YData: Introduction to Data Science

Lecture 25: The bootstrap and confidence intervals continued

Overview

Interpreting confidence intervals

Another example of the bootstrap

Connections between confidence intervals and hypothesis tests

If there is time: measures of central tendency

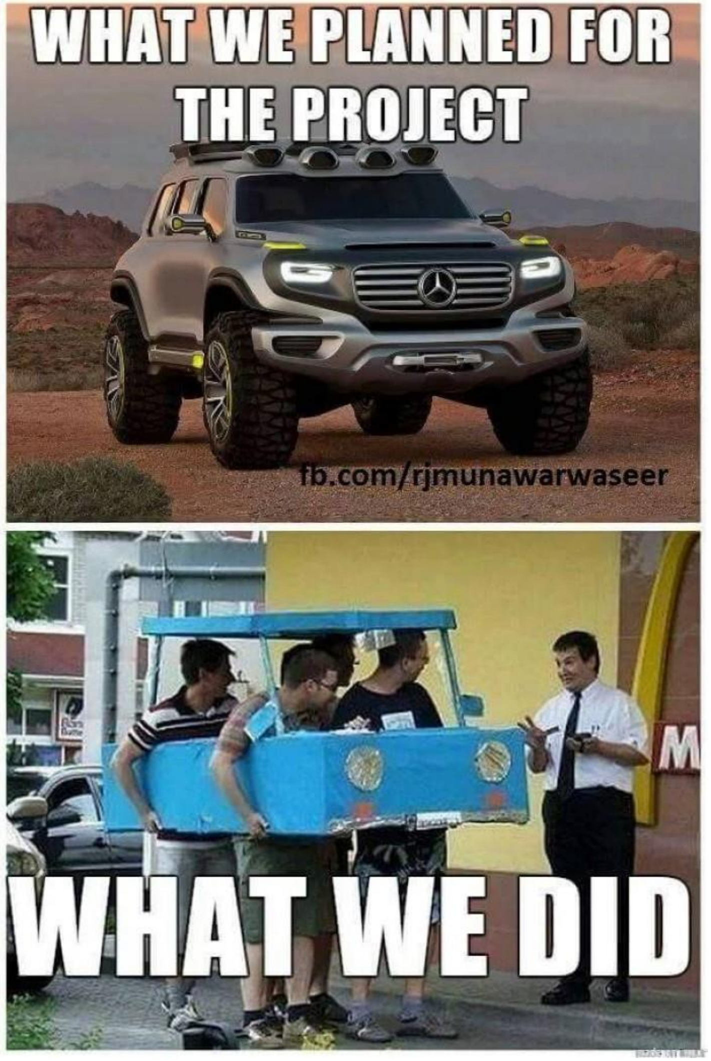
Announcements

Project 2 has been posted!

- It is due on Friday, April 8th

Homework 7 is optional and will not be turned in

- This is to give you more time to work on your project
- Homework 8 (next week's homework) is not optional, so try to finish project 2 early!



Estimation

Inference: Estimation

We want to know the value of a population parameter

We only have a random sample from the population

- Use a statistic as a point estimate of the parameter
    - Best guess at the parameter value

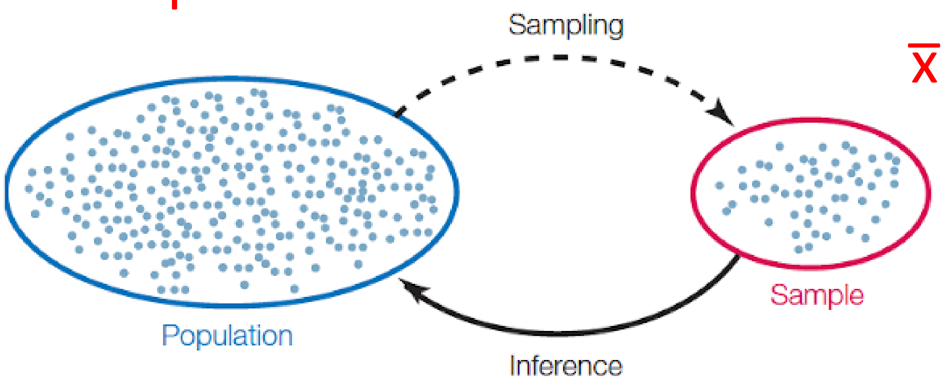
Question: is our statistic a good estimate of the population parameter?

