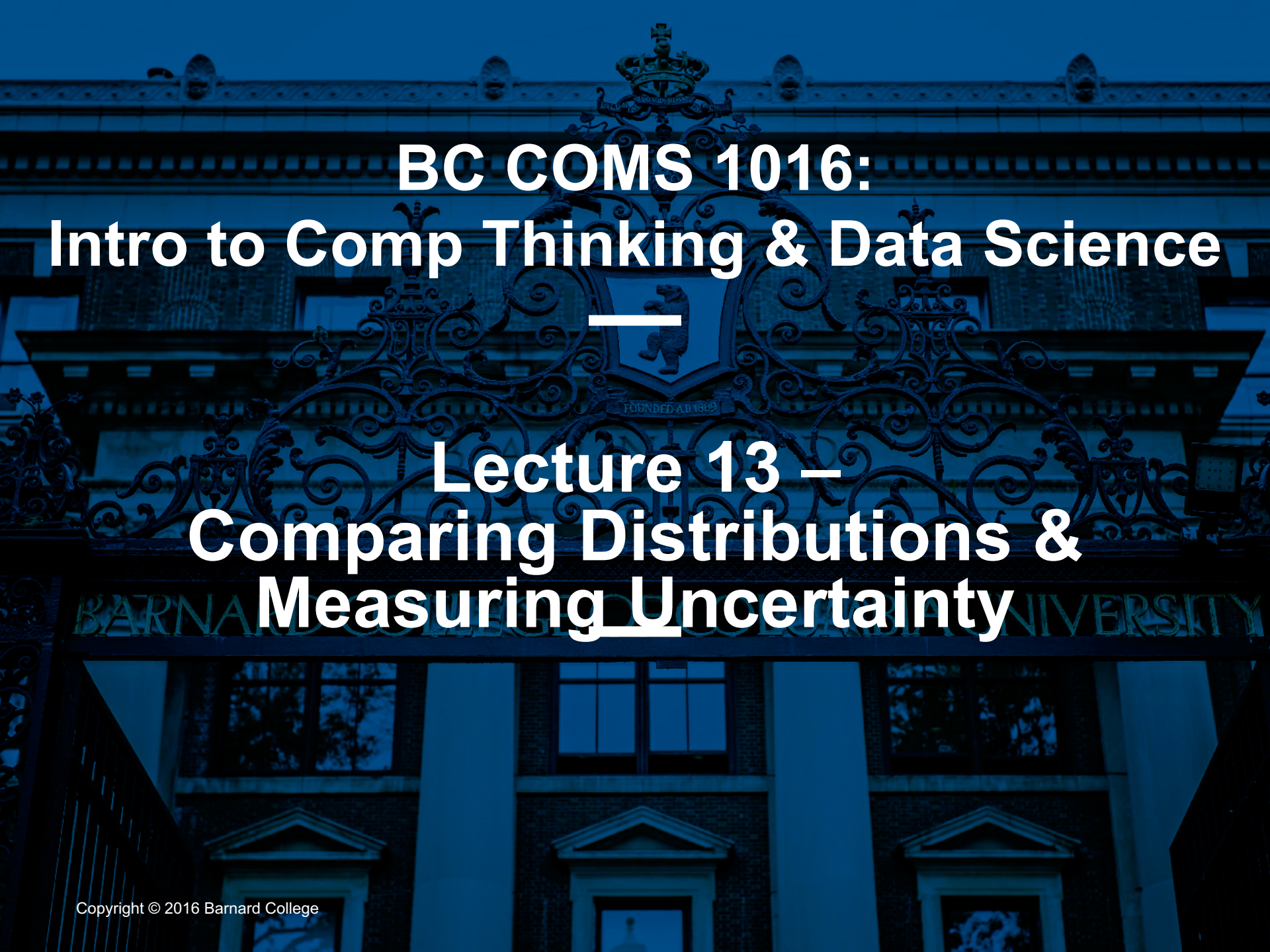


BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 13 – Comparing Distributions & Measuring Uncertainty

Announcements - update

- Thursday Office Hours change:
  - ~~9-10pm~~ 5-6pm
- Lab 05 - Assessing Models: Examining the Therapeutic Touch
  - Optional – recommended but optional
- Lab 06 - Inference and the Death Penalty
  - Due Friday 11/20
- HW05 - Probability, Simulation, Estimation, and Assessing Models
  - Due Thursday 11/19
- Checkpoint/Project 1:
  - Due Wednesday 11/18
- Checkpoint/Project 2 (midterm):
  - Released Thursday 11/19, due Tuesday 11/24

On Asking for Help

Before asking for help:

- Read the entire assignment/question at least twice
  - Don’t skip portions of the assignments that are not “*questions*”
- Read the textbook
- Consult the python reference page
- Read and try understanding the error message

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

Discussion Question – choosing a statistic

In each of (a) and (b), choose a statistic that will help you decide between the two viewpoints.

