COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 13: Statistical Significance

Oct 29, 2025

Update

- No class on Monday (election woo!!)
- We won’t get to A/B testing in time for HW5
    - SKIP THE ENTIRETY OF QUESTION 3 ON HW5

Final Project Groups

- If you want us to match you to a group: Link
- If you want to be in a specific group: Link
  - Previously this form had many questions… It is now shorter
  - We only ask group member names and the lab you're attending Dec 3/4 (and that you email the teaching staff to give us a heads up)
- Deadline for forming groups is (tentatively) end of next week (**Nov 7**)
  - This is not graded but we will bug you if you haven't filled out either of the forms
    - If you don't have a preference for groups, fill out the first one!

Lecture Outline

- Testing with numerical data
- Statistical significance (p-value)

Comparing Distributions

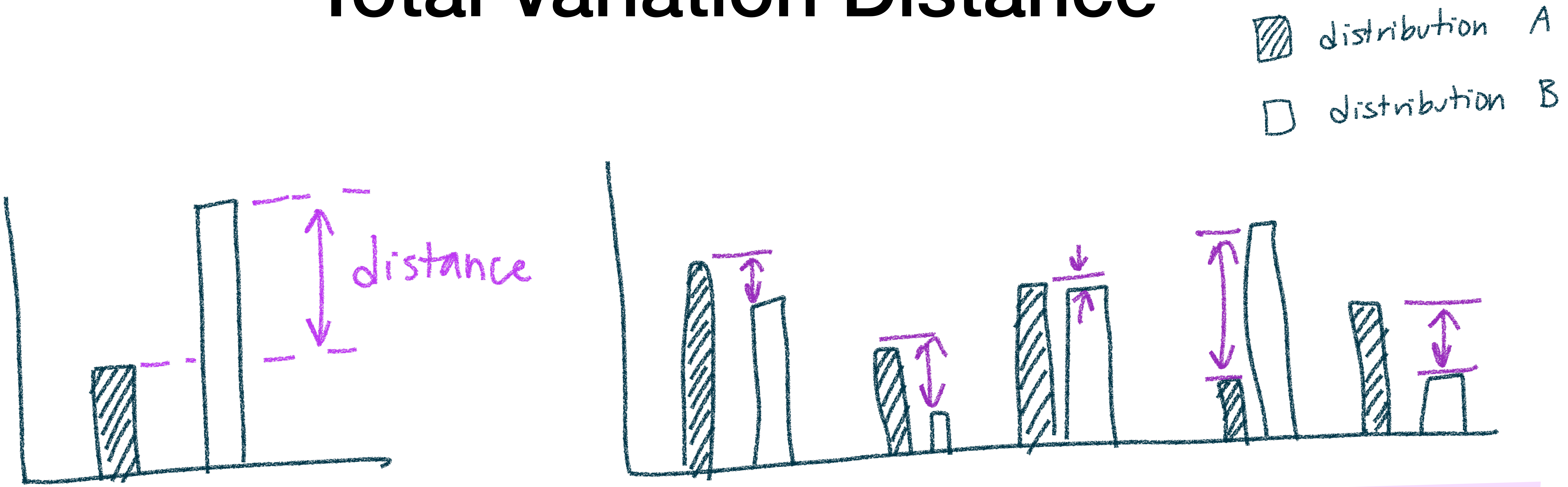
- Mendel example we computed the distance of random samples from the model

