DATA 8

Spring 2020

Lecture 23

Confidence Intervals

Announcements

- Online tutoring sections sign ups start on Tuesday
  - Officially commence after Spring Break
- Midterm: if you had answers erased when submitting
  - Good News:
    - We were able to recover you work!

Weekly Goals

- **Today**
  - Estimation
  - Bootstrap
  - Confidence intervals
- Wednesday
  - Interpreting confidence intervals
- Friday
  - Describing a distribution
  - Center and spread

Percentiles

Computing Percentiles

The Xth percentile is first value on the sorted list that is at least as large as X% of the elements.

Example: `s = [1, 7, 3, 9, 5]`

`s_sorted = [1, 3, 5, 7, 9]`

Percentile

Data set

`percentile(80, s)` is 7

The 80th percentile is ordered element 4: `(80/100) * 5`

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 **smallest 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

(Demo)

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
```

(Demo)

Estimation

Inference: Estimation

- How do we calculate the value of 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
- Big question:
  - How different would it be if we did it again?

(Demo)

Quantifying Uncertainty

- The estimate is usually not exactly right:

Estimate = Parameter + Error

- How accurate is the estimate, usually?
- How big is a typical error?
- When we have a census, we can do this by simulation

(Demo)

Where to Get Another Sample?

- We want to understand errors of our estimate
- Given the **population**, we could simulate
  - ...but we only have the **sample**!
- 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

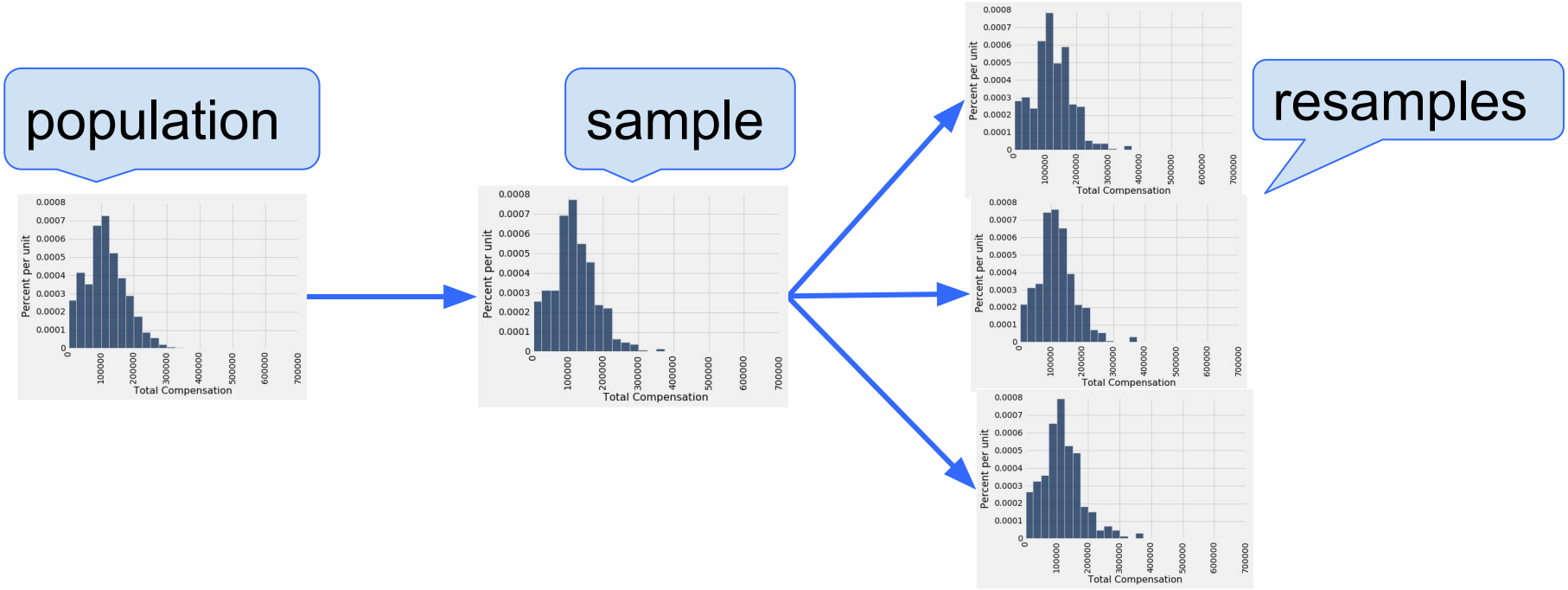
The Bootstrap

- 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

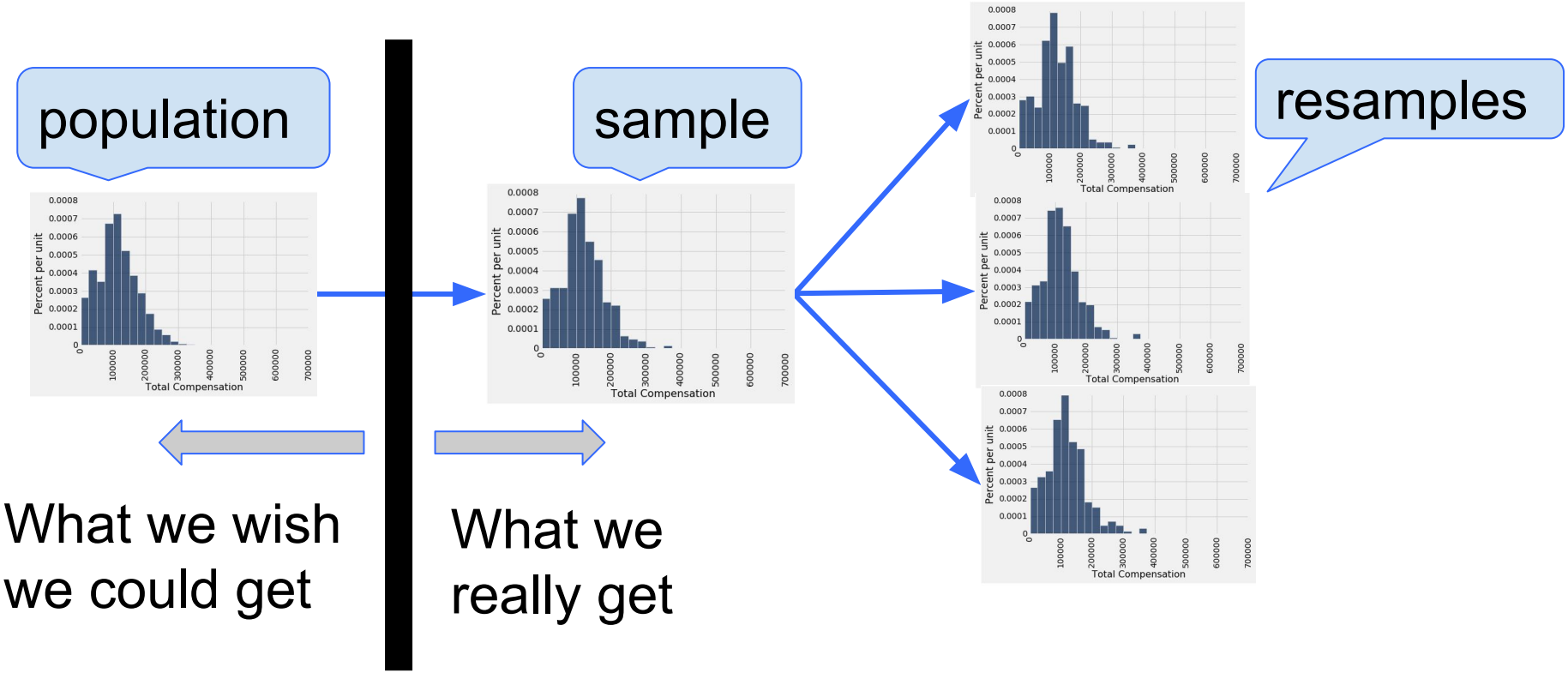
- So we sample at random from the original sample!

Why the Bootstrap Works



All of these look pretty similar, most likely.

Why We Need the Bootstrap



Real World vs. Bootstrap World

Real world:

- True probability distribution (**population**)
  - → Random sample 1
    - → Estimate 1
  - → Random sample 2
    - → Estimate 2
  - …
  - → Random sample 10000
    - → Estimate 10000

Bootstrap world:

- Empirical distribution of original sample (**“population”**)
  - → Bootstrap sample 1
    - → Estimate 1
  - → Bootstrap sample 2
    - → Estimate 2
  - ...
  - → Bootstrap sample 1000
    - → Estimate 1000

Hope: these two scenarios are analogous

The Bootstrap Principle

- The bootstrap principle:
  - **Bootstrap-world** sampling ≈ **Real-world** sampling
- Not always true!
  - … but reasonable if sample is large enough
- We hope that:
  a. Variability of bootstrap estimate
  b. Distribution of bootstrap errors

  ...are similar to what they are in the real world

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

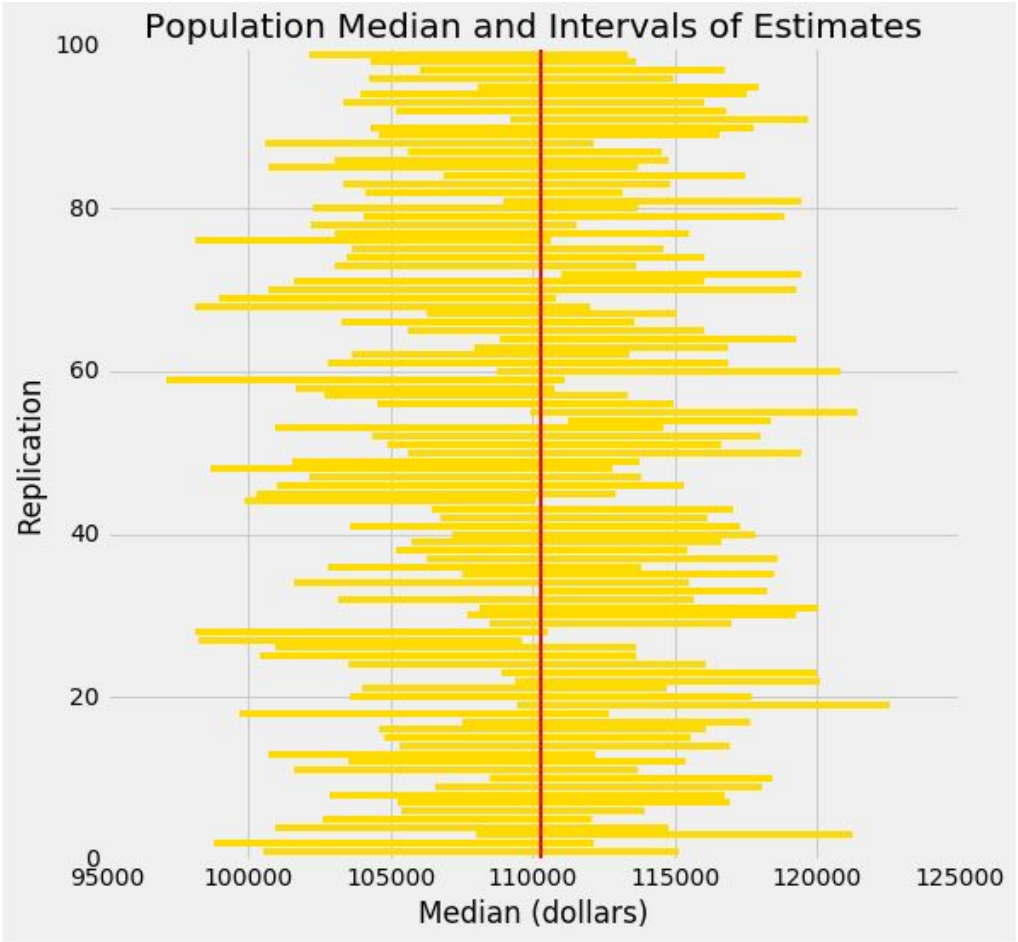
(Demo)

Confidence Intervals

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

(Demo)



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