parameter

μ

statistic

$\overline{x}$



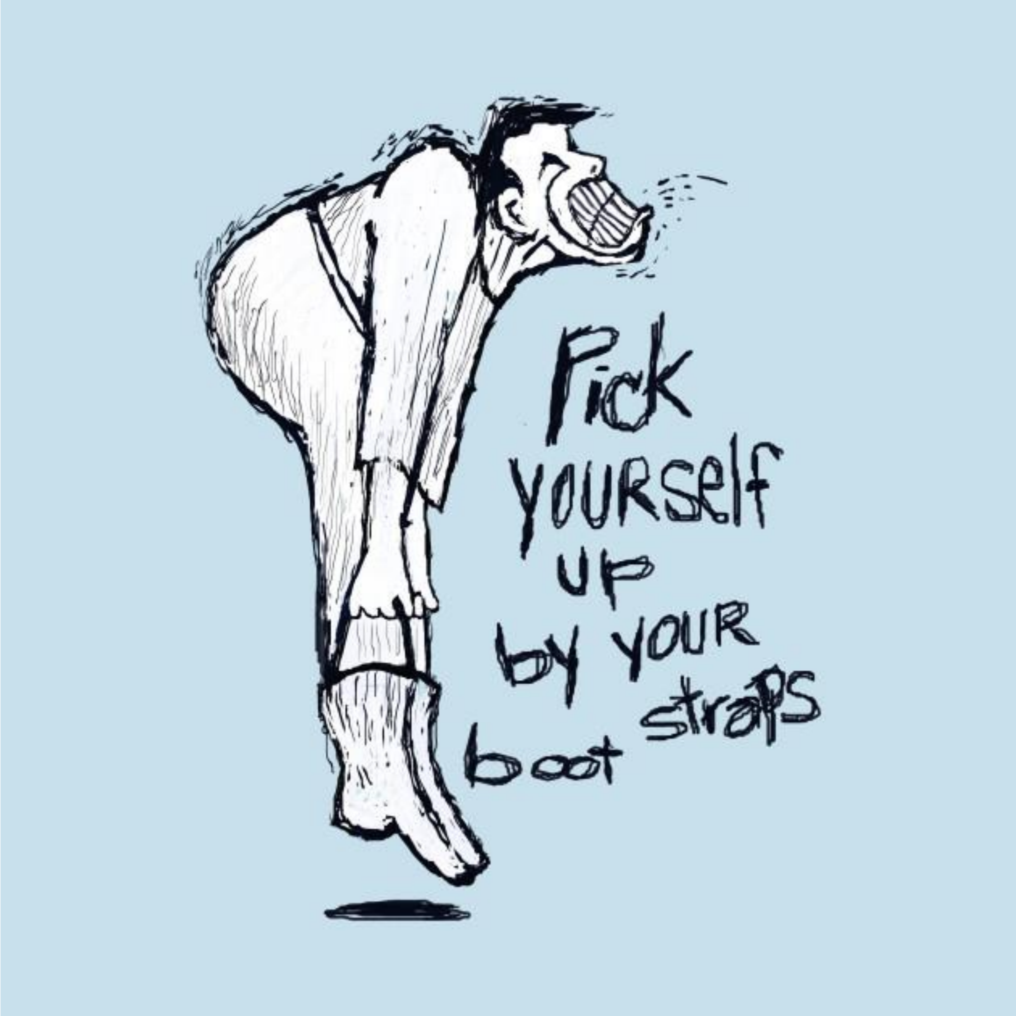
Where to get another sample?

