Confidence Intervals and the Bootstrap

LECTURE 23

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Announcements

- Homework 9 due Thursday
- Nb9 session tomorrow (with our TA): 5pm-6pm via Zoom
- Quiz 8 Friday;
    - Scope: HW 8, nb8, Lessons 19 & 20
- We’re over halfway through!
    - You’ve got this!

Today’s Roadmap

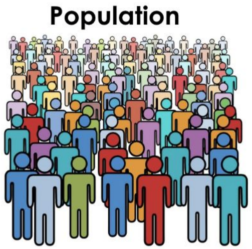
CSCI 3022

- Parameter Estimation
- Confidence Intervals
- Bootstrapping

Estimation

Inference: Estimation

- How can we figure out the value of an **unknown parameter**?
  - Example: *Average GPA of all students*



- If you have a census: **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
    - Example: Statistic = *Average GPA of the Sample*

(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)

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
- A “good” interval is one that contains the parameter
- The **confidence is in the process** that creates the interval:
  - It generates a “good” interval about 95% of the time.

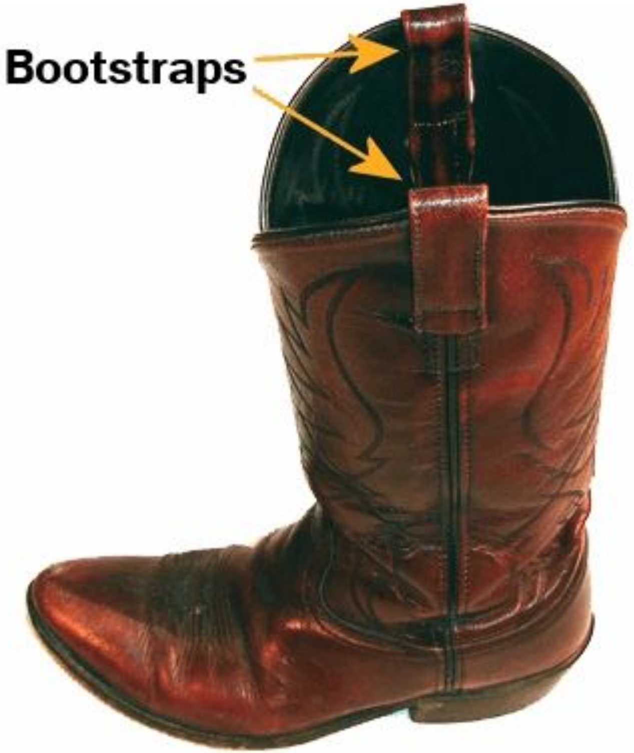
(Demo)

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?

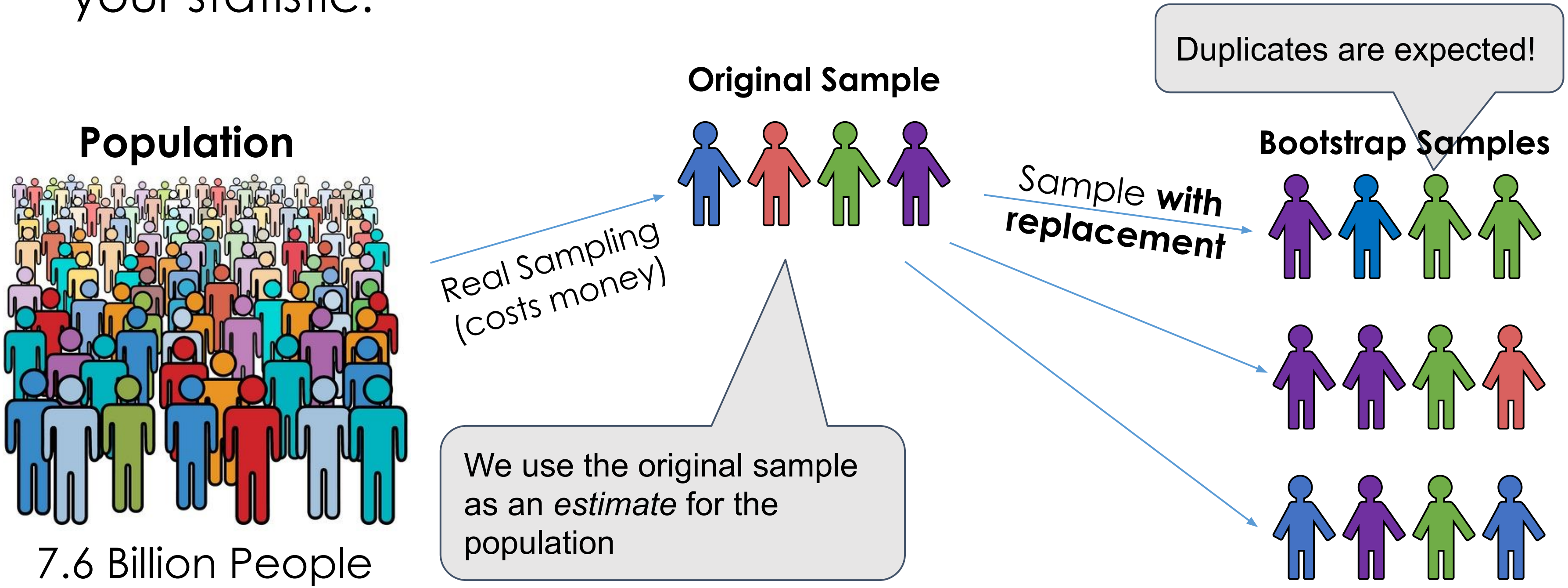
“Pull Yourself Up By Your Bootstraps”

