DATA 8

Fall 2016

Lecture 22, October 17

The Bootstrap

Slides created by Ani Adhikari and John DeNero

Announcements

- The class did fine on the midterm. The score distribution, answers, regrading policy, and graded exams will be released by Tuesday evening.
- Labs meet as usual this week.
- No homework due this week.
- Homework will be assigned on Friday.
- We're being filmed today. If you'd prefer not to appear, try sitting near the back.

Percentiles

- The data: a collection of numerical values
- The *p*th percentile of the data is:
  - the smallest value in the collection
  - that is at least as large as
  - *p*% of all the values.
- The median (50th percentile) of 4, 7, 9, 10, 15:
  - 9

`percentile`

- Function:

  `percentile(p, array)`

- `p` is between 0 and 100

- Returns the *p*th percentile of the array

Inference: Estimation

- How big is 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 → One estimate

- But the random sample could have come out differently.
- Then the estimate would 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.

Where to Get Another Sample?

- One sample → One estimate
- To get another value of the estimate, need another random sample.
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?

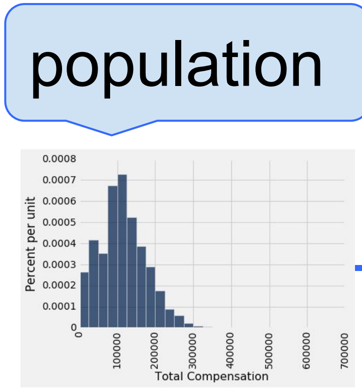
The Bootstrap

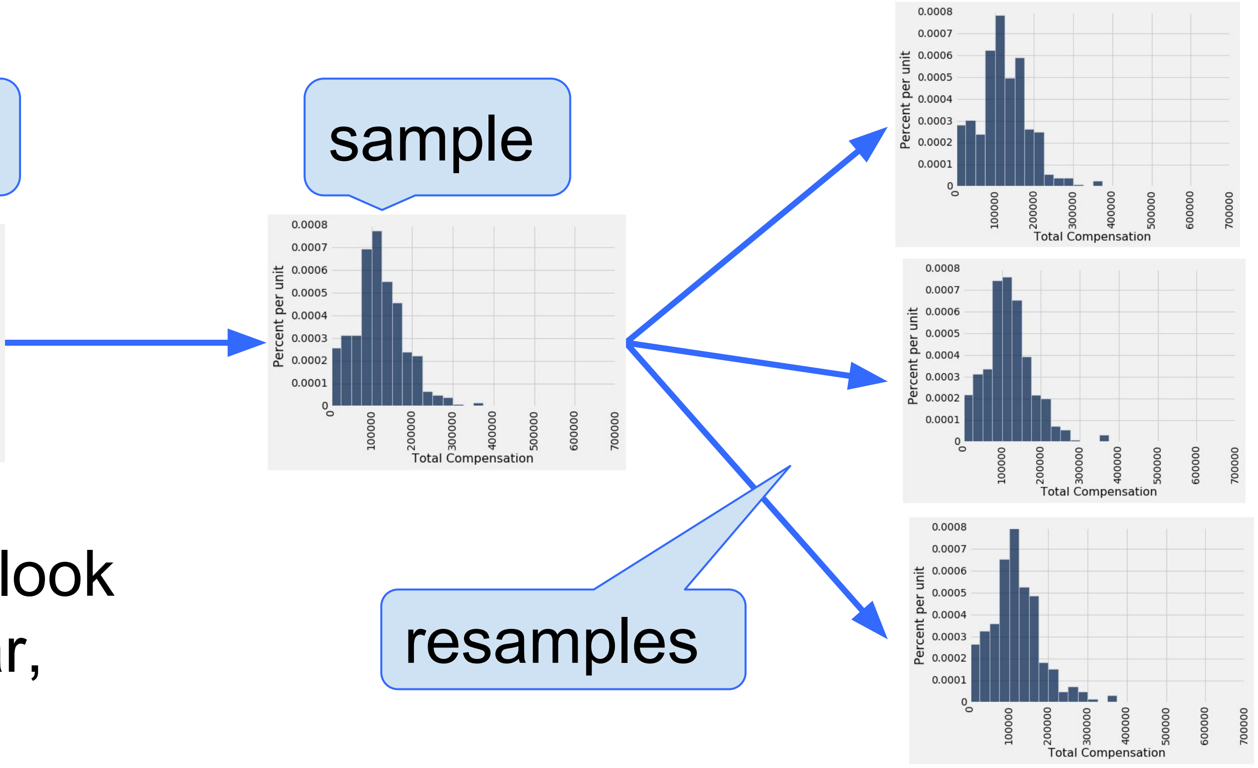
- Need another random sample that looks like the population

- All that we have is the original sample
  - which is large and random.
  - It's a good bet that it resembles the population.

