DATA 8

Spring 2022

Lecture 17

Comparing Distributions

Weekly Goals

- **Today**
  - Comparing distributions
  - Hypothesis tests
- Wednesday
  - Making decisions when visualizations don’t suffice
  - Comparing numerical data
- Friday
  - A/B testing
  - Permutation Test

Announcements

- Homework 6 due this Thursday
  - Turn in on Wednesday for a bonus point
  - Please start this one early!
- Lab 6 due Friday

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 Assessment

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts.
- We can then compare the model's predictions (simulations) to the data that were observed.
- If the data and the model's predictions are not consistent, that is evidence against the model.

Two Viewpoints

Model and Alternative

- **Jury selection:**
  - **Model:** The people on the jury panels were selected at random from the eligible population
  - **Alternative viewpoint:** No, they were biased against black men
- **Genetics:**
  - **Model:** Each plant has a 75% chance of having purple flowers
  - **Alternative viewpoint:** No, it doesn't

Steps in Assessing a Model

- **Choose a statistic** to measure “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 Questions

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.”

“Fair”

For both (a) and (b),

- The percent of heads in the 400 tosses is a good starting point, but might need adjustment
- A percent of heads around 50% suggests “fair”

Answers

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

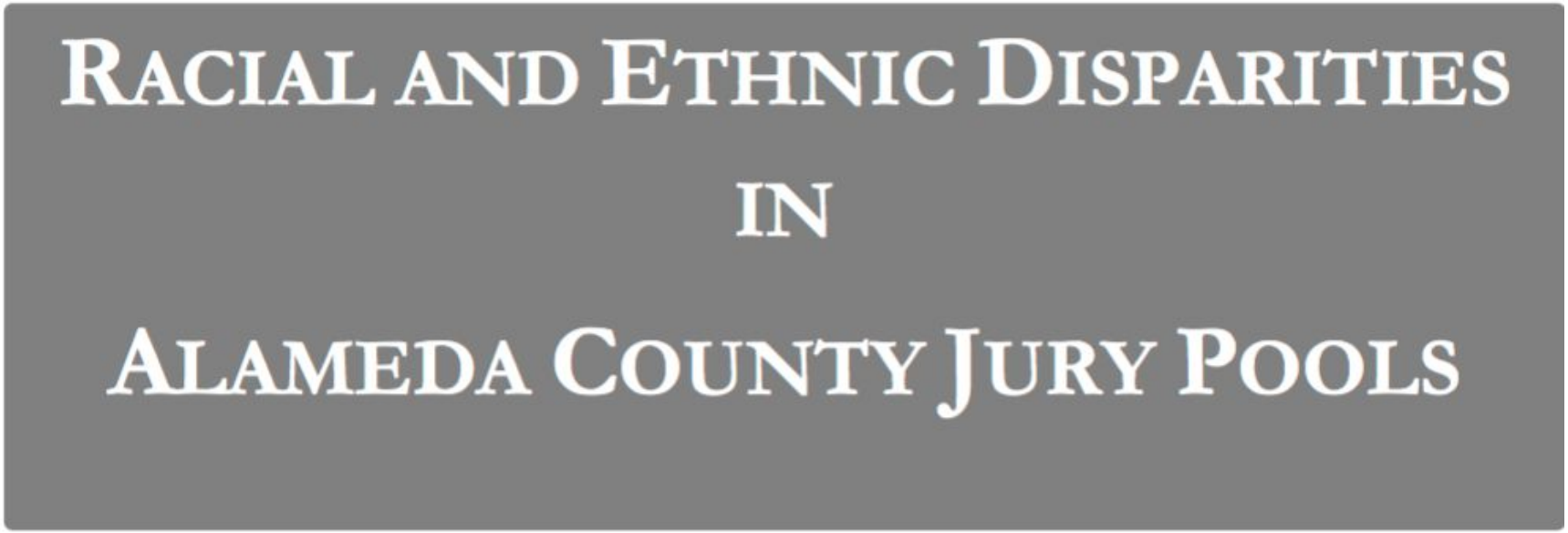
- The **distance** between percent of heads and 50% is the key
- Statistic: | percent of heads − 50% |
- Large values of the statistic suggest “not fair”

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

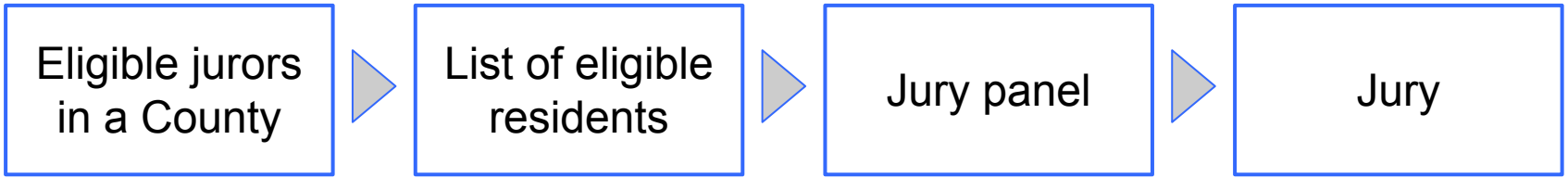
- Statistic: percent of heads

Comparing Distributions

Jury Selection in Alameda County



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."

(Demo)

A New Statistic

Distance Between Distributions

- People on the panels are of multiple ethnicities
- Distribution of ethnicities is categorical

- To see whether the 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

(Demo)

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

(Demo)

Summary of the Method

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

- Use TVD as the statistic because it 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
  - Actual TVD from the sample in the study

Testing Hypotheses

Testing Hypotheses

- A test chooses between two views of how data were generated
- The views are called **hypotheses**
- The test picks the hypothesis that is better supported by the observed data

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 “high”. Try to avoid “both high and low”.

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays 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 distribution, then the test favors the alternative (“data is consistent with the alternative”)

Whether a value is consistent with a distribution:

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