The Bootstrap



Bootstrap the Distribution of a Statistic

Simulation method to estimate the sample distribution of your statistic.



Duplicates are expected!

Original Sample

Population

Bootstrap Samples

Sample **with replacement**

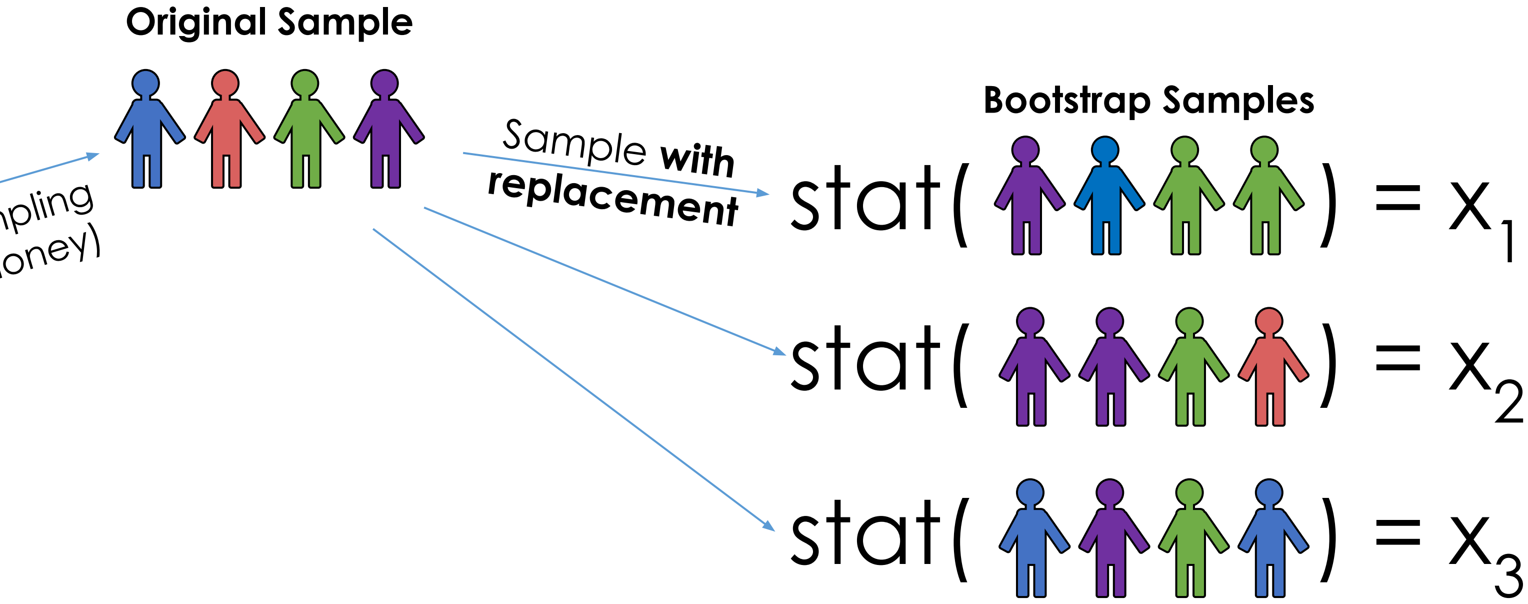
Real Sampling (costs money)

We use the original sample as an *estimate* for the population

7.6 Billion People

Bootstrap the Distribution of a Statistic

Simulation method to estimate the sample distribution of your statistic.



