YData: An Introduction to Data Science

Lecture 23: Confidence Intervals and the Bootstrap

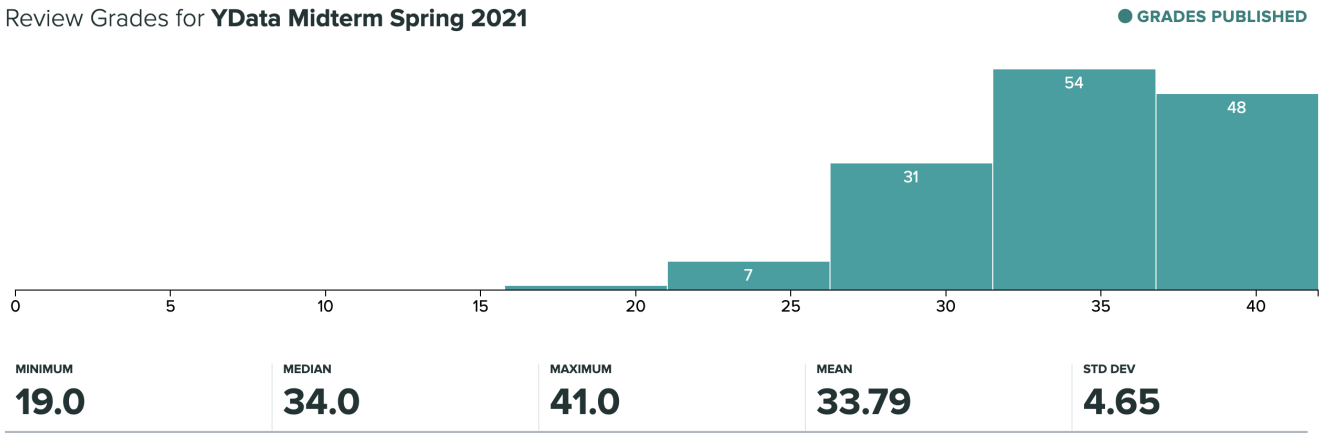
Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

- Midterm scores posted
- Assignment 07 due on Thursday
- Normal OH resume this week

Midterm



- You did really well!
- We will post sample solutions later
- Questions about grading to instructors via email, please
- Scores will be curved upwards

High level view

Intro, Cause and Effect	Lectures 1–2
Python, Tables, Visualization	Lectures 3–13
Probability and Distributions	Lectures 14–17
Hypothesis Testing and Causality	Lectures 18–20
Midterm exam	—
Confidence and the Normal Distribution	Lectures 23–28
Regression and Classification	Lectures 29–37
Final exam	—

Percentiles

Computing Percentiles

The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

The 80th percentile is ordered element 4:

`(80/100)`	*	5
Percentile		Size of set

For a percentile that does not exactly correspond to an element, take the next greater element instead

The `percentile` Function

- The pth percentile is the value in a set that is at least as large as p% of the elements in the set
- Function in the `datascience` module:

```
percentile(p, values)
```

- `p` is between 0 and 100
- Returns the pth percentile of the array

Discussion Question

Which are True, when s = [1, 7, 3, 9, 5]?

```
percentile(10, s) == 0

percentile(39, s) == percentile(40, s)

percentile(40, s) == percentile(41, s)

percentile(50, s) == 5
```

(DEMO)

Estimation

Inference: Estimation

- What is the value of an unknown parameter?
- If you have a census (that is, the whole population):
  - Just calculate the parameter and you're done
- If you don't have a census:
  - Take a random sample from the population
  - Use a statistic as an **estimate** of the parameter

(DEMO)

Variability of the Estimate

- One sample $\longrightarrow$ One estimate
- But the random sample could have come out differently
- And so the estimate could have been different
- Main question:
  - **How different could the estimate have been?**
- The variability of the estimate tells us something about how accurate the estimate is:

  estimate = parameter + error

Where to Get Another Sample?

- One sample $\longrightarrow$ One estimate
- To get many values of the estimate, we needed many random samples
- Can't go back and sample again from the population:
  - No time, no money
- Stuck?