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

- 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



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

```python
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

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

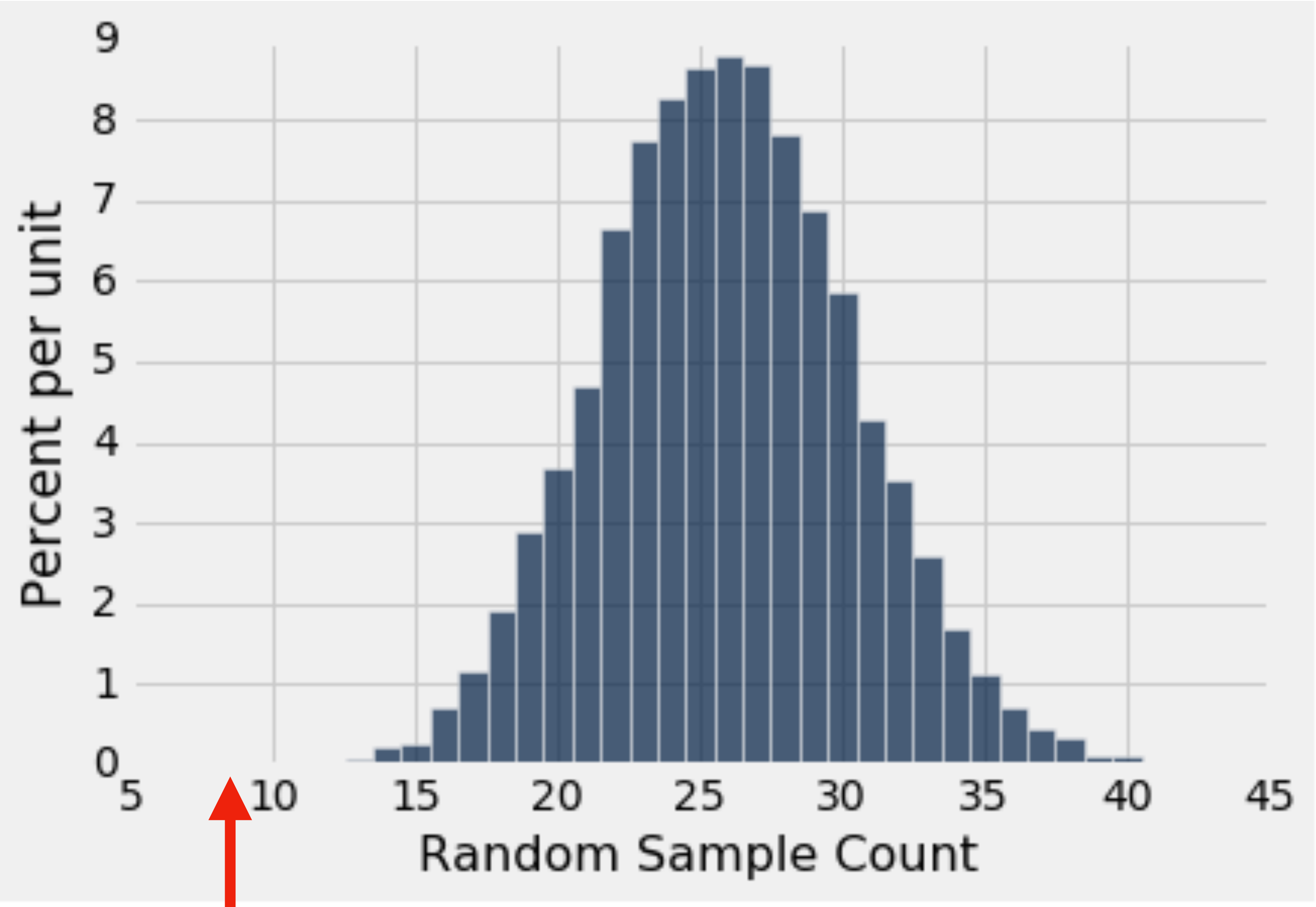
Statistic: average score per 27 students

Exam Notebook Demo

Statistical Significance

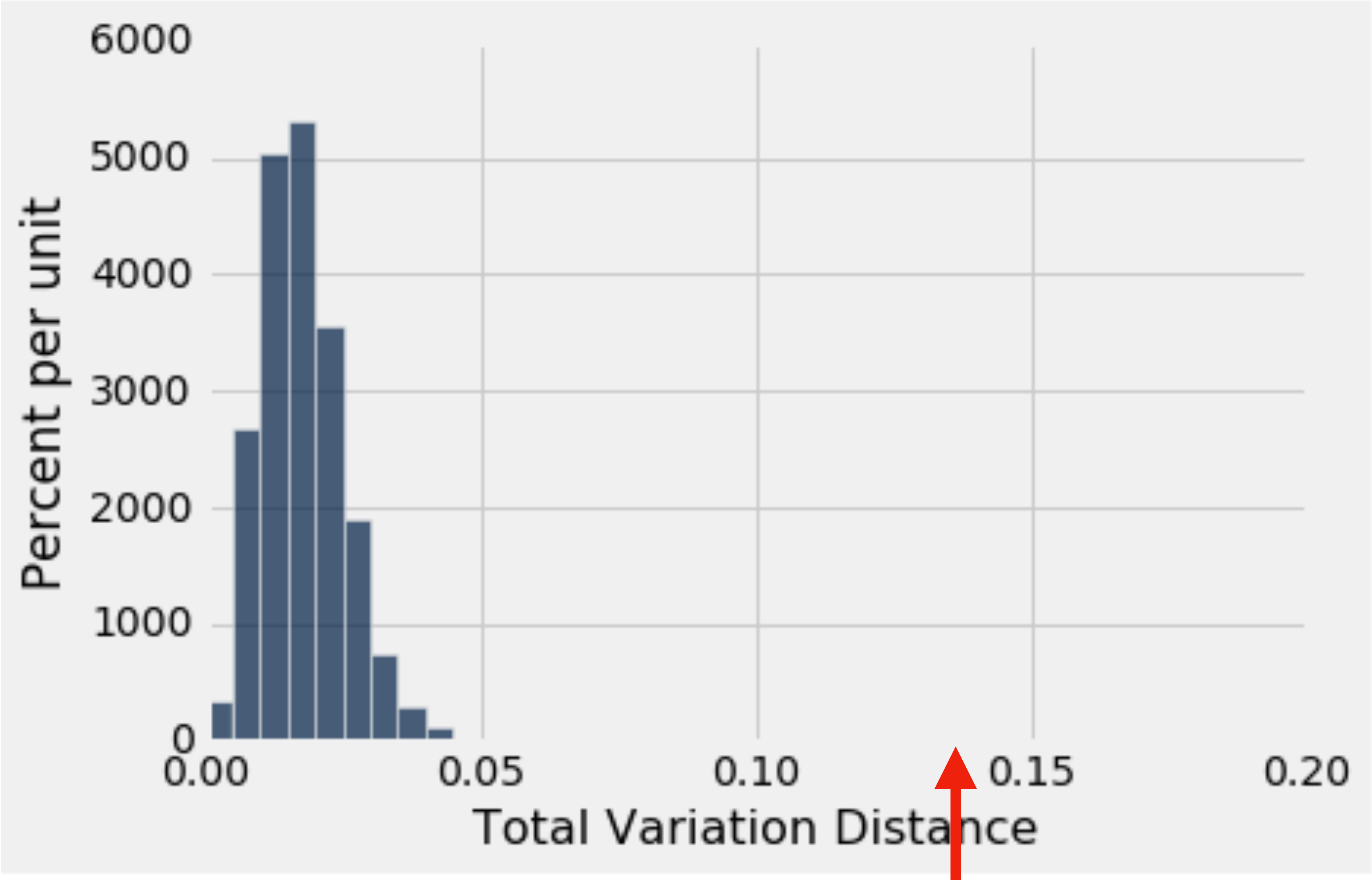
Our Examples So Far

Swain v Alabama



