

# YData: An Introduction to Data Science

## Lecture 23: Confidence Intervals

Elena Khusainova & John Lafferty  
Statistics & Data Science, Yale University  
Spring 2021

Credit: [data8.org](https://data8.org)

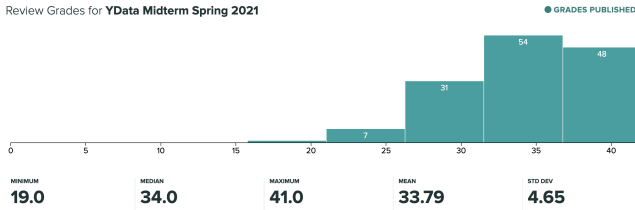


# Announcements

- Midterm scores posted
- Assignment 07 due on Thursday
- Normal OH resume this week

# Midterm

Review Grades for YData Midterm Spring 2021



- You did really well!
- We will post sample solutions later
- Questions about grading to instructors via email, please
- Scores will be curved upwards

# High level view

Intro, Cause and Effect	Lectures 1–2
Python, Tables, Visualization	Lectures 3–13
Probability and Distributions	Lectures 14–17
Hypothesis Testing and Causality	Lectures 18–20
Midterm exam	—
Confidence and the Normal Distribution	Lectures 23–28
Regression and Classification	Lectures 29–37
Final exam	—

# Percentiles

# Computing Percentiles

The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For  $s = [1, 7, 3, 9, 5]$ , `percentile(80, s)` is 7

The 80th percentile is ordered element 4:

$$\begin{array}{ccc} (80/100) & * & 5 \\ \text{Percentile} & & \text{Size of set} \end{array}$$

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

- What is 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

(DEMO)

# Variability of the Estimate

- One sample  $\longrightarrow$  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:  
$$\text{estimate} = \text{parameter} + \text{error}$$

# Where to Get Another Sample?

- One sample  $\longrightarrow$  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?

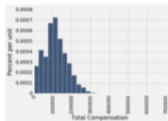
# The Bootstrap

# The Bootstrap

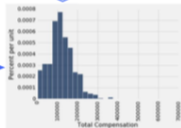
- A technique for simulating repeated random sampling
- All that we have is the original sample
  - ... which is large and random
  - Therefore, it probably resembles the population
- So we sample at random from the original sample!

# Why the Bootstrap Works

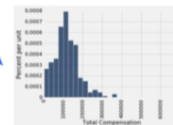
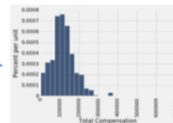
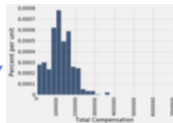
population



sample



resamples



All of these look pretty similar, most likely.

# 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)



## REGULAR UNCERTAINTY

OUR STUDY FOUND  
THE DRUG WAS 74%  
EFFECTIVE, WITH A  
CONFIDENCE INTERVAL  
FROM 63% TO 81%.



## EPISTEMIC UNCERTAINTY

OUR STUDY FOUND THE  
DRUG TO BE 74% EFFECTIVE.  
HOWEVER, THERE IS A 1 IN 4  
CHANCE THAT OUR STUDY  
WAS MODIFIED BY GEORGE  
THE DATA TAMPERER, WHOSE  
WHIMS ARE UNPREDICTABLE.



[xkcd.com/2440](http://xkcd.com/2440)