BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 16 – Significant Testing (P-values) & A/B Testing

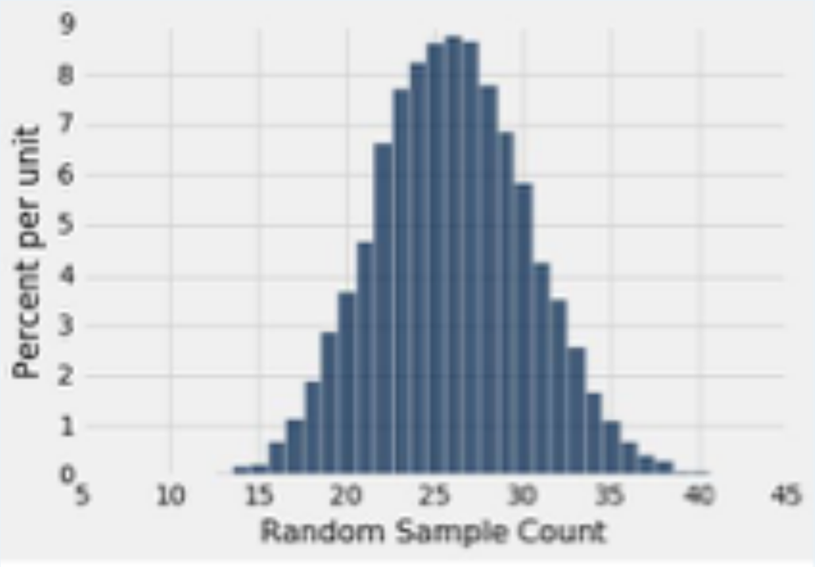
Announcements

- Lab 06 - Inference and the Death Penalty
  - Due Monday 03/28
- HW06 - Testing Hypotheses
  - Due Thursday 03/31

