BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 15 – P-values & Comparing Two Samples

Announcements

- Today's office hours – cancelled but I'll still around a bit

- Lab 06 - Inference and the Death Penalty
  - Due Monday 03/28

- HW06 - Testing Hypotheses
  - Due Thursday 03/31

Mid semester survey

- Thanks for your feedback!!!
  - Form is staying open
- Homeworks will be released more timely
  - HW06, HW07, HW08 are all already posted
- Moved HW deadlines to Thursday
  - Might move hw07 back, depending on our progress in class
- Speak up in lecture:
  - No news means good news
  - If you have a question or are confused, other people are too

HW – readjusting autograder

HW04 question 1.4

> **Question 1.4.** Shoumik wants to see how Columbia did against every opponent during the 2019 season. Using the `final_scores` table, assign `results` to an array of `True` and `False` values that correspond to whether or not Columbia won. Add the `results` array to the `final_scores` table, and assign this to `final_scores_with_results`. Then, respectively assign the number of wins and losses Columbia had to `cu_wins` and `cu_losses`.

If your code printed out the correct line: 3 wins and 7 losses but you got points off the autograder because you used strings and not Booleans, let us know and we'll fix the grade

HW02 question 1.6

If the autograder failed because the order of the tables is wrong, let us know

Review: Assessing Models

Models

- A model is a set of assumptions about the data
- In data science, many models involve assumptions about processes that involve randomness:
  - "Chance models"
- **Key question:** does the model fit the data?

Approach to Assessing Models

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts

- We can compare the model's predictions (simulations) to the observed data
  - Here, "observed data" == what actually happened

- If the data and the model's predictions are not consistent, that is evidence against the model

Steps in Assessing a Model

- Choose a statistic to measure the “discrepancy” between model and data
- Simulate the statistic under the model’s assumptions
- 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
- If the observed statistic is far from the histogram, that is evidence against the model

Comparing Distributions

Comparing Distributions

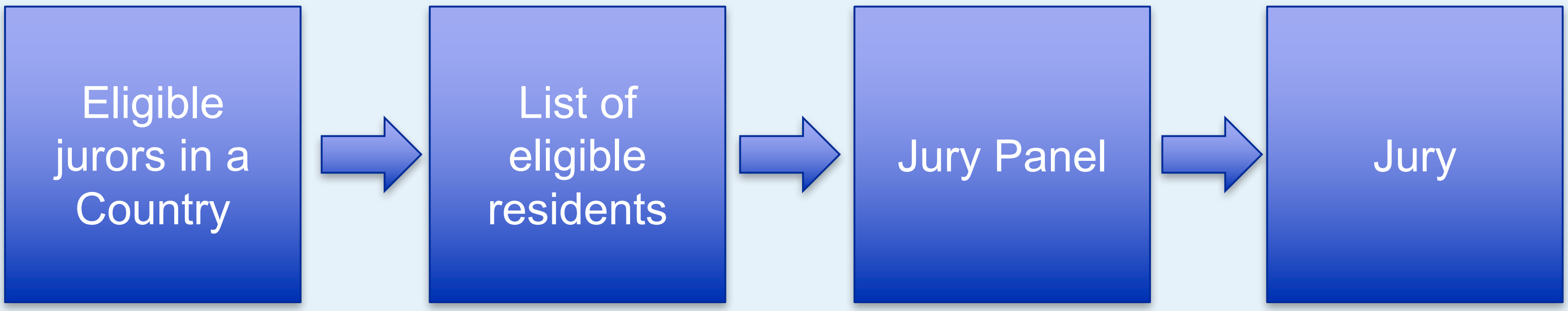
Jury Selectin in Alameda County

RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California — October 2010

https://www.aclunc.org/sites/default/files/racial_and_ethnic_disparities_in_alameda_county_jury_pools.pdf

Jury Panels



Section 197 of California's Code of Civil Procedure says, "All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court."

Model and Alternative

- Model:
  - The people on the jury panels were selected at random from the eligible population
- Alternative viewpoint:
  - No, they weren’t chosen at random
- What are we comparing here?

A New Statistic

Distance Between Distributions

- People on the panels are of multiple ethnicities
- Distribution of ethnicities is:
  - categorical or numerical?

- To see whether the distribution of ethnicities of the panels is close to that of the eligible jurors, we have to measure the distance between two categorical distributions