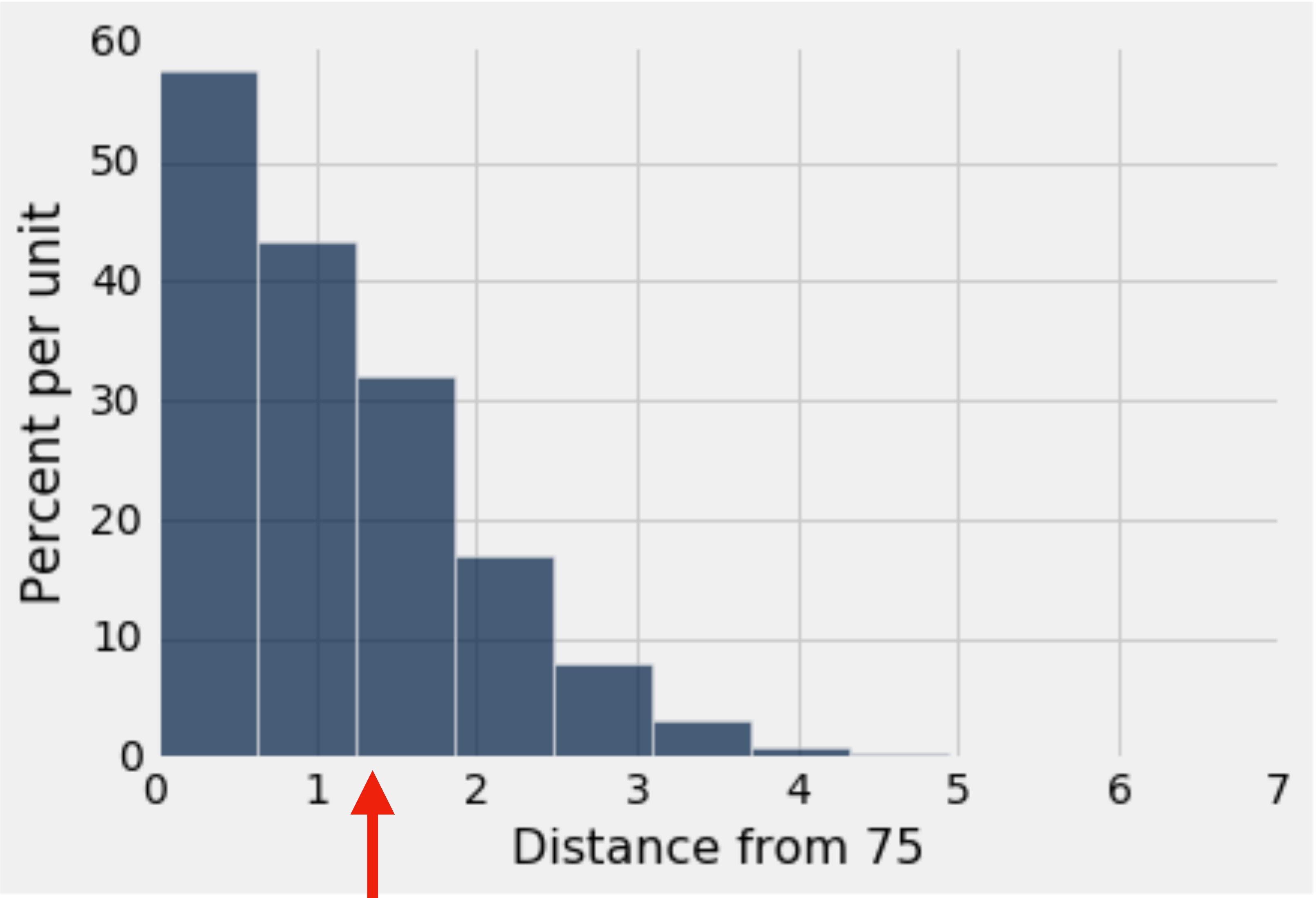
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

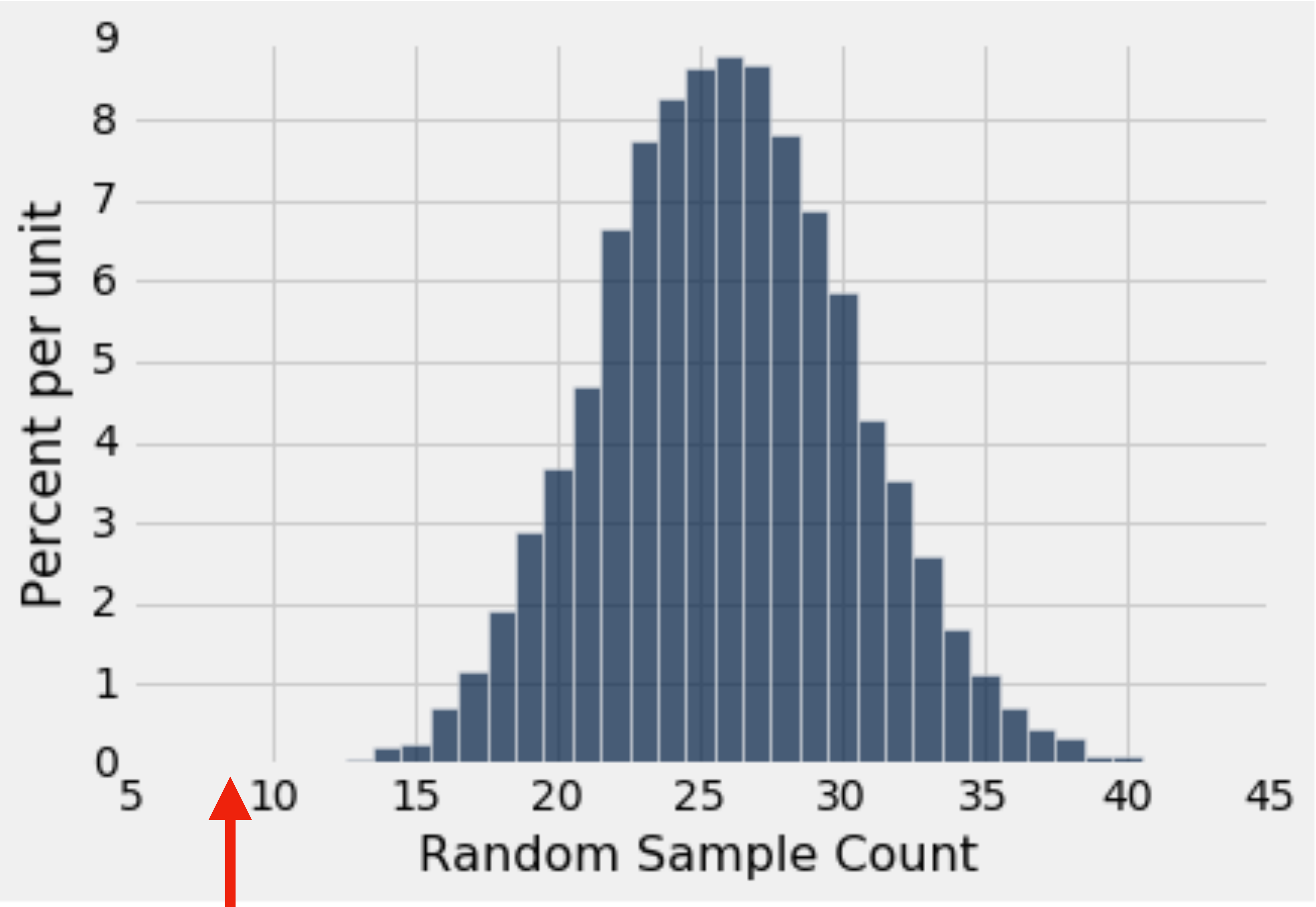
Pea Plants



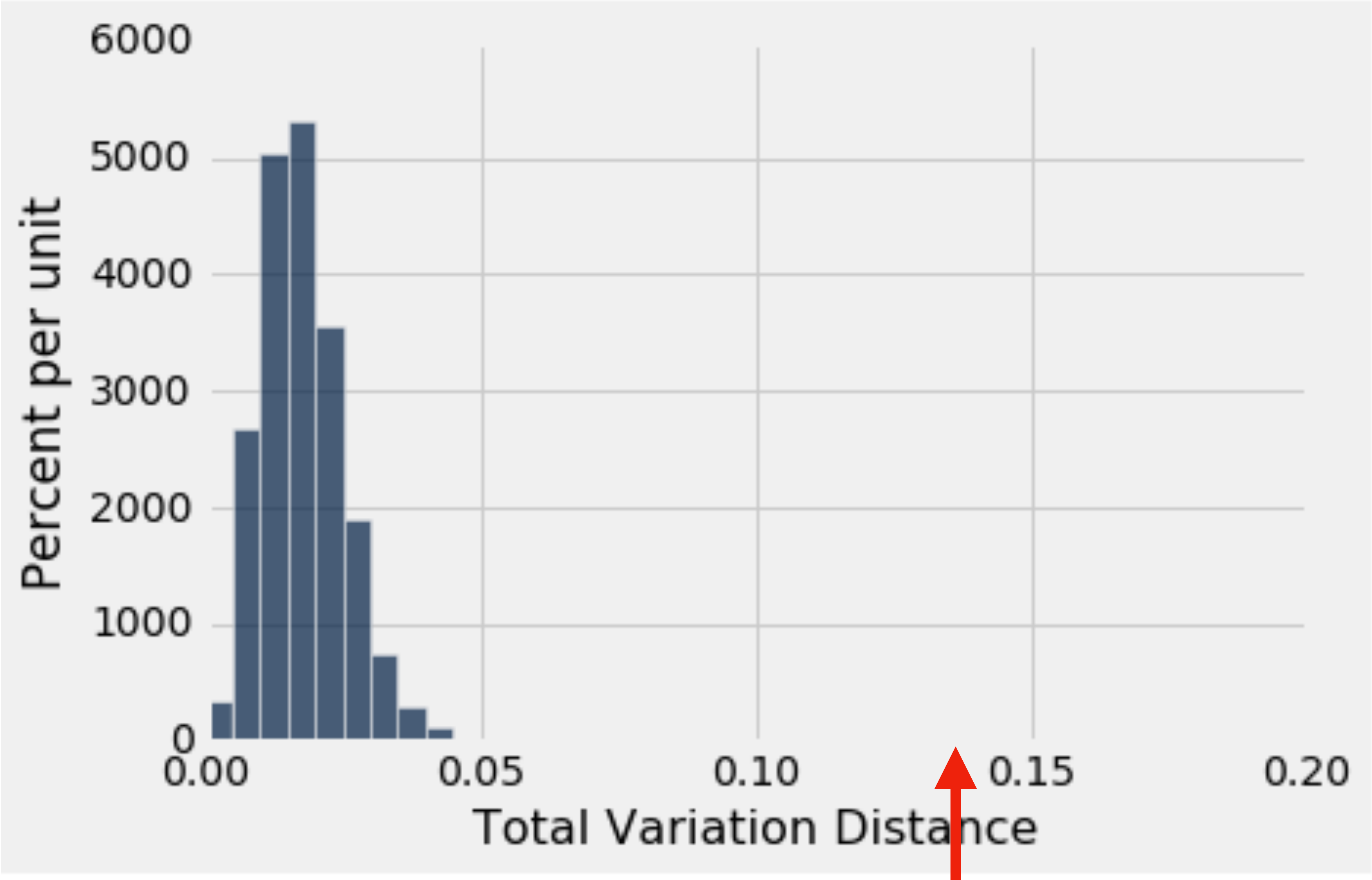
Observed Distance (1.32)

Our Examples So Far

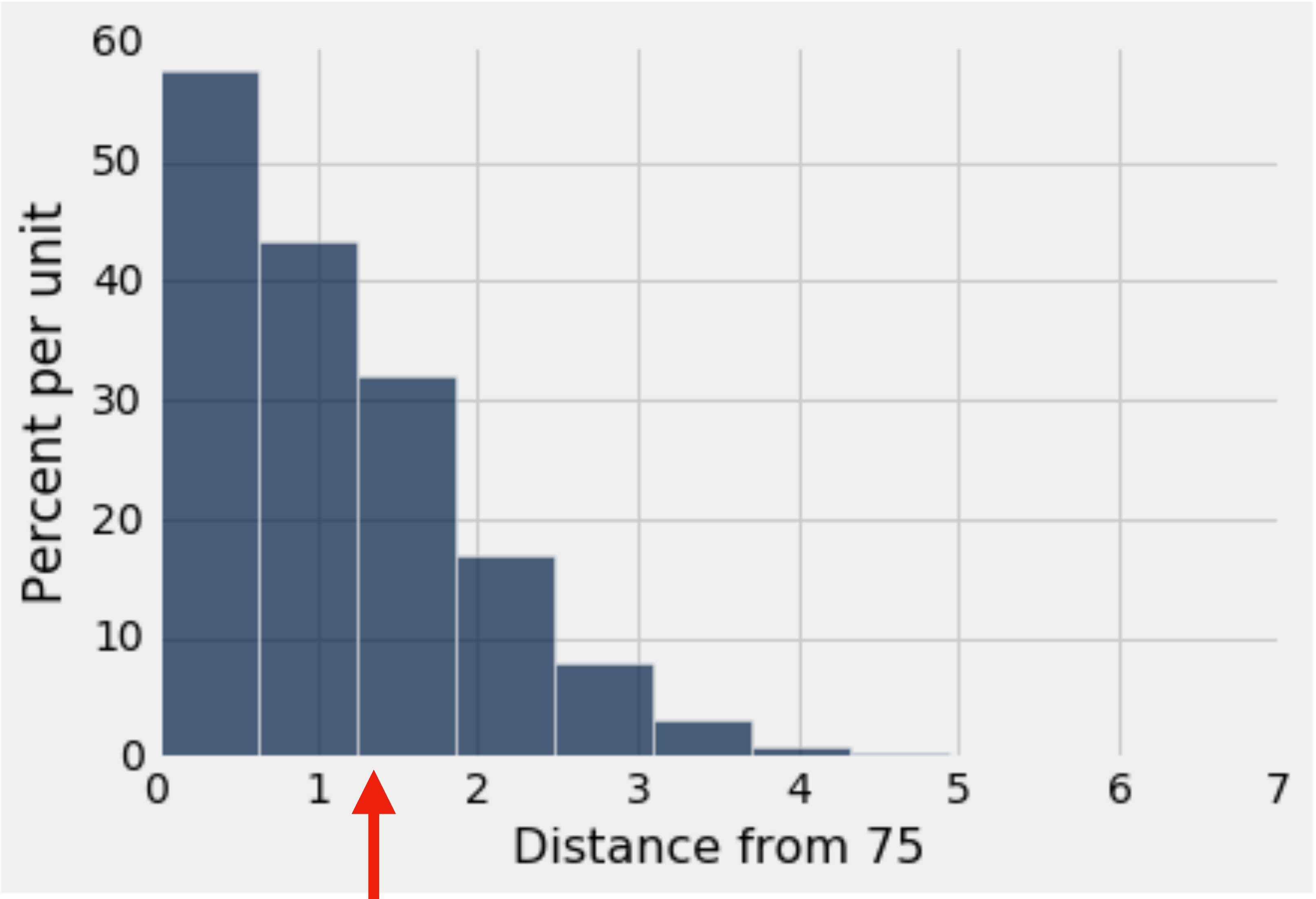
Swain v Alabama



Alameda Jury



Pea Plants



Observed Number (8)

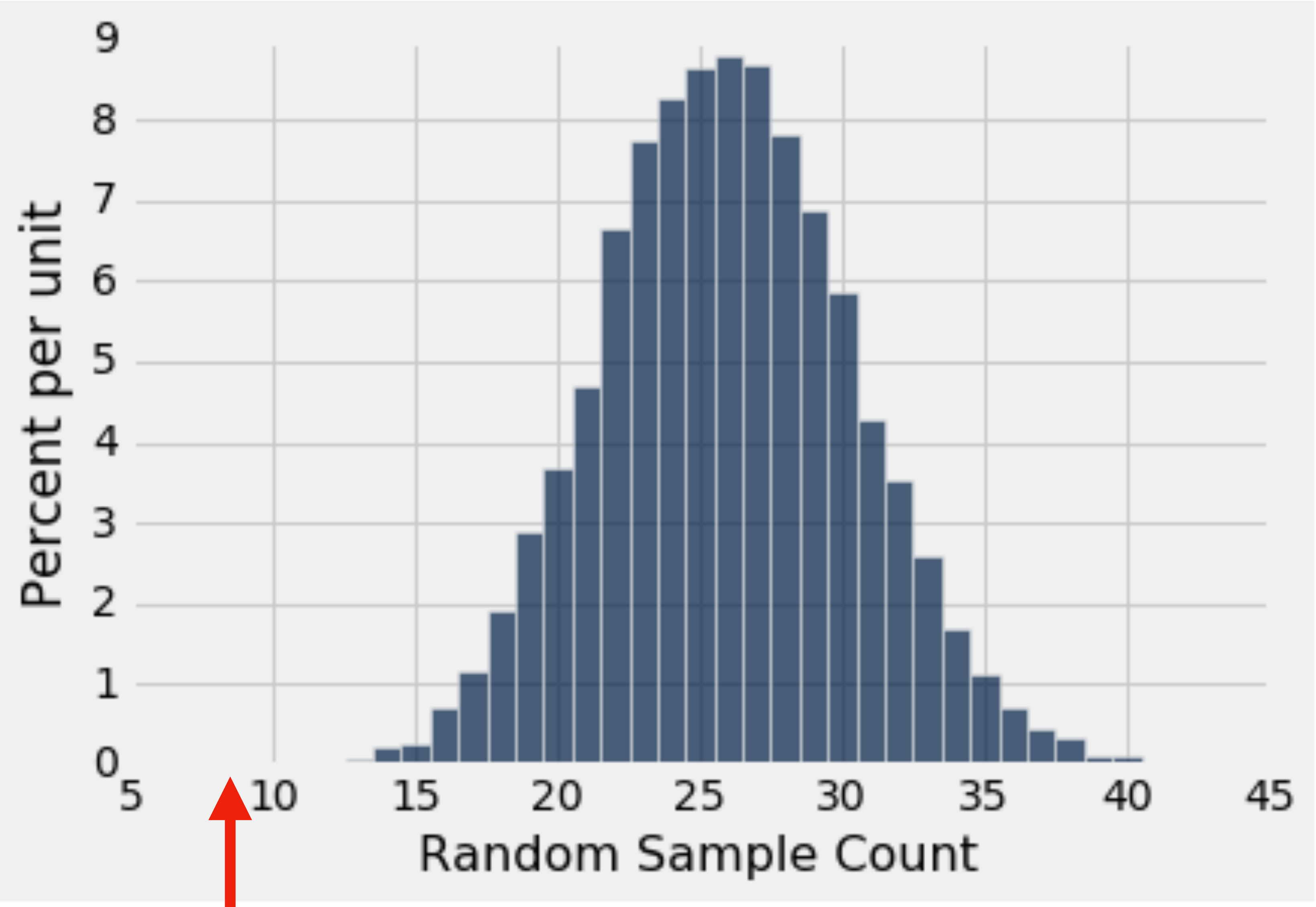
Observed TVD (0.14)

Observed Distance (1.32)

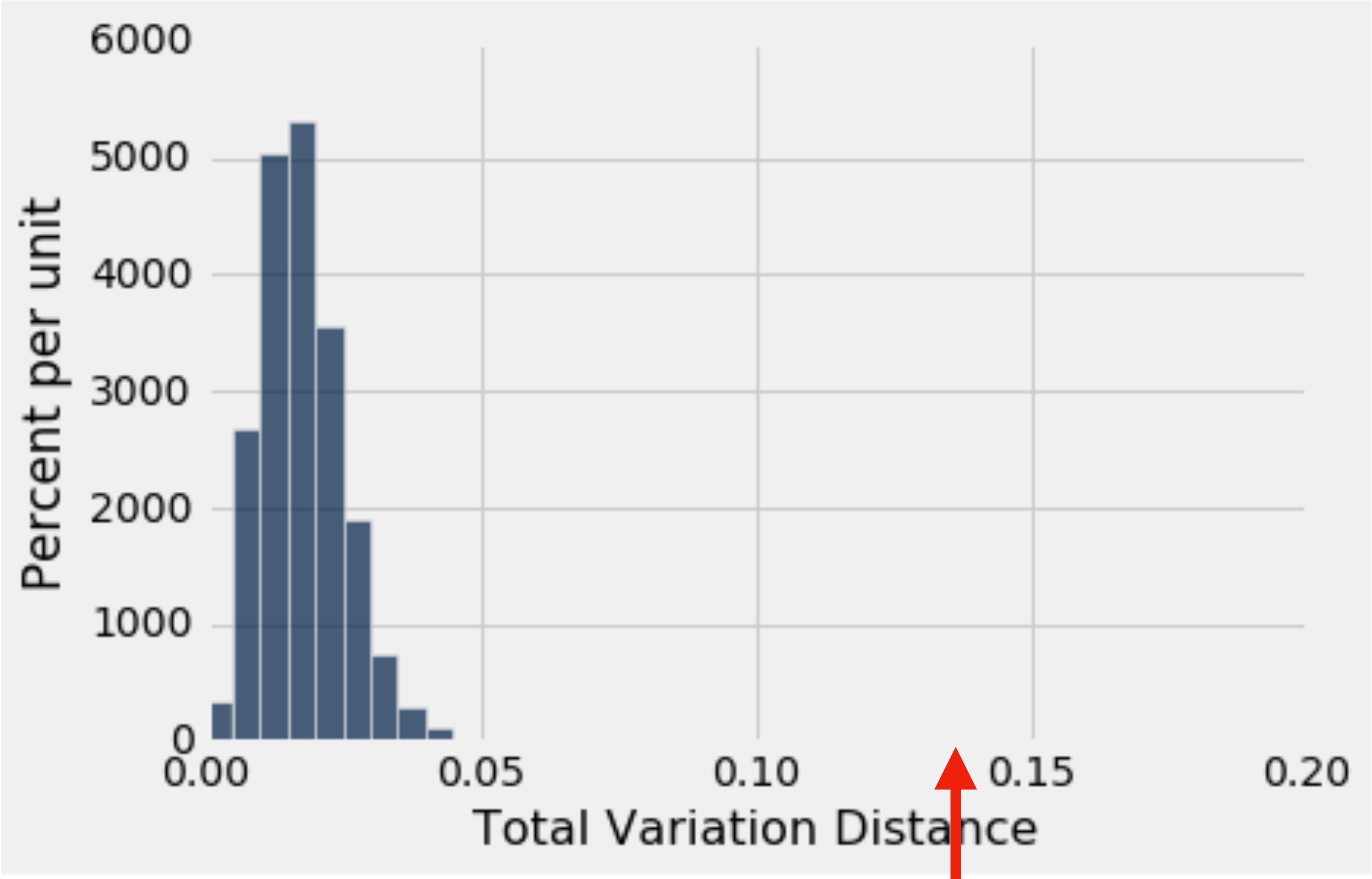
