DSFA

Spring 2020

Lecture 18

Confidence Intervals

Estimation (Review)

Inference: Estimation

- What is the value of a population 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:

  estimate = parameter + error

(Demo)

Where to Get Another Sample?

- One sample ➜ 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

The Bootstrap

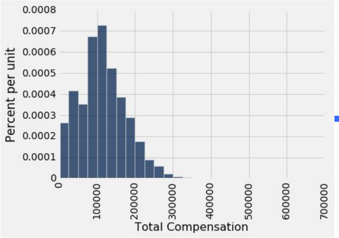
- A technique for simulating repeated random sampling

- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population

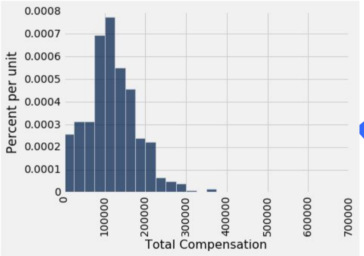
- So we sample at random from the original sample!

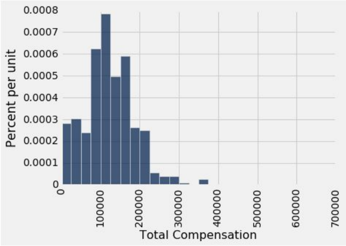
Why the Bootstrap Works

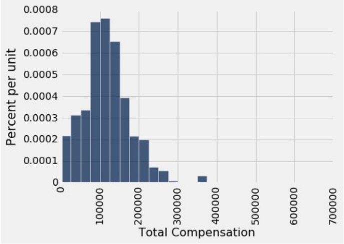
population

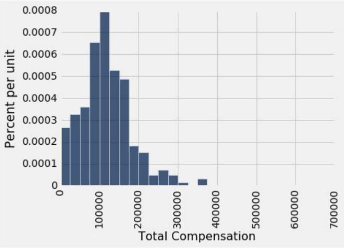


sample









resamples

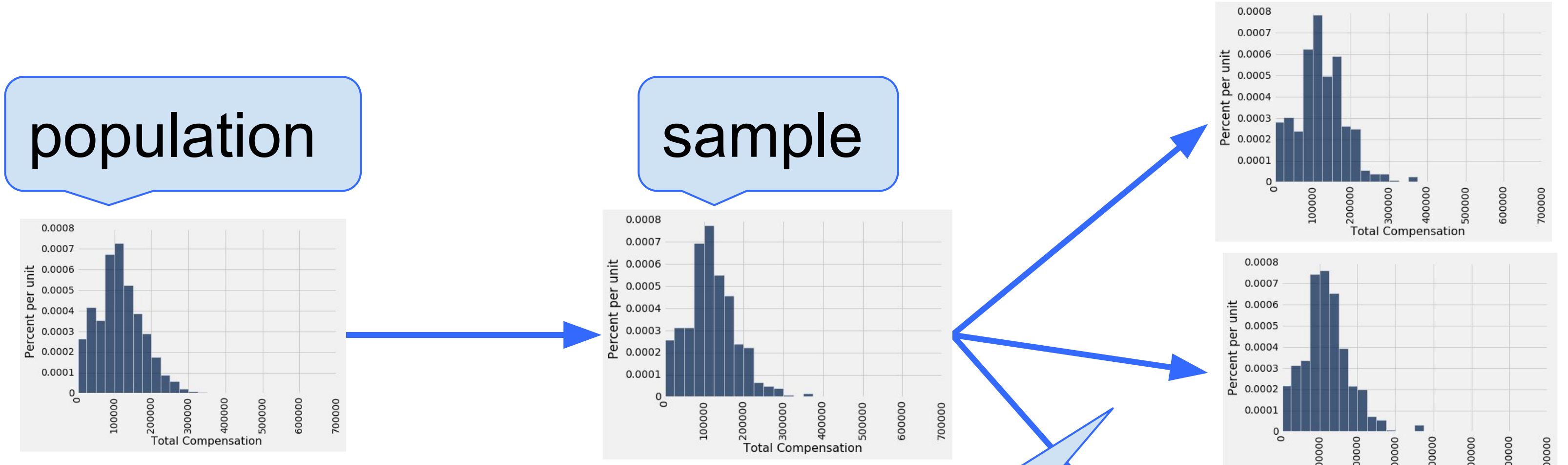
All of these look pretty similar, most likely.