Total Variation Distance

Every distance has a computational recipe

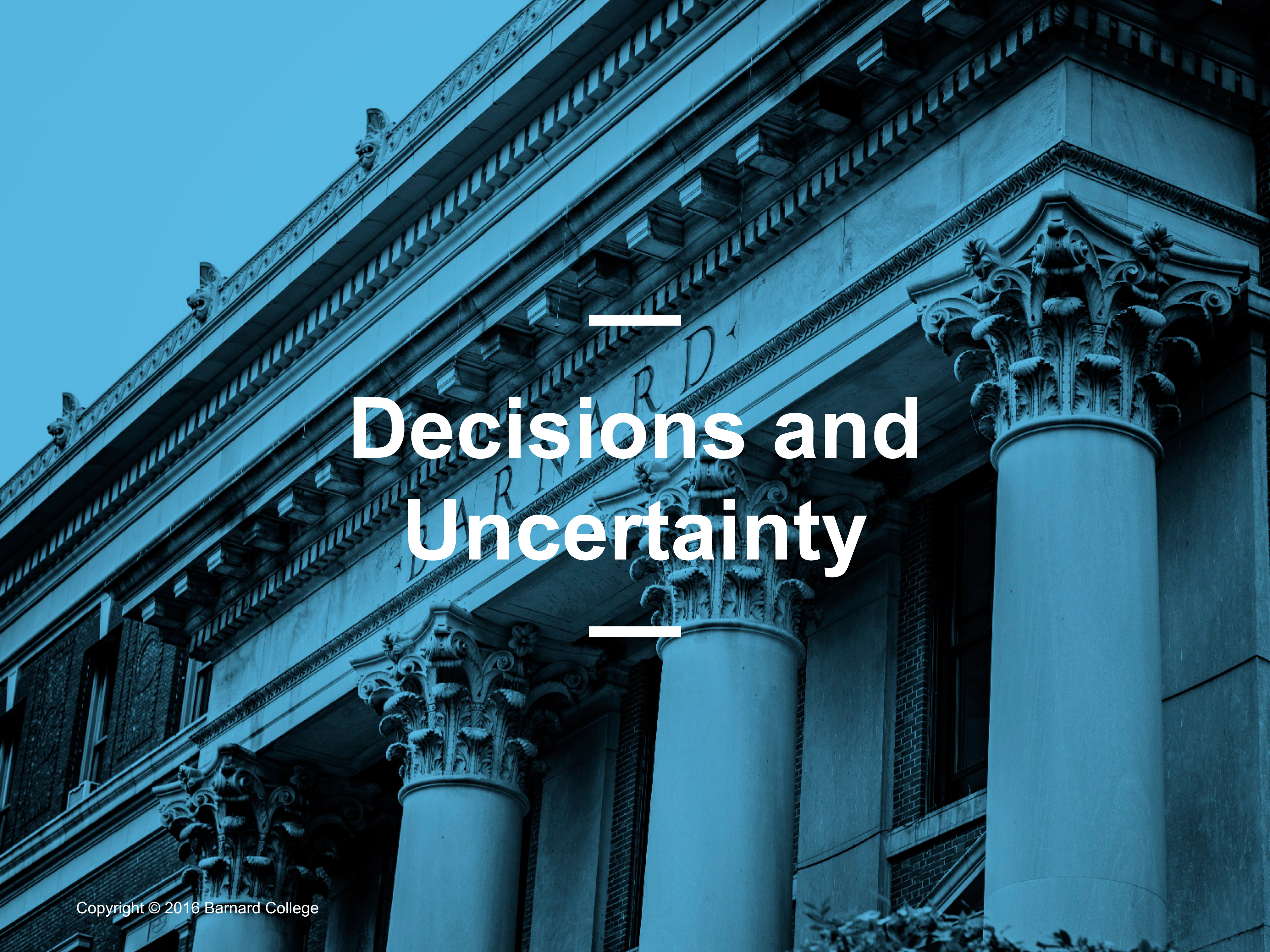
Total Variation Distance (TVD):

- For each category, compute the difference in proportions between two distributions
- Take the absolute value of each difference
- Sum, and then divide the sum by 2

Summary of the Method

To assess whether a sample was drawn randomly from a known categorical distribution:

- Use Total Variation Distance as the statistic:
  - TVD measures the distance between categorical distributions
- Sample at random from the population and compute the TVD from the random sample; repeat numerous times
- Compare:
  - Empirical distribution of simulated TVDs with
  - Actual TVD from the sample in the study



Decisions and Uncertainty

Incomplete Information

- We are trying to choose between two views of the world, based on data in a sample.
- It is not always clear whether the data are consistent with one view or the other.
- Random samples can turn out quite extreme. It is unlikely, but possible

Terminology

Null and Alternative

The method only works if we can simulate data under one of the hypotheses.

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model
    - "Under the null hypothesis"
- **Alternative hypothesis:**
  - A different view about the origin of the data

Test Statistic

- The statistic that 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?
  - Preferably, 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 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
  - Also how likely they are (**if the null hypothesis is true)**
- The probabilities are approximate, because we can't generate all the possible 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
  - “data is more consistent with the alternative”

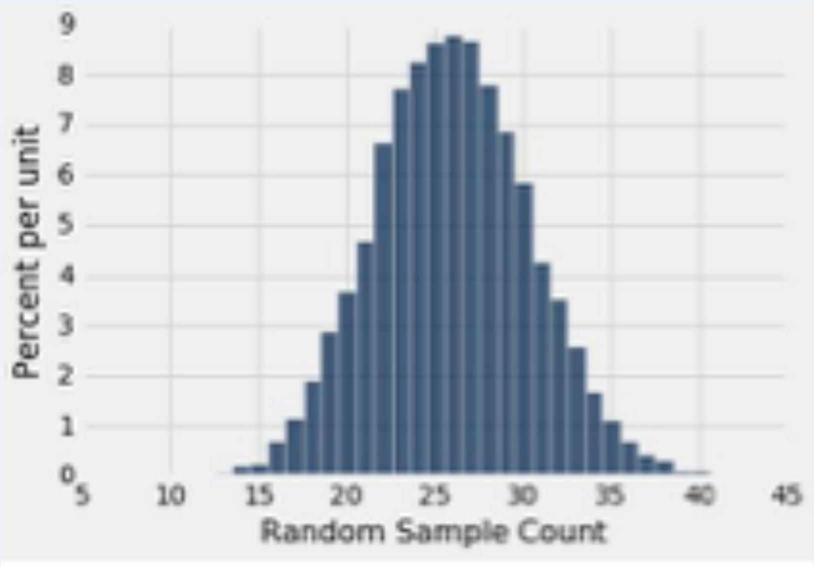
Whether a value is consistent with a distribution:

- A visualization may be sufficient
- If not, there are conventions about “consistency”

