DATA 8

Spring 2018

Lecture 16

Assessing Models

Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@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.

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

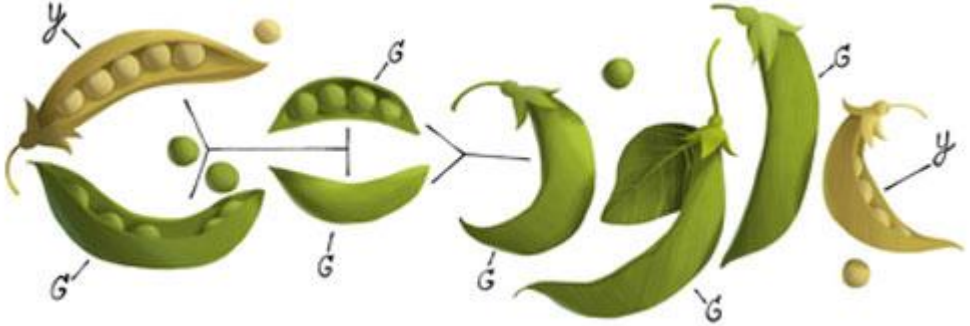
(Demo)

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)