Data: the results of 400 tosses of a coin

(a):

- "This coin is fair."
- "No, it's not."

(b):

- "This coin is fair."
- "No, it's biased towards heads."

What is “fair”?

For both (a) and (b),

- The number of heads in the 400 tosses is a good starting point, but might need adjustment
- A number of heads around 200 suggests “fair”

Answers

(a) Very large or very small values of the number of heads suggest “not fair.”

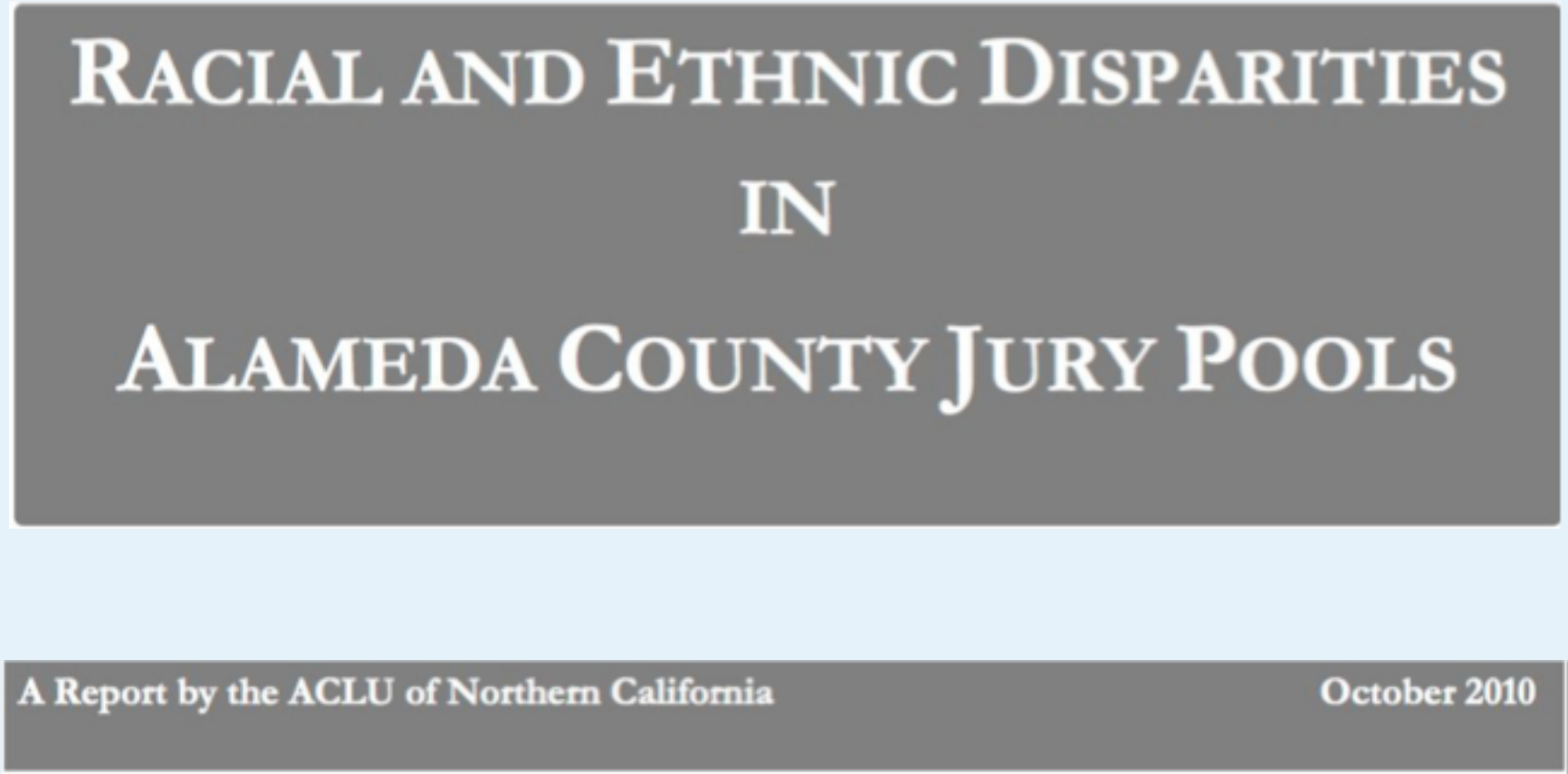
- The **distance** between number of heads and 200 is the key
- Statistic: | number of heads − 200 |
- Large values of the statistic suggest “not fair”