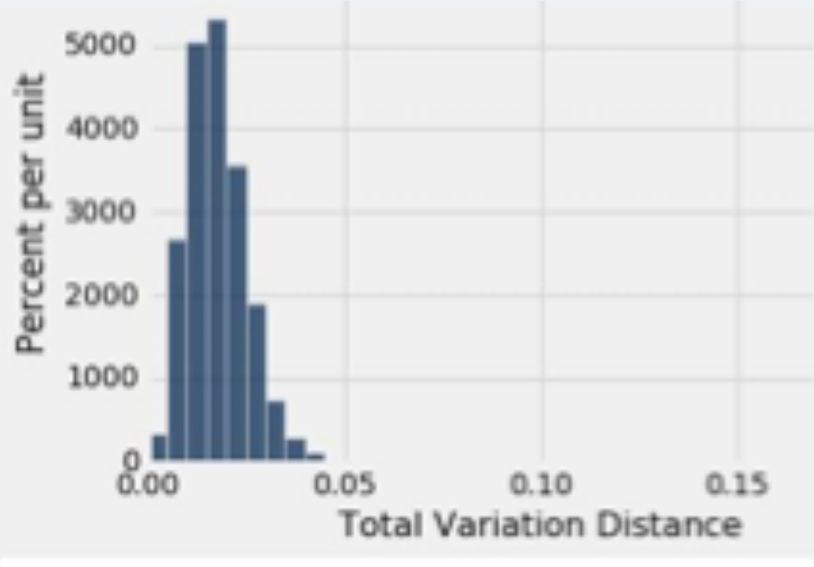
Statistical Significance

Tail Areas

Alabama Jury

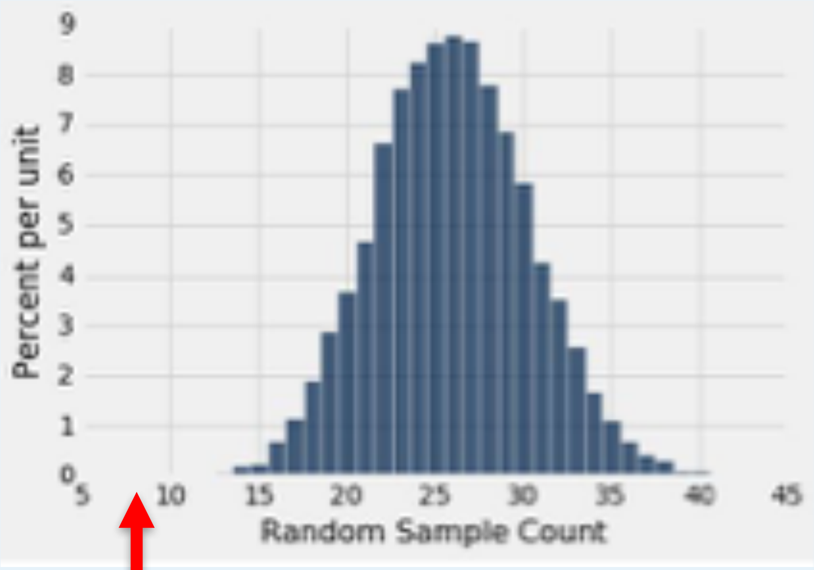


Alameda Jury



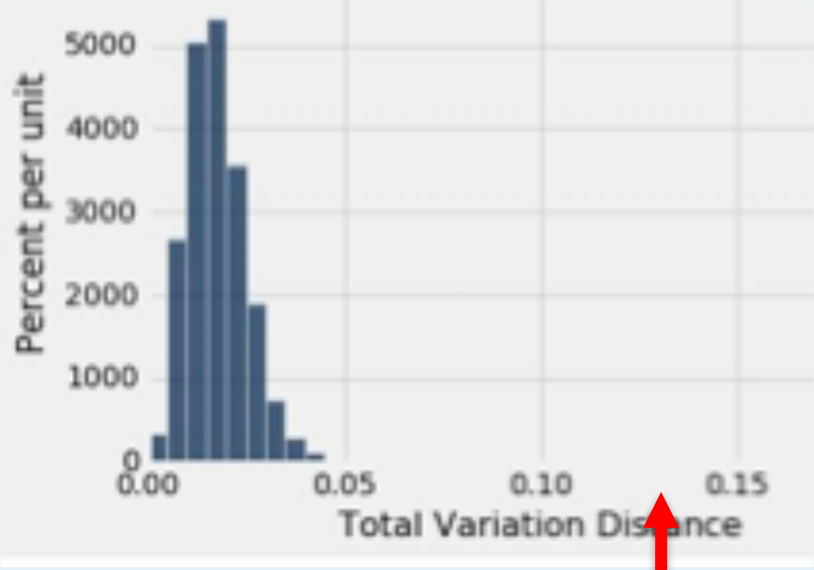
Tail Areas

Alabama Jury



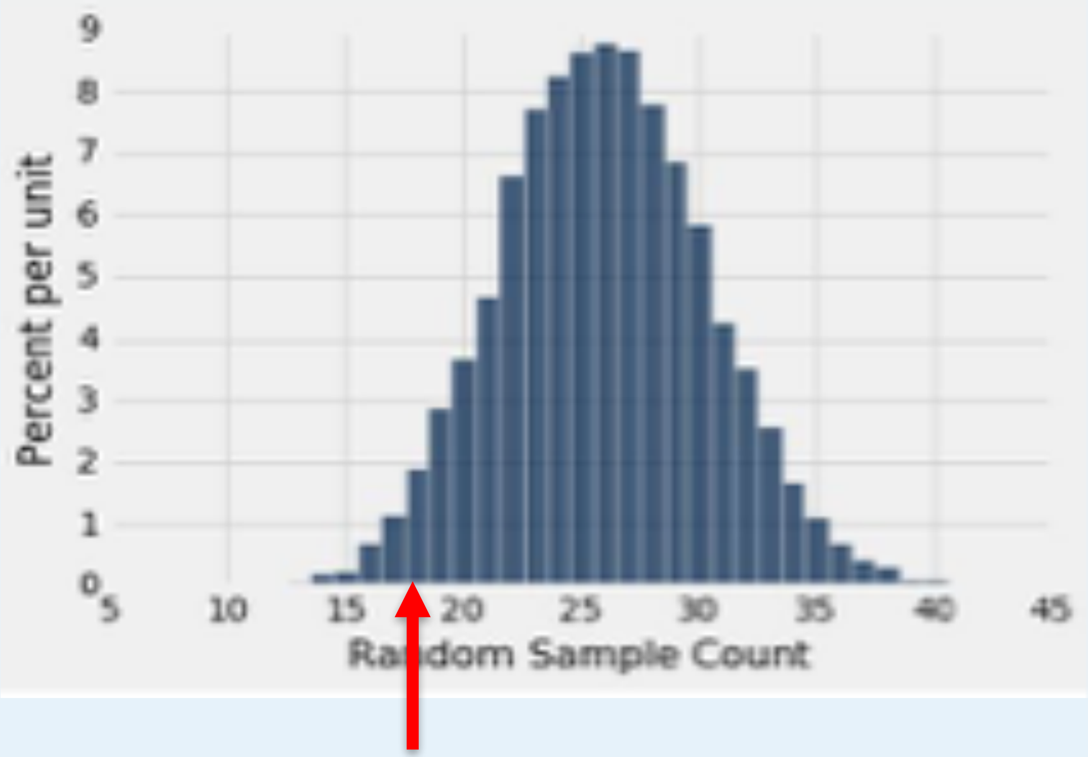
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Not so clear example

