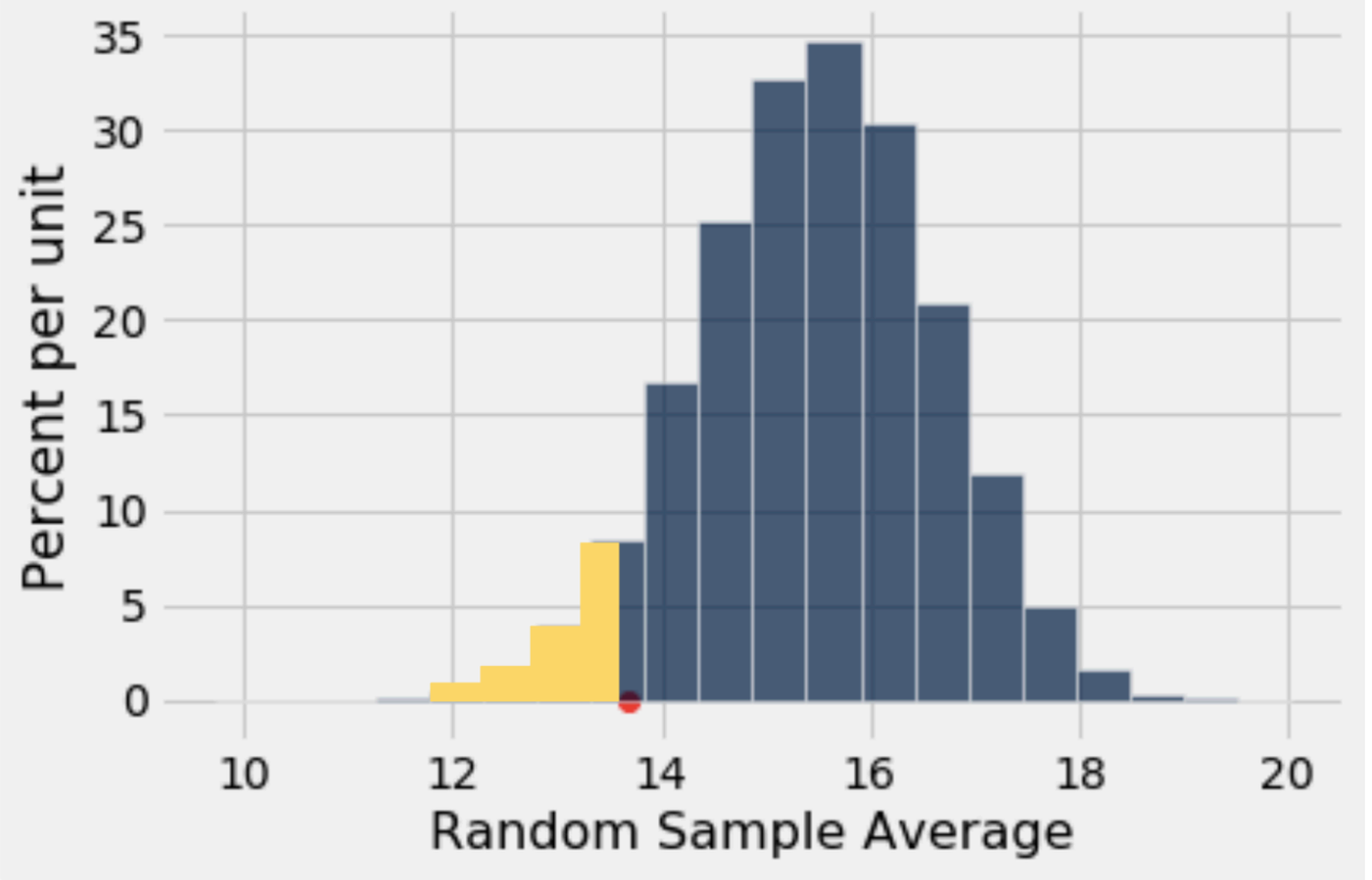
Discussion Questions

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



