190F Foundations of Data Science

Spring 2020

Lecture 13

Confidence Intervals

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 *p*th 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 *p*th 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
```

Estimation (Review)

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

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

Where to Get Another Sample?

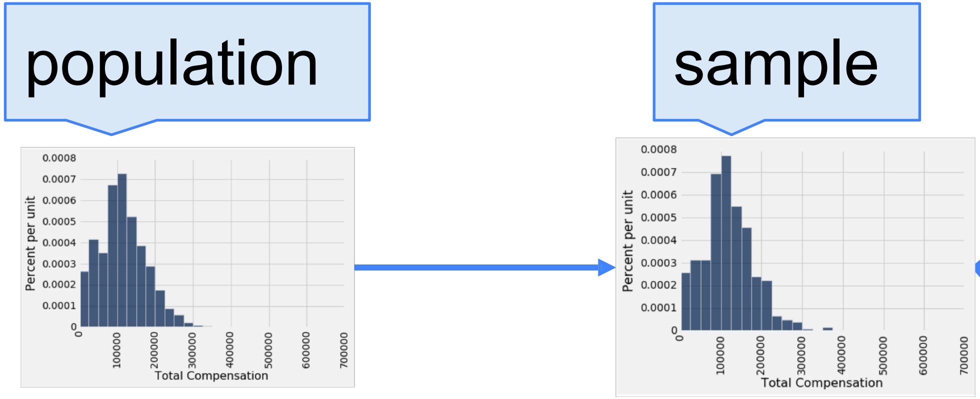
- One sample ➜ One estimate
- To get many values of the estimate, we needed many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?

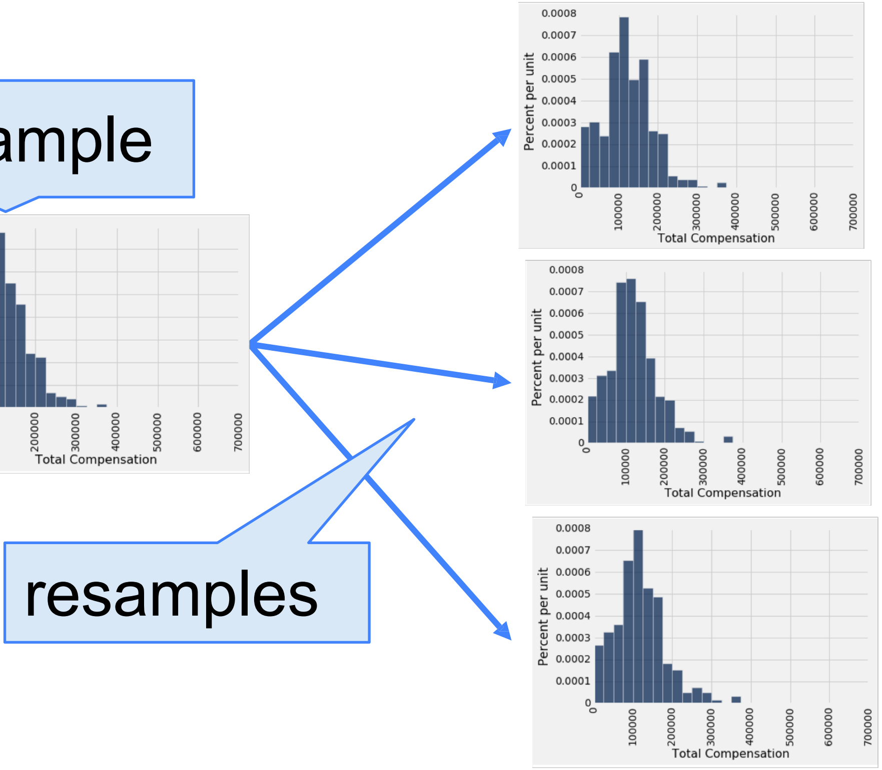
The Bootstrap

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





All of these look pretty similar, most likely.