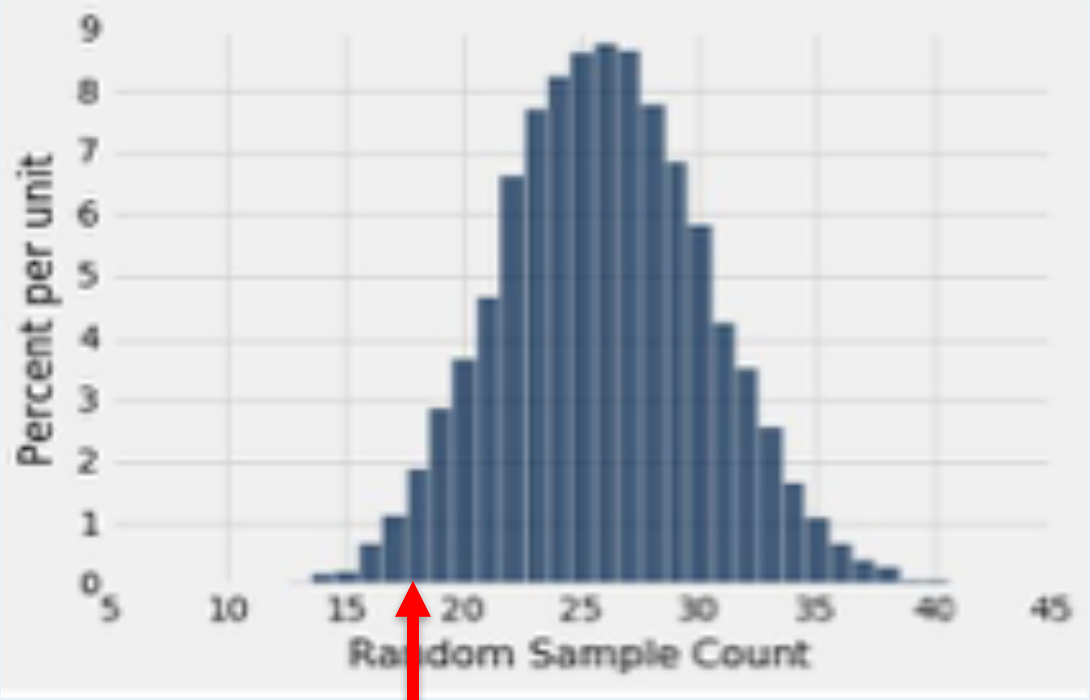


Observed Number (18)

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:

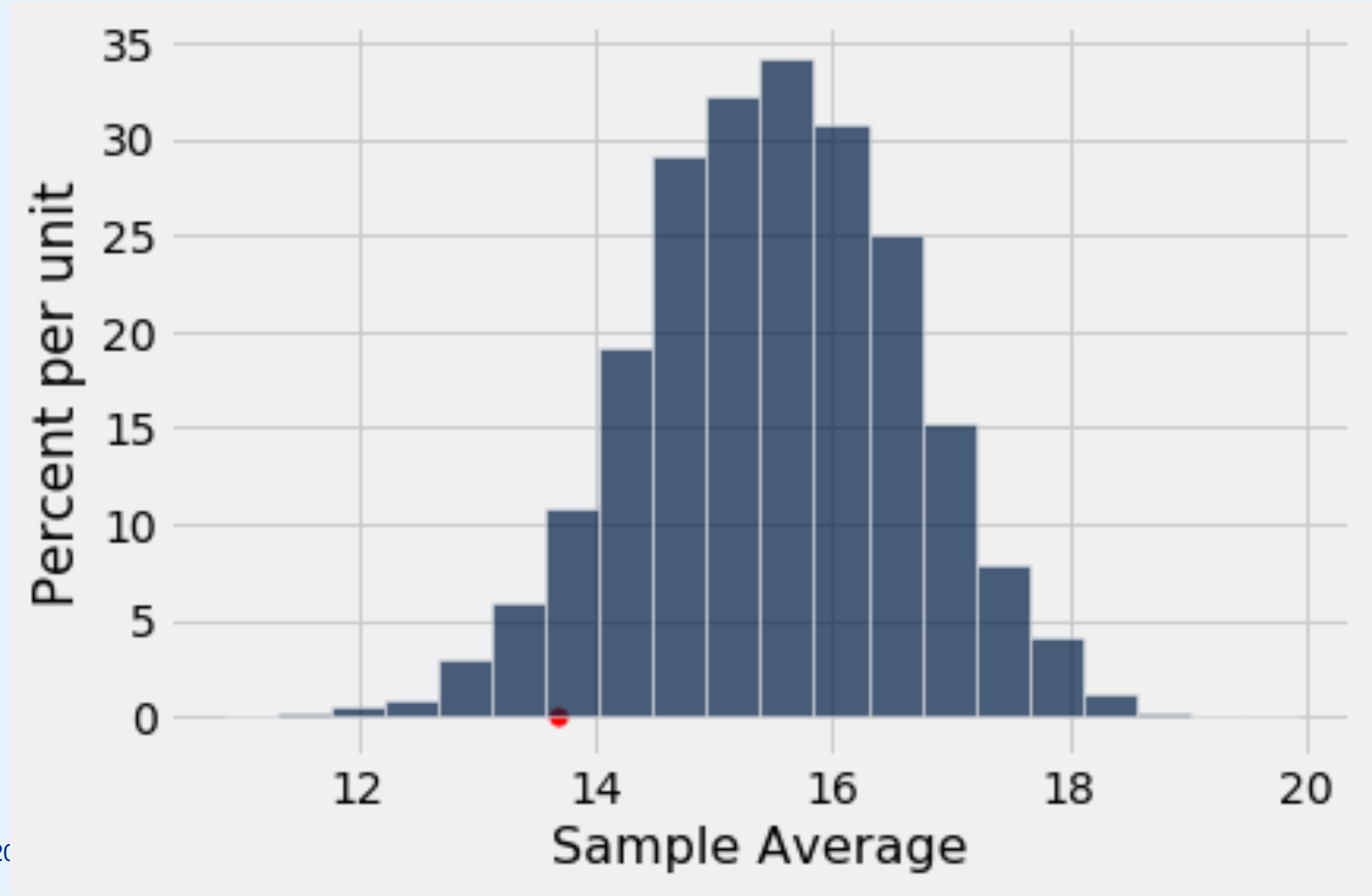
The average score per section (27 students)

