DATA 8

Spring 2022

Lecture 23

Confidence Intervals

Announcements

Percentiles

Computing Percentiles

The *p*th percentile is first value on the sorted list that is at least as large as *p*% of the elements.

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

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

Percentile Data array

`percentile(80, s)` is 7

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

If *p*% does not exactly correspond to an element (e.g. 85th percentile), take the next greater element (9).

The `percentile` Function

- The *p*th percentile of a set of numbers is the **smallest value** in the set that is **at least as large as *p*%** of the elements in the set
- Function in the `datascience` module:

  `percentile(p, values_array)`

- **`p`** is between 0 and 100
- Evaluates to the *p*th percentile of the array

(Demo)

Discussion Question

Which are `True`, when `s = [1, 5, 7, 3, 9]`?

```
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 can we figure out 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
- How accurate is the estimate, usually?
- If we already have a census, we can check this by comparing the estimate and the parameter

(Demo)

Where to Get Another Sample?

- We want to understand variability 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

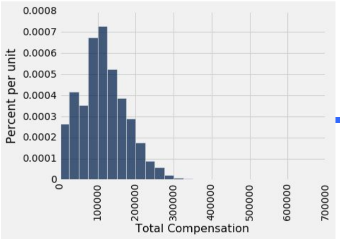
- A technique for simulating repeated random sampling

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

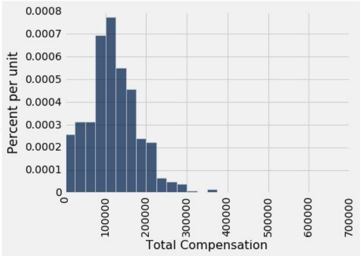
- So we sample at random from the original sample!

The Problem

population



sample

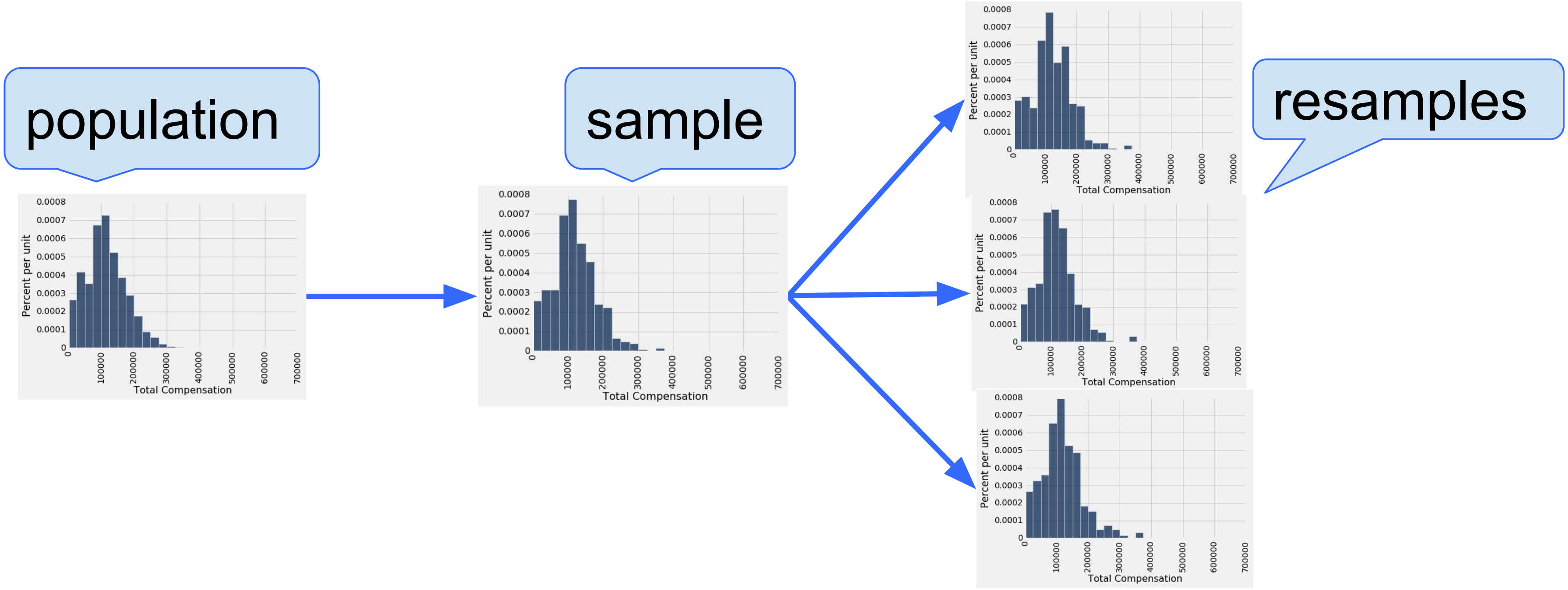


What we wish we could see

What we get to see

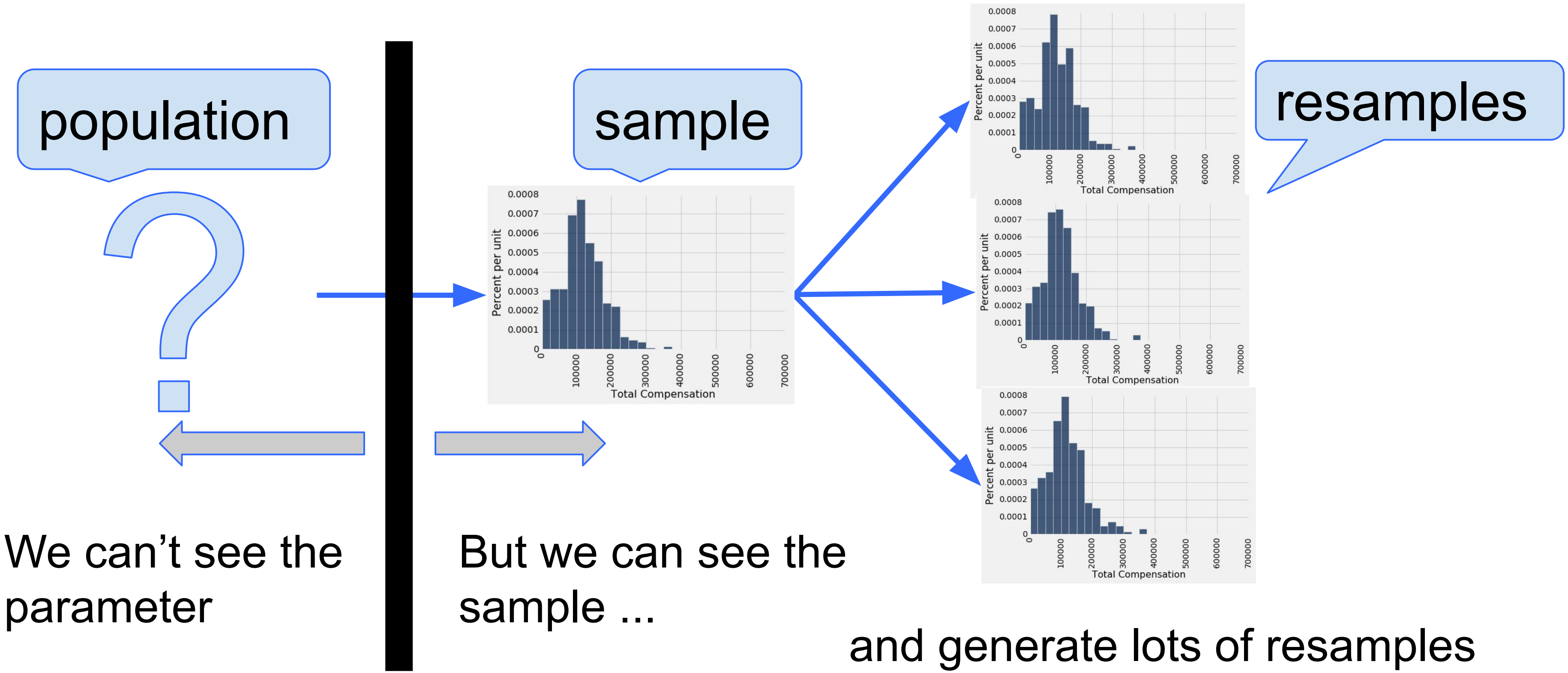
- All we have is the random sample
- We know it could have come out differently
- We need to know how different, to quantify the variability in estimates based on the sample
- So we need to create another sample … or two … or more

Why the Bootstrap Works



All of these look pretty similar, most likely.

Why We Need the Bootstrap



The Bootstrap Principle

- The bootstrap principle:
  - **Re**-sampling from the original random sample

    ≈ Sampling from the population
  - with high probability

- Doesn't always hold
  - … but reasonable for estimating many parameters if the original random sample is large enough

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

The Bootstrap Process

One Random Sample

- True but unknown distribution (**population**)
  - → Random sample (the original sample)

Bootstrap:

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

(Demo)

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 creates the interval:
  - It generates a “good” interval about 95% of the time.

(Demo)

Confidence Intervals