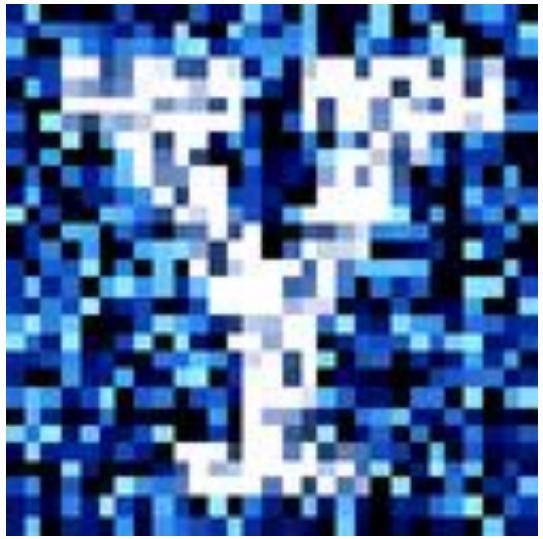
YData: Introduction to Data Science



Lecture 21: The bootstrap

Overview

Hypothesis test continued

- Causality continued

Percentiles

Estimation

If there is time: The Bootstrap

Announcements

Please fill out the mid-semester feedback on Canvas

Note: all classes have been recorded, so you can review the recordings in the media library on Canvas

Causality

Causality

Recall from class 2:

- **An association** is the presence of a reliable relationship between the treatments an outcome
- **A causal relationship** is when changing the value of a treatment variable influences the value outcome variable

Is there an association and/or causal relationship for:

- The example of smoking mothers and baby weights?
- Deflategate?

What are some confounding variables?

Randomized Controlled Experiment

Sample A: control group

Sample B: treatment group

The treatment and control groups are selected at random; this allows causal conclusions!

Any difference in outcomes between the two groups could be due to:

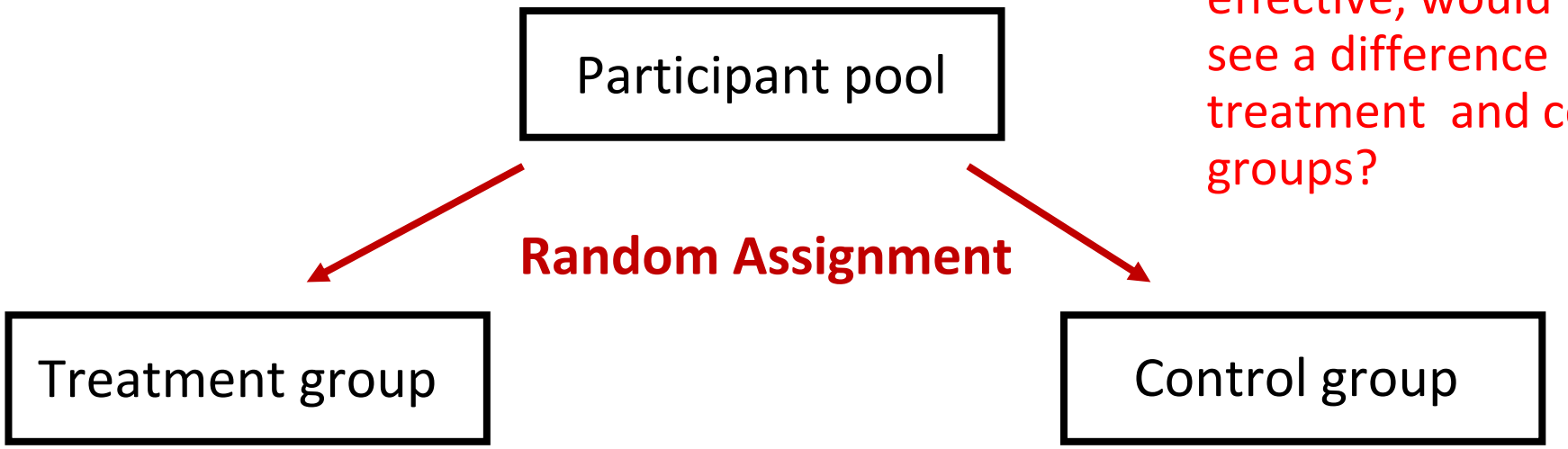
- Chance
- The treatment

Randomized Controlled Experiment

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get chocolate
- Half in a *control group* where they get a fake chocolate (placebo)
- See if there is more improvement in the treatment group compared to the control group