In a histogram, the area of each bar is the percent of individuals in the corresponding bin

Our Examples So Far: Tail Area

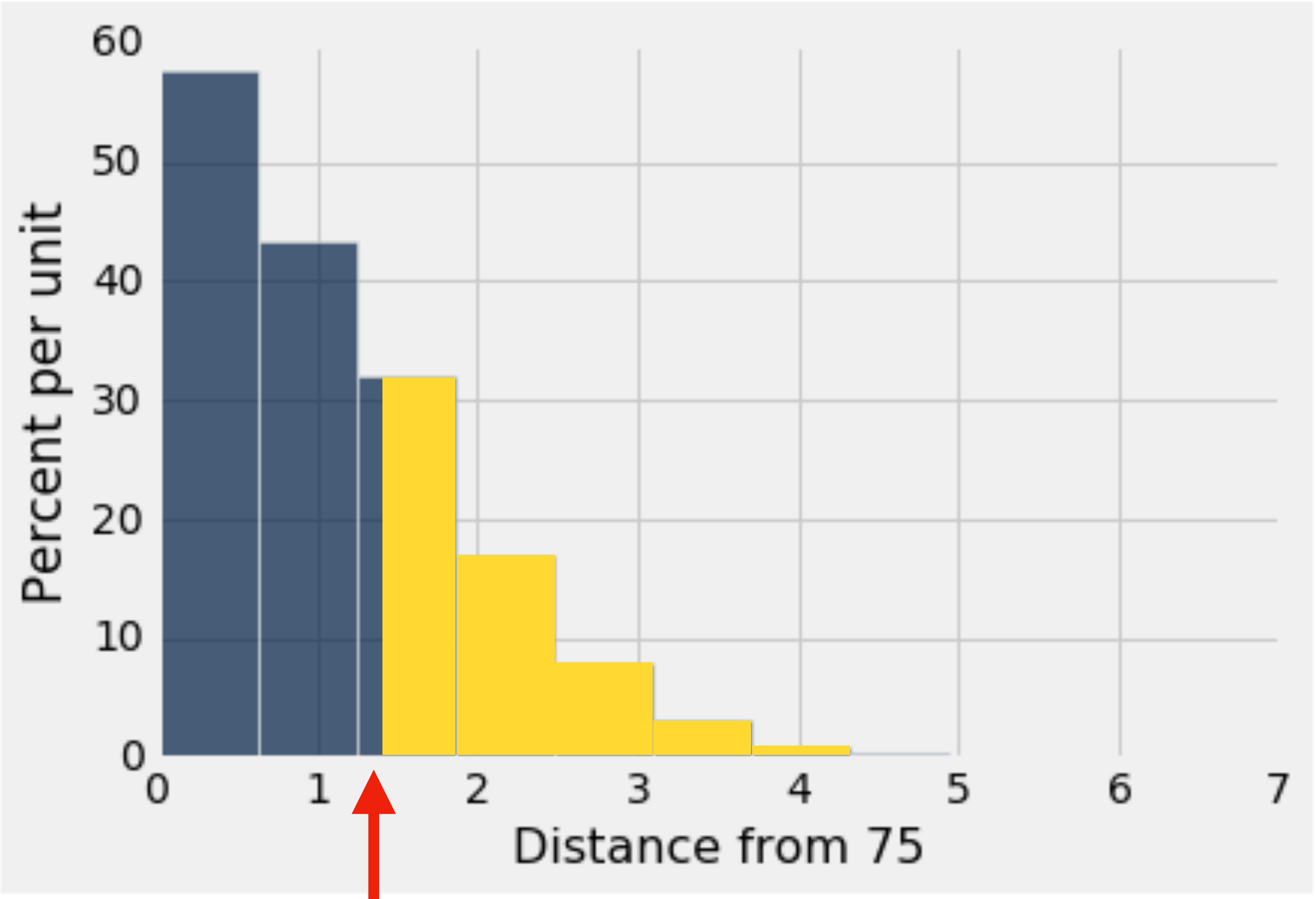
Swain v Alabama



Alameda Jury



Pea Plants



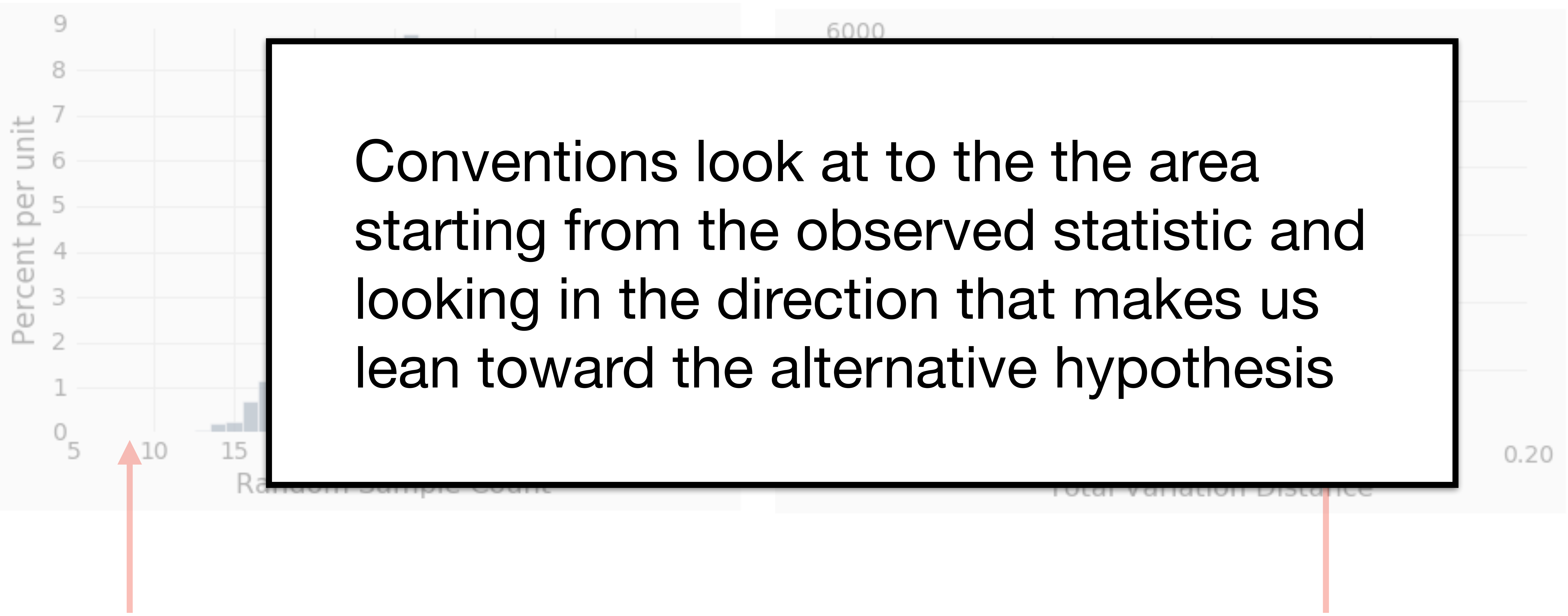
Observed Number (8)

Observed TVD (0.14)

Observed Distance (1.32)

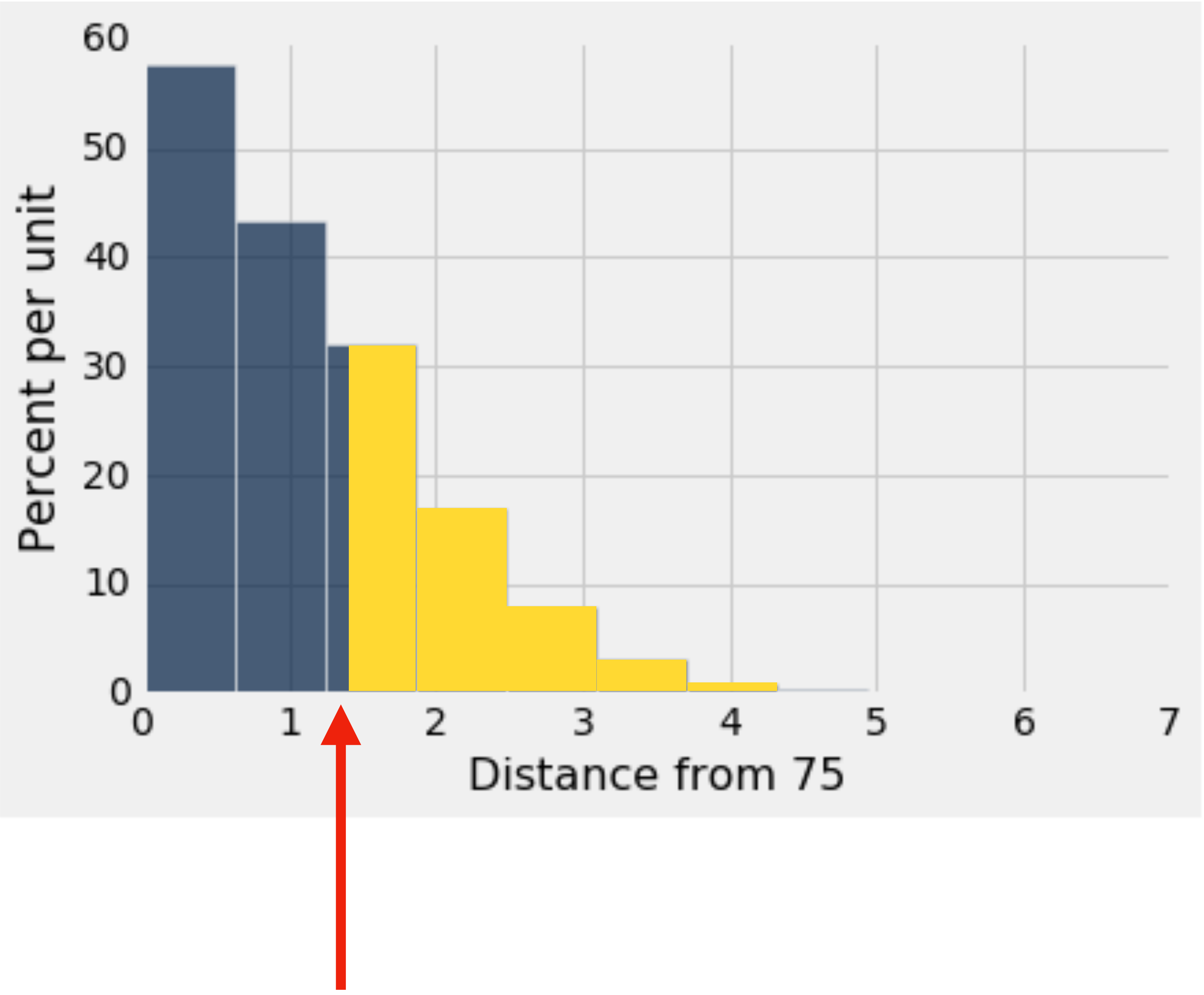
In a histogram, the area of each bar is the percent of individuals in the corresponding bin

Our Examples So Far: Tail Area



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