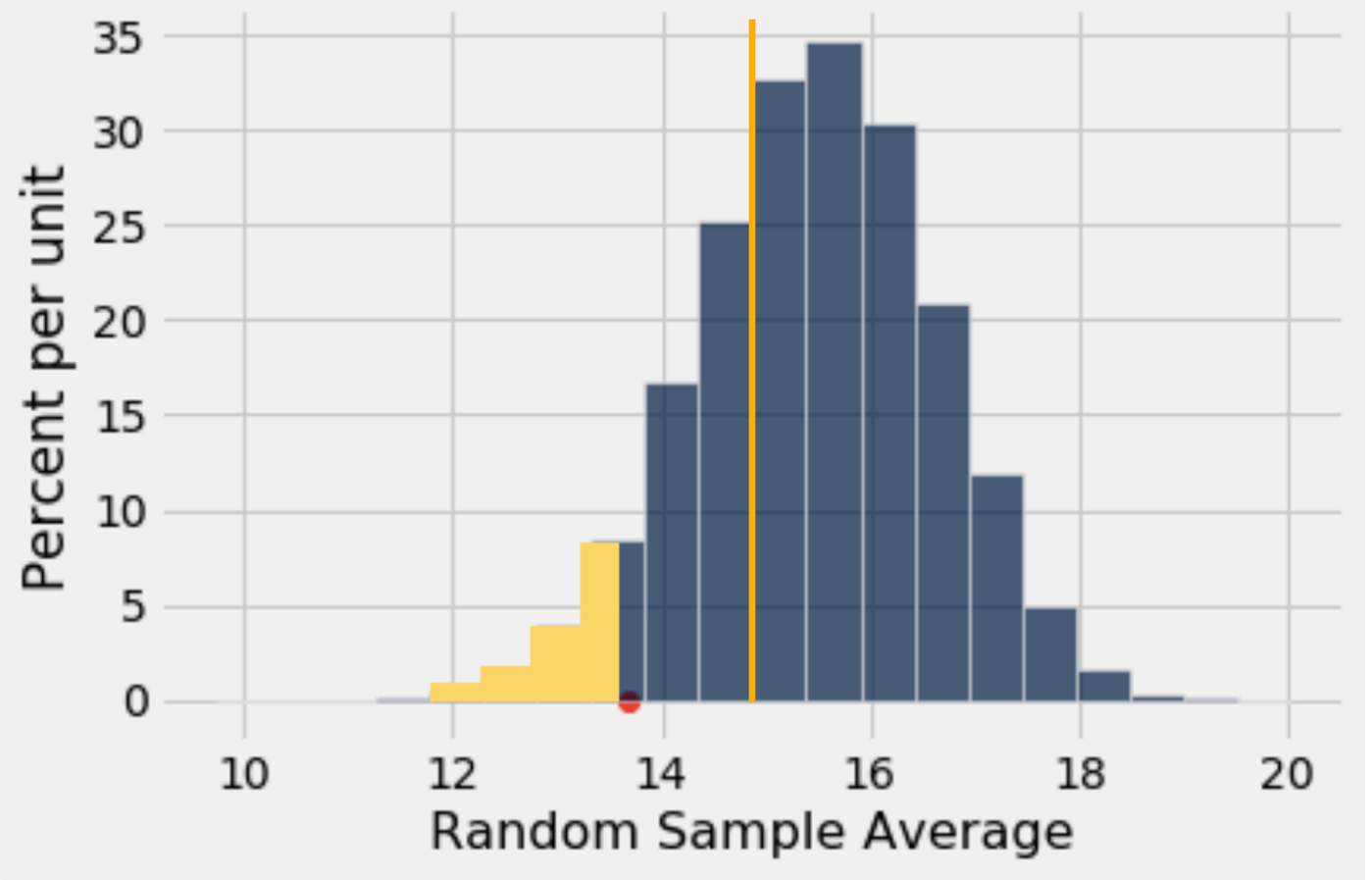
Discussion Questions

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



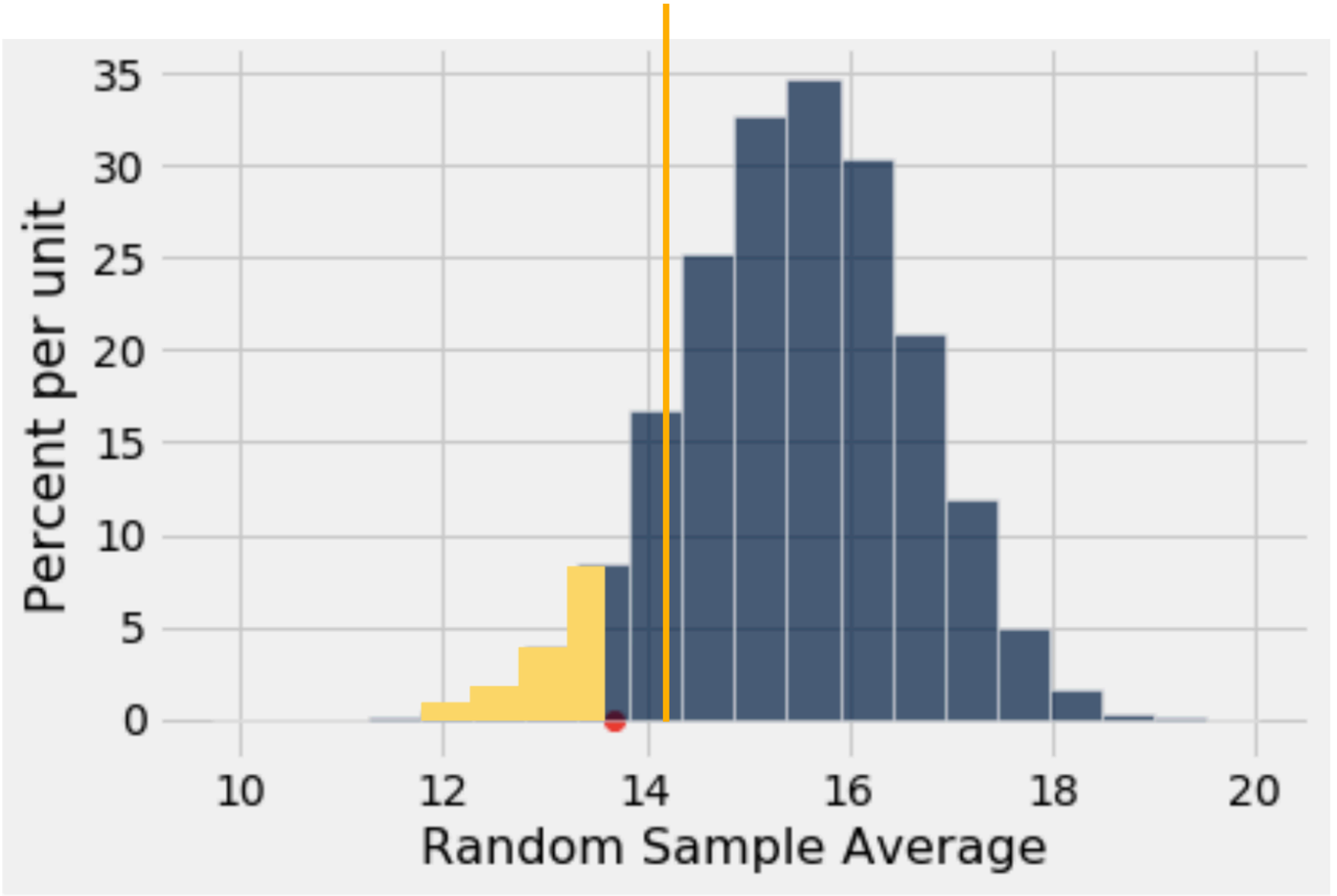
Discussion Questions

Imagine that the yellow vertical bar at ~14 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