(b) Large values of the number of heads suggest “biased towards heads”

- Statistic: number of heads

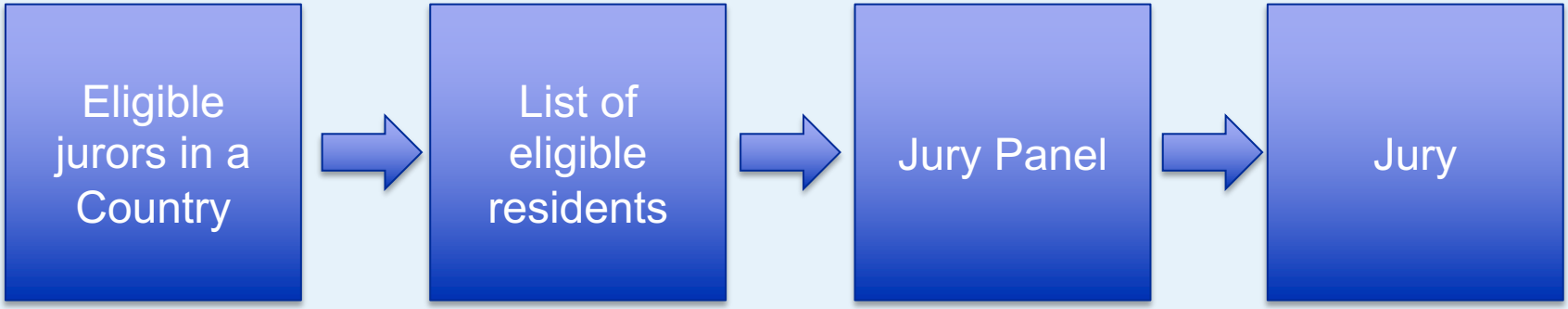
Comparing Distributions

Jury Selectin in Alameda County



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

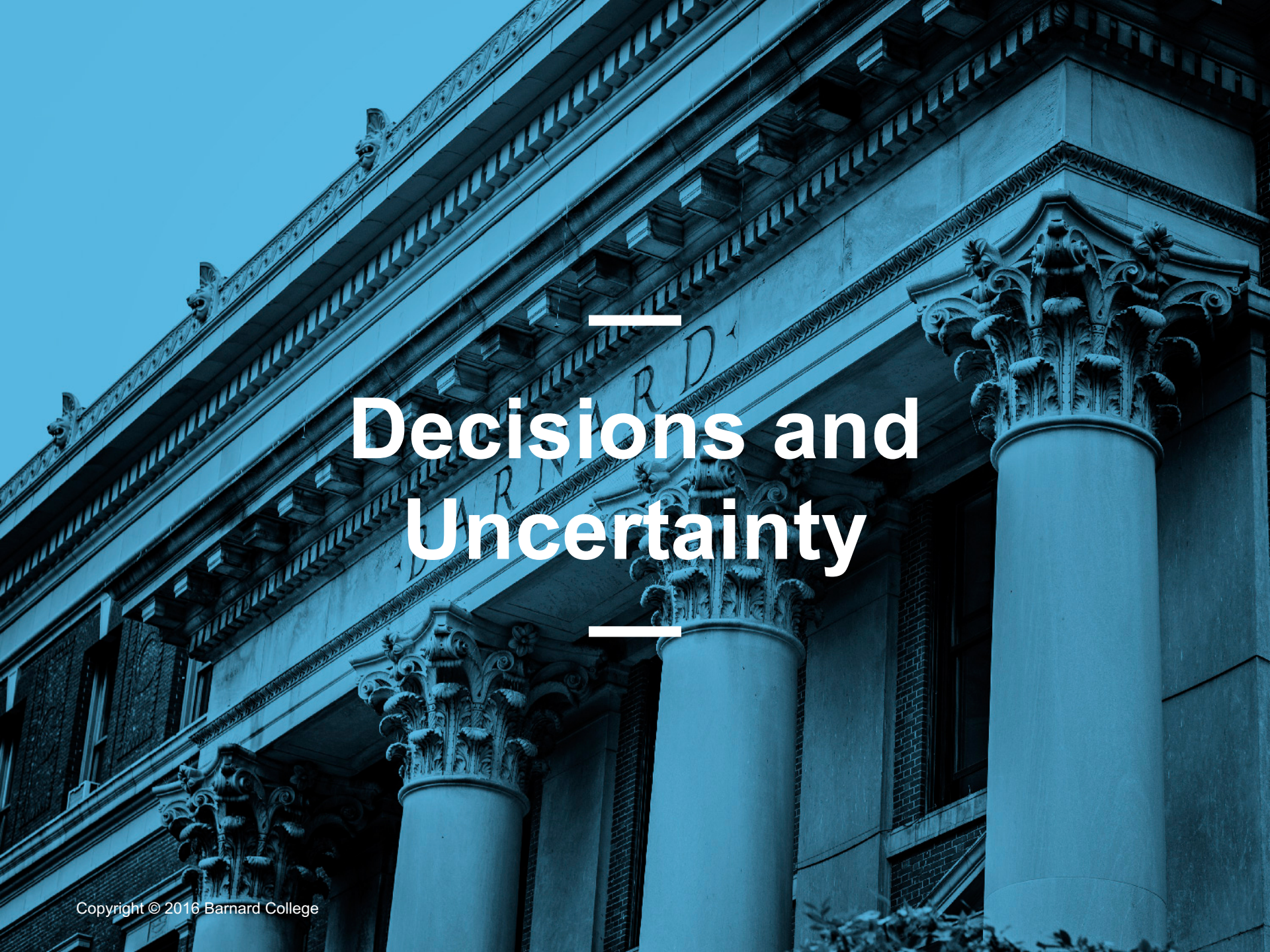
Total Variation Distance (TVD):