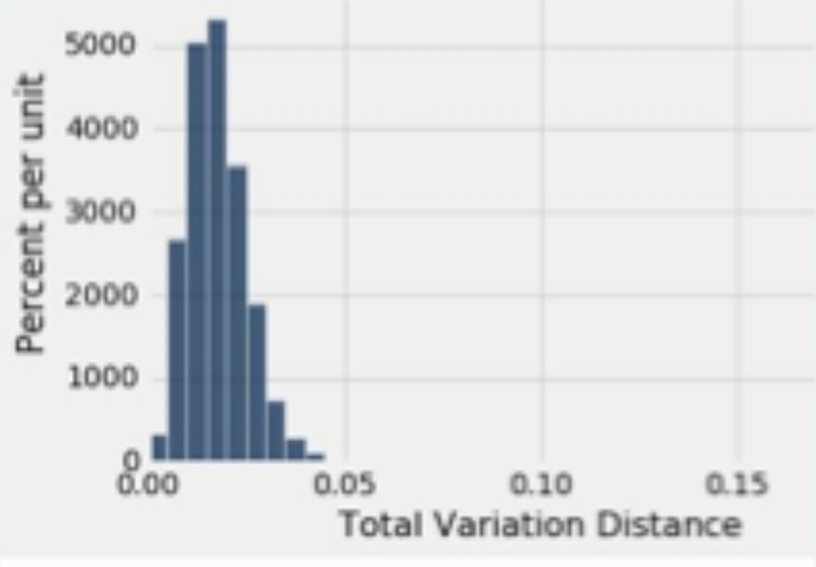
Statistical Significance

Tail Areas

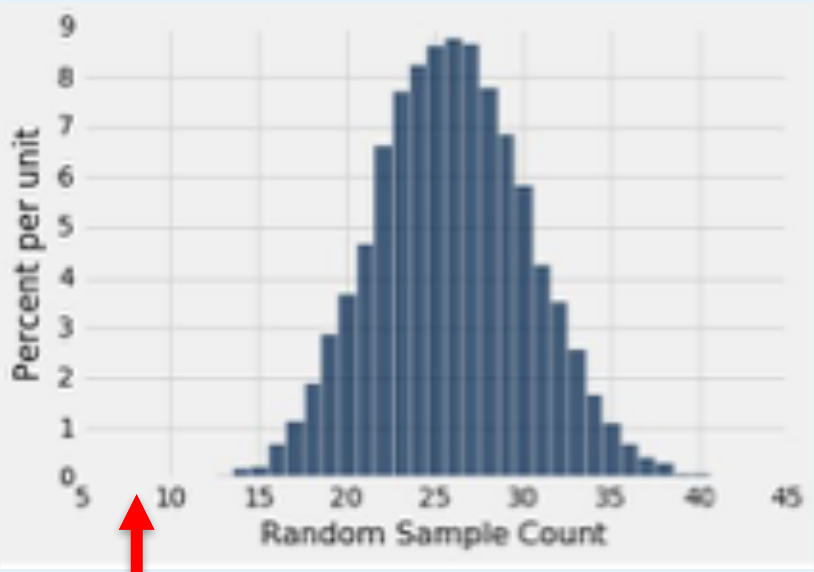
Alabama Jury



Alameda Jury

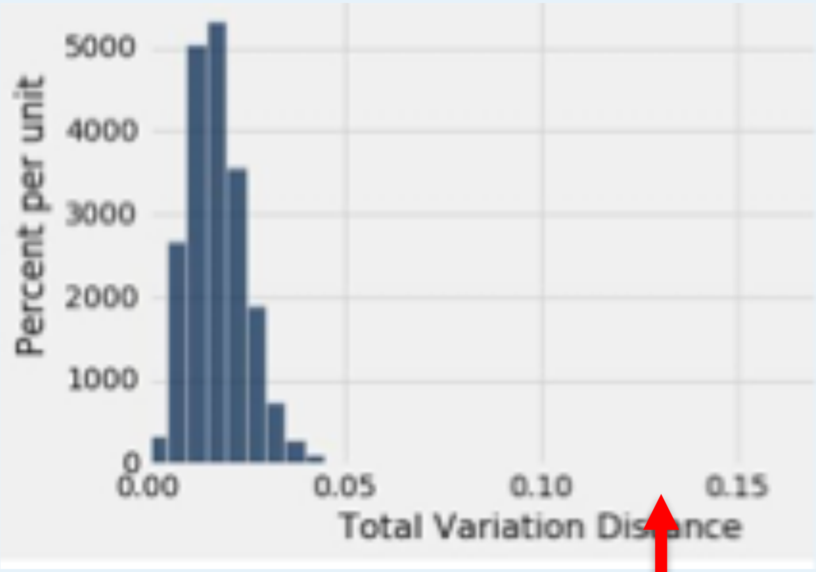


Alabama Jury



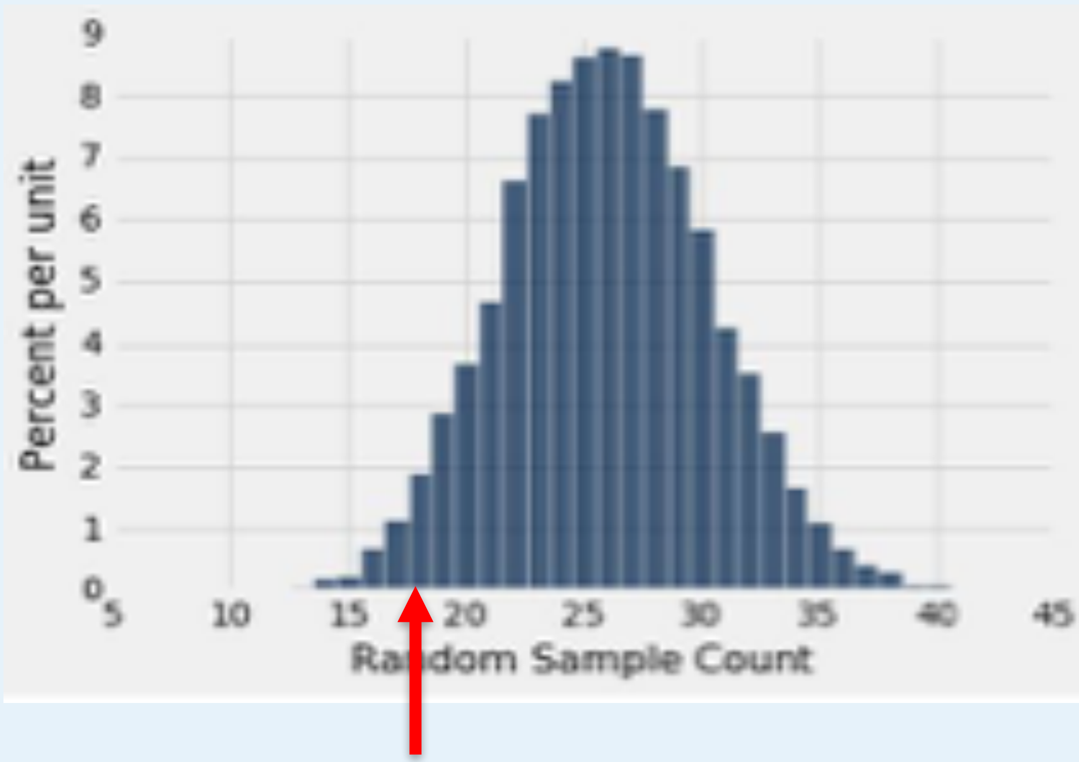
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Not so clear example