If we had many statistics (from many samples), we could get a sense of whether any given statistic is a good estimate of the parameter

- i.e., if the statistics vary a lot from sample to sample, then any one statistic would be a poor estimate

Too costly to sample repeatedly from the population

We just have to pick ourselves up by the bootstraps!



Plug-in principle

Suppose we get a sample from a population of size *n*

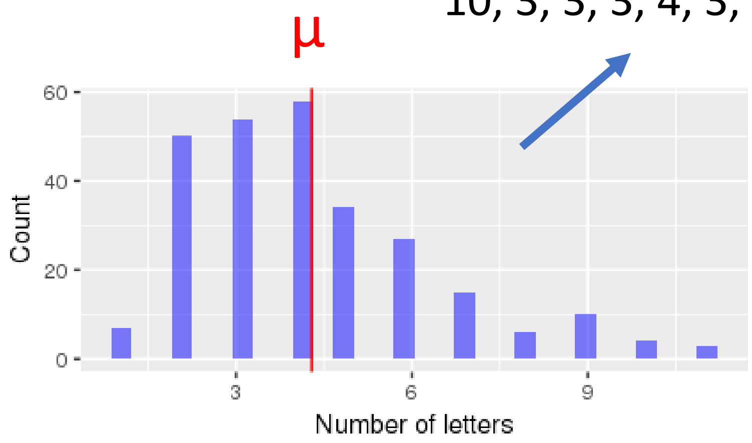
We pretend that *the sample is the population* (plug-in principle)

1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

2. We repeat this process 1000's of times and get a ***bootstrap distribution***

3. We can calculate percentiles of the bootstrap distribution to create confidence intervals
    - i.e., confidence intervals are ranges of values that usually contain the population parameter value.

Bootstrap distribution illustration



The sample (n = 10)
10, 3, 3, 3, 4, 3, 2, 6, 4, 5

3, 3, 3, 5, 3,
4, 5, 2, 2, 10

$\overline{x}^* = 4$

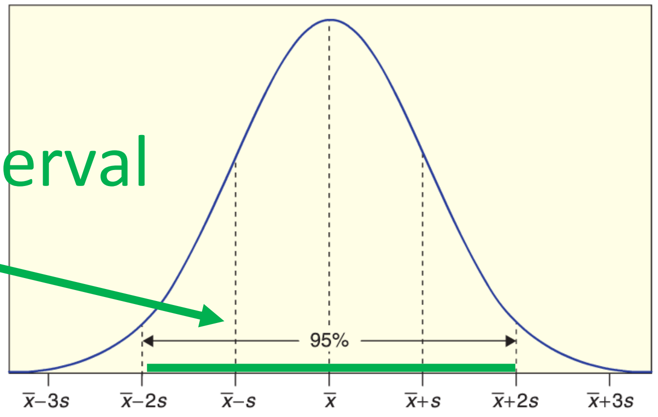
3, 3, 2, 3, 6,
4, 6, 5, 3, 6

$\overline{x}^* = 4.1$

5, 3, 2, 3, 3,
3, 10, 3, 4, 3

$\overline{x}^* = 3.9$

Confidence interval



Bootstrap distribution

Let's explore this in Jupyter!

