DSC 10, Spring 2018
Lecture 16

Bootstrapping and Confidence Intervals

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Inference: Estimation

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
- 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

Where to Get Another Sample?

- One sample ➜ One estimate
- To get many values of the estimate, we needed many random samples
- What if we 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
  - Therefore, it probably resembles the population
- So we sample at random from the original sample!
- A technique for simulating repeated random sampling

The Bootstrap

Questions

What should be the size of your new sample?

A. 25% of the original sample
B. 50% of the original sample
C. 75% of the original sample
D. 100% of the original sample
E. Depends on the problem

How should we obtain this new sample?

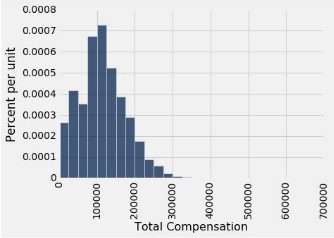
A. **with** replacement
B. **without** replacement
C. Depends on the problem

Key to Resampling

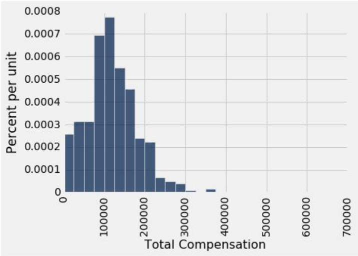
- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained
- The size of the new sample has to be the same as the original one, so that the two estimates are comparable

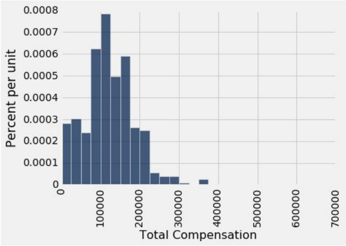
Why the Bootstrap Works

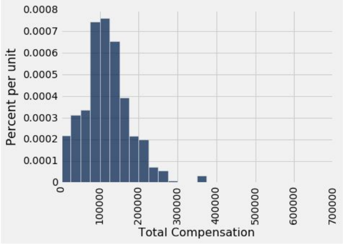
population

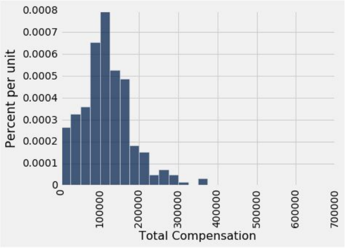


sample









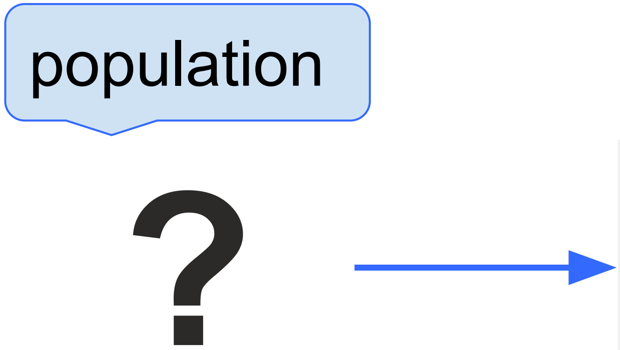
resamples

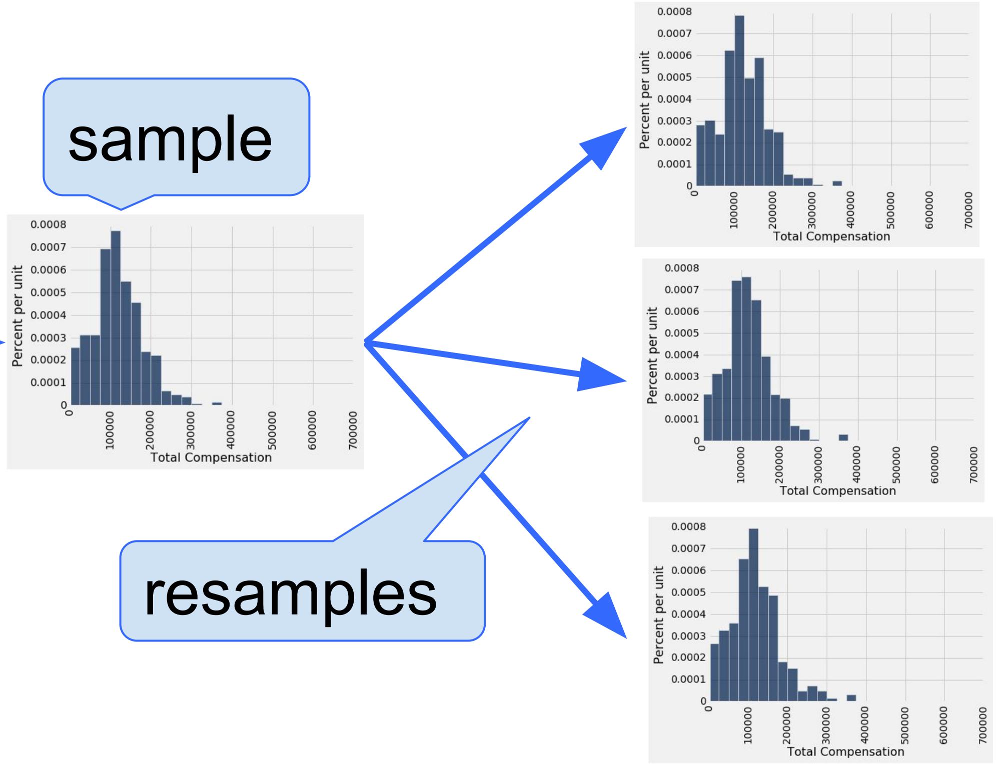
All of these look pretty similar, most likely.

(Demo)

Confidence Intervals

Inference Using the Bootstrap





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 means wider intervals
- The **confidence is in the process** that generated the interval:
  - It generates a “good” interval about 95% of the time.

Important Note

- “It generates a “good” interval about 95% of the time”
  - Which means 95% of the samples will result in the “good interval”
  - Not resamples!
- If my original sample was way off, your interval will be way off…
  - ..even if you keep bootstrapping

(Demo)

Use Methods Appropriately

When *Not* to Use The Bootstrap

- If you're trying to estimate very high or very low percentiles, or min and max
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
- If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)
- If the original sample is very small

Can You Use a C.I. Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

A: True
B: False
C: I'm lost

(Demo)