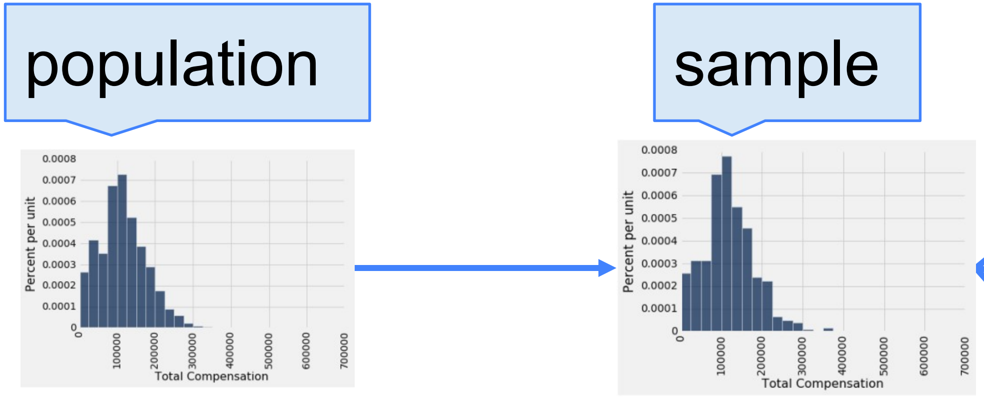
Key to Resampling

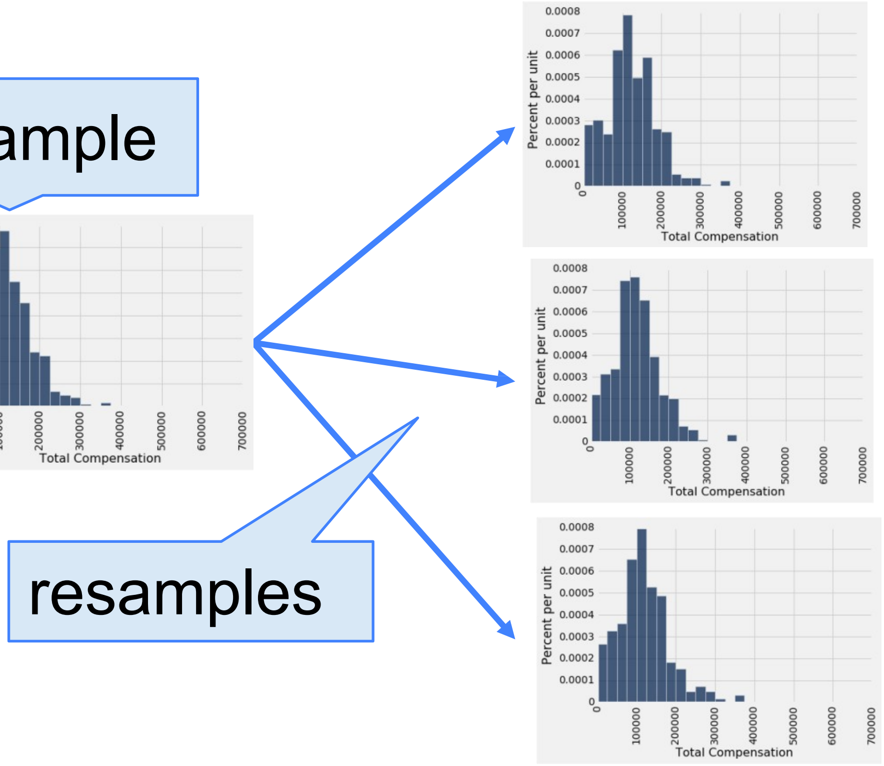
- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained

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

(Demo)

Why the Bootstrap Works





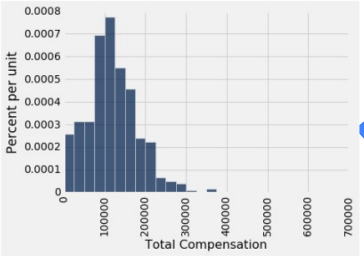
All of these look
pretty similar,
most likely.