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

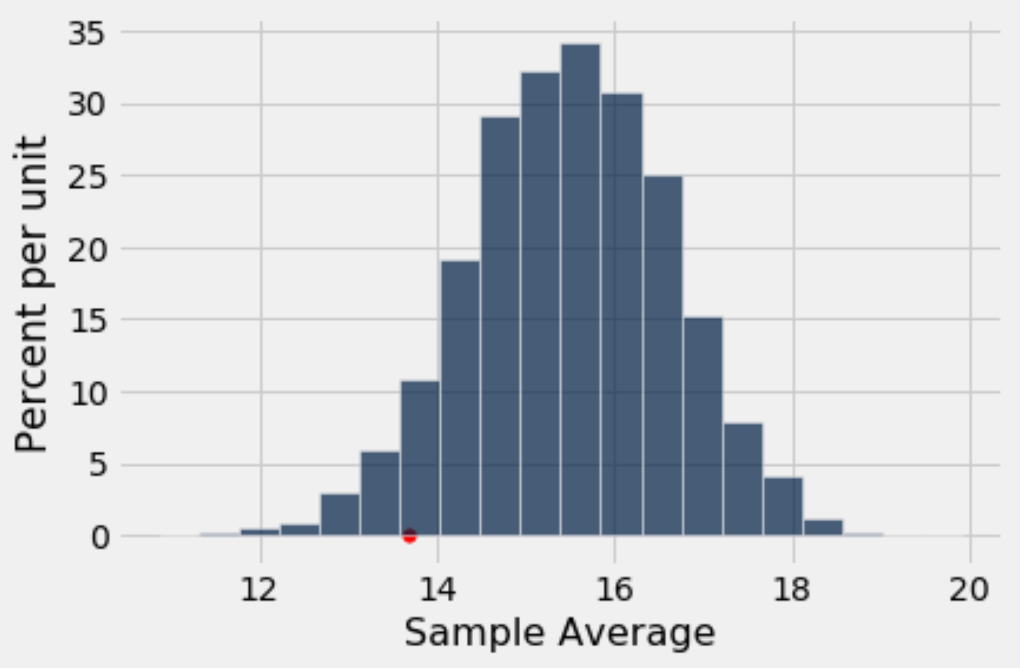
Is the observed statistic consistent with the histogram?



