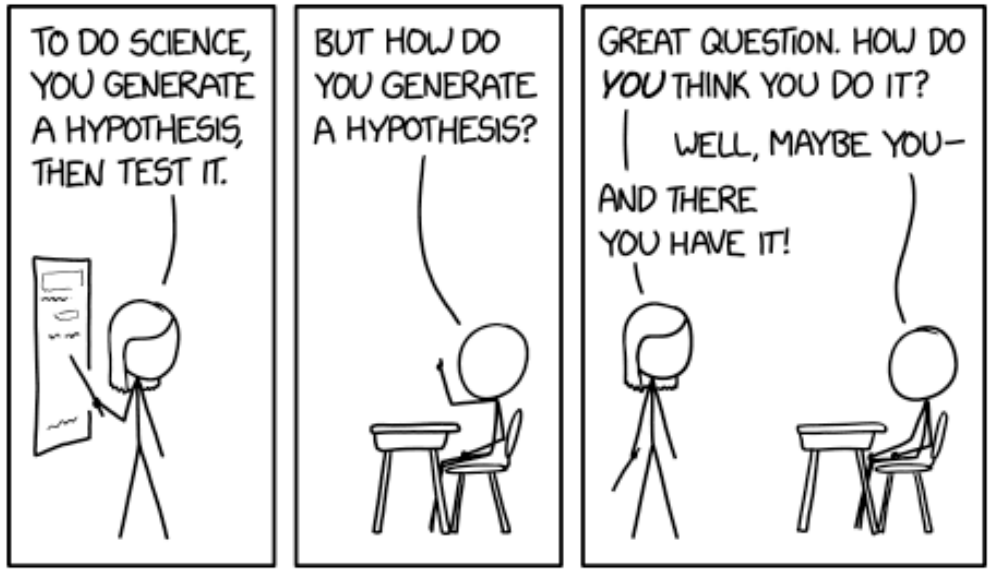
Hypothesis testing (2)

Introduction to data analysis: Lecture 8(a)

Ori Plonsky

Spring 2023



Source: https://xkcd.com/2569/

Hypothesis testing using models

We want to know whether our model is correct, but we only have a sample!

The truth (how the world behaves) → Observations → Our sample (data from the world) → Statistic

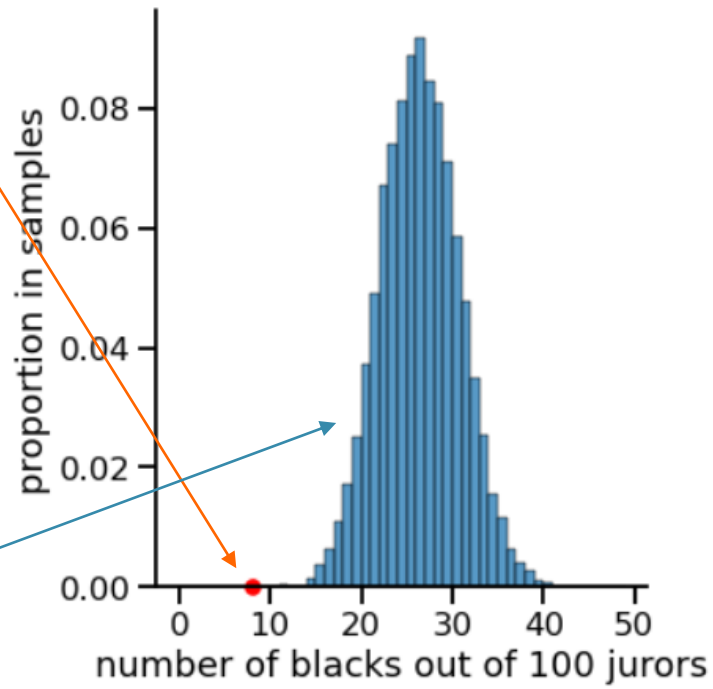
[B, W, W, W, B, ..., W]

