Yale NUS College

YSC2239 Lecture 9

Recap

- Steps for statistical tests of hypotheses
  - Null hypothesis and Alternative hypothesis
  - The test statistic and observed value of the test statistic
  - Distribution of test statistic by simulation under null hypothesis
  - Conclusion: reject or not reject (using p-value)
- p-value: the probability of the observed value or even more extreme results if null hypothesis is true. (in short: p-value if the probability of null hypothesis being true.
- Significant level (also called alpha level)

Today’s class

- A/B Testing
- Confidence Intervals

- Reading: Chapter 12 and 13

A/B Testing

Comparing Two Samples

- Compare values of sampled individuals in Group A with values of sampled individuals in Group B.

- Question: Do the two sets of values come from the same underlying distribution?

- Answering this question by performing a statistical test is called **A/B testing**.

(Demo)

The Groups and the Question

- Random sample of mothers of newborns. Compare:
  - (A) Birth weights of babies of mothers who smoked during pregnancy
  - (B) Birth weights of babies of mothers who didn’t smoke
- Question: Could the difference be due to chance alone?

Hypotheses

- Null:
  - In the population, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance.)
- Alternative:
  - In the population, the babies of the mothers who smoked weighed less, on average, than the babies of the non-smokers.

Test Statistic

- Group A: smokers
- Group B: non-smokers

- Statistic: Difference between average weights

  Group A average - Group B average

- Smaller values of this statistic favor the alternative

Simulating Under the Null





...


Non-smoker	Non-smoker	Smoker	Non-smoker	...	Smoker
120 oz	113 oz	128 oz	136 oz		108 oz

Simulating Under the Null





…


Smoker	Non-smoker	Non-smoker	Smoker	…	Non-smoker
120 oz	113 oz	128 oz	136 oz		108 oz

Shuffling Rows

Random Permutation

- `tbl.sample(n)`
  - Table of n rows picked randomly with replacement
- `tbl.sample()`
  - Table with same number of rows as original `tbl`, picked randomly with replacement
- `tbl.sample(n, with_replacement = False)`
  - Table of n rows picked randomly without replacement
- `tbl.sample(with_replacement = False)`
  - All rows of tbl, in random order
  - This is what we'll use for A/B testing

(Demo)

Simulating Under the Null

- If the null is true, all rearrangements of labels are equally likely
- Plan:
  - Shuffle all group labels
  - Assign each shuffled label to a birth weight
  - Find the difference between the averages of the two shuffled groups
  - Repeat

(Demo)

A/B Tests are Hypothesis Tests

- Determine the 2 models (Null Hypothesis and Alternative Hypothesis)
  - Ex. Null hypothesis: In the population, the distributions of the birth weights of the babies in the two groups are the same.
- Determine a test statistic that gives evidence for the alternative model
  - Test statistic is often (but not always) the difference or absolute difference between group means
- Simulate the test statistic under the null hypothesis many times and store those values in an array
  - Simulated by shuffling the labels column of the table
- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- Draw a conclusion comparing the p-value to the p-value cutoff

Percentiles

Computing Percentiles

Sort the numerical set in increasing order. The 80th percentile is first value on the sorted list that is at least as large as 80% of the elements in the set