Pea Plants



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

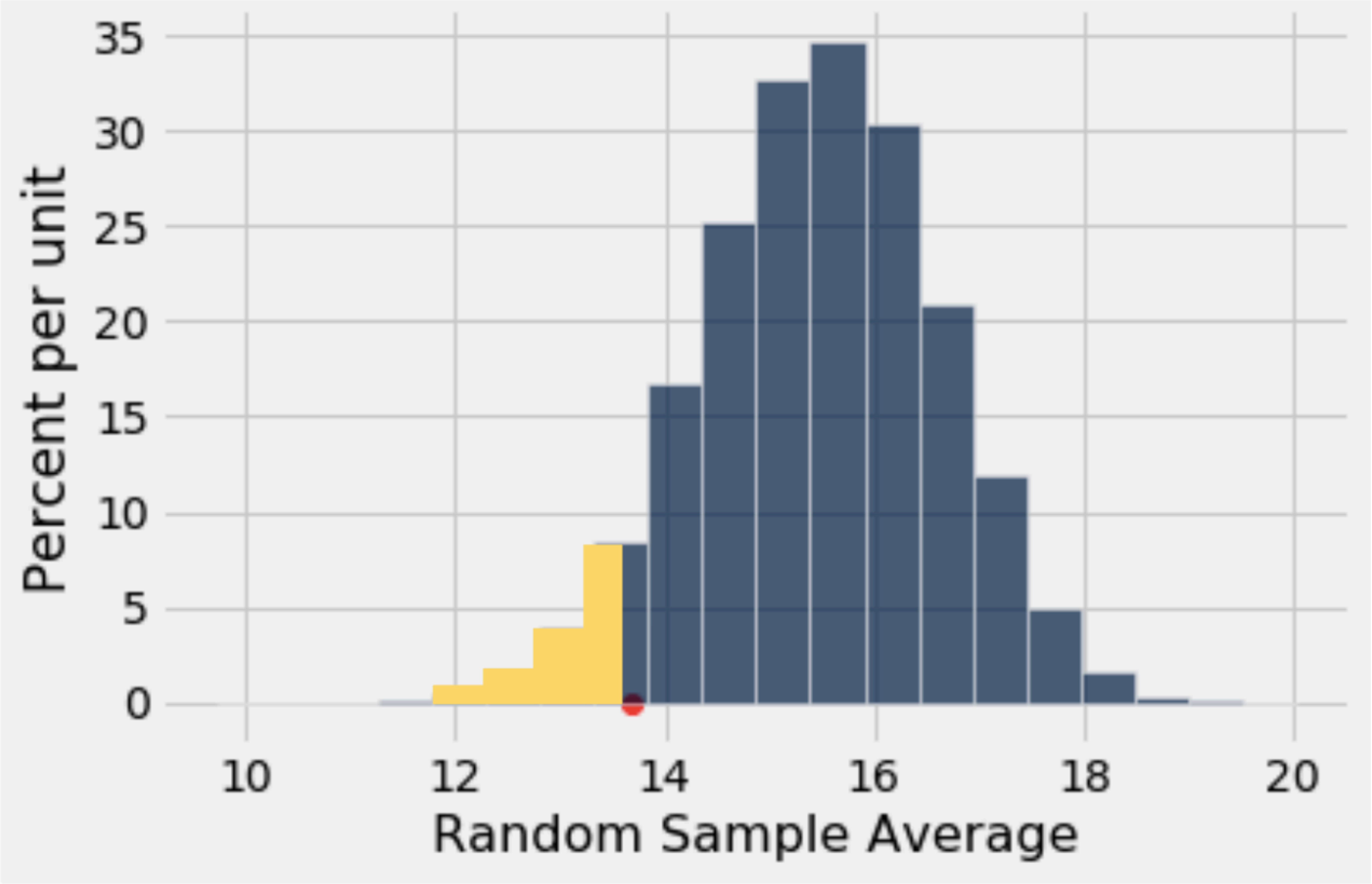
- **"Inconsistent with the null"**: test statistic is in the tail of the empirical distribution under the null hypothesis
- **"In the tail"**
    - **< 5% (Area in the tail is less than 5%)**
        - The result is "statistically significant"
    - **< 1% (Area in the tail is less than 1%)**
        - 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



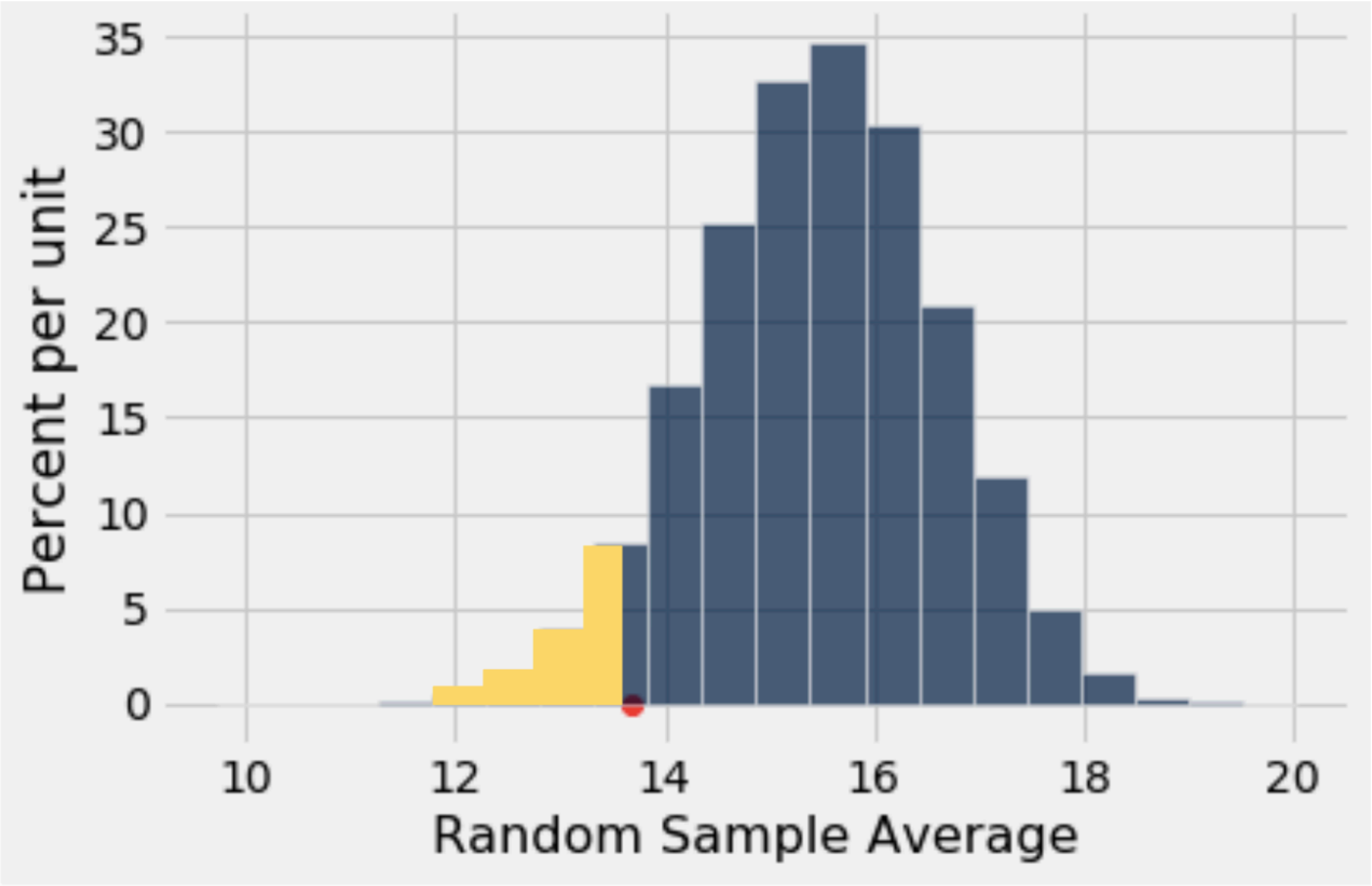