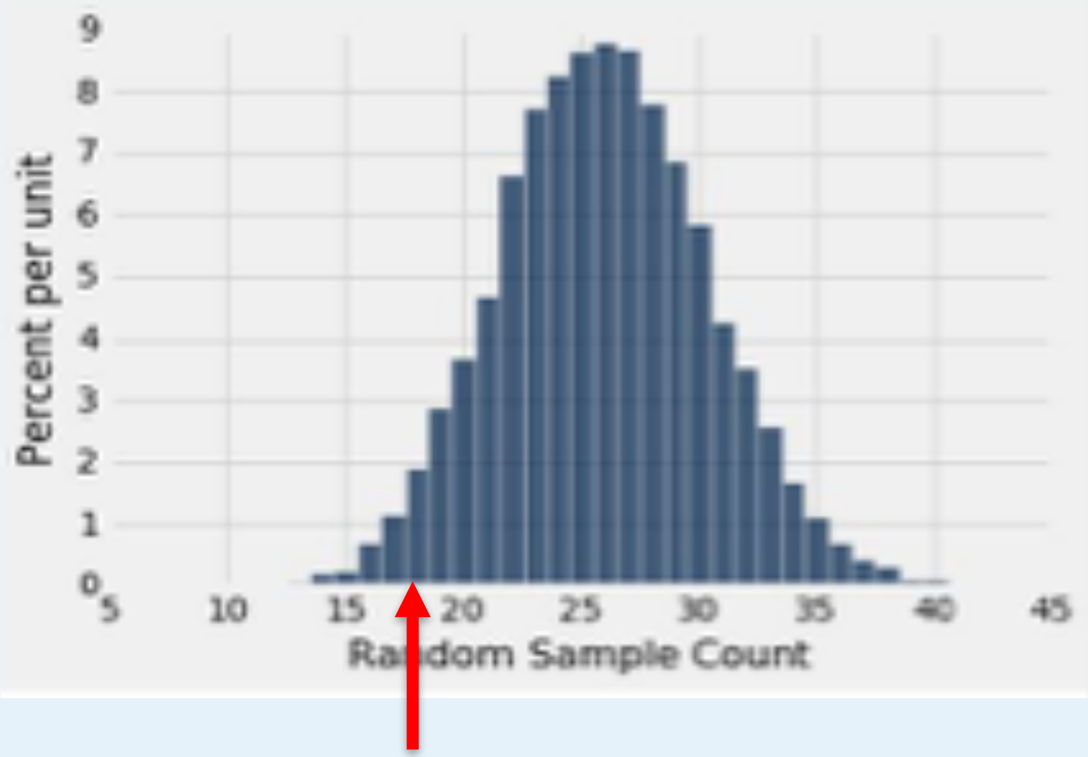


Observed Number (18)

Conventions About Inconsistency

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

Not so clear example



Observed Number (18)

Conventions About Inconsistency

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

- **“In the tail,” first convention:**
  - The area in the tail is less than 5%
  - The result is “statistically significant”

- **”In the tail,” second convention:**
  - The area in the tail is less than 1%
  - The result is “highly statistically significant”

Observed significance level (*aka* P-value)

Formal name: **observed significance level**

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 tail

Example

Scenario: After the midterm, students in a MW lab (of 27 students) noticed that their scores were on average lower than the rest of the class.

Question:

~~Why did the section do worse than others?~~

Potential Answers:

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

Alternative Hypothesis: No, the average is too low

Example

Scenario: After the midterm, students in a MW lab noticed that their scores were on average lower than the rest of the class.

Question:

Did the 27 students do lower by chance?

Potential Answers:

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

Alternative Hypothesis: No, the average is too low

Statistic to measure:

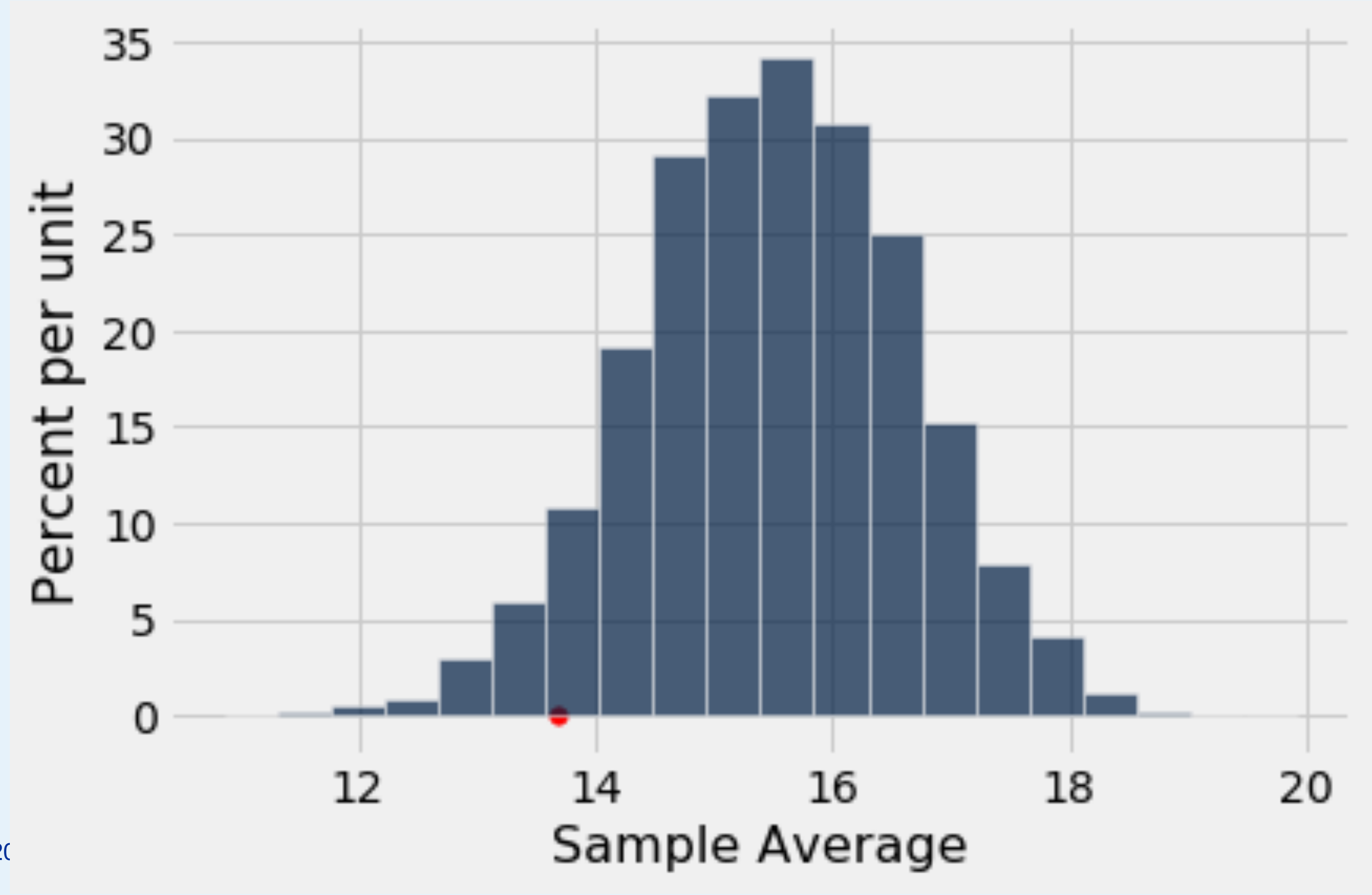
The average score per section (27 students)

Assessing a Model

- Choose a statistic to measure the “discrepancy” between model and data
  - Average score per 27 students
- Simulate the statistic under the model’s assumptions
  - `np.average(scores_only.sample(27, with_replacement=False))`
- Compare the data to the model’s predictions:
  - Draw a histogram of simulated values of the statistic
  - Compute the observed statistic from the real sample

Histogram of simulated values & observed statistic

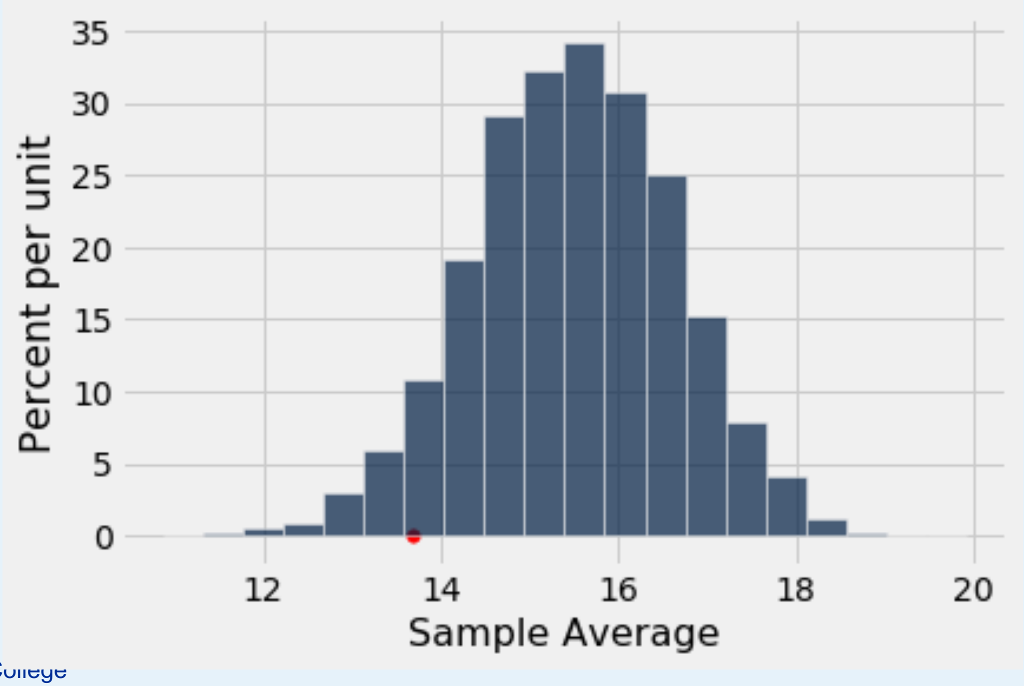
Is the observed statistic consistent with the histogram?