Percentile

For `s = [1, 7, 3, 9, 5]`, `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 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
```

(Demo)

Estimation

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

(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

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

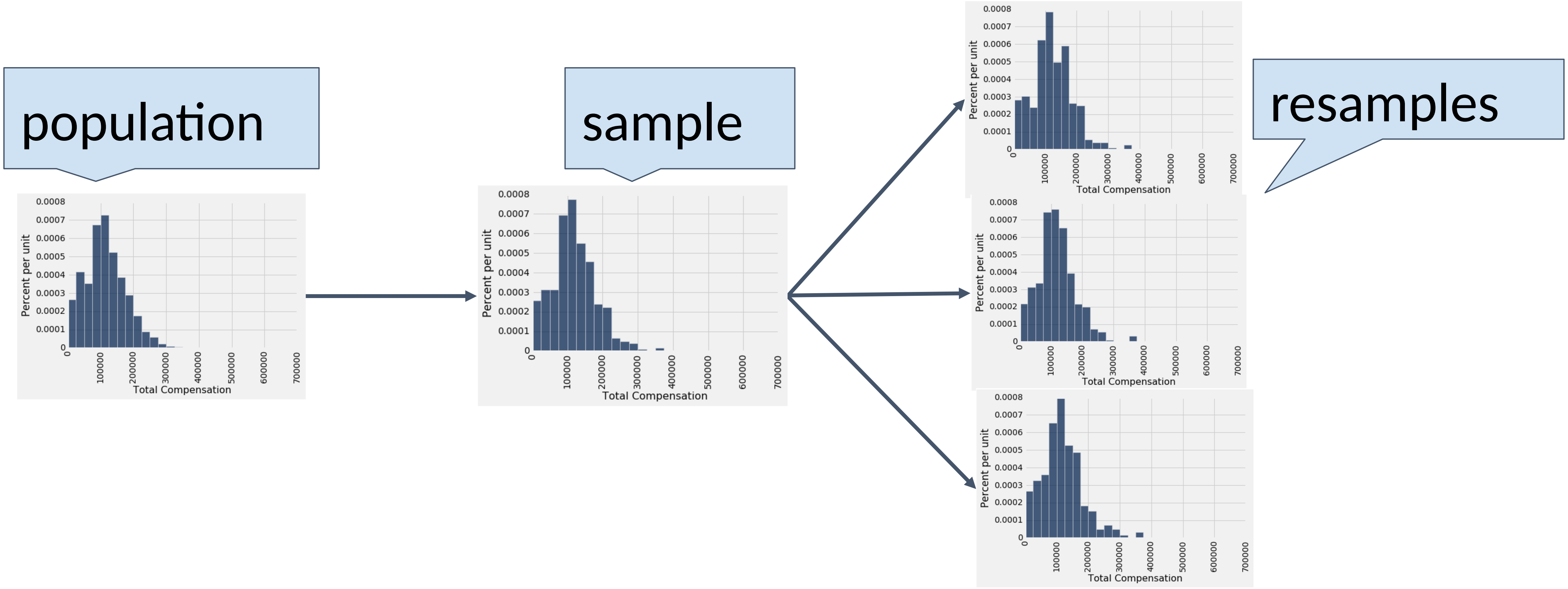
The Bootstrap

- 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

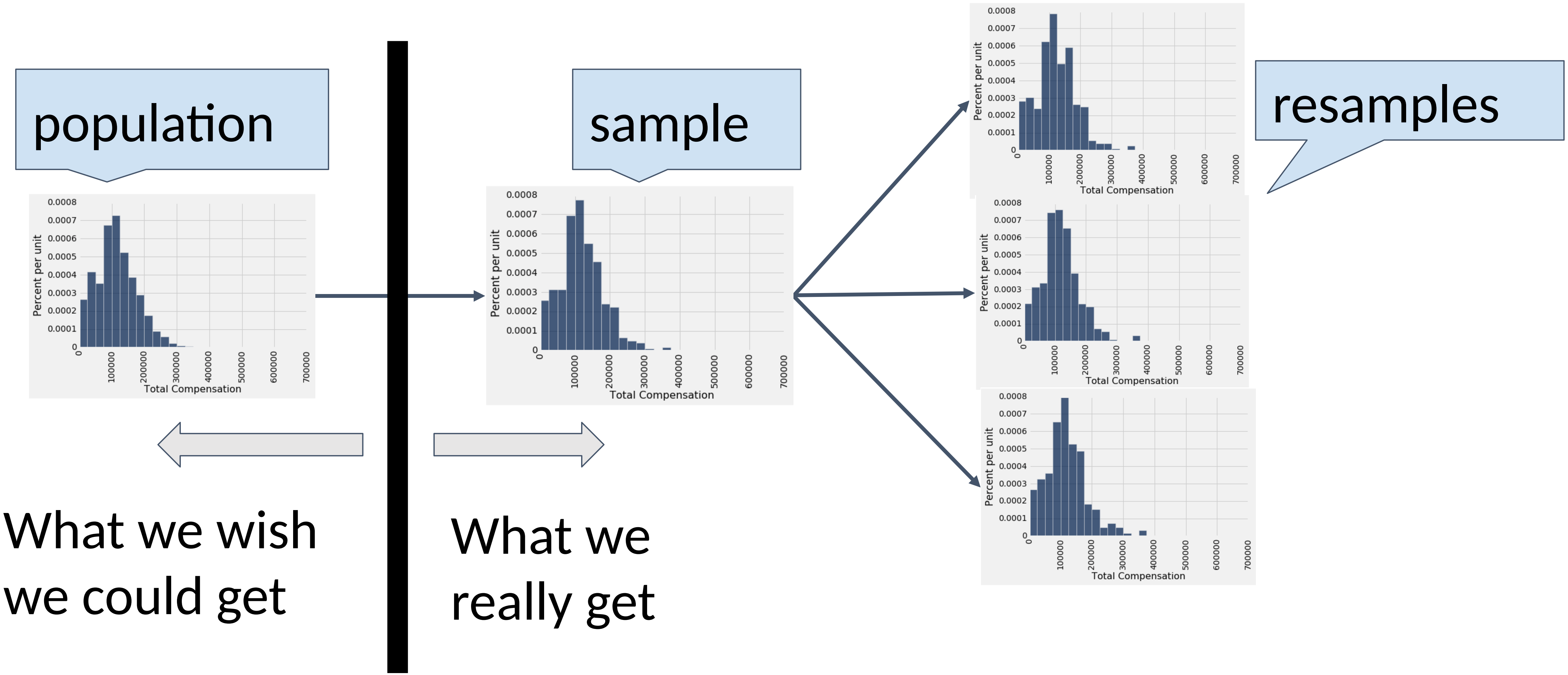
- So we sample at random from the original sample!

Why the Bootstrap Works



All of these look pretty similar, most likely.

Why We Need the Bootstrap



Key to Resampling