Confidence intervals

Confidence Intervals

A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

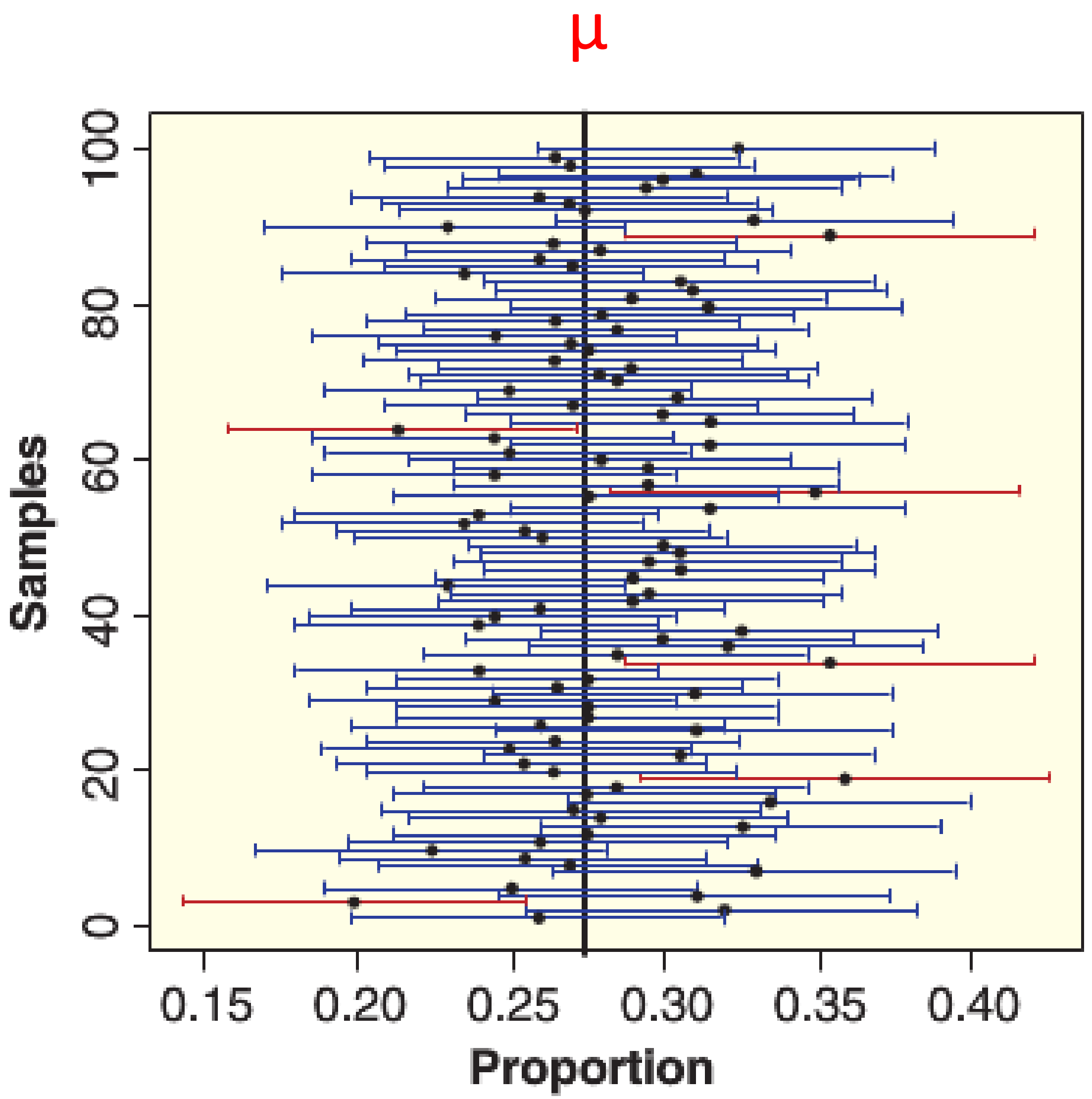
- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

The **confidence level** is the percent of all intervals that contain the parameter

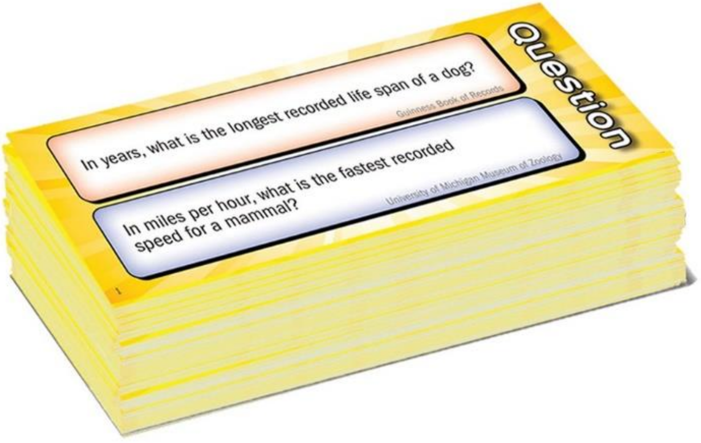
Confidence Intervals

For a **confidence level** of 95%...

