DATA 8

Summer 2018

Lecture 15

Assessing Models

Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)
Contributions by Fahad Kamran (fhdkmrn@berkeley.edu) and Vinitra Swamy (vinitra@berkeley.edu)

Announcements

A Statistic

Terminology

- **Statistical Inference**

  Making conclusions based on data in random samples
- **Parameter**
  - A number associated with the population
- **Statistic**
  - A number calculated from the sample

A statistic can be used to **estimate** a parameter, or to **test hypotheses** about the process that generated the data

Simulating a Statistic

- Figure out the code to generate *one* value of the statistic
- Create an empty array in which you will collect all the simulated values
- For each repetition of the process:
  - Simulate one value of the statistic
  - Append this value to the collection array
- At the end of all the repetitions, the collection array will contain all the simulated values

(Demo)

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

Testing Hypotheses

Choosing One of Two Viewpoints

- Based on data
  - “Chocolate has no effect on cardiac disease.”
  - “Yes, it does.”
  - “This jury panel was selected at random from eligible jurors.”
  - “No, it has too many people with college degrees.”

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”

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 predictions to the data that were observed.
- If the data and the model's predictions are not consistent, that is evidence against the model.

How to Compare Predictions to Data

- When we simulate data according to the assumptions of the model, we need an easy representation of this data
- We calculate a statistic on each simulated sample
- We then calculate the statistic on our observed data
- Check whether our observed statistic and our simulated statistics are consistent
- The difficulty is in picking the statistic
  - Either large statistics or small statistics should be evidence against your model
  - The statistic chosen must help us distinguish between our model and any other alternative viewpoint

Jury Selection

Swain vs. Alabama, 1965

- Talladega County, Alabama
- Robert Swain, black man convicted of crime
- Appeal: one factor was all-white jury
- Only men 21 years or older were allowed to serve
- 26% of this population were black
- Swain’s jury panel consisted of 100 men
- 8 men on the panel were black

Supreme Court Ruling

- About disparities between the percentages in the eligible population and the jury panel, the Supreme Court wrote:

"... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes"

- The Supreme Court denied Robert Swain's appeal

Sampling from a Distribution

- Sample at random from a categorical distribution

```
sample_proportions(sample_size, pop_distribution)
```

- Samples at random from the population
  - Returns an array containing the distribution of the categories in the sample

(Demo)

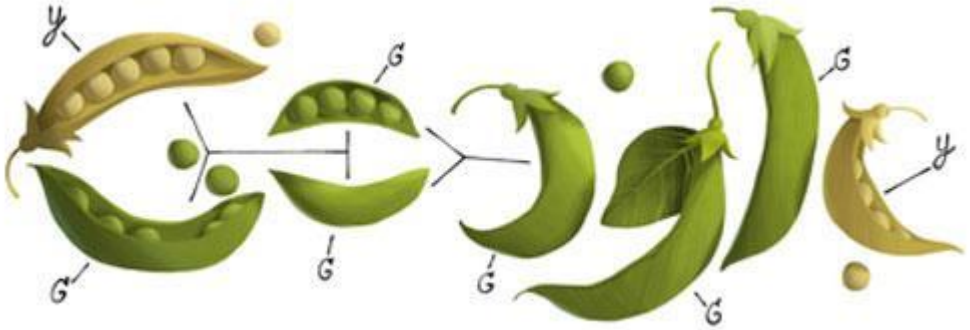
Break

A Genetic Model

Steps in Assessing a Model

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

Gregor Mendel, 1822-1884



A Model

- Pea plants of a particular kind
- Each one has either purple flowers or white flowers

- Mendel's model:
  - Each plant is purple-flowering with chance 75%,
  - regardless of the colors of the other plants

- Question:
  - Is the model good, or not?

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:

  | sample percent of purple-flowering plants - 75 |

- If the statistic is large, that is evidence against the model

(Demo)

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 tails.”

“Fair”

For both (a) and (b),

- The number of heads in the 400 tosses is a good starting point, but might need adjustment
- A number of heads around 200 suggests “fair”

Answers

(a) 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: | number of heads - 200 |
- Large values of the statistic suggest “not fair”

(b) Small values of the number of heads suggest “biased towards tails”

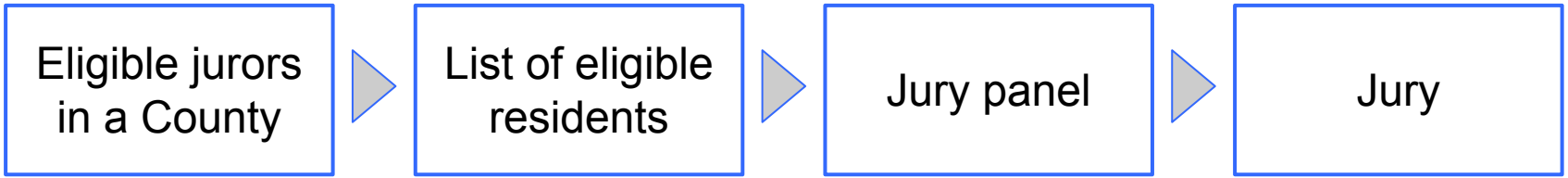
- Statistic: number of heads

Comparing Distributions

Jury Selection in Alameda County

A Report by the ACLU of Northern California October 2010

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