8%

A model → Simulations → Samples (many) → Estimate for the parameter, under the model

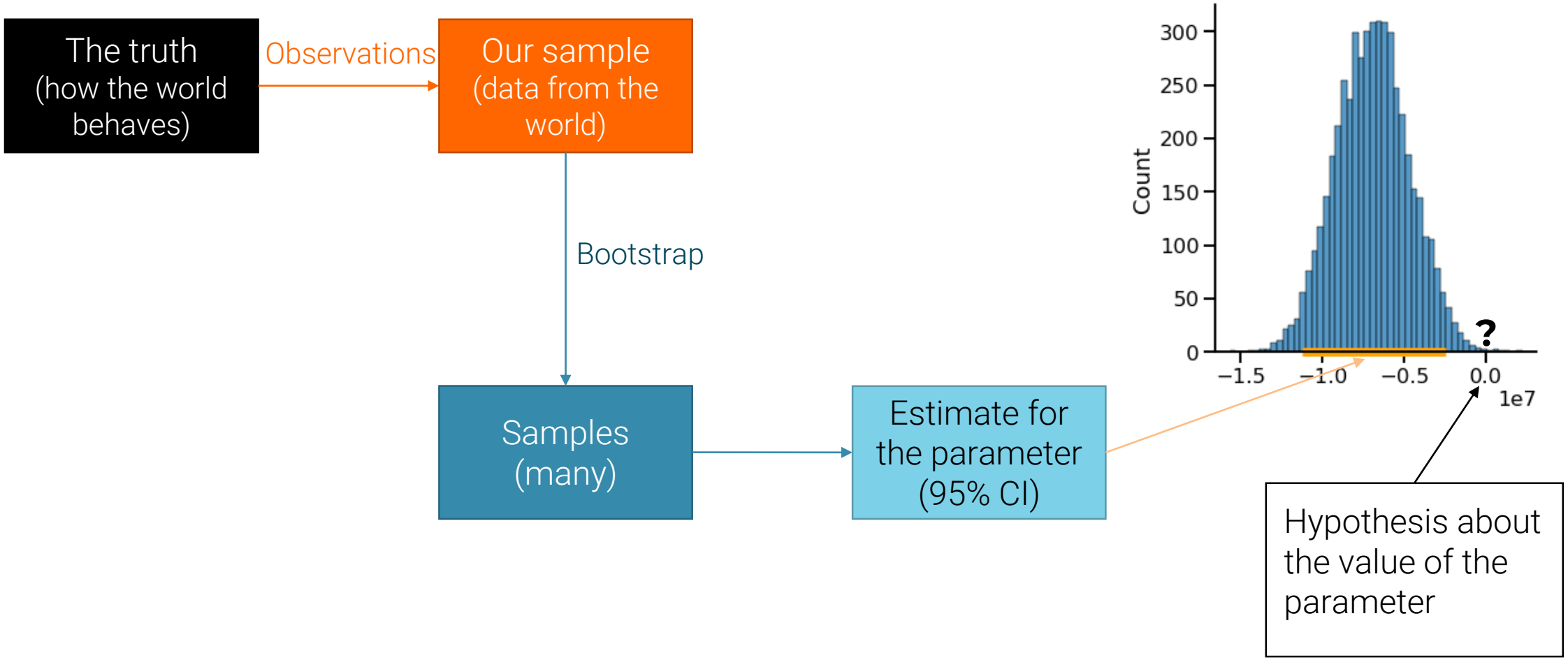
[W, W, W, B, B, ..., W]
[W, B, W, B, B, ..., B]
[W, W, B, B, W, ..., W]
...
[B, W, B, W, B, ..., W]

16%
27%
22%
...
25%



Hypothesis testing using bootstrap

We want to know something about a parameter, but we only have a sample!



CI for testing difference in population means

- Null hypothesis: Difference between population averages = *0*
- Alternative hypothesis: Difference between population averages ≠ *0*
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the difference between population averages
  - If 0 is not in the interval, reject the null
  - If 0 is in the interval, can't reject the null

More jury panels...

