190F Foundations of Data Science

Fall 2018

Lecture 15

Comparing Distributions

Announcements

Testing Hypotheses

Choosing One of Two Viewpoints

- Based on data, choose between two different viewpoints (or hypotheses) about the process that generated the data :
  - “Chocolate has no effect on cardiac disease” vs “Yes, it does.”
  - “This jury panel was selected at random from eligible jurors” vs “No it wasn’t.”

Models

- A *model* is a set of assumptions about the data.
- To test a hypothesis, we need to *formalize* the assumptions it implies using a model.
- In data science, many models involve assumptions about processes that involve *randomness*.

Assessing a Model

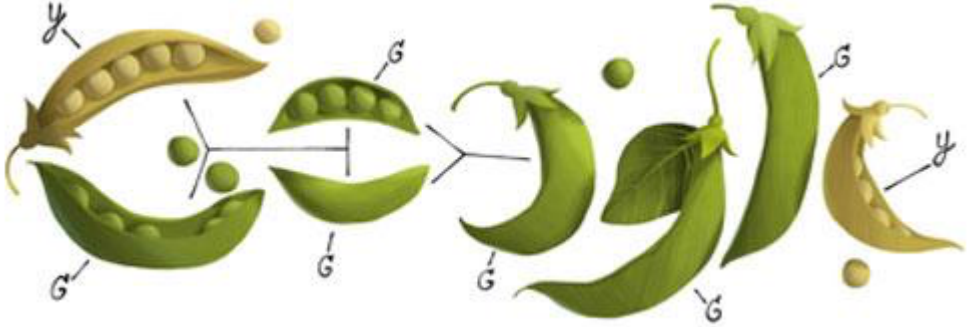
- If we can simulate data according to the assumptions of a model, we can learn what the model predicts.
- We can then compare the predictions to the data that were observed.
- If the data and the model's predictions are not consistent, that is evidence against the model.

Steps in Assessing a Model

1. Come up with a statistic that will help you decide whether the data support the model or an alternative view.
2. Simulate the statistic under the assumptions of the model.
3. Draw a histogram of the simulated values. This is the model's prediction for how the statistic should come out.
4. Compute the observed statistic from the sample of data.
5. Compare this value with the histogram.
6. If the two are not consistent, that's evidence against the model.

A Genetic Model

Gregor Mendel, 1822-1884



A Model

- Pea plants of a particular kind either have purple or white flowers.
- Mendel's model: Each plant is purple-flowering with chance 75%, independent of other plants.
- **Questions:** Is this model good?

25%

75%

Choosing a Statistic

- Start with percent of purple-flowering plants in sample
- If that percent is much larger or much smaller than 75, that is evidence against the model

- *Distance* from 75% is the key.

- Statistic: abs(sample % of purple-flowering plants − 75)

- If the statistic is large, the sample % deviates from the model, which is evidence that the model is wrong.

Discussion Questions

In each of (a) and (b), choose a statistic that will help you decide between the two viewpoints about a possibly biased coin. **Data:** the results of 400 tosses of the coin.

(a) “This coin is fair” vs “No, it’s not.”

(b) “This coin is biased towards heads” vs “No, it’s not.”

Answer (a)

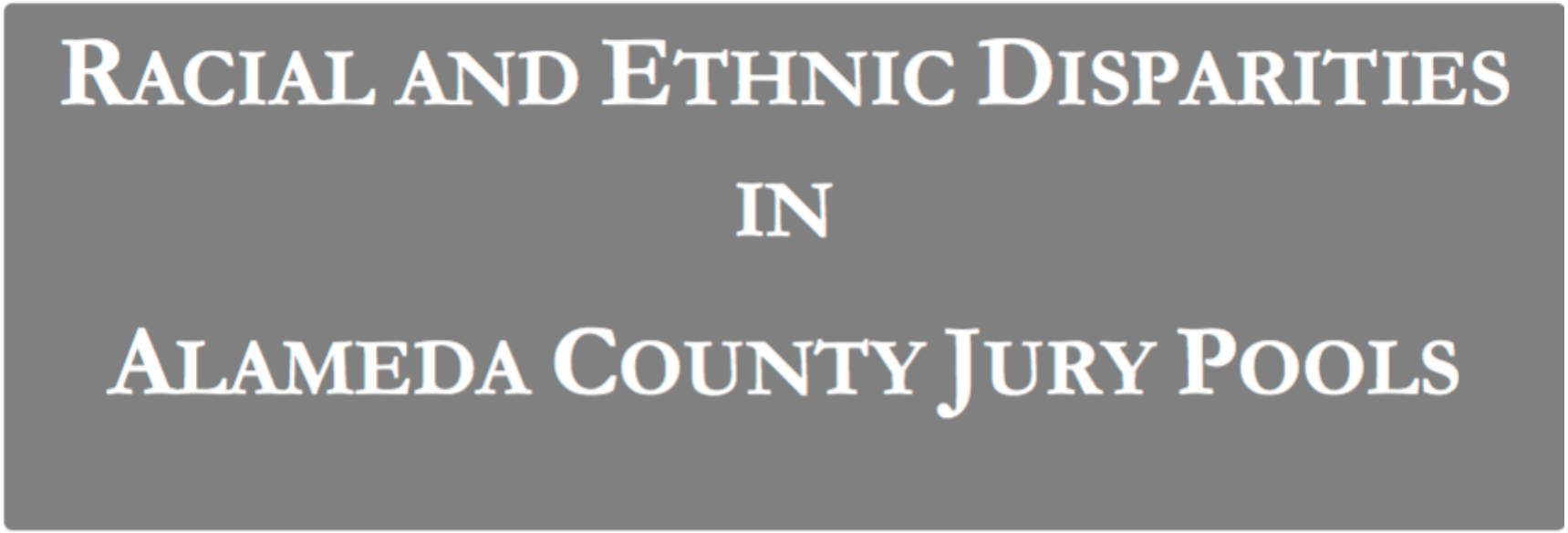
- A number of heads around 200 (50%) suggests “fair.”
- Very large or very small values of the number of heads suggest “not fair.”
  - The **distance** between number of heads and 200 is the key.
  - Statistic: abs(number of heads − 200)
  - Large values of the statistic suggest “not fair”

Answer (b)

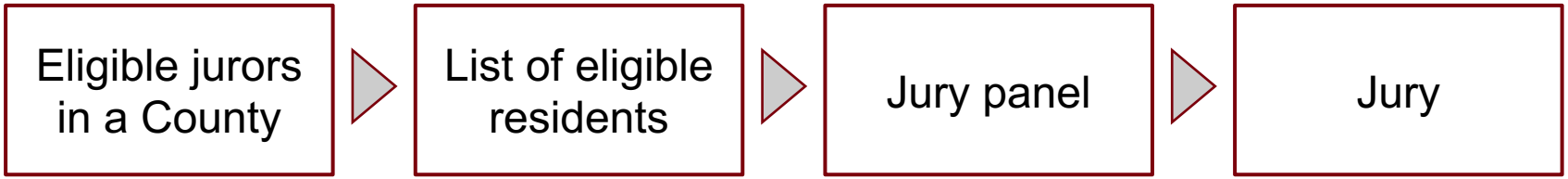
- Large values of the number of heads suggest “biased towards heads”
- Statistic: number of heads.
- Importantly, the alternative to “biased towards heads” isn’t that the coin is fair, it’s that the coin is fair **or** biased towards tails.

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."

Model and Alternative

- **Model:** The people on the jury panels were selected at random from the eligible population

- **Alternative viewpoint:** No, they weren’t

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

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