Original Sample

Sample **with replacement**

Bootstrap Samples

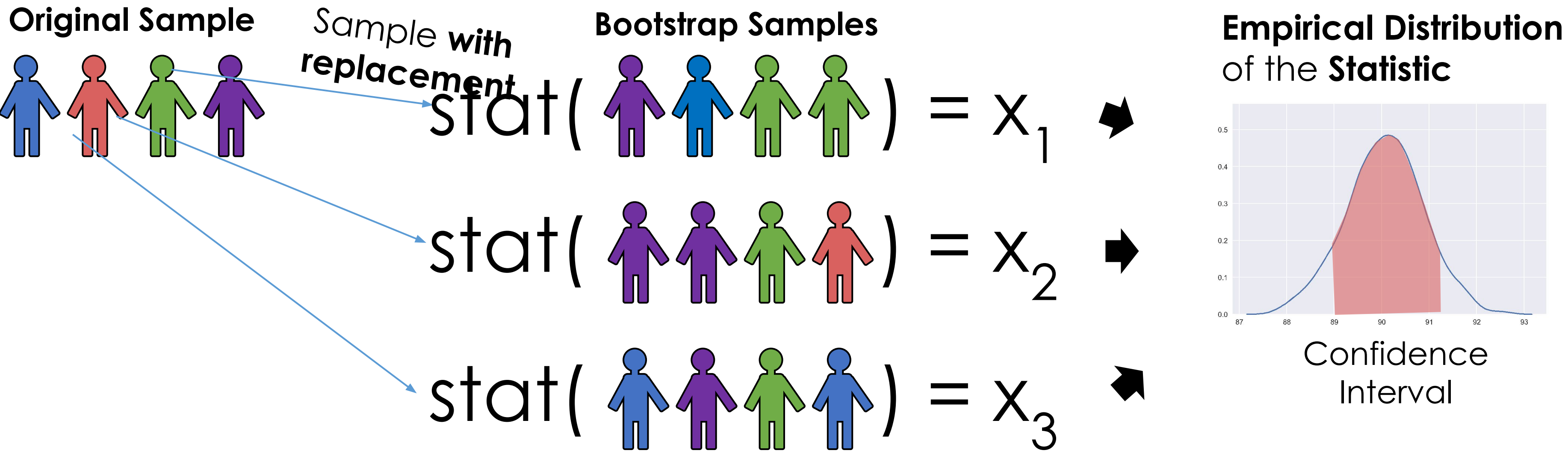
stat( ) = x_1

stat( ) = x_2

stat( ) = x_3

Bootstrap the Distribution of a Statistic

Simulation method to estimate the sample distribution of your statistic.