RACIAL AND ETHNIC DISPARITIES

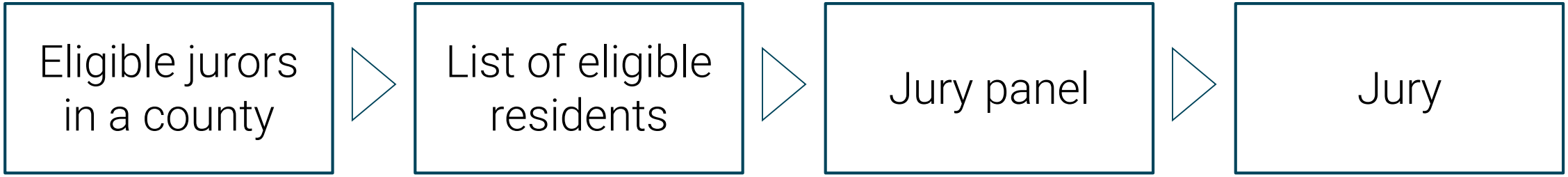
IN

ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California

October 2010

Jury selection in Alameda County



Section 197 of California's Code of Civil Procedure:
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.

Jury selection in Alameda County

- Alameda County, California, 2010
- American Civil Liberty Union (ACLU) of Northern California
- Report: some ethnic groups are underrepresented in jury panels
- Though the process should be random
- Sample: 1453 jury panelists
- Check ethnic composition: Asian, Black, Latino, White, Other
- And compare to Alameda County ethnic composition

Hypotheses?

Hypotheses

- H_0: The people on the jury panels were selected at random from the eligible population
- H_1: No, they weren't
- How to decide?

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 under the null hypothesis
  - It shows all the likely values of the statistic
  - And how likely they are **(if the null hypothesis is true)**
- The probabilities are approximate, because we can't generate all the possible random samples

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 under the null hypothesis
  - It shows all the likely values of the statistic
  - And how likely they are **(if the null hypothesis is true)**
- The probabilities are approximate, because we can't generate all the possible random samples

- What is the test statistic?

Comparing distributions

Previously, we saw:

- how to compare an empirical proportion to a theoretical one
  - e.g. number of black jurors in a jury panel
- How to compare an empirical mean to a theoretical one
  - e.g. the average IMDB score of a movie
- (and also how to compare between means/medians via bootstrap)

Now, we need to compare two categorical distributions:
theoretical and empirical racial composition

Examine how “close” they are to one another

Distance between distributions

- People on the jury panels are of multiple ethnicities
- Distribution of ethnicities is categorical
  - Asian, Black, Latino, White, Other
- To see whether the distribution of ethnicities of the panel jurors is “close” to that of the eligible jurors, we need to measure the **distance between two categorical distributions**
- This distance will be our test statistic

Total Variation Distance

Total Variation Distance (TVD):

- For each category, compute the difference in proportions between two distributions
- Take the absolute value of each difference
- Sum over categories, and divide the sum by 2

$$TVD(P,Q) = \frac{\sum_i |p_i - q_i|}{2}$$

Total Variation Distance

Total Variation Distance (TVD):

- For each category, compute the difference in proportions between two distributions
- Take the absolute value of each difference
- Sum over categories, and divide the sum by 2

$$TVD(P,Q) = \frac{\sum_i |p_i - q_i|}{2}$$

- What is the minimal TVD possible?
- What is the maximal TVD possible?

Summary of the method

Is a sample 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 many times
- Compare:
  - Empirical distribution of simulated TVDs
  - Actual TVD from the sample in the study

(notebook)

Where do hypotheses come from?

- Declared processes
  [e.g., “jury panels are selected at random from eligible population]
- Prior experiments
  [e.g. observing a phenomena, hypothesizing, testing with new data]
- Scientific understanding
  [e.g., genetics, physics]
- ...