Compute the p-value

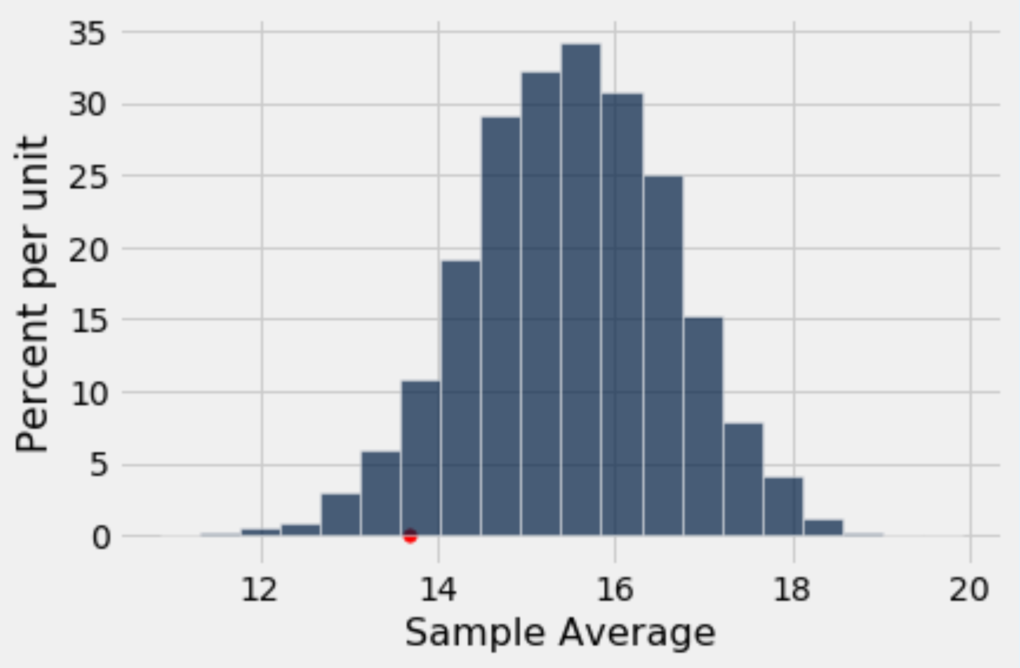
The *P*-value is the chance,

- 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 tail



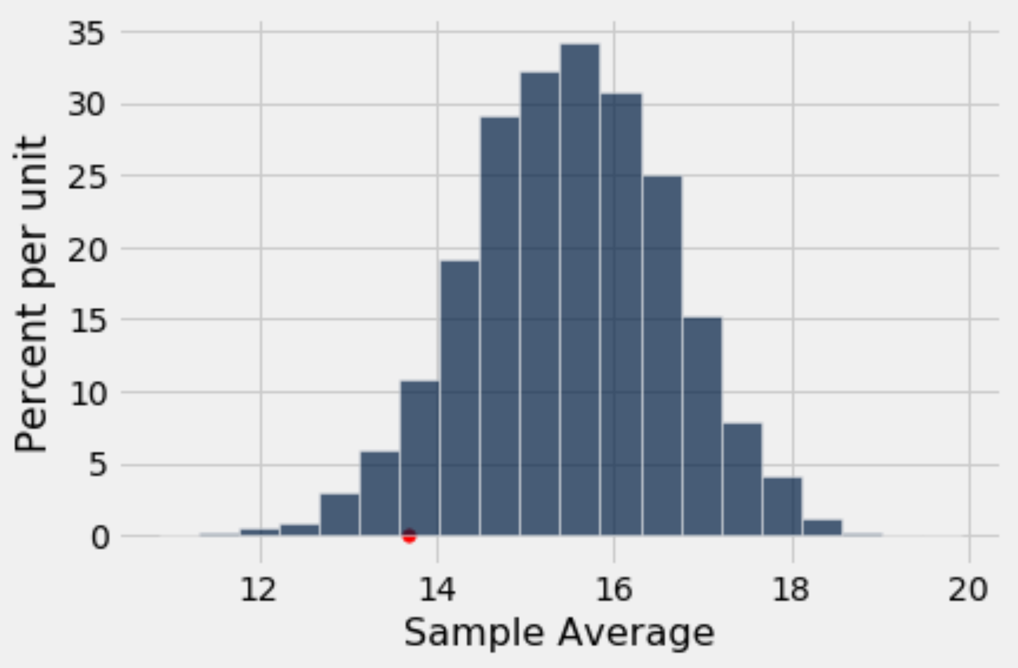
Compute the p-value

$$\text{Probability (A)} = \frac{\textit{number of outcomes that make A happen}}{\textit{total number of outcomes}}$$