- 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

Testing Hypotheses

- A test chooses between two views of how data are generated
- What are these views called?
  - Answer**: hypotheses**
- The test picks the hypothesis that is better supported by the observed data
- What is the method for choosing the hypotheses?
  - Simulate data under one of the hypotheses
  - Compare the simulation results and the observed data
  - Pick one of the hypotheses based on whether the simulated results and observed data are consistent

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”

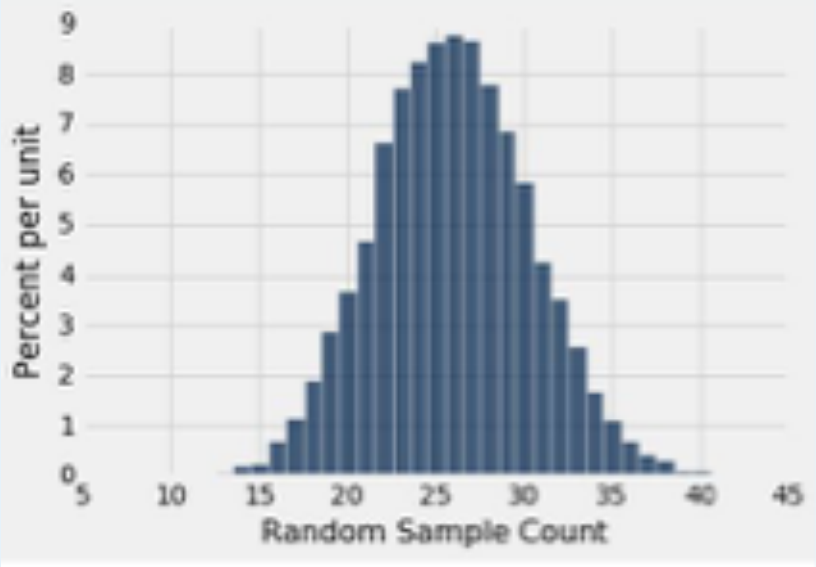
Whether a value is consistent with a distribution:

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

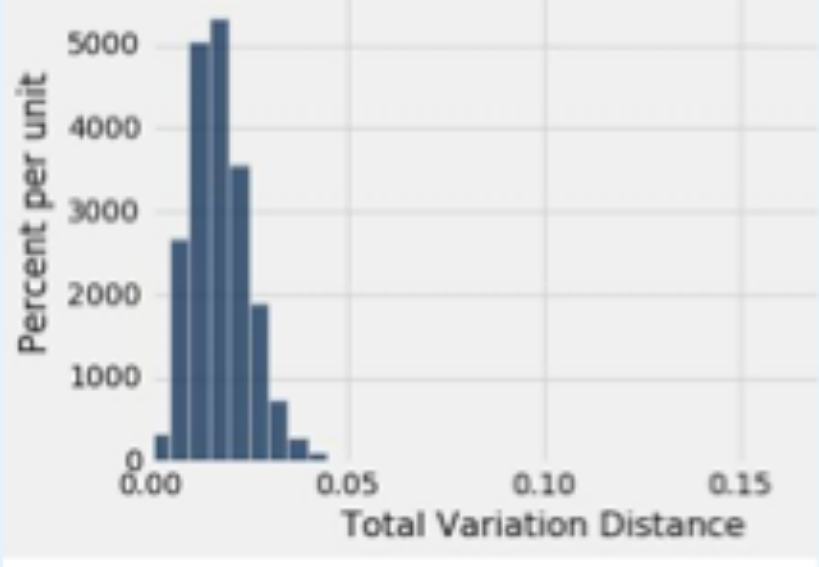
Statistical Significance

Tail Areas

Alabama Jury



Alameda Jury



Conventions About Inconsistency

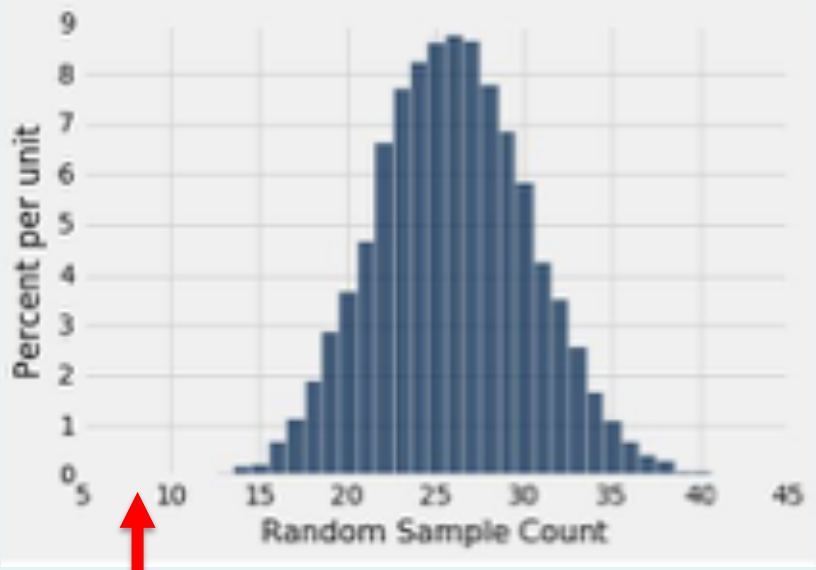
- **"Inconsistent with the null"**: The test statistic is in the tail of the empirical distribution under the null hypothesis

- **"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"

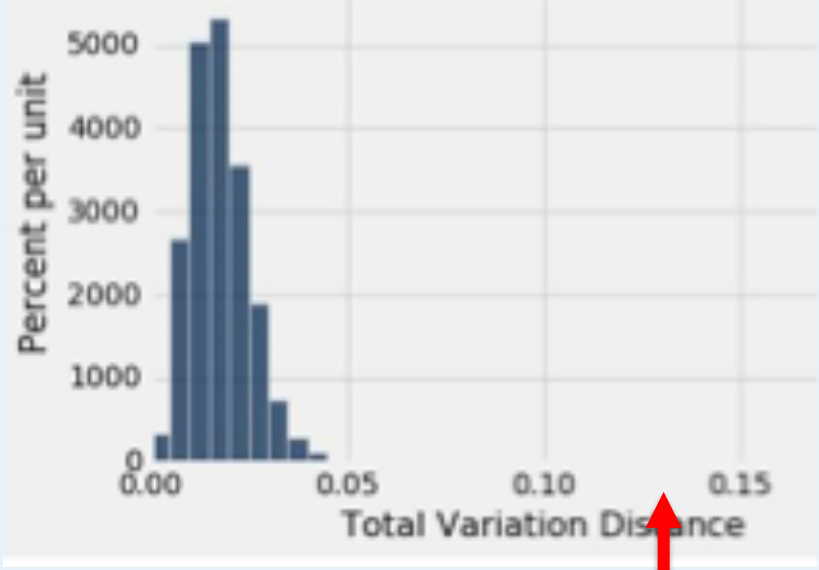
Tail Areas

Alabama Jury



Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Definition of the 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