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

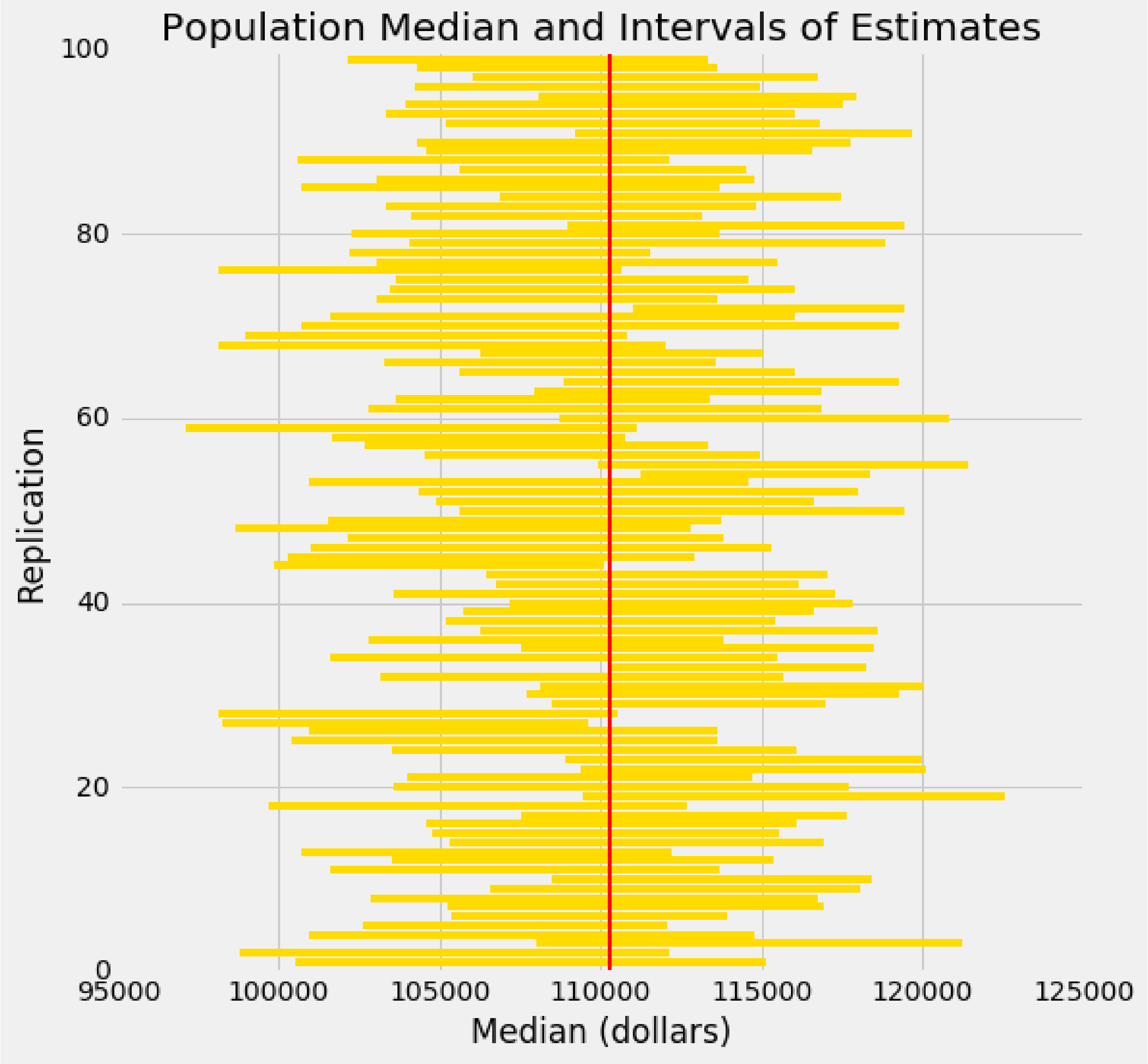
- The size of the new sample has to be the same as the original one, so that the two estimates are comparable

(Demo)

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 generated 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

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.

Is This What a CI Means?

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.** The average age of the mothers in the population is unknown but it's a constant. It's not random. No chances involved.

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)

To-do

- Assignment 4