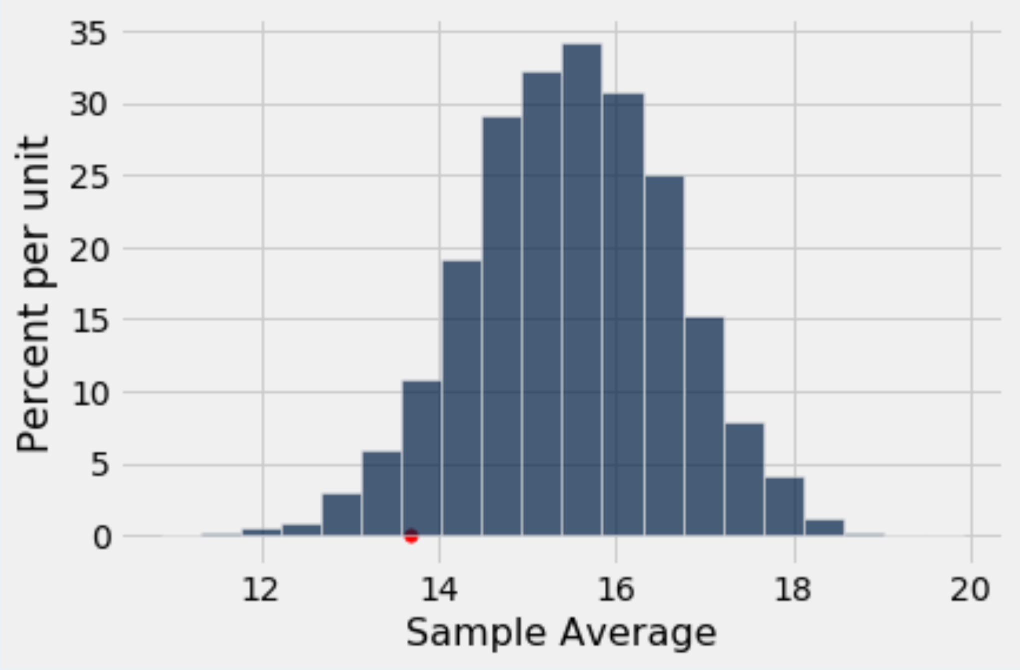
Compute the p-value

A = the sampled statistic was less than or equal to the observed statistic



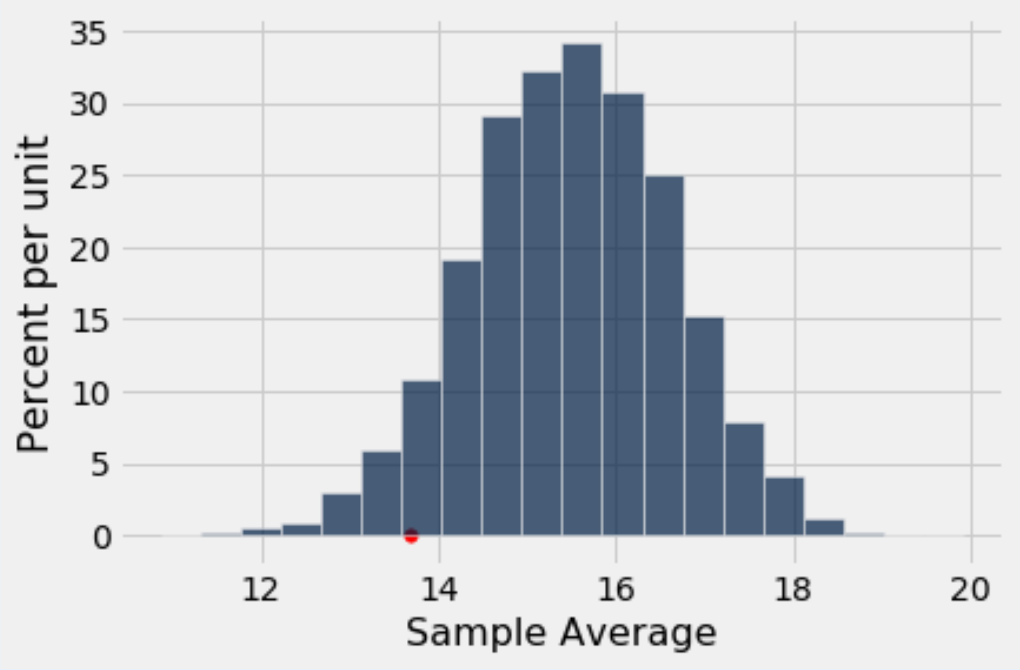
Compute the p-value

P(A) = (the number of times the sampled statistic was less than the observed statistic) divided by the number of samples