95% of the **confidence intervals** will have the parameter in them



Wits and Wagers:
90% confidence interval estimator


Question
In years, what is the longest recorded life span of a dog?
Guinness Book of Records
In miles per hour, what is the fastest recorded speed for a mammal?
University of Michigan Museum of Zoology

I will ask 10 questions that have numeric answers

Please come up with a range of values that contains the true value in it for 9 out of the 10 questions

- i.e., be a 90% confidence interval estimator

Tradeoff between interval size and confidence level

There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Let's explore this in Jupyter!

Interpreting confidence intervals

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.

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

Confidence Intervals for Testing

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

Let's explore this in Jupyter!

Measures of central tendency

Questions

How can we quantify natural concepts like "center" and "variability" of data?

Why do many of the empirical distributions that we generate come out bell shaped?

How is sample size related to the accuracy of an estimate?

The average

The average (or mean)

Data: 2, 3, 3, 9 Average = (2+3+3+9)/4 = 4.25

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Does not need to be a value in the collection

Does not need to be an integer even if the data are integers

Somewhere between min and max, but not necessarily halfway in between

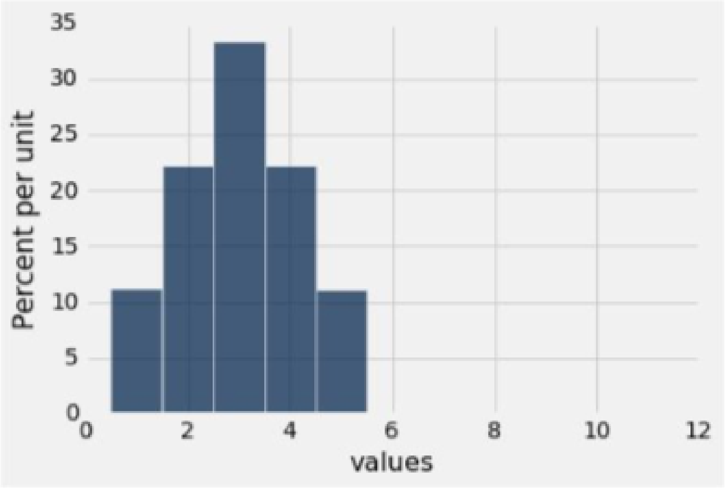
Same units as the data

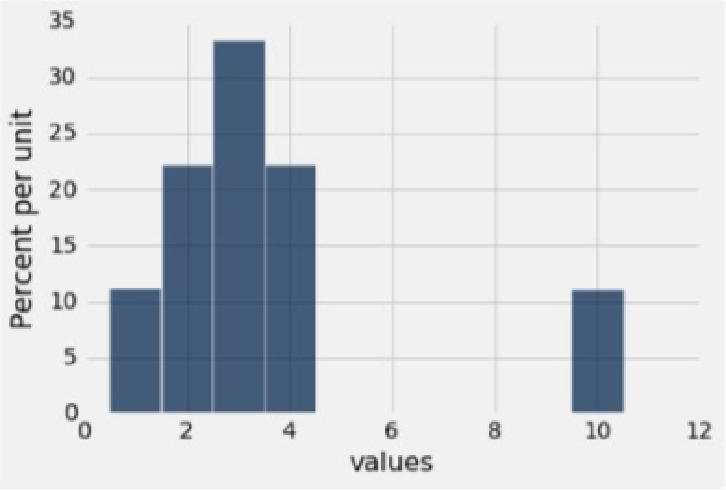
Discussion question

1. Are the medians of these two distributions the same or different?

2. Are the means of these two distributions the same or different?

   - If you say "different", then say which one is bigger?