Compute the p-value

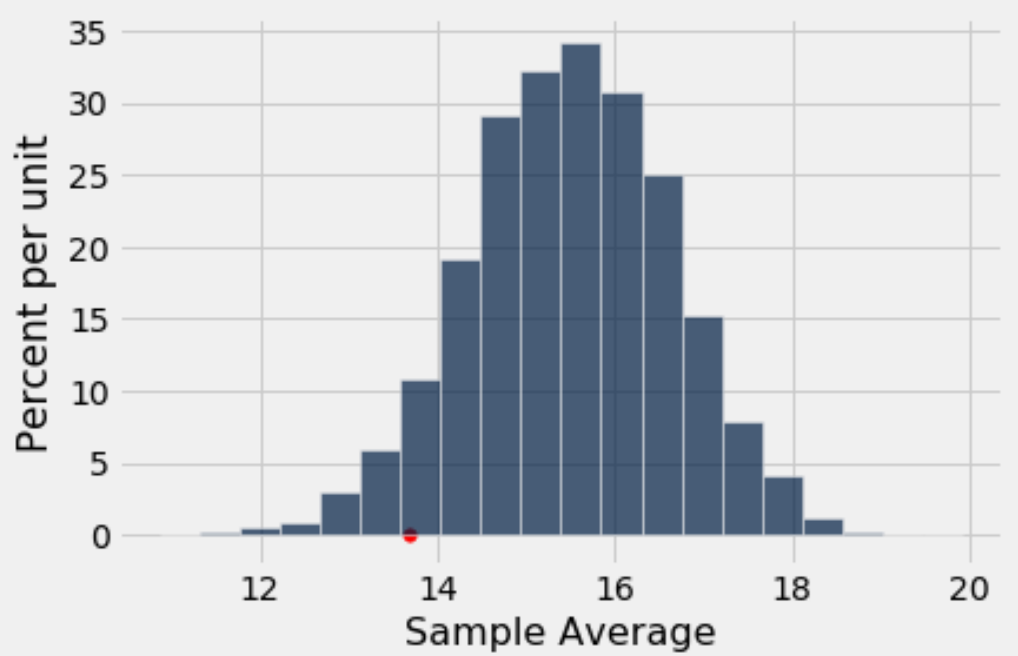
The *P*-value is the chance,

- 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



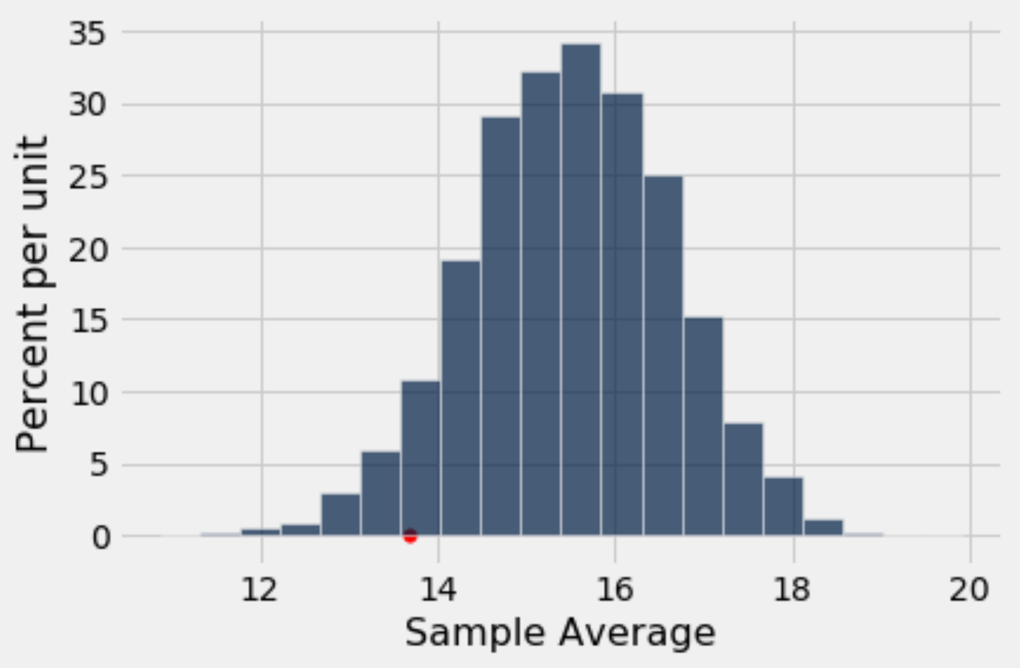
Compute the p-value

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



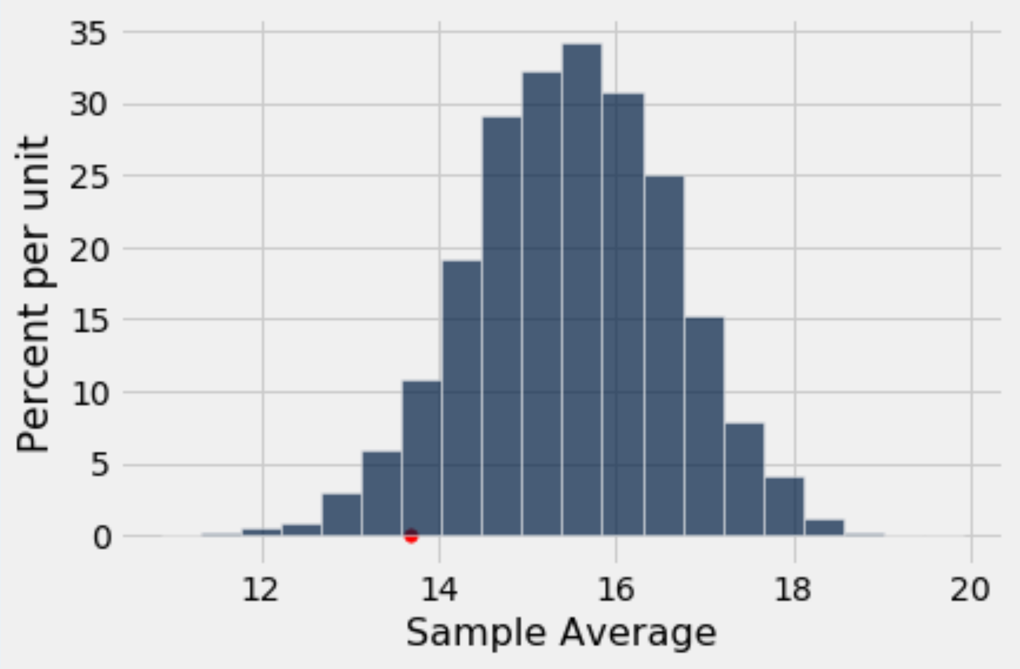
Compute the p-value

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



Compute the p-value

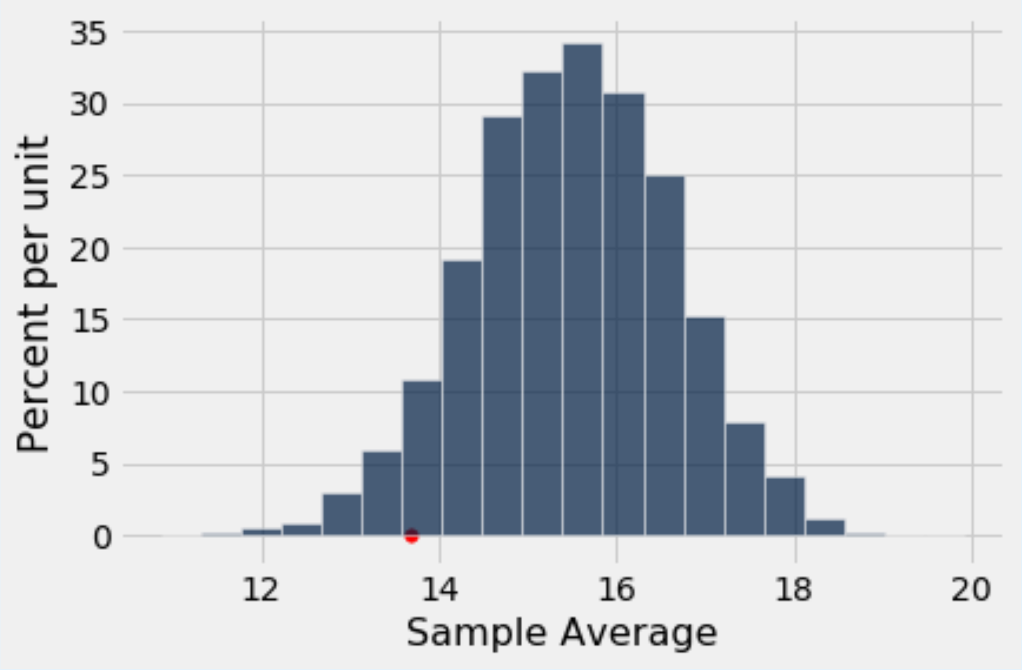
P(A) = (the number of times the sampled statistic was less than the observed statistic) divided by the number of samples



