YData: An Introduction to Data Science

Lecture 17: Comparing Distributions

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

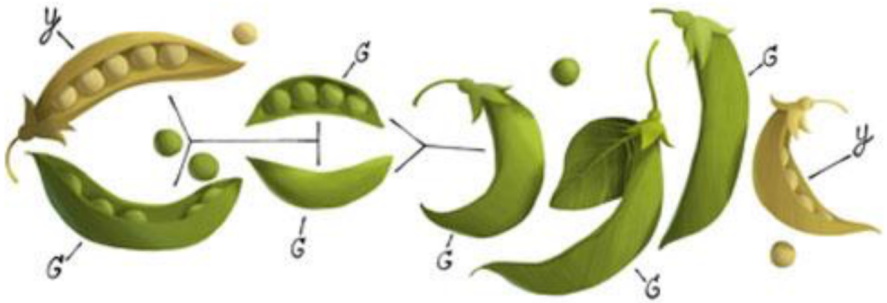
- Project 1 is due tonight
- Hw 06 has been posted
- Practice 4b on table manipulation has been posted

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
- Mendel grew 929 plants and 705 out of them had purple flowers

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

RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

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."

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

- There are 1453 people on the panels
- People on the panels are of multiple ethnicities
- Distribution of ethnicities is categorical
- To see whether 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

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