Discussion Questions

What does the red dot represent?

A. Our p-value

B. Our expected outcome

C. Our observed outcome



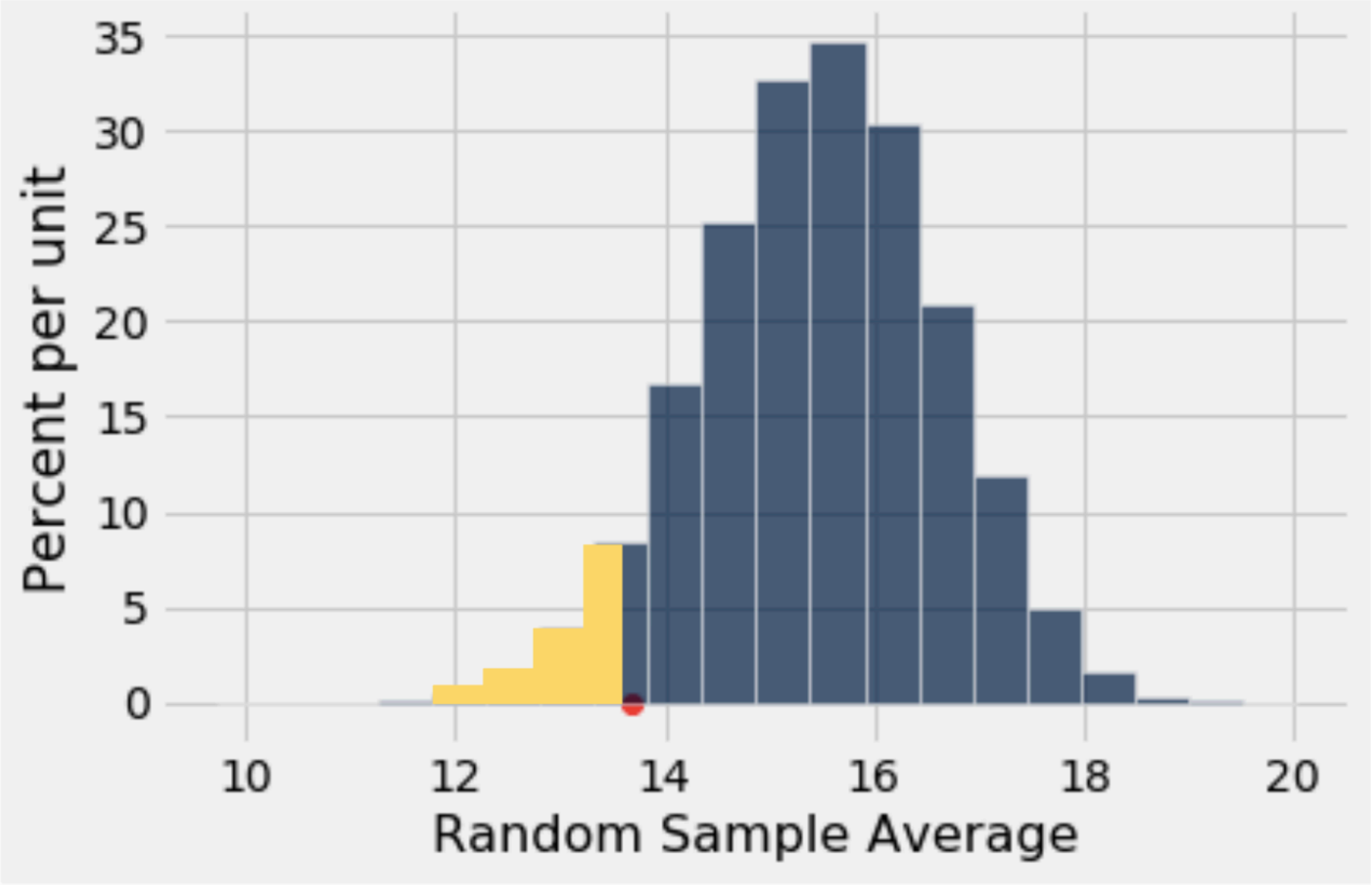
Discussion Questions

What does the red dot represent?

A. Our p-value

B. Our expected outcome

C. Our observed outcome



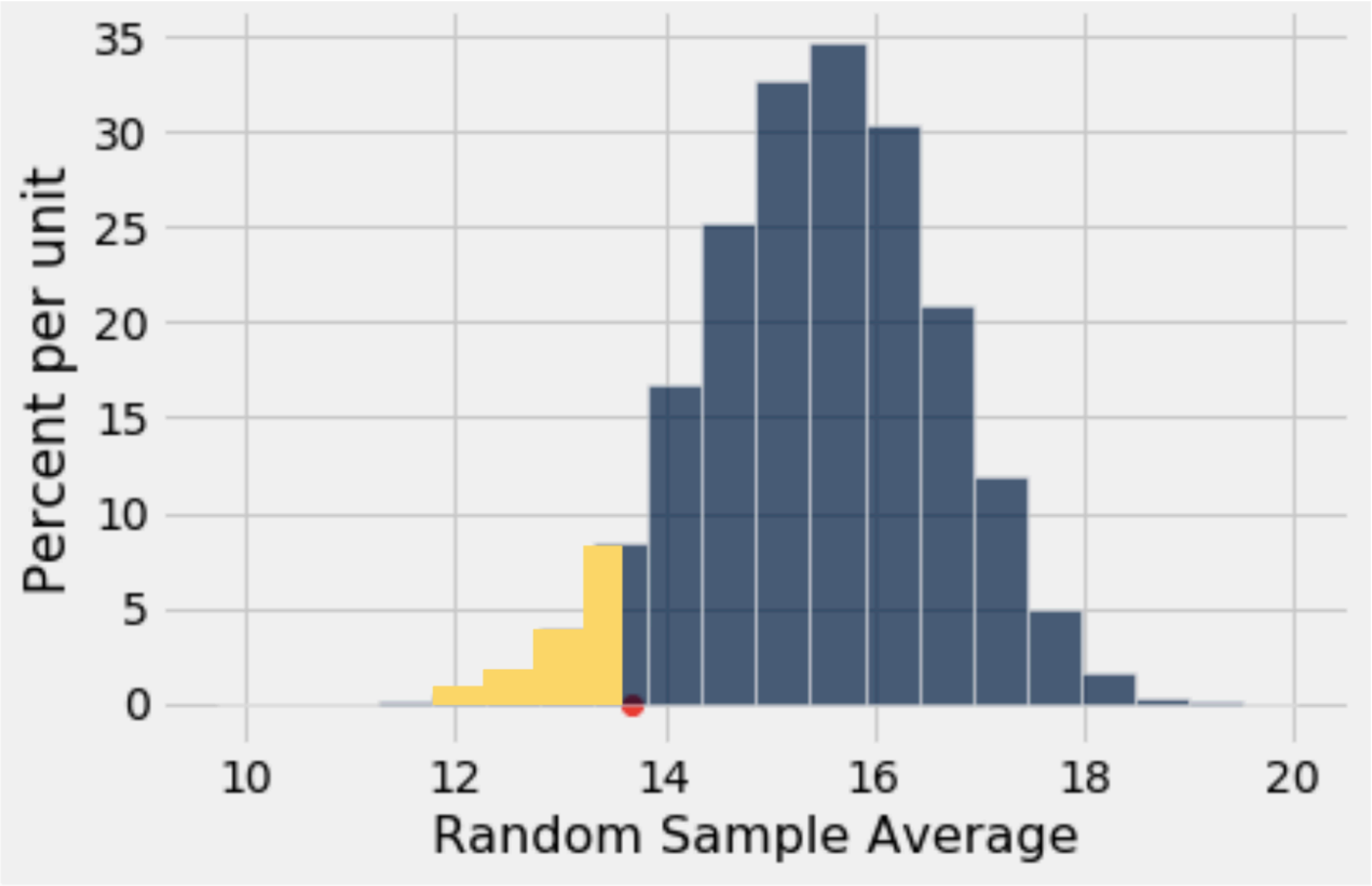
Discussion Questions

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



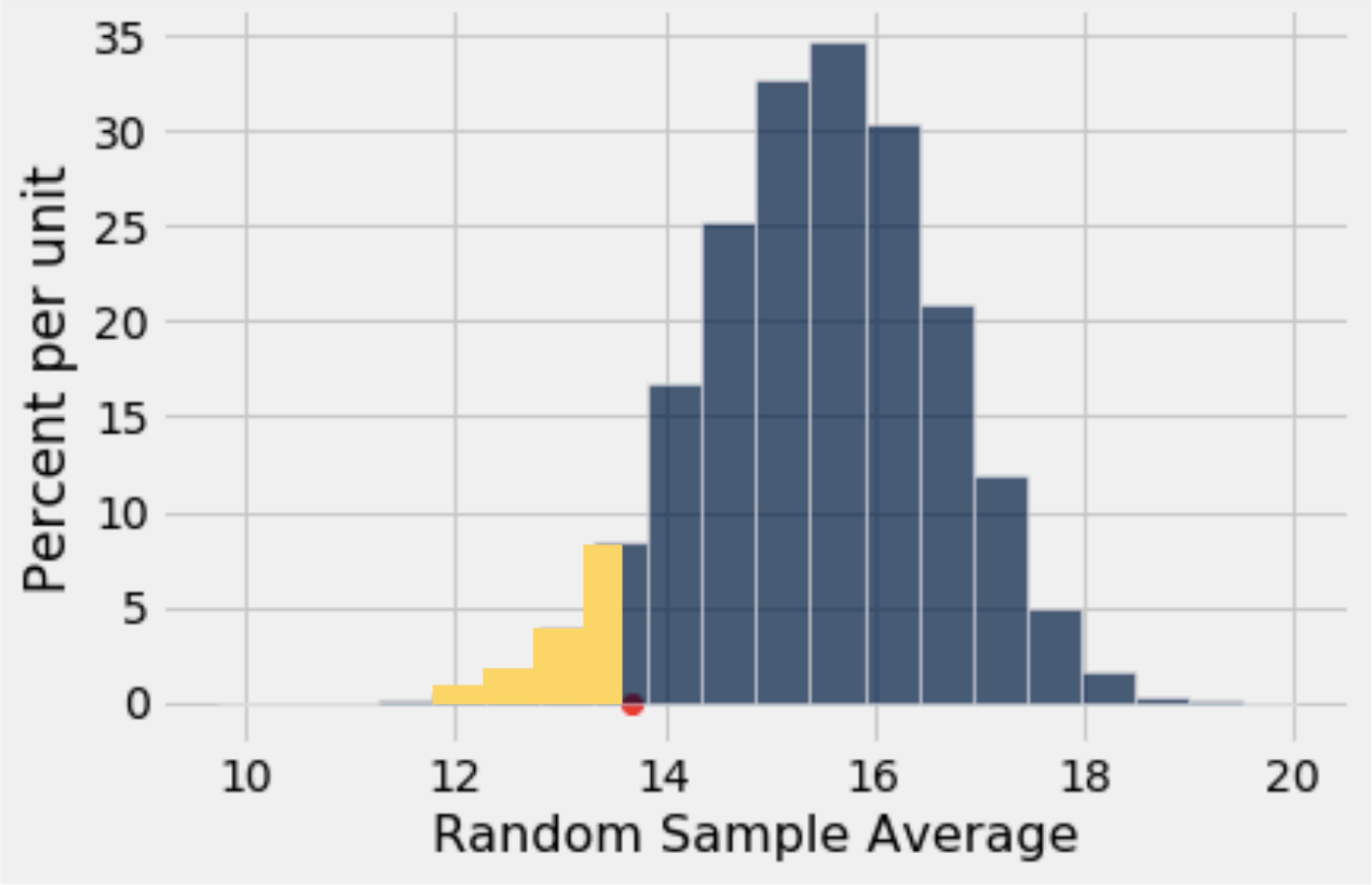
Discussion Questions