Original Sample

Sample **with replacement**

Bootstrap Samples

stat( ) = x_1

stat( ) = x_2

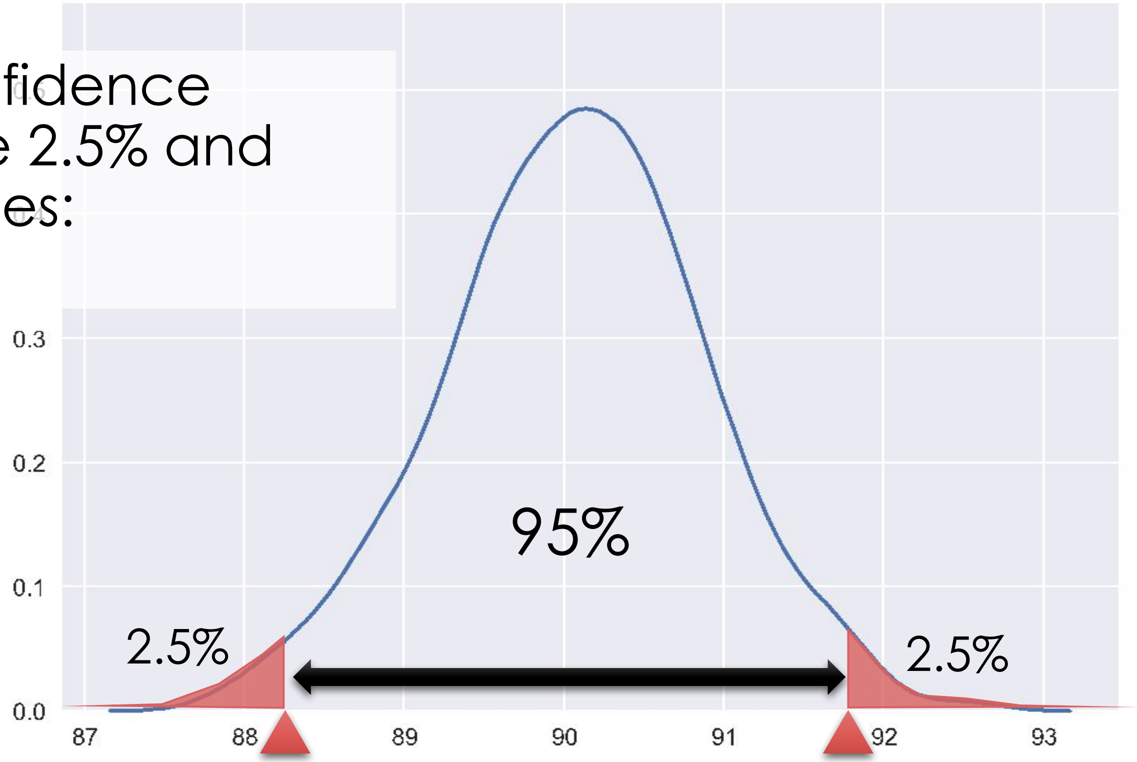
stat( ) = x_3

Empirical Distribution of the **Statistic**

Confidence Interval

Bootstrap Confidence Interval

Construct a 95% confidence interval by taking the 2.5% and (100 - 2.5)% percentiles:



The Bootstrap in words

- From the original sample,
  - draw at random
  - **with replacement**
    - Otherwise you would always get the same sample
  - **Use the same sample size** as the original sample
    - The size of the new sample has to be the same as the original one, so that estimates are comparable
- For each sample, **compute the statistic**
- Compute **empirical distribution of** the **statistics**

(Demo)