Compute the p-value

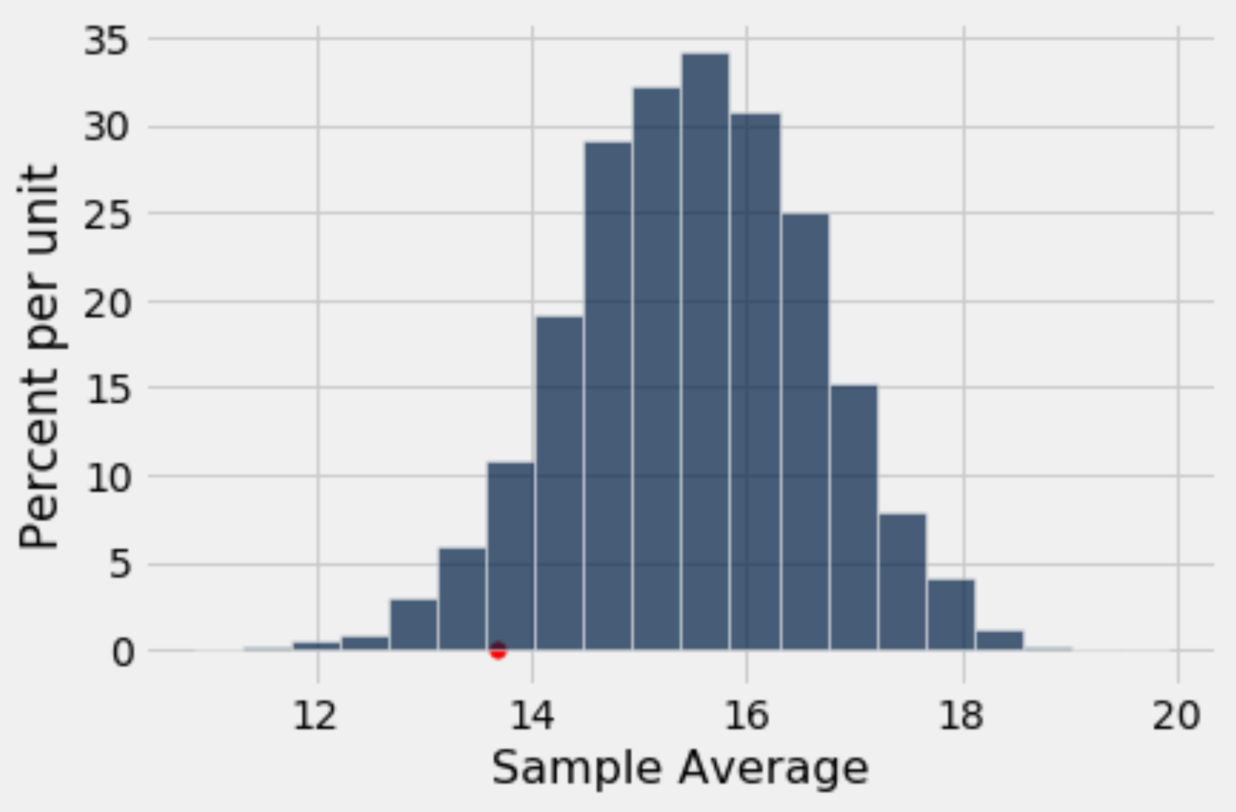
P(A) =

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

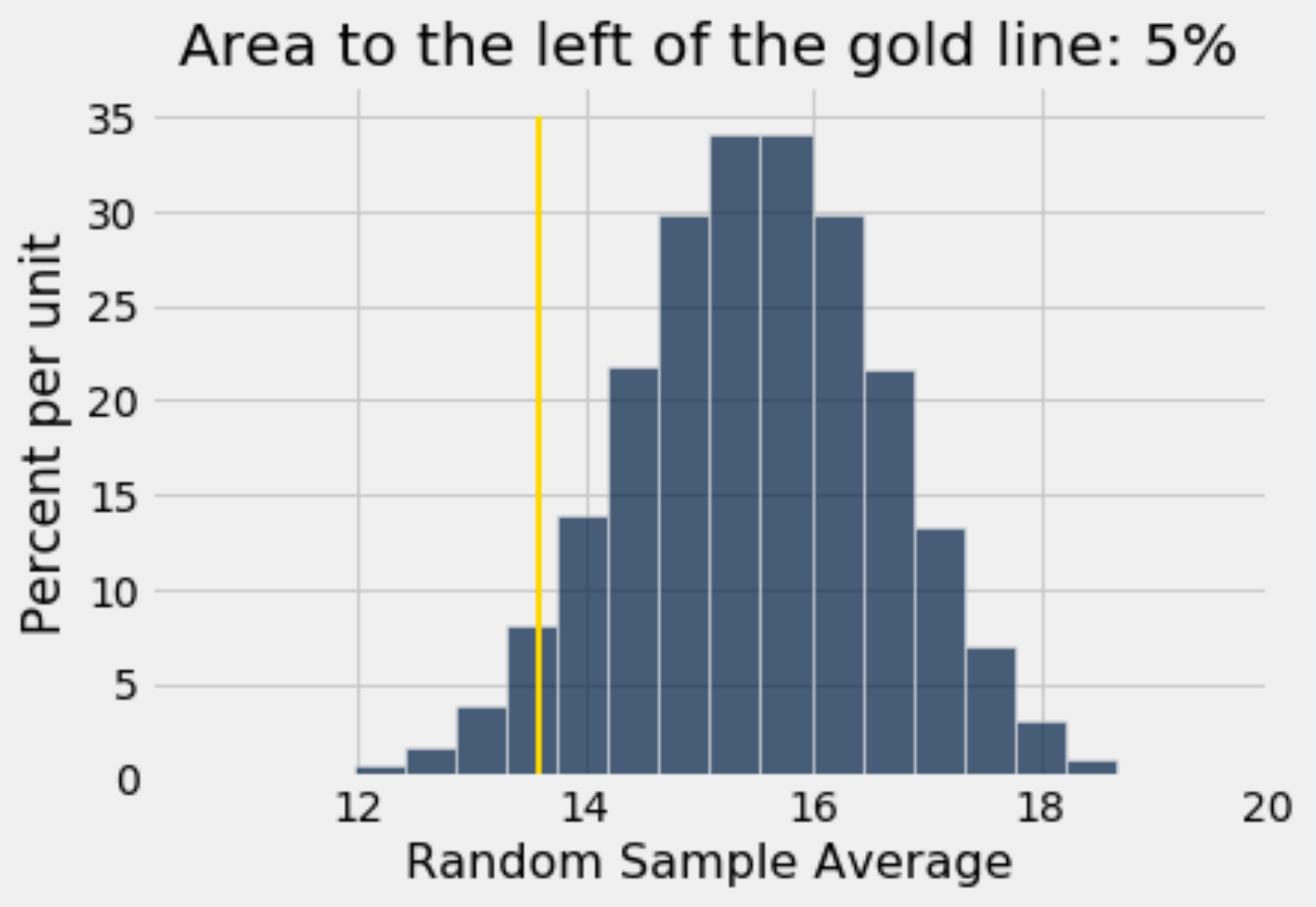


Compute the p-value

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



