DATA 8

Fall 2016

Lecture 23, October 19

Confidence Intervals

Slides created by Ani Adhikari and John DeNero

Announcements

- Exams, solutions, score summary, and regrading policy have been released. See Gradescope and Piazza.
- Labs meet as usual this week.
- No homework due this week.
- Homework will be assigned on Friday.
- Later this week I will post a note about courses to take if you are interested in learning more about data science.
- As yet there is no clear timetable for a Data Science major or minor. But we’re working on it.

Variability of an 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.

The Bootstrap

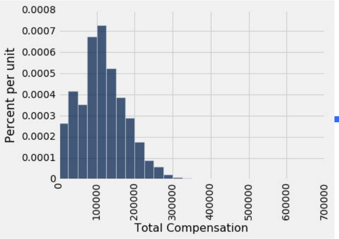
- Need another random sample that looks like the population

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

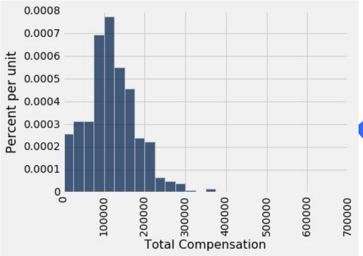
- So **sample at random from the original sample**!

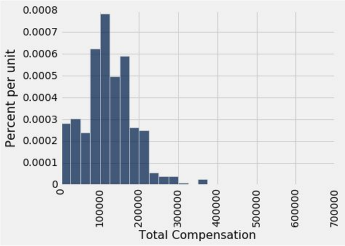
Why the Bootstrap Works

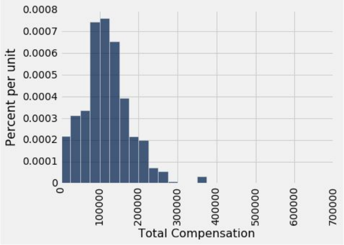
population

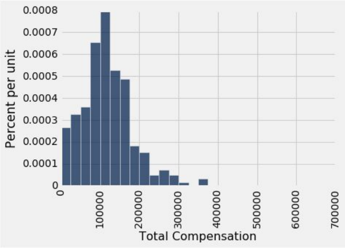


sample









resamples

All of these look pretty similar, most likely.

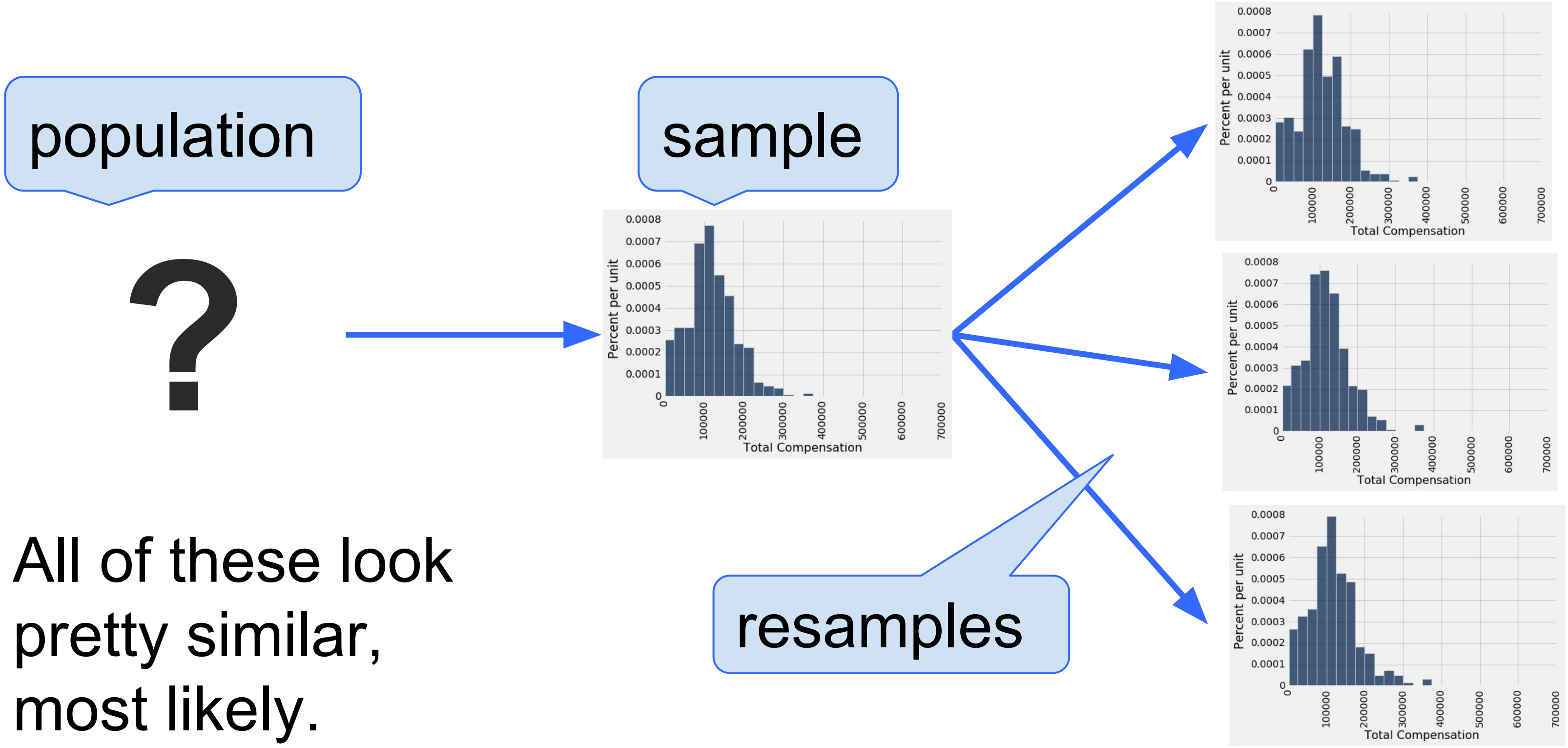
Key to Resampling

- 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



population

sample

resamples

All of these look pretty similar, most likely.

95% Confidence Interval

- Interval of **estimates of a parameter**
- Based on random sampling
- 95% is called the **confidence level**
  - Could be any percent between 0 and 100
  - Bigger is better
- The **confidence is in the process** that generated the interval:
  - It generates a “good” interval about 95% of the time.

Bootstrap Percentile Method

- For constructing a confidence interval for an unknown parameter
- Starting point: one large random sample
- One replication:
  - Bootstrap the sample to get a “resample”
  - Get an estimate based on the resample
- Repeat several thousand times (10,000 recommended)
- For an approximate 80% confidence interval, take the 10th and 90th percentiles of all the bootstrap estimates

(Demo)