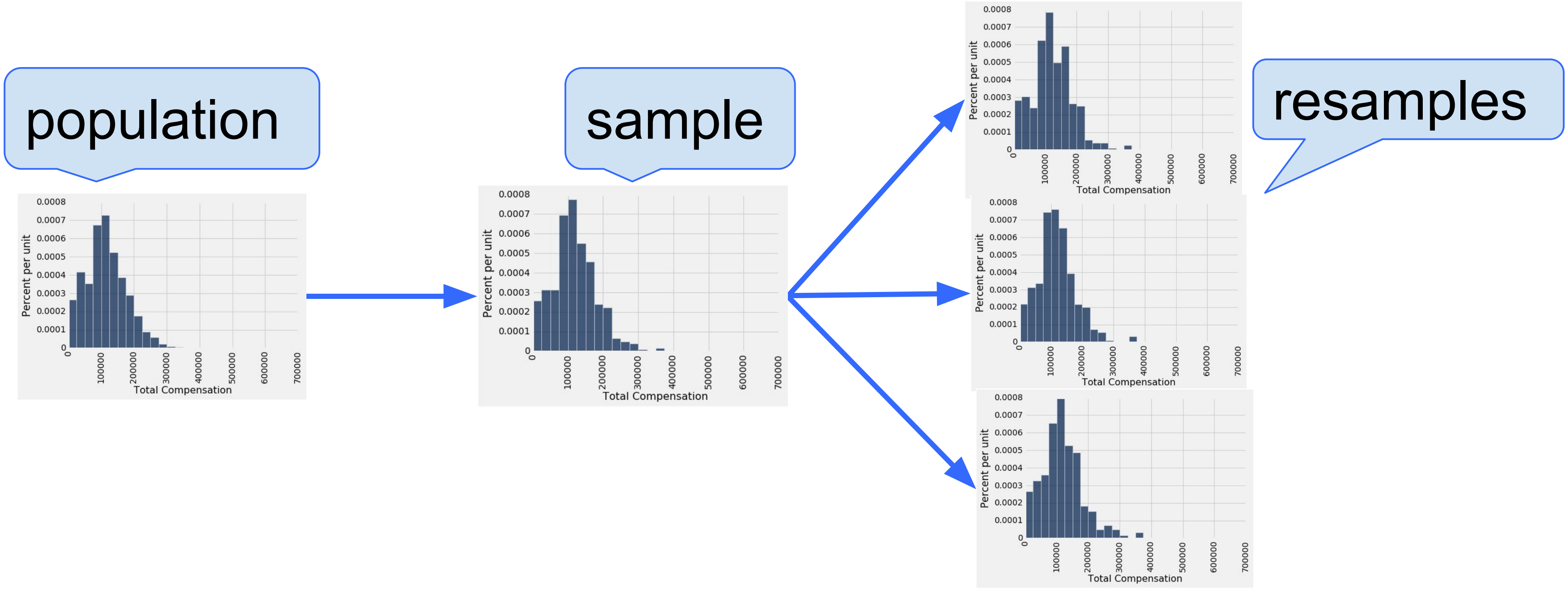
Bootstrap Resampling

- To determine the properties (e.g. variance) of the sampling distribution of an estimator, we'd need to have access to the population.
  - We would have to consider all possible samples, and compute an estimate for each sample.
- But we don't, we only have one random sample from the population.

Idea: Treat our random sample as a "population", and resample from it.

- Intuition: a random sample resembles the population, so a random resample resembles a random sample.

Why the Bootstrap Works



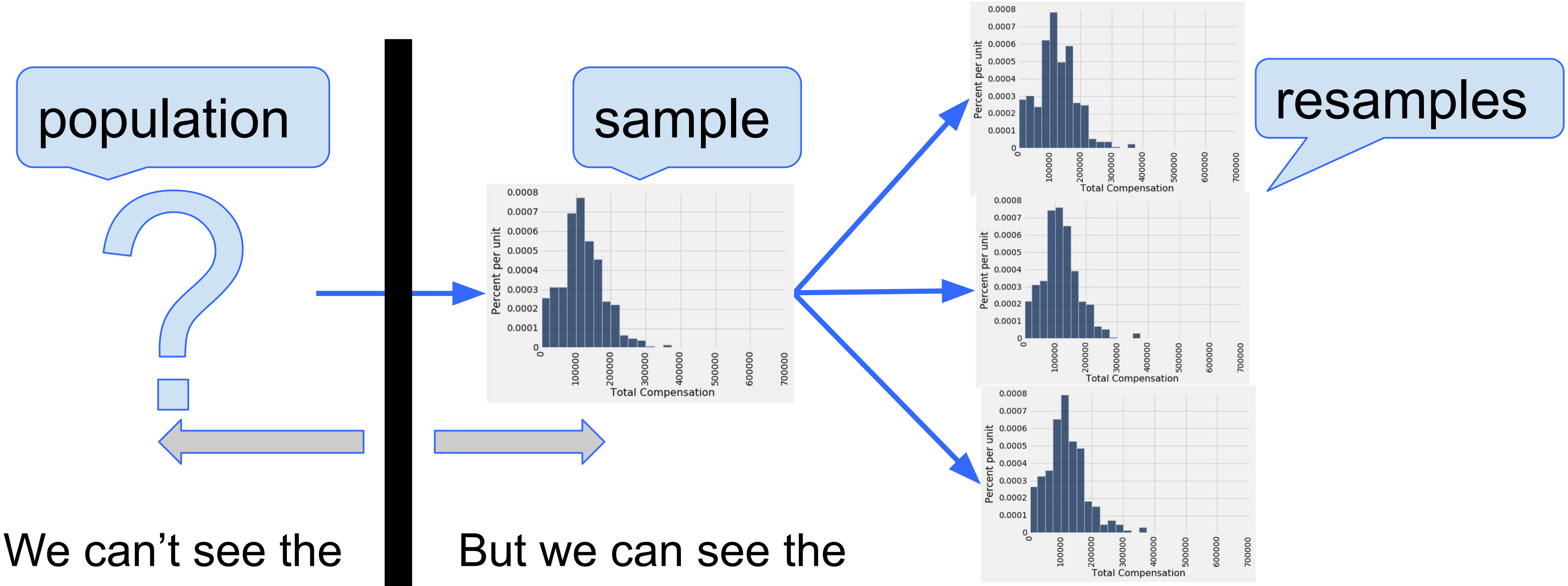
population

sample

resamples

All of these look pretty similar, most likely.