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



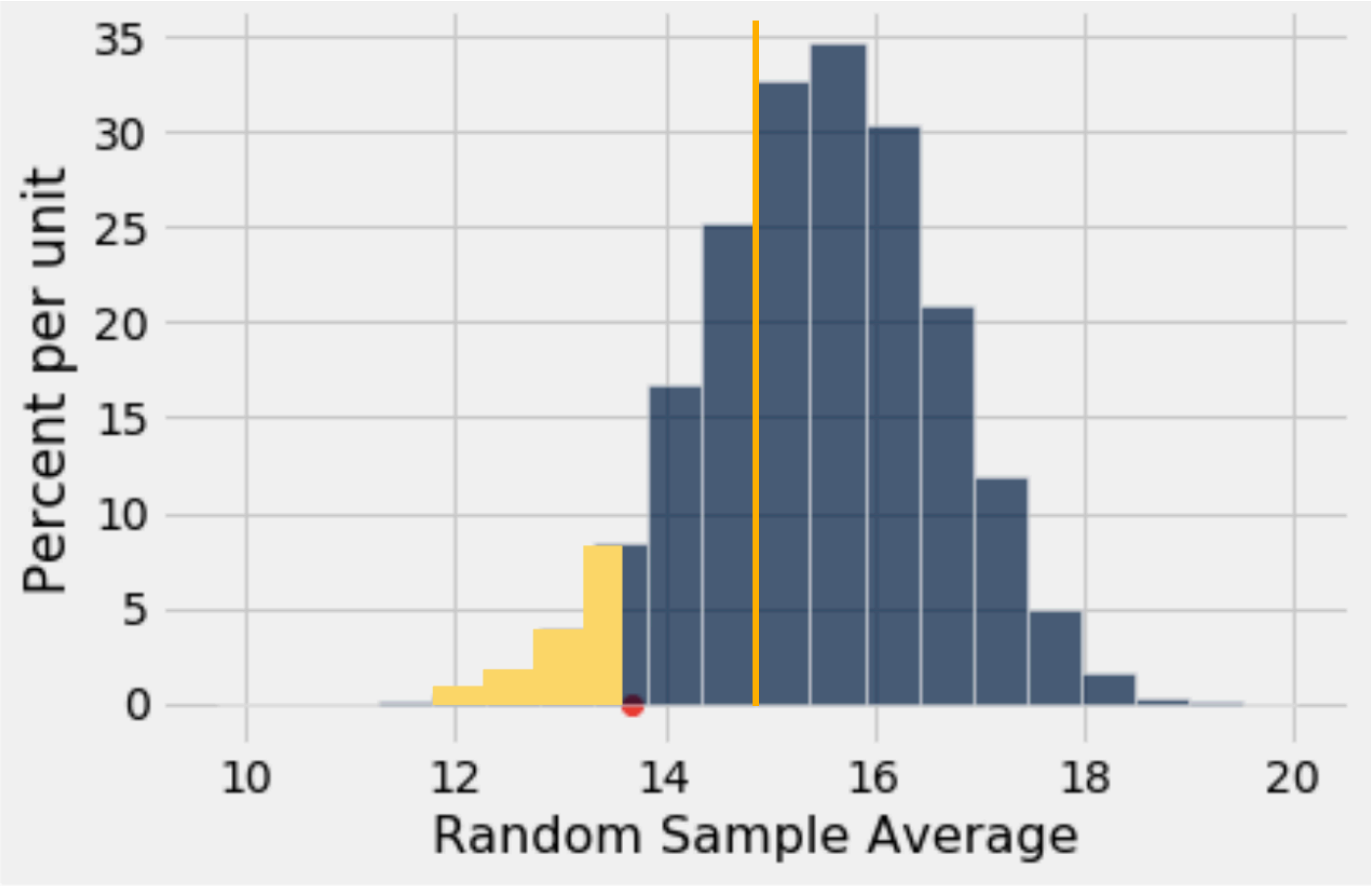
Discussion Questions

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



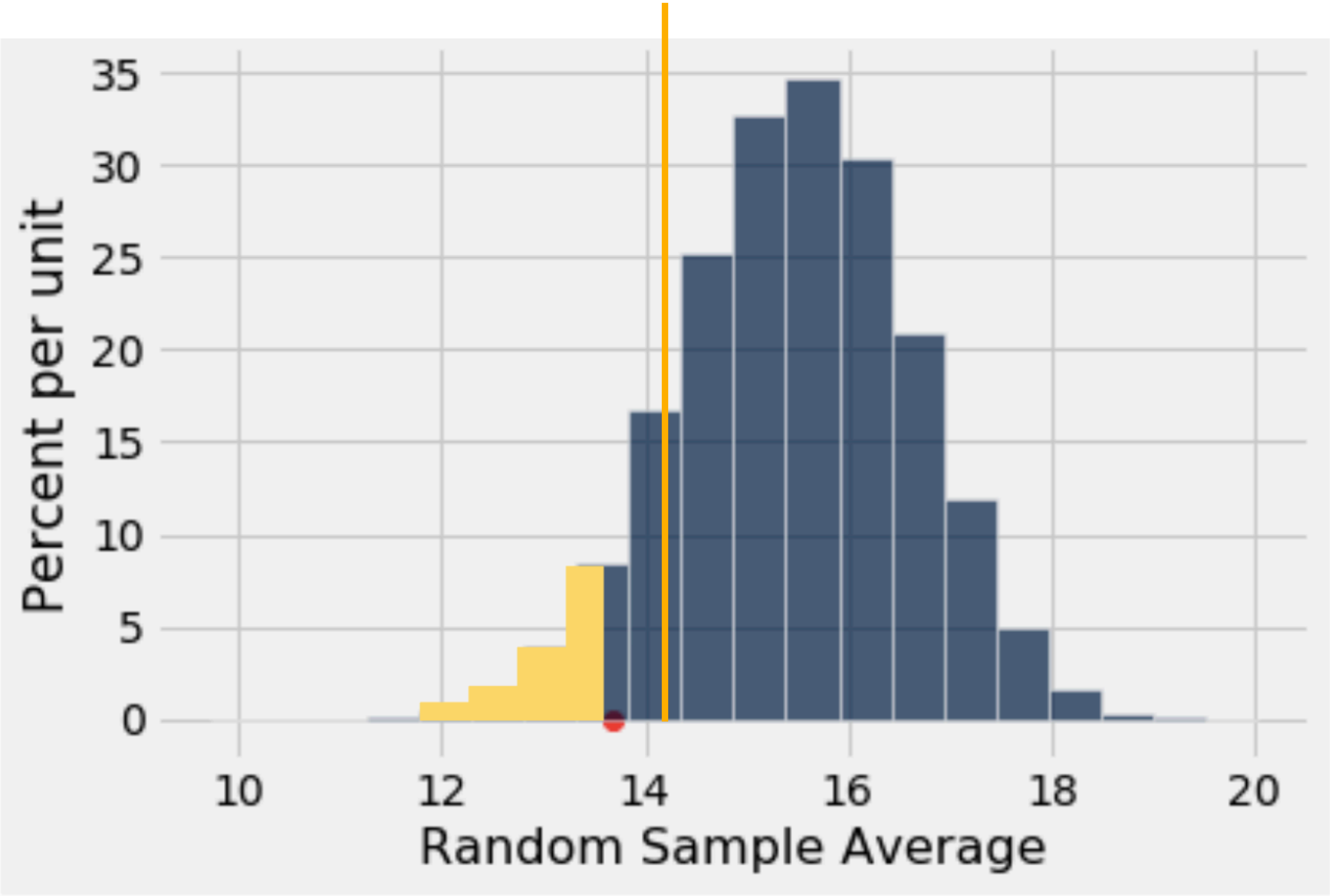
Discussion Questions

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



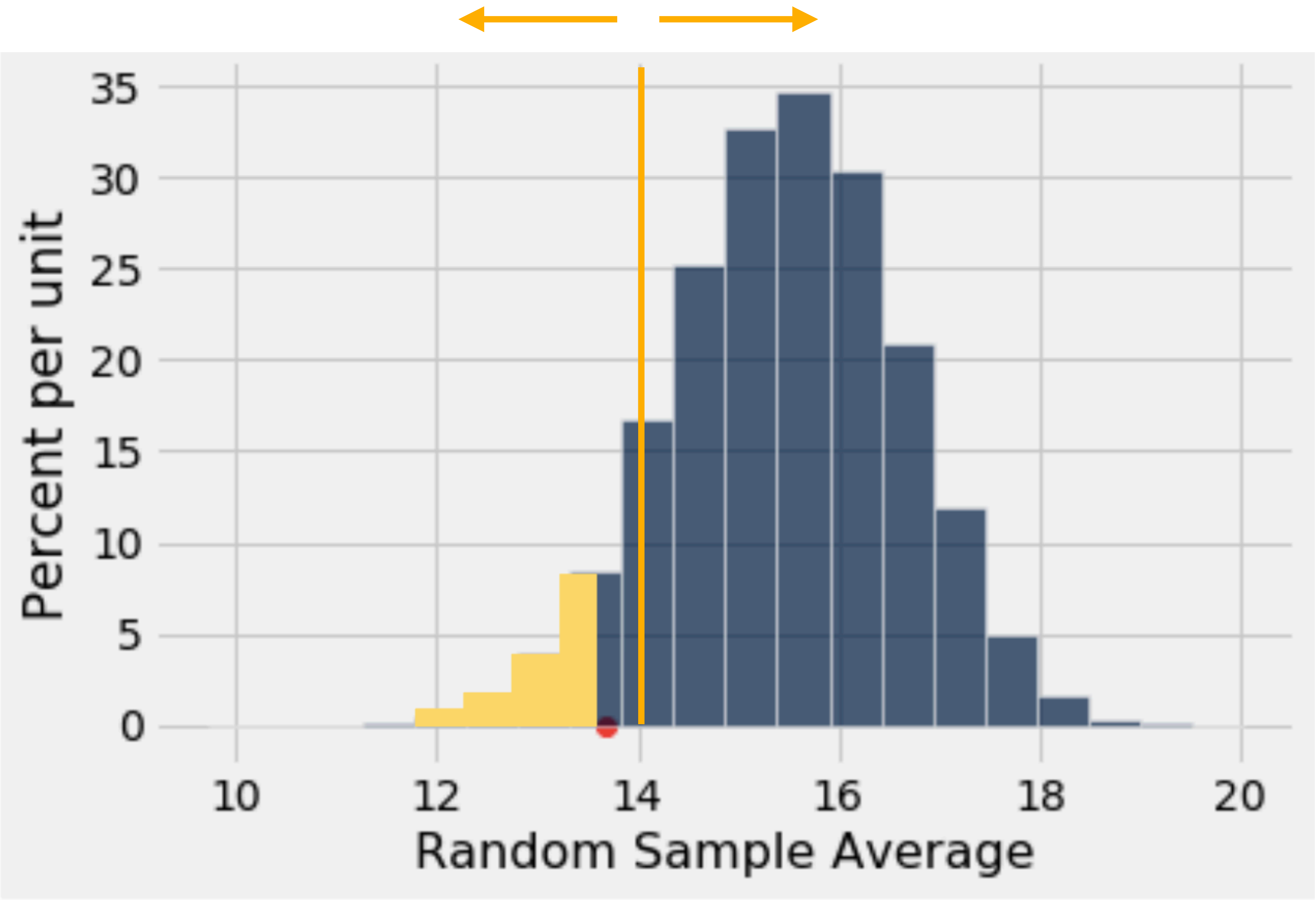
Discussion Questions

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



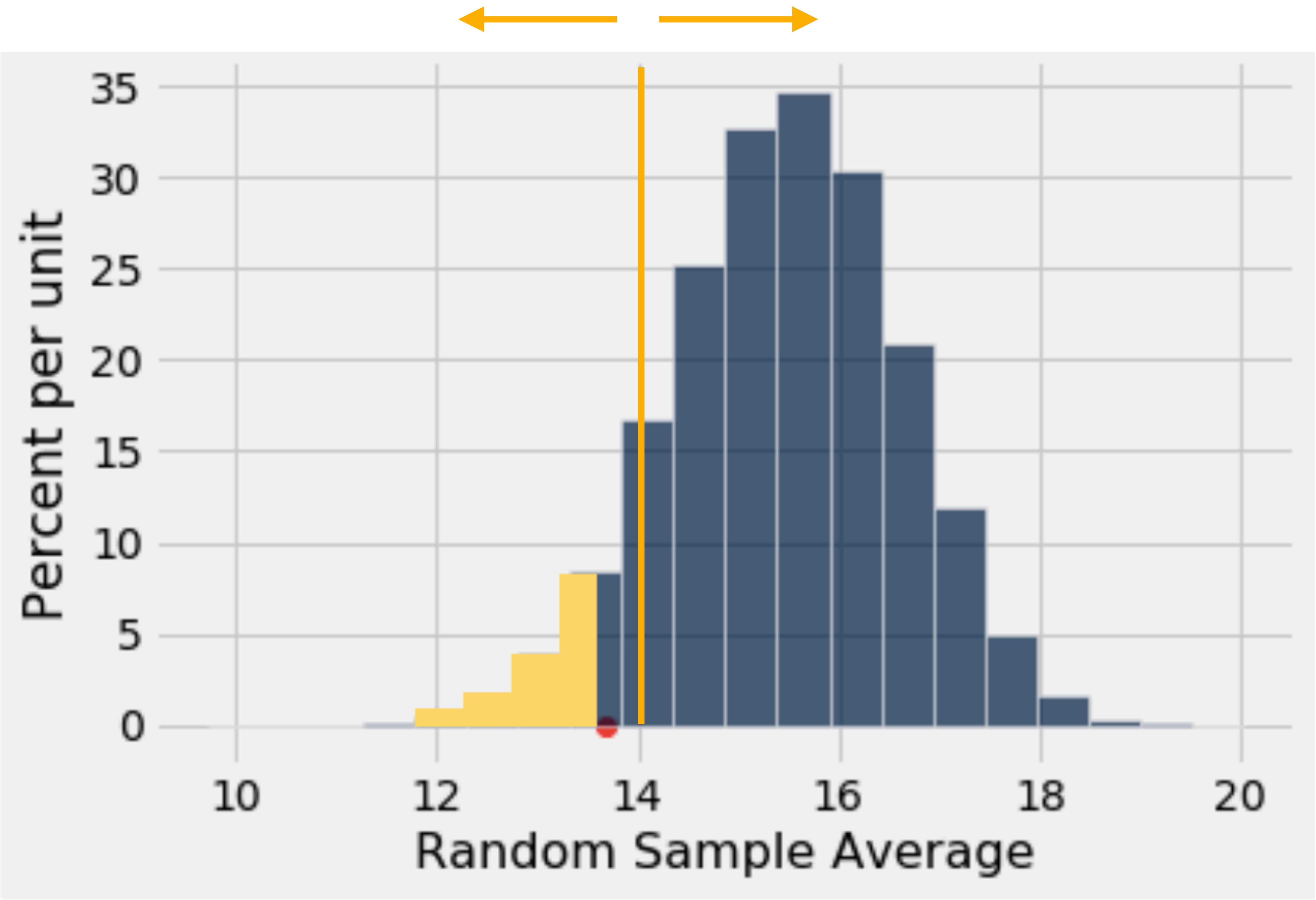
Discussion Questions

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