190F Foundations of Data Science

Fall 2018

Lecture 14

Statistics

Announcements

Probability & Simulation

Calculation

Roll a fair die 4 times.

What is P(get at least one 6)?

Calculation

Roll a fair die 20 times. What is P(get at least one 6)?
Three ways to compute it:

- **Calculation:** Use math.
- **Enumeration:** Count all outcomes.
- **Estimation:** Randomly sample outcomes. Estimate.

Statistical Inference & Simulation

Terminology

- **Statistical Inference:** Making conclusions based on data in random samples
- **Parameter:** A number associated with a population.
- **Statistic:** A number calculated from a sample drawn at random from a population.

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 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 consists of:
    - 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.*

(Demo)

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 people on the panel were black (8%)

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 [blacks]"

- 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

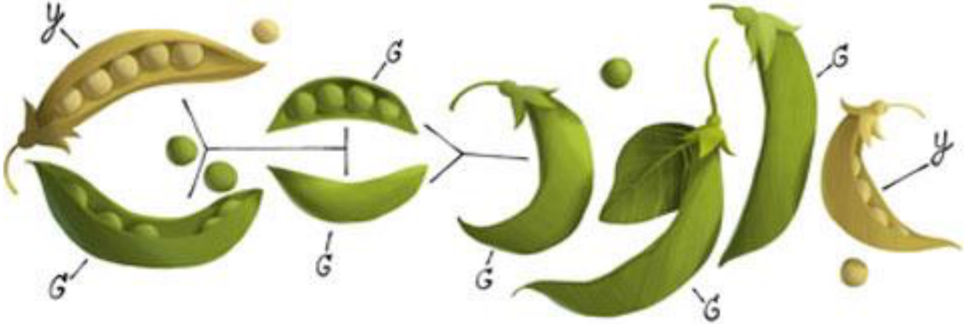
(Demo)

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)

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.”

Estimation

How many enemy planes?

Assumptions

- Planes have serial numbers 1, 2, 3, …, N.
- We don't know N.
- We would like to estimate N based on the serial numbers of the planes that we see.

The main assumption

- The serial numbers of the planes that we see are a uniform random sample drawn with replacement from 1, 2, 3, …, N.

Discussion question

If you saw these serial numbers, what would be your estimate of N?

```
170  271  285  290  48
235   24   90  291  19
```

One idea: 291. Just go with the largest one.

The largest number observed

- Is it likely to be close to N?
  - How likely?
  - How close?

Option 1. We could try to calculate the probabilities and draw a probability histogram.

Option 2. We could simulate and draw an empirical histogram.

(Demo)

Verdict on the estimate

- The largest serial number observed is likely to be close to N.
- But it is also likely to underestimate N.

Another idea for an estimate:
Average of the serial numbers observed ~ N/2

New estimate: 2 times the average

(Demo)