Compute the p-value

$$P(A) = \frac{sum(sample\ averages \leq observed\ averages)}{50K}$$

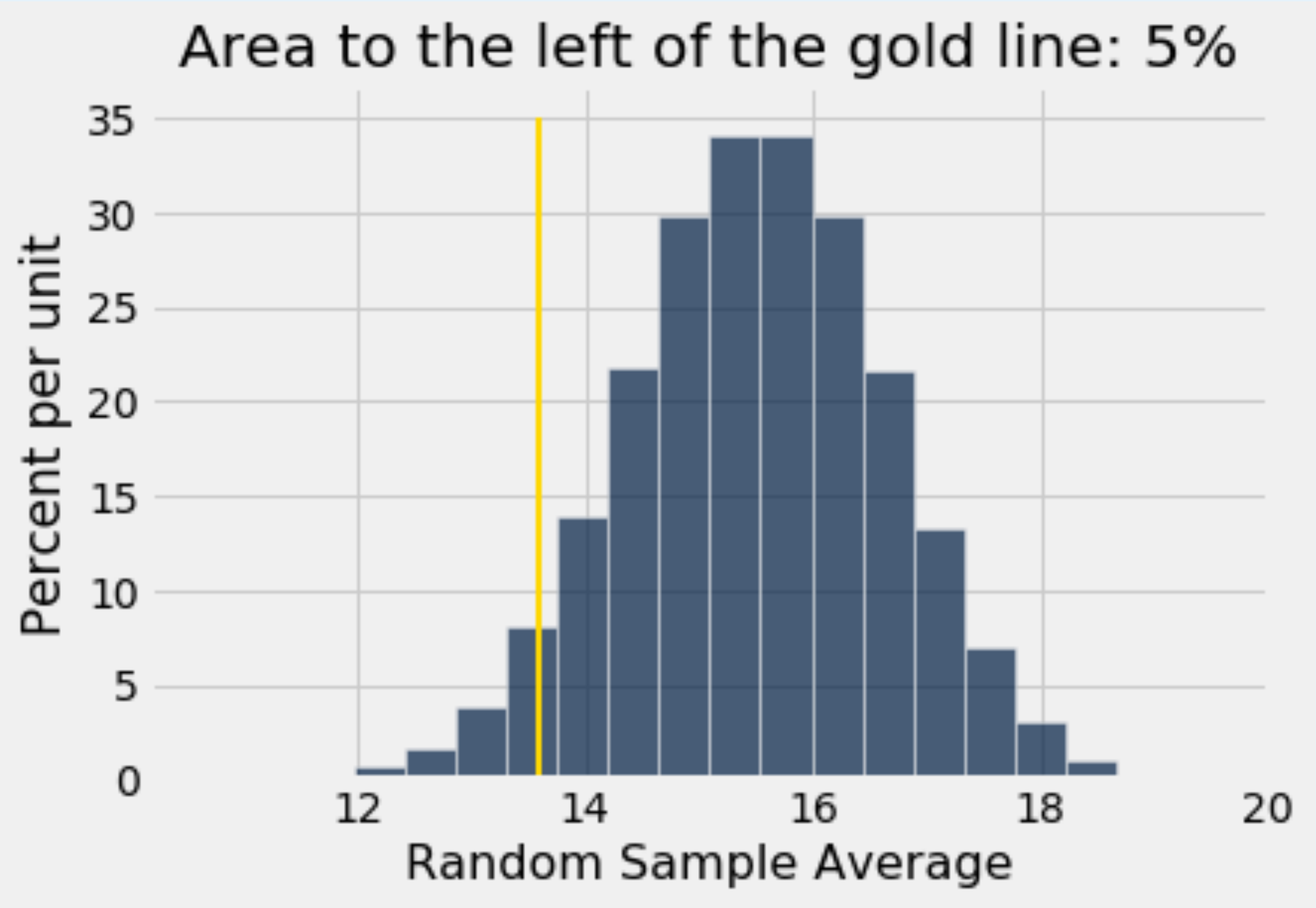


Compute the p-value

$$P(A) = 0.05682 \approx 5\%$$

$$P(A) = 0.05682 \approx 5\%$$

Compute the p-value



Comparing Two Samples
A/B Testing

Terminology

- Compare values of sampled *individuals* in **Group A** with values of sampled *individuals* in **Group B**.
- Question: Do the two sets of values come from the same underlying distribution?
- Answering this question by performing a statistical test is called **A/B testing**.

The Groups and the Questions

- Random sample of mothers of newborns. Compare:
  A. Birth weights of babies of mothers who smoked during pregnancy
  B. Birth weights of babies of mothers who didn’t smoke

- Question: Could the difference be due to chance alone?

Hypotheses

Null Hypothesis:

- In the population, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance.)