Comparing the mean and the median

Mean: Balance point of the histogram

Median: Half-way point of data; half the area of histogram is on either side of median

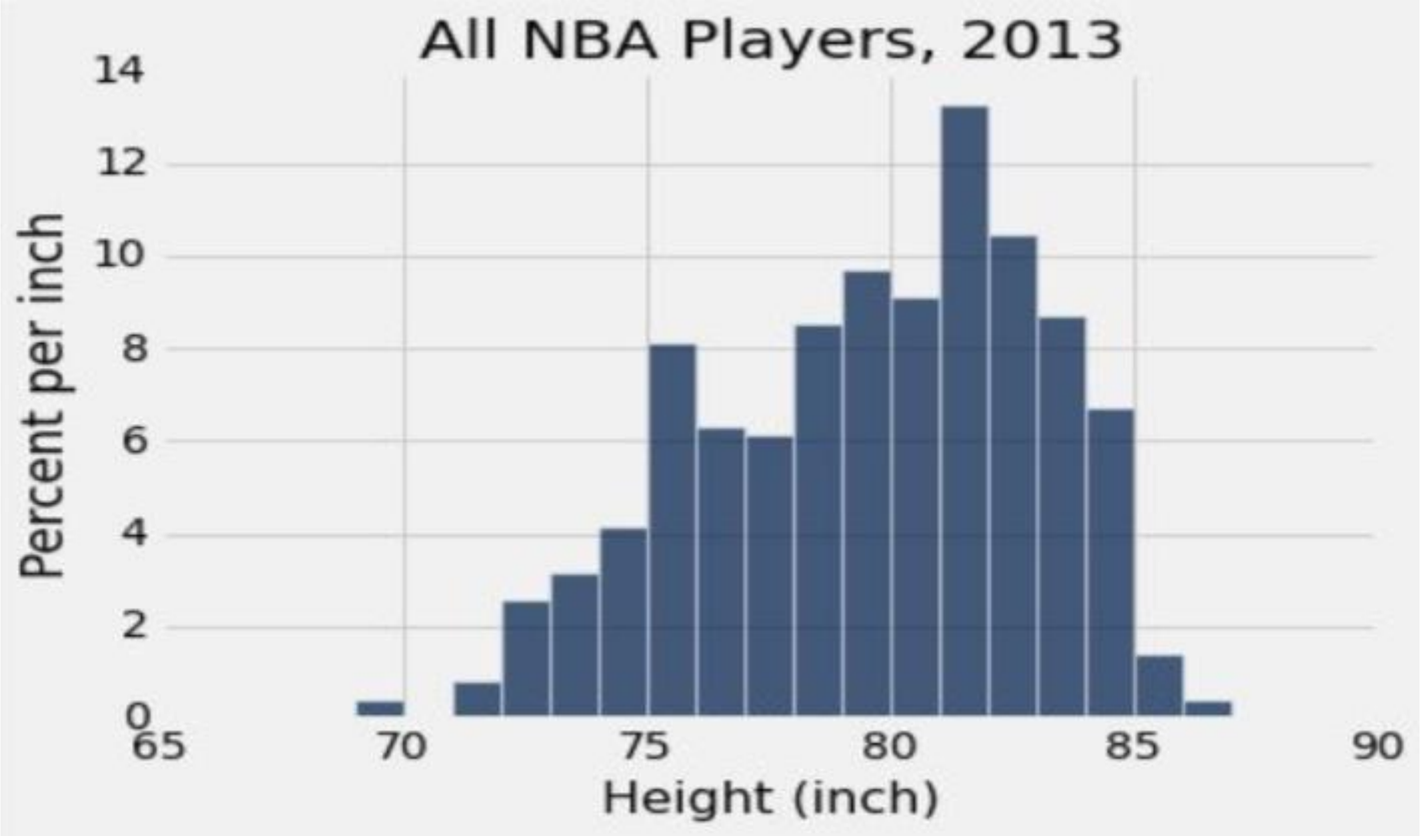
If the distribution is symmetric about a value, then that value is both the average and the median.

If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.

Discussion question

Which is bigger?

- A. The mean
- B. The median



Let's explore this in Jupyter!