The Bootstrap

International Prize in Statistics Awarded to Stanford's Bradley Efron

2 JANUARY 2019 995 VIEWS NO COMMENT

The International Prize in Statistics has been awarded to Bradley Efron, professor of statistics and biomedical data science at Stanford University, in recognition of the "bootstrap," a method he developed in 1977 for assessing the uncertainty of scientific results that has had extraordinary impact across many scientific fields.

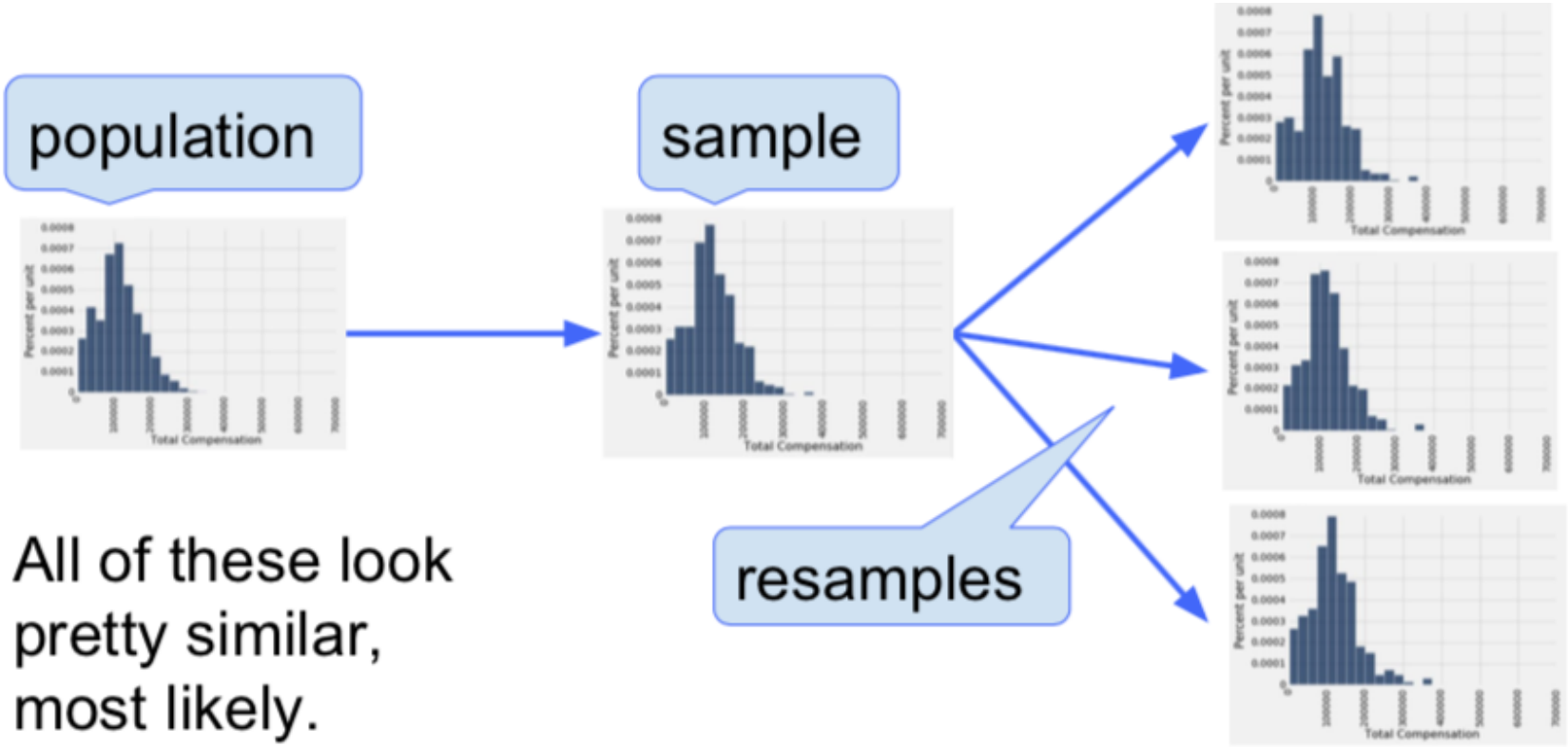
With the bootstrap, it is possible to simulate a potentially infinite number of data sets from an original data set and—in looking at the differences—measure the uncertainty of the result from the original data analysis.

