DATA 8

Fall 2016

Lecture 17, October 3

Total Variation Distance

Slides created by Ani Adhikari and John DeNero

Announcements

- Project is due 5 pm tomorrow Tuesday October 4.
- Homework due this week as usual.
- Midterm is on Friday Oct 14, less than two weeks away. No computers or calculators on the midterm.
- No alternate dates for the midterm.

Statistic

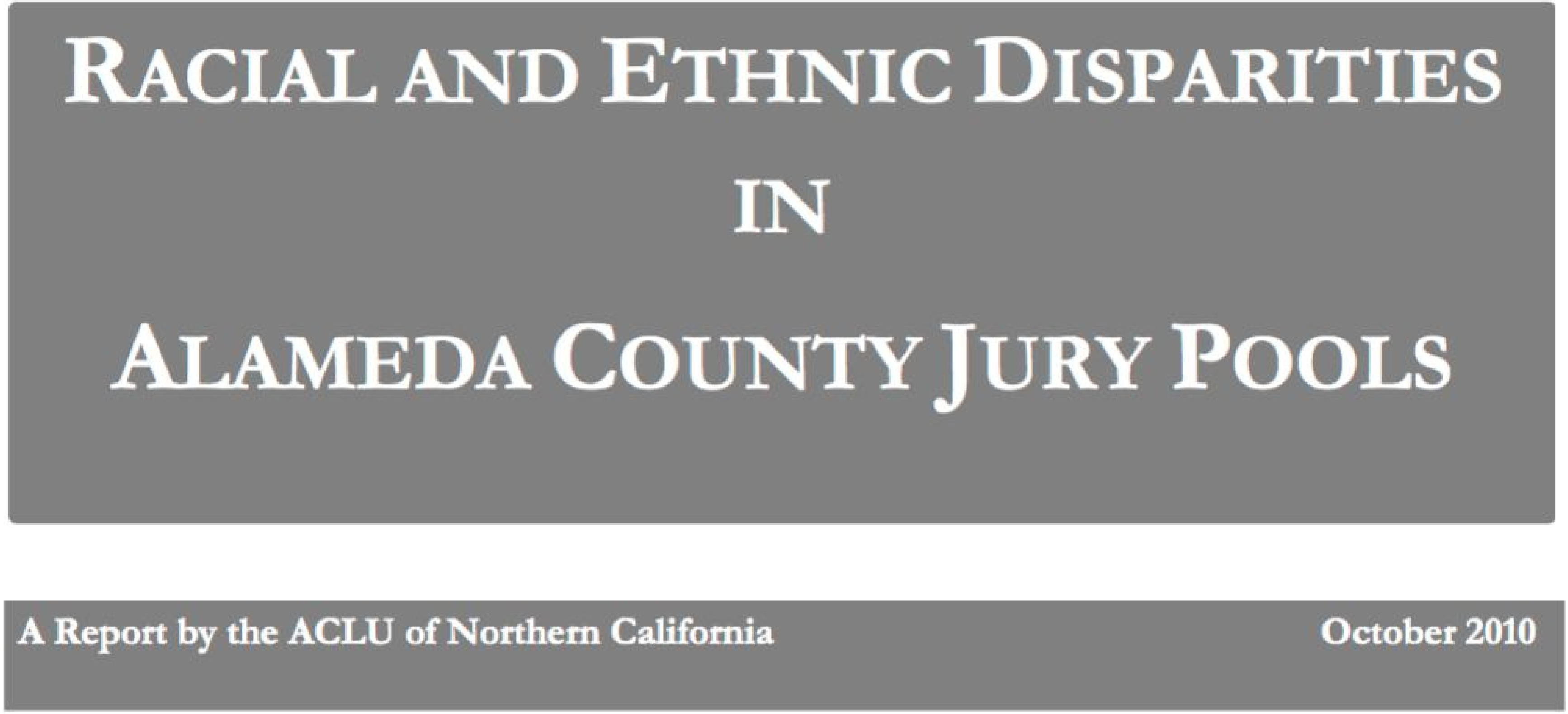
A number calculated from a sample

Simulating a Statistic

Fix a sample size and choose your statistic.

1. Simulate the statistic once:
   a. Draw a random sample of the size you fixed.
   b. Calculate the statistic and keep a record of the value
2. Repeat Step 1 numerous times (as many times as you have patience for; thousands are good).
3. You now have one value of the statistic for each repetition. Visualize the results.

Jury Selection in Alameda County



RACIAL AND ETHNIC DISPARITIES

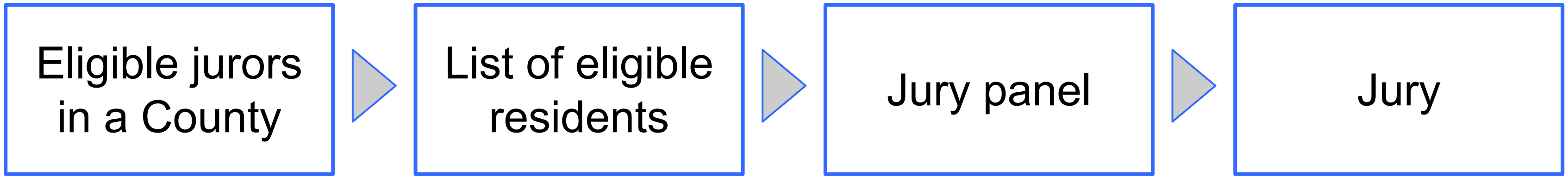
IN

ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California

October 2010

Jury Panels



Eligible jurors in a County

List of eligible residents

Jury panel

Jury

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)

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 divide by 2

(Demo)

Sampling from a Distribution

`proportions_from_distribution`

- Arguments:
  - Table name
  - Label of column containing distribution from which to draw
  - Sample size
- Returns new table: the old table augmented with column `Random Sample` consisting of proportions that appear in a random sample from the given distribution

Summary

Assessing if a sample was drawn randomly from a known population:

- Decide on a statistic that measures the distance between distributions
- Compute the statistic from the sample; that is, the distance between distributions of sample and known population
- Sample at random and from the population and compute the statistic from the random sample; repeat numerous times
- Compare