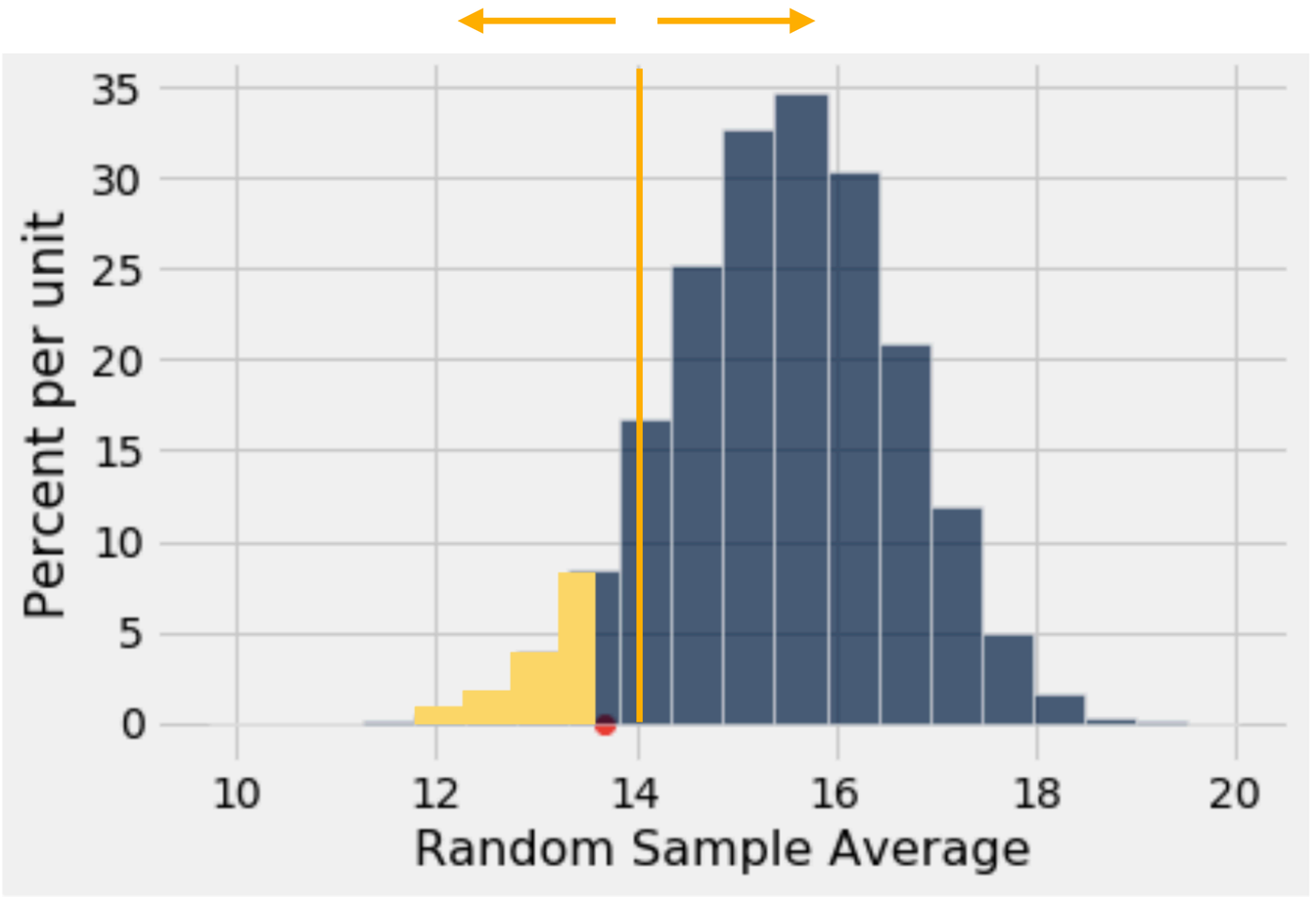
Discussion Questions

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



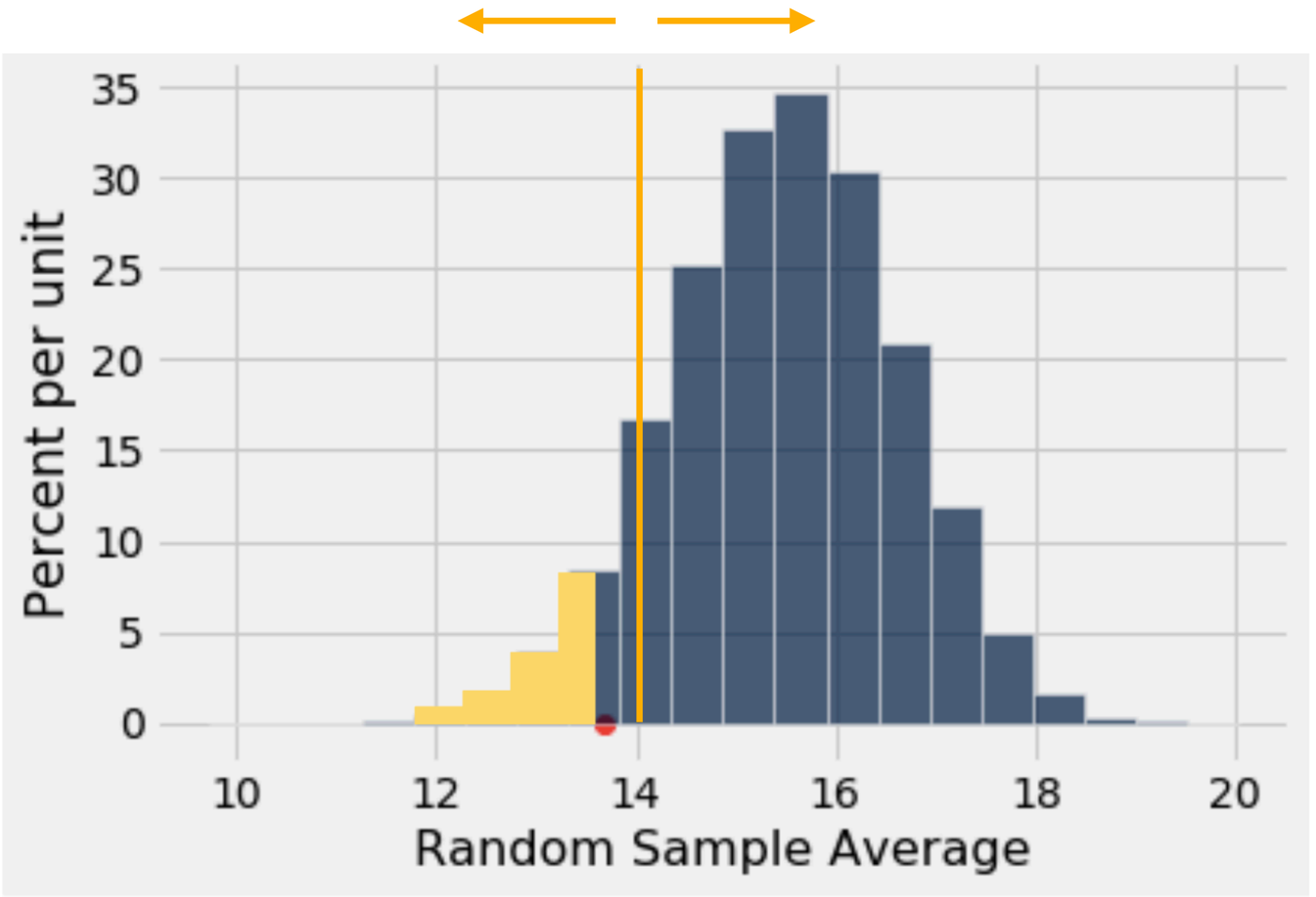
Discussion Questions

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



Summary of Hypothesis Testing so Far

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)`

Next time

- A/B Testing