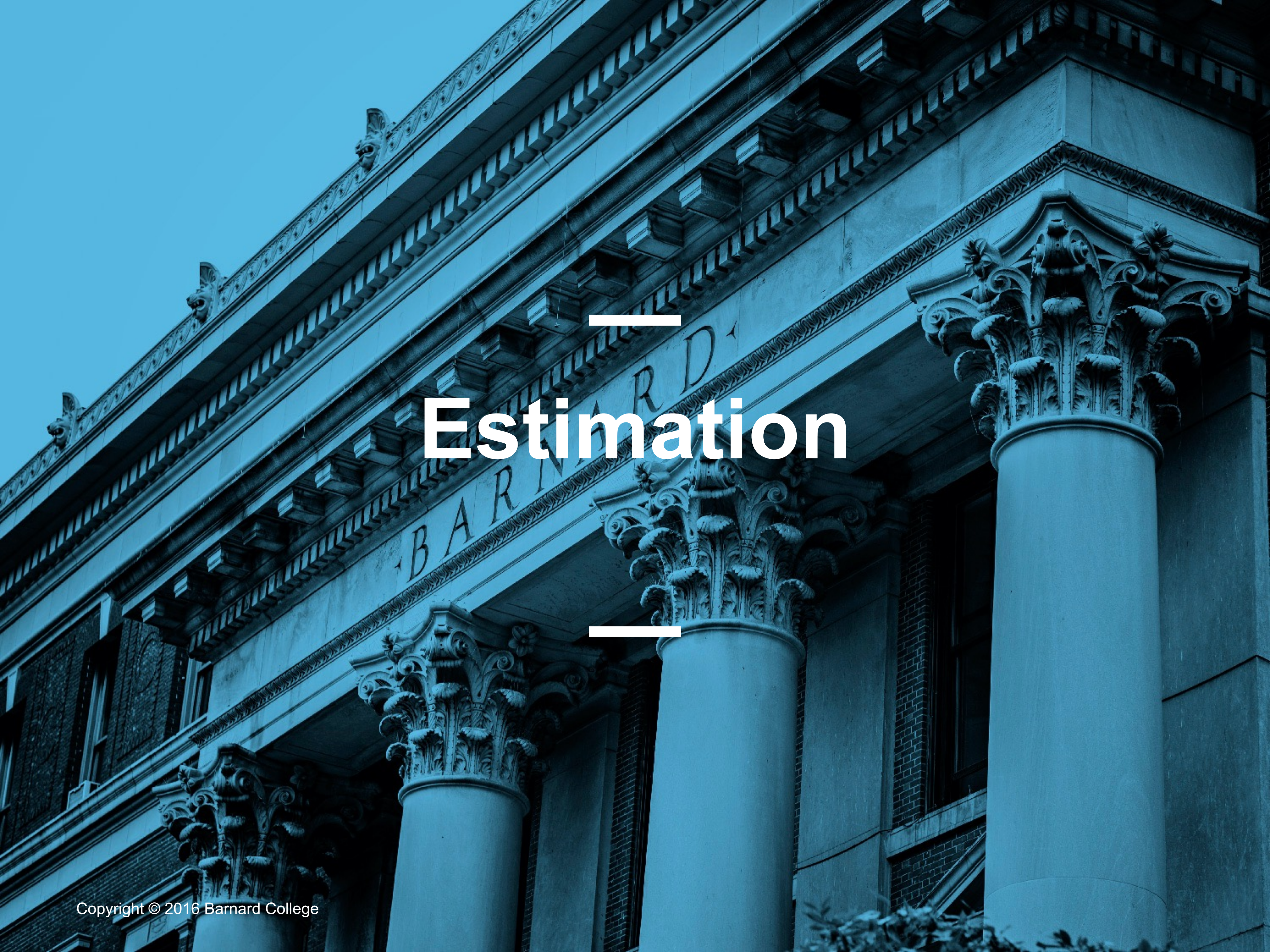
BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 18 – Estimation Variability Bootstrap Confidence Interval

Announcements - update

- HW06 - [Testing Hypotheses]
  - Due Thursday 03/31
- Project 2
  - Released due Friday 04/15
  - Currently fixing autograder bug
- Lab07 – Normal Distribution and Variance of Sample Means
  - Due Monday 04/04
- HW 07
  - Due Thursday 04/07



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

Estimation Variability

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 estimated again?

Quantifying Uncertainty

- The estimate is usually not exactly right.
- Variability of the estimate tells us something about how accurate the estimate is:

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

Where to Get Another Sample?

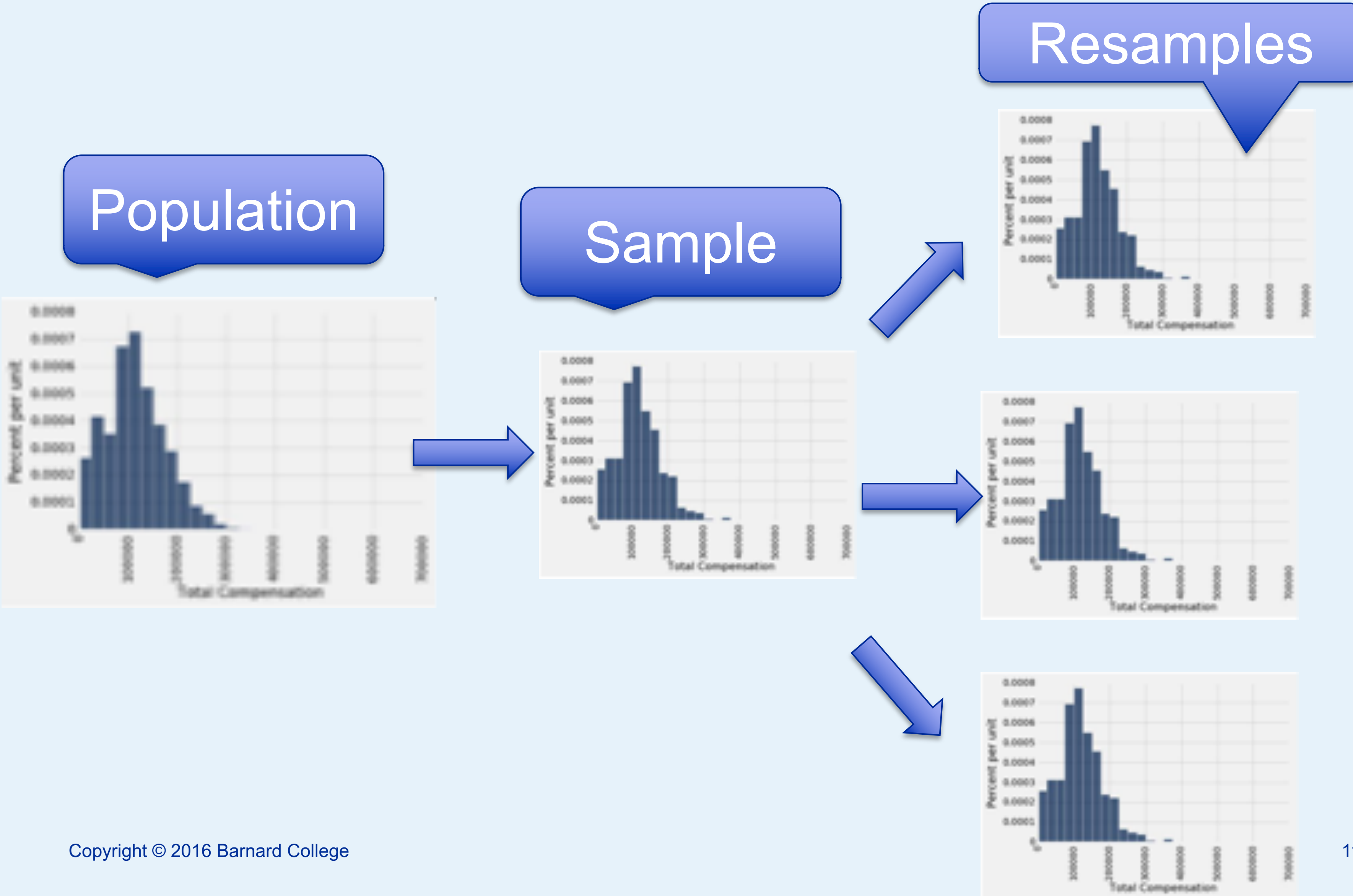
- We want to understand errors of our estimate
- Given the **population**, we could simulate
  - ...but we only have the **sample**!
- To get many values of the estimate, we needed many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?

The Bootstrap

The Bootstrap

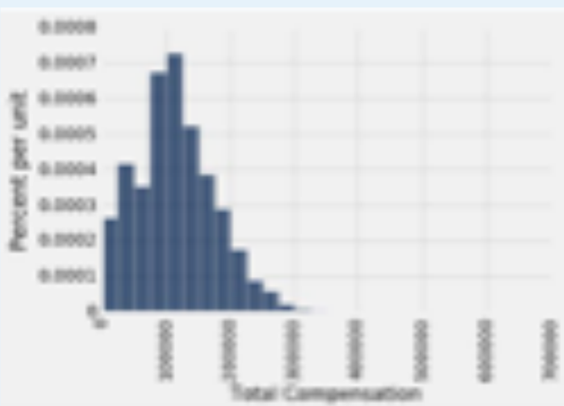
- A technique for simulating repeated random sampling
- All that we have is the original sample
  - ... which is large and random
  - Therefore, it probably resembles the population
- So we sample at random from the original sample!

How the Bootstrap works



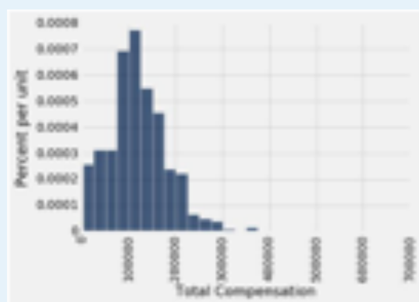
Why the Bootstrap works

Population



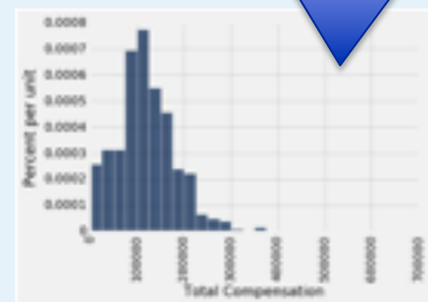
What we wish we could get

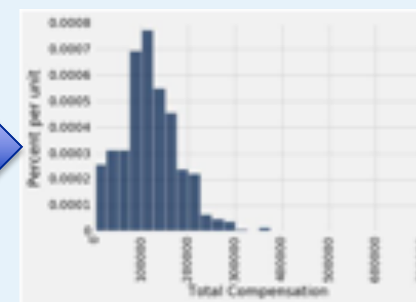
Sample

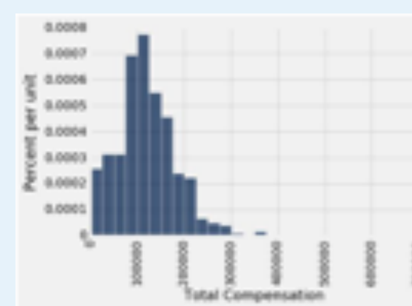


What we actually can get

Resamples







Real World vs Bootstrap World

Real World

- True probability distribution (population)
  - Random sample 1
    - Estimate 1
  - Random sample 2
    - Estimate 2
  - …
  - Random sample 1000
    - Estimate 1000

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

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

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

Using a CI for Testing

- Null hypothesis: **Population average = *x***
- Alternative hypothesis: **Population average ≠ *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

Data Science in this course

- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)

- Inference
  - Make reliable conclusions about the world
  - Statistics is useful

- **Prediction**
  - **Informed guesses about unseen data**