YaleNUSCollege

YSC2239 Lecture 7

Recap on previous class

- Chance / Probability: Multiplication Rule and Addition Rule
- Python:
  - Join tables
  - Booleans: True or False, comparison operators
  - Control statements (if else/ for)
  - Append array: np.append
  - Random choice: np.random.choice

Today’s class

- Sampling
- Assessing Models
- Comparing Distributions

- Reading: Chapter 10, 11

Sampling

Probability Samples

- Deterministic sample:
  - Sampling scheme doesn’t involve chance

- Probability sample:
  - Before the sample is drawn, you have to know the selection probability of every group of people in the population
  - Not all individuals have to have equal chance of being selected

Sample of Convenience

- Example: sample consists of whoever walks by
- Just because you think you're sampling "at random", doesn't mean you are.
- If you can't figure out ahead of time
  - what's the population
  - what's the chance of selection, for each group in the population

  then you don't have a random sample

(Demo)

Distributions

Probability Distribution

- Random quantity with various possible values
- “Probability distribution”:
  - All the possible values of the quantity
  - The probability of each of those values
- If you can do the math, you can work out the probability distribution can without ever simulating the random quantity

Empirical Distribution

- “Empirical”: based on observations
- Observations can be from repetitions of an experiment
- “Empirical Distribution”
  - All observed values
  - The proportion of times each value appears

(Demo)

Large Random Samples

Law of Averages

If a chance experiment is repeated many times,
independently and under the same conditions,
then the proportion of times that an event occurs
gets closer to the theoretical probability of the event

As you increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6

(Demo)

Empirical Distribution of a Sample

If the sample size is large,

then the empirical distribution of a uniform random sample

resembles the distribution of the population,

with high probability

A Statistic

Terminology

- **Parameter**
  - A number associated with the population
- **Statistic**
  - A number calculated from the sample

A statistic can be used as an **estimate** of a parameter, or to **test hypotheses** about how the data were generated

(Demo)

Inference

- **Statistical Inference:**
  Making conclusions based on data in random samples

- **Example**:
  Use the data to guess the value of an unknown number

fixed

depends on the random sample

Create an **estimate** of the unknown quantity

Probability Distribution of a Statistic

- Values of a statistic vary because random samples vary
- “Sampling distribution” or “probability distribution” of the statistic:
  - All possible values of the statistic,
  - and all the corresponding probabilities
- Can be hard to calculate
  - Either have to do the math
  - Or have to generate all possible samples and calculate the statistic based on each sample

Empirical Distribution of a Statistic

- Empirical distribution of the statistic:
  - Based on simulated values of the statistic
  - Consists of all the observed values of the statistic,
  - and the proportion of times each value appeared

- Good approximation to the probability distribution of the statistic
  - if the number of repetitions in the simulation is large

Assessing Models

Models

- A model is a set of assumptions about the data

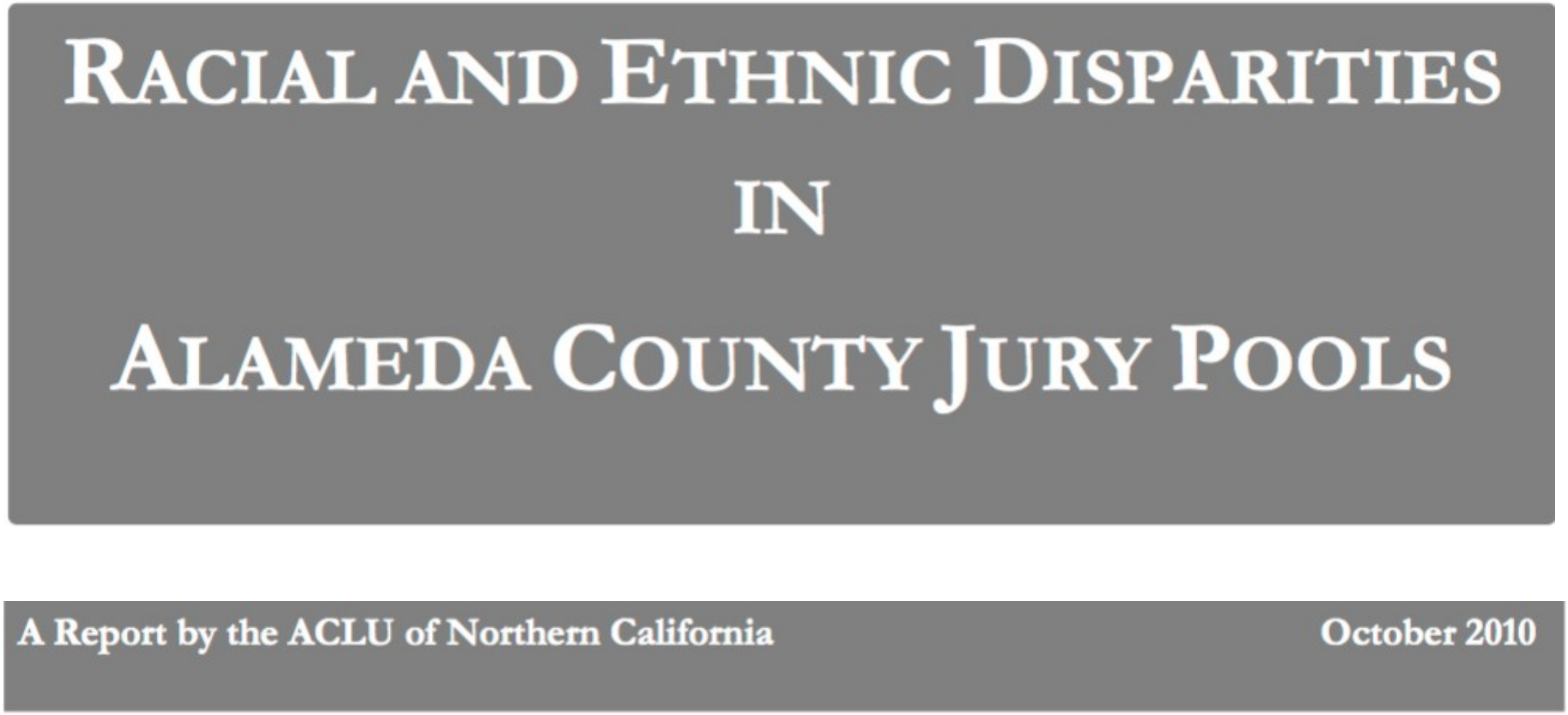
Models

- A model is a set of assumptions about the data

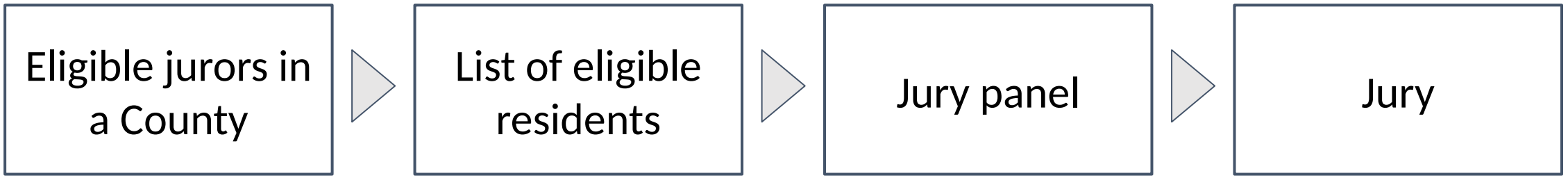
- In data science, many models involve assumptions about processes that involve randomness
  - “Chance models”

Jury Selection

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)

Two Viewpoints

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

Reminders

- Assignment 3