Made possible by computing, the bootstrap powered a revolution that placed statistics at the center of

The Bootstrap

- A technique for estimating confidence by simulating repeated random sampling
- All that we have is the original sample
  - ... which is large and random
  - Therefore, it probably resembles the population
- So we sample at random from the original sample!

How the Bootstrap Works



All of these look pretty similar, most likely.

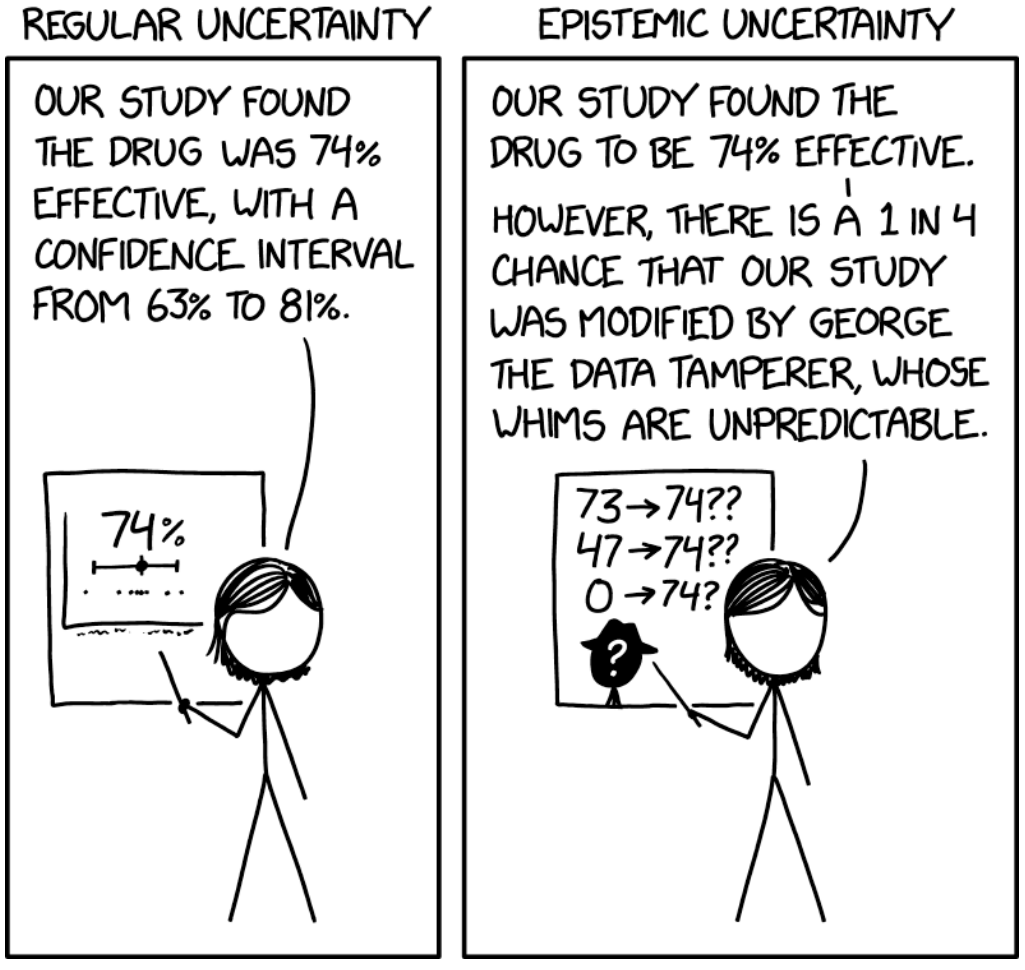
Key to Resampling

- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained

- *The size of the new sample has to be the same as the original one*, so that the two estimates are comparable

(DEMO)

Wednesday: More on Confidence Intervals



xkcd.com/2440