Why We Need the Bootstrap



population

sample

resamples

We can’t see the parameter

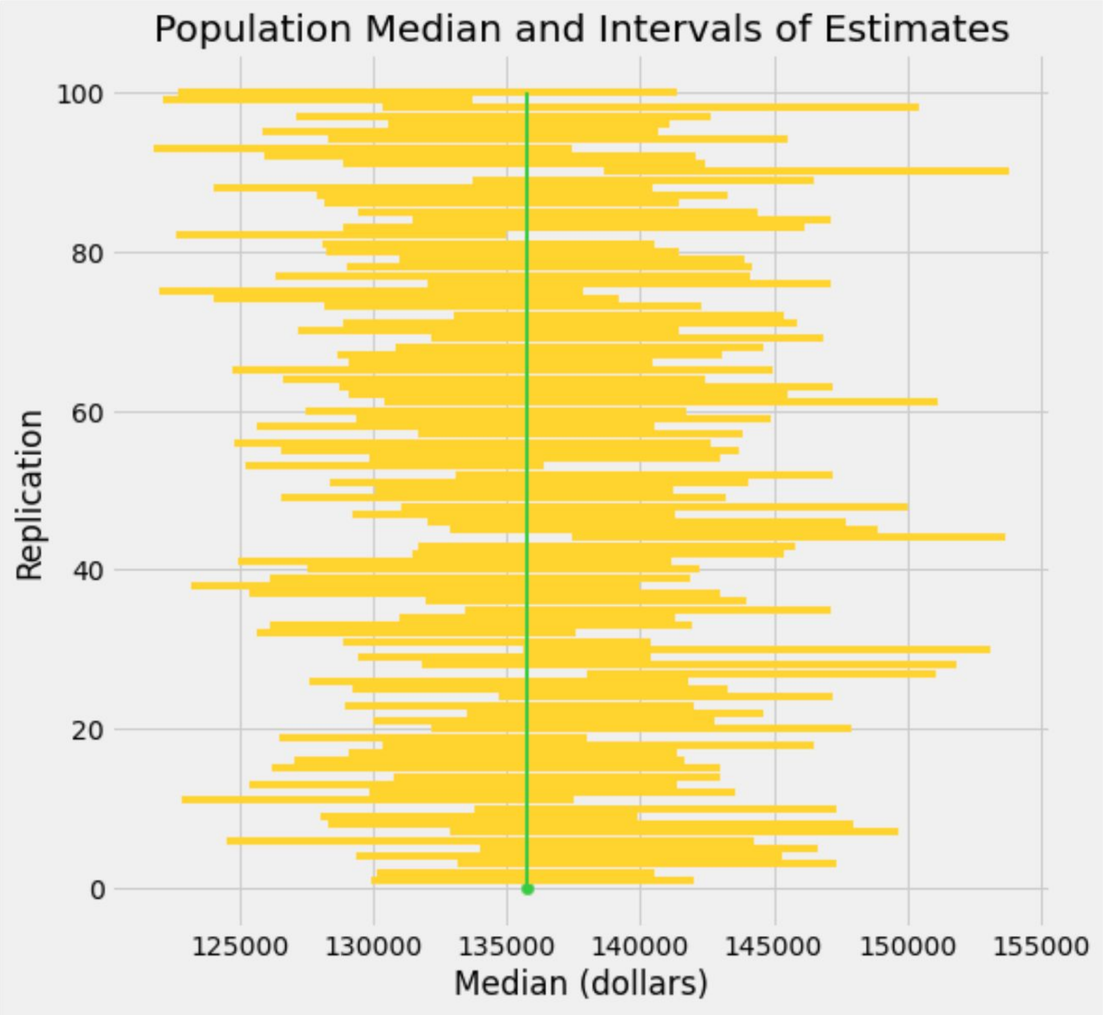
But we can see the sample ...

and generate lots of resamples

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



The Meaning of 95% confidence

The green line is the parameter value.
It is fixed and unknown.
(For this demo we we had access to the population but you won't in practice.)

Each yellow line is a 95% **confidence interval** based on a **fresh sample** from the population

There are **100 intervals**. We expect **roughly 95** to contain the parameter.

Bootstrapping Pseudocode

```
collect random sample of size n (called the bootstrap population)
initiate list of estimates
repeat 10,000 times:
    resample with replacement n times from bootstrap population
    apply estimator f to resample
    store in list
list of estimates is the bootstrapped sampling distribution of f
```

Why **must** we resample **with replacement**?

Bootstrap Discussion

The **bootstrapped sampling distribution of an estimator** does not exactly match the **sampling distribution of that estimator**.

- The center and spread are both wrong (but often close).

The center of the bootstrapped distribution is the estimator applied to our original sample.

- We have no way of recovering the estimator's true expected value.

The variance of the bootstrapped distribution is often close to the true variance of the estimator.

The quality of our bootstrapped distribution depends on the quality of our original sample.

- If our original sample was not representative of the population, bootstrap is next to useless.