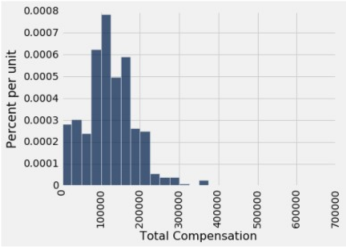
Inference Using the Bootstrap

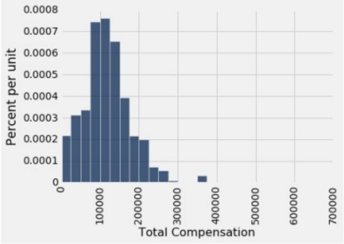
population

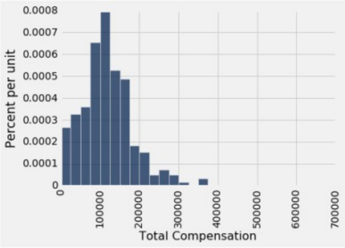
?

sample



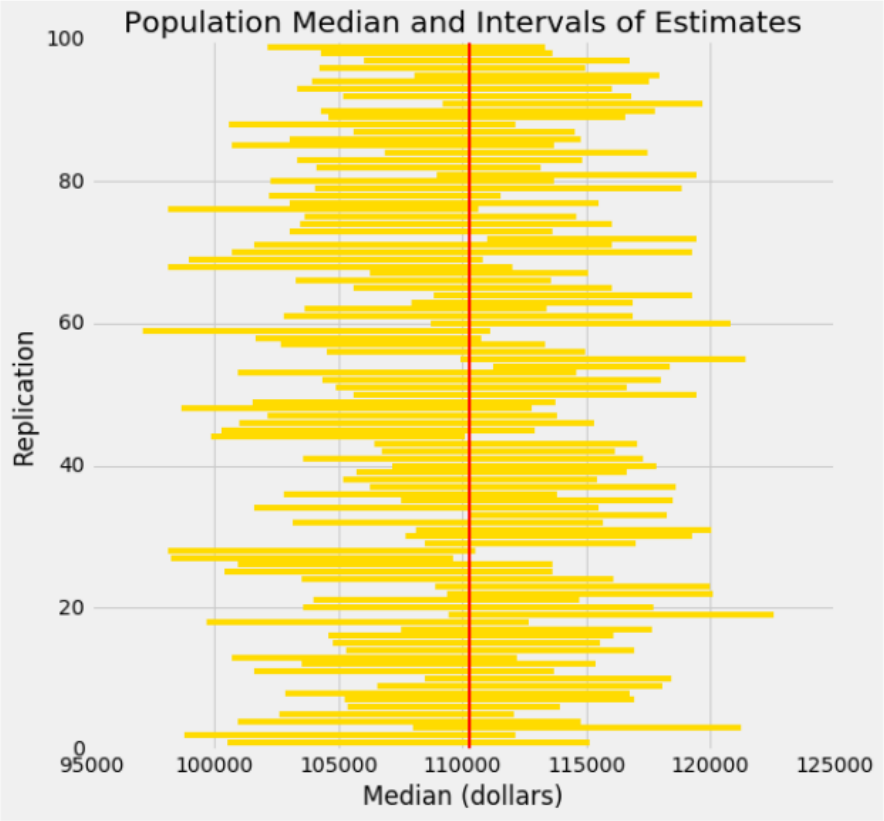






resamples

All of these look pretty similar, most likely.



Each line here is a confidence interval from a fresh sample from the population

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.

(Demo)

Use Methods Appropriately

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 95% of the time.

Is This What a CI Means?

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:

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

Answer: False. It's not a probability. It's either true or false that the average age of mothers is in the range (26.9, 27.6)

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

(Demo)

Confidence Interval Tests

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.

(Demo)

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 average
  - If *x* is not in the interval, reject the null
  - If *x* is in the interval, can't reject the null