Alternative Hypothesis:

- In the population, the babies of the mothers who smoked weigh less, on average, than the babies of the non-smokers

Test Statistic

Group A: non-smokers

Group B: smokers

Statistic:

- Difference between average weights:
  - Group B average - Group A average

Negative values of this statistic favor the alternative

ating Under the Null

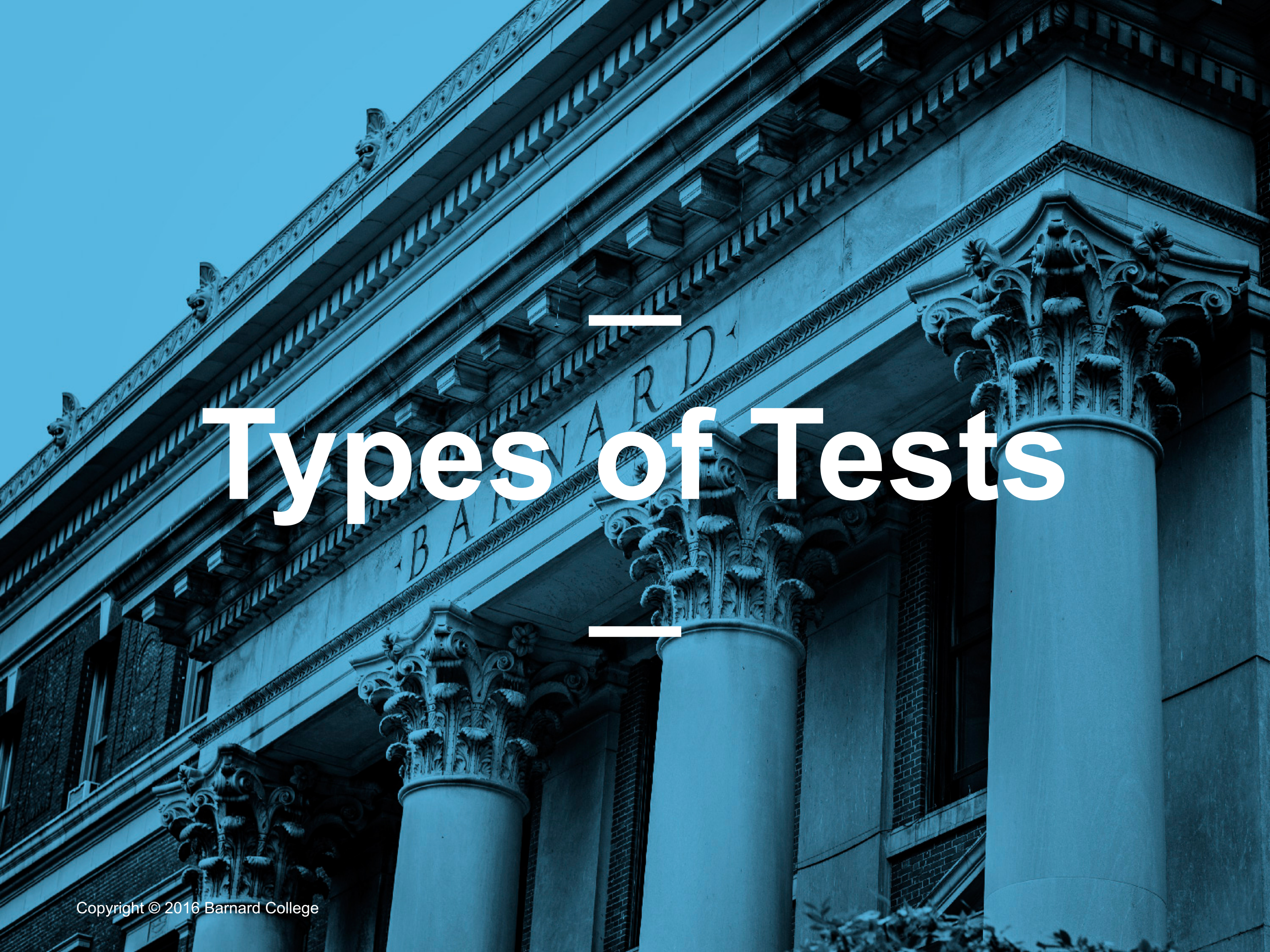
If the null is true, all rearrangements of labels are equally likely

Permutation Test:

- Shuffle all birth weights
- Assign some to Group A and the rest to Group B
  - Key: keep the sizes of Group A and Group B that same from before
- Find the difference between the two shuffled groups
- Repeat

Random Permutations

- tbl.sample(n)
  Table of n rows picked randomly with replacement
- tbl.sample()
  - Table with same number of rows as original **tbl**,
- 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



Types of Tests

Hypothesis Testing Review

1 Sample: One Category *(e.g. percent of black male jurors)*

- Test Statistic: empirical_percent, abs(empirical_percent - null_percent)
- How to Simulate: sample_proportions(n, null_dist)

1 Sample: Multiple Categories *(e.g. ethnicity distribution of jury panel)*

- Test Statistic: tvd(empirical_dist, null_dist)
- How to Simulate: sample_proportions(n, null_dist)

1 Sample: Numerical Data *(e.g. scores in a lab section)*

- Test Statistic: empirical_mean, abs(empirical_mean - null_mean)
- How to Simulate: population_data.sample(n, with_replacement=False)

2 Samples: Numerical Data *(e.g. birth weights of smokers vs. non-smokers)*

- Test Statistic: group_a_mean - group_b_mean,
  - group_b_mean - group_a_mean, abs(group_a_mean - group_b_mean)
- How to Simulate: empirical_data.sample(with_replacement=False)