Q: If the treatment was not effective, would we expect to see a difference between the treatment and control groups?



Case study

RCT to study Botulinum Toxin A (BTA) as a treatment to relieve chronic back pain

- 15 patients in the treatment group (received BTA)
- 16 in the control group (normal saline)

Trials were run double-blind: neither doctors nor patients knew which group they were in.

Results

- 2 patients in the control group had relief from pain (outcome=1)
- 9 patients in the treatment group had relief.

Can this difference be just due to chance?

The hypotheses

Null:

- BTA does not lead to an increase in pain relief
  - i.e., if many people were to get BTA and saline, the proportion of people who experienced pain relief would be the same in both groups.

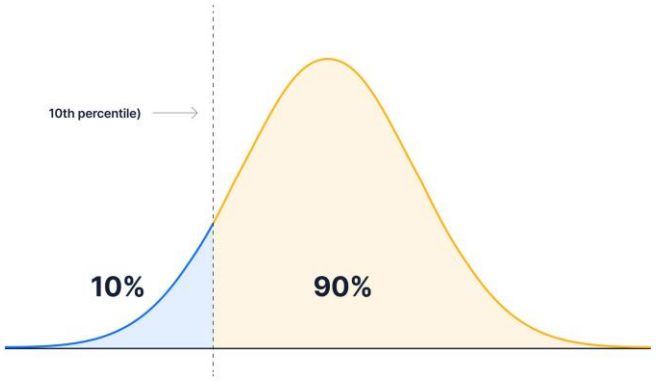
Alternative:

- BTA leads to an increase in pain relief
  - i.e., if many people were to get BTA and saline, the proportion of people who experienced pain relief would be higher for those who received BTA

Let's explore this in Jupyter!

Percentiles

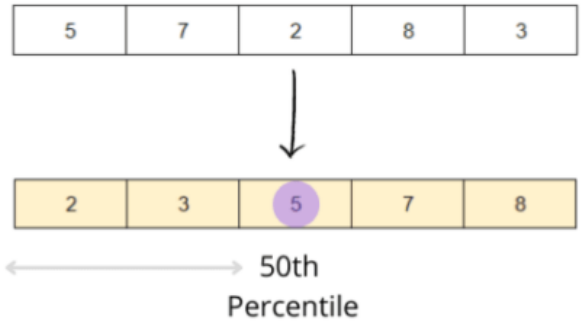
The percentile function



The p^{th} percentile is the smallest value in a set that is as large or larger than p% of the elements in the set

Function in the datascience module: percentile(p, values)

- p: a number between 0 and 100
- values: an array or list of values



For a percentile that does not exactly correspond to an element, take the next greater element instead

- sidenote: percentile functions can be defined slightly differently, but this is the definition used in the datascience package

Computing percentiles

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

Suppose we have a list: s = [1, 7, 3, 9, 5]

What is the 80th percentile? percentile(80, s)

If we order the elements in s, the 80th percentile is the 4th element:

(80/100) * 5

Percentile * Size of set

The 4th elements of [1, 3, 5, 7, 9] is 7

Discussion question

Which are True, when s = [1, 7, 3, 9, 5]?

- The sorted elements are: [1, 3, 5, 7, 9]

- percentile(10, s) == 1
- percentile(19, s) == 1
- percentile(20, s) == 1
- percentile(21, s) == 1

Let's explore this in Jupyter!

Estimation

Inference: Estimation