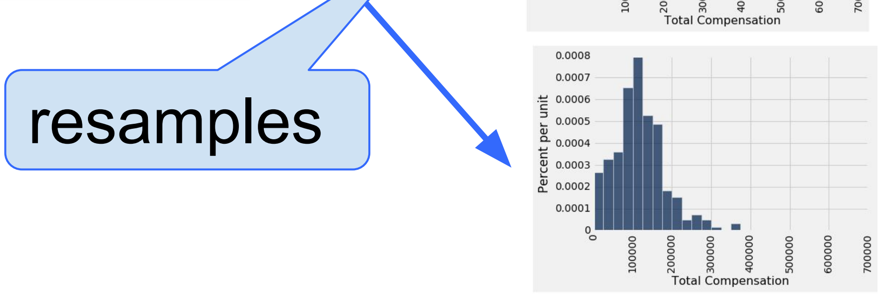
Key to Resampling

- From the original sample,
  - draw at random
  - with replacement
  - 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

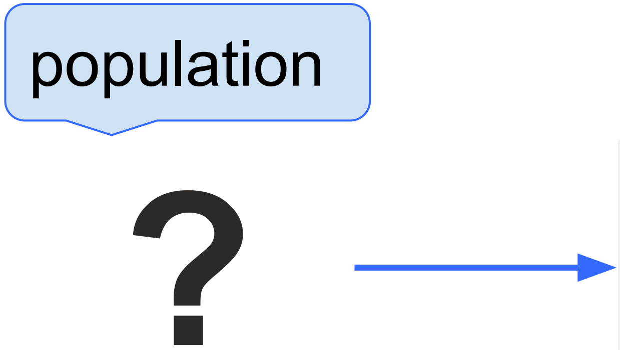
Why the Bootstrap Works

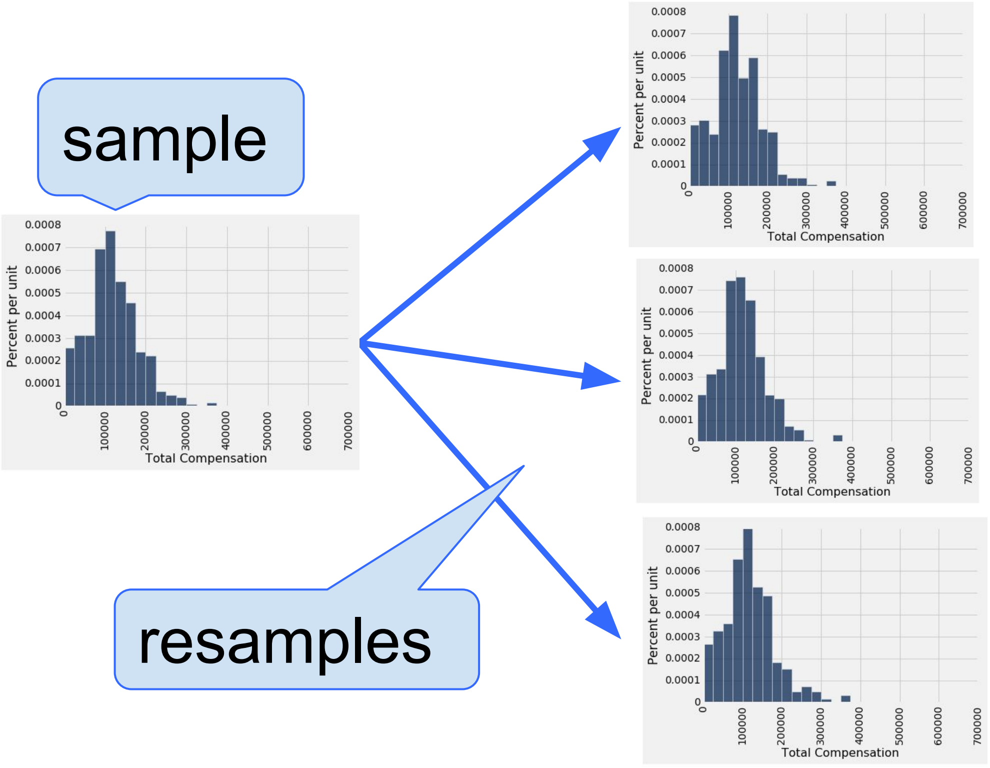




All of these look pretty similar, most likely.

Inference Using the Bootstrap





All of these look pretty similar, most likely.

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 (~15)
- Be sure to take lots of resamples! (10,000)

95% Confidence Interval

- Interval of **estimates of a parameter**
- Based on random sampling
- Confidence level: typically 95%
    - Could be any percent between 0 and 100
    - Bigger means wider intervals
- The interval contains the parameter about 95% of the time **in repeated sampling**

(Demo)

Can You Use a CI 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.

Answer: False. We're estimating that their **average age** is in this interval.

Is This What a CI Means?

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

True or False:

- There is a 0.95 probability that the average age of mothers in the population is in the range 26.9 to 27.6 years.

Answer: False. It's not a probability. Either the population average is in the interval or it isn’t!

Confidence Interval Tests

Using a CI for Testing

- Null hypothesis: **Population mean = *x***
- Alternative hypothesis: **Population mean ≠ *x***
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the population statistic
  - If *x* is not in the interval, reject the null
  - If *x* is in the interval, can't reject the null