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

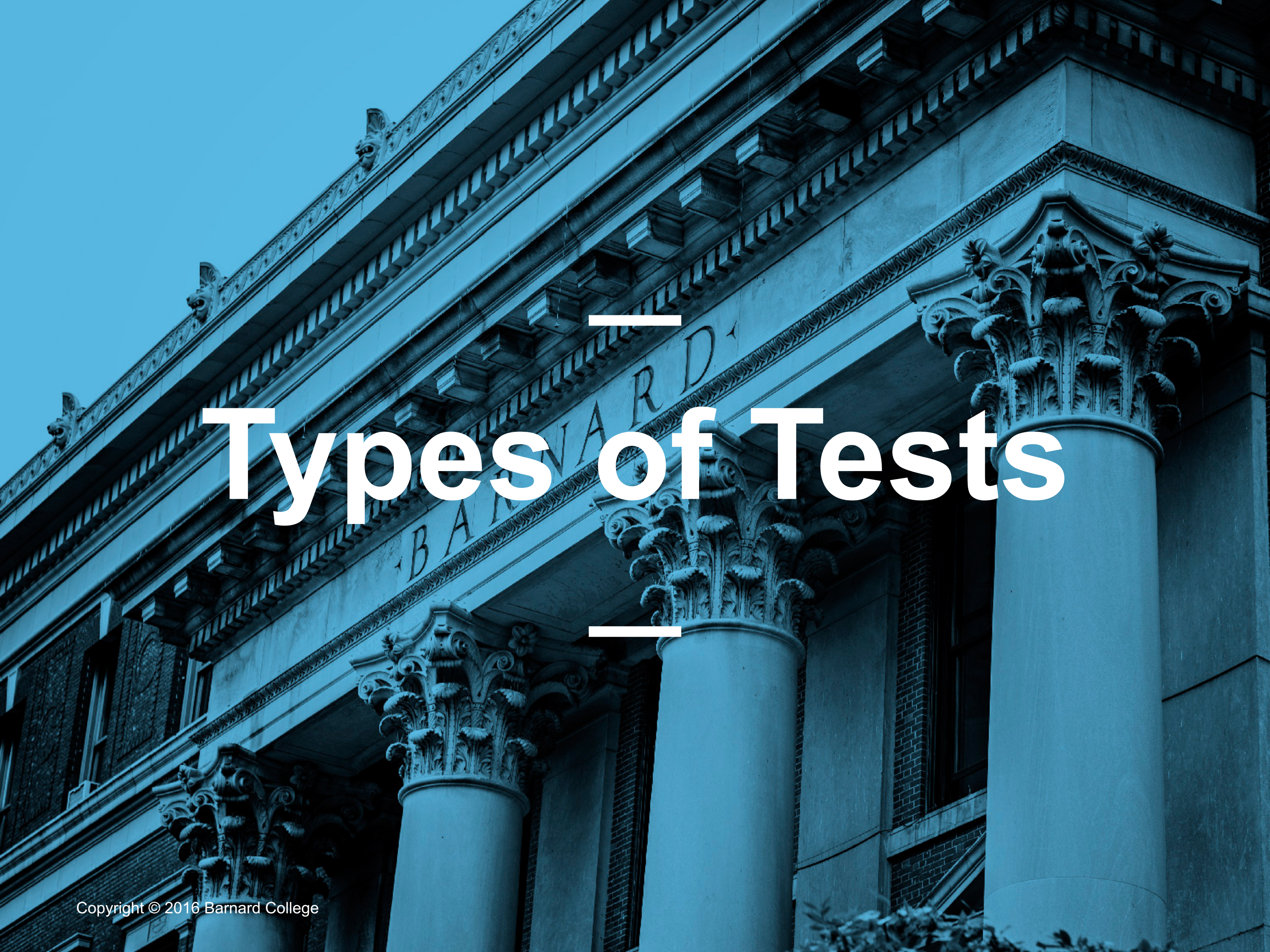
If the null is true, all rearrangements of labels are equally likely

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)