- So **sample at random from the original sample**!

Why the Bootstrap Works





All of these look pretty similar, most likely.

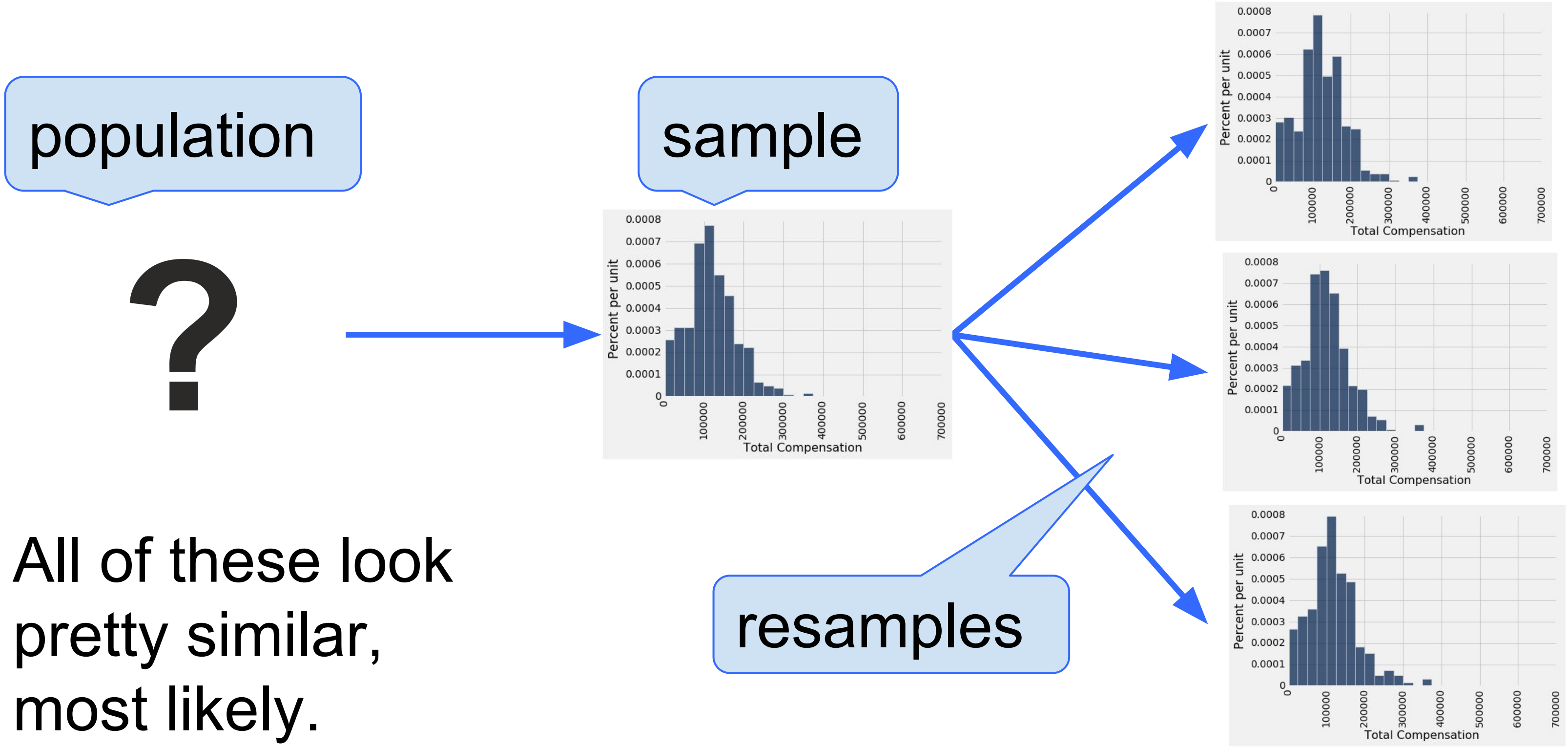
Key to Resampling

- From the original sample,
  - draw at random
  - **with** replacement
  - the **same number of times** as the original sample size.

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

(Demo)

Inference Using the Bootstrap



All of these look pretty similar, most likely.