Data 8, Project 2 Lab

The Bootstrap and Confidence Intervals

Hubert Luo

Spring 2020

3 April 2020

UNIVERSITY OF CALIFORNIA

Agenda

1. The Bootstrap
2. Confidence Intervals

Parameters and Statistics

- **Population Parameter:** A metric about a population
  - Fixed and not random!
  - Example: Average GPA of all Cal students
- **Sample Statistic:** A metric about a sample of that population
  - Different for each sample of the population!
  - Example: Average GPA of Cal students who are taking Data 8

Why Bootstrap?

- We need a sample to estimate a population parameter!
- In order to evaluate the **variability of the statistic**, we need multiple samples
- If the original sample is large and selected at random, it is close enough to the population that we can resample from our original sample instead of from the population
  - This is extremely helpful in real life: it saves us time and money!

The Bootstrap

- If the original sample is large and selected at random, it likely resembles the population
- Instead of getting entirely new samples from the population, we resample from the original sample
- Resample same number of individuals **with replacement**
  - If we sample without replacement, we will always get the sample original sample back!

Confidence Interval

- Interval of estimates of a population parameter
- A 95% confidence interval means if we create 100 confidence intervals from different samples, we expect 95 of them to contain the true population parameter
- It does NOT mean that there is a 95% probability the true population parameter is in a confidence interval!!
  - The population parameter is a constant – it's either in an interval, or it isn't so there's nothing we can conclude about probability

Confidence Interval Facts

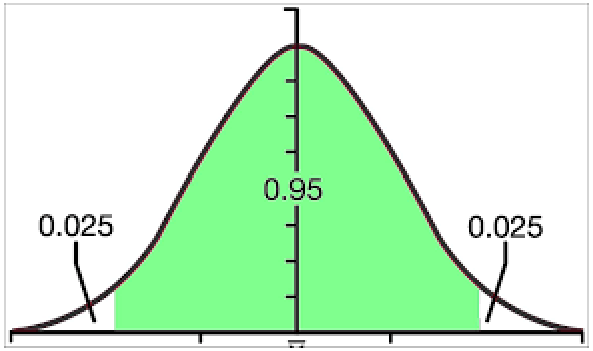
- For a 95% confidence interval, it is true that:
  1. Out of 100 confidence intervals, we expect 95 of them to contain the true population parameter
  2. There is a 95% probability the confidence interval contains the population parameter
     1. Note that the confidence interval is random here so we can make this statement
- It is **not** true that there is a 95% probability a population parameter is in a confidence interval.
  - This is because the population parameter is fixed so we can't so anything probabilistic about it.

Confidence Interval Facts (Cont'd)

- Given a fixed confidence interval, statement #2 from the previous slide no longer applies
  - The confidence interval is fixed so we can't make any probabilistic conclusions about it
- Example: It is **not true** that there is a 95% probability the confidence interval [2, 5] contains the population parameter
  - The population parameter fixed AND
  - The confidence interval [0.439, 0.5] is fixed
  - Therefore, there is **no probability involved here**

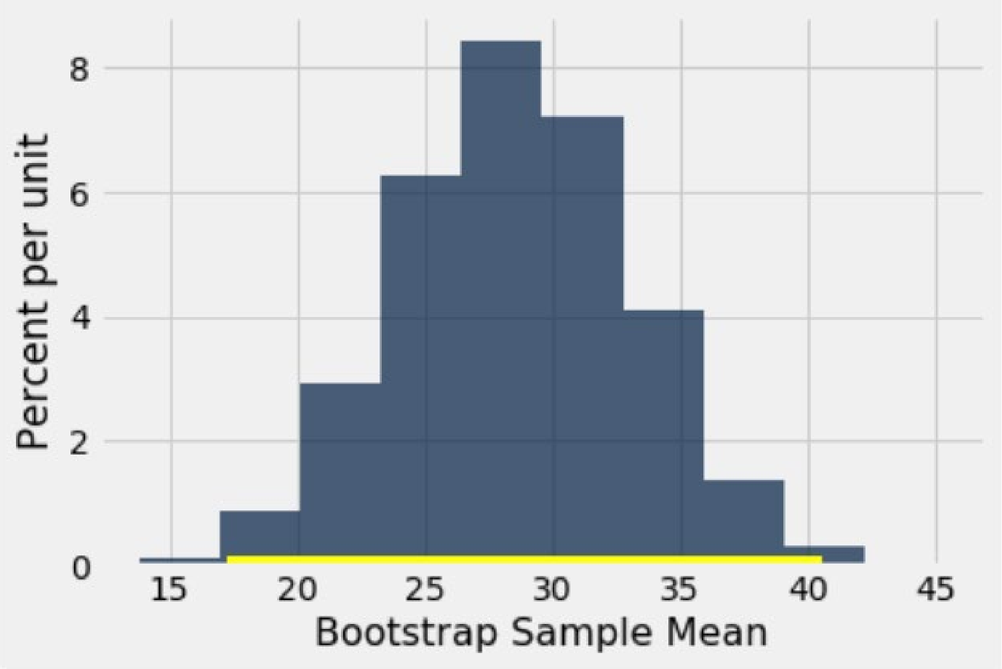
Creating Confidence Intervals

- We first need to bootstrap to get an array of sample statistics
- To get the values of a confidence interval, we use the percentile of those sample statistics!
- Example: A 95% confidence interval will be created from the middle 95% of sample statistics
  - Symmetrical so $\frac{5\%}{2} = 2.5\%$ of the sample statistics fall outside the confidence interval on both sides
  - Lower bound: 2.5th percentile
  - Upper bound: 97.5th percentile



Berkeley
UNIVERSITY OF CALIFORNIA

Visualizing Confidence Intervals



Visualizing Confidence Intervals

https://rpsychologist.com/d3/CI/

Announcements

- HW8 is due Thursday 4/2
- Project 2:
  - Checkpoint 1 is due this Friday 4/3
  - Checkpoint 2 is due next Friday 4/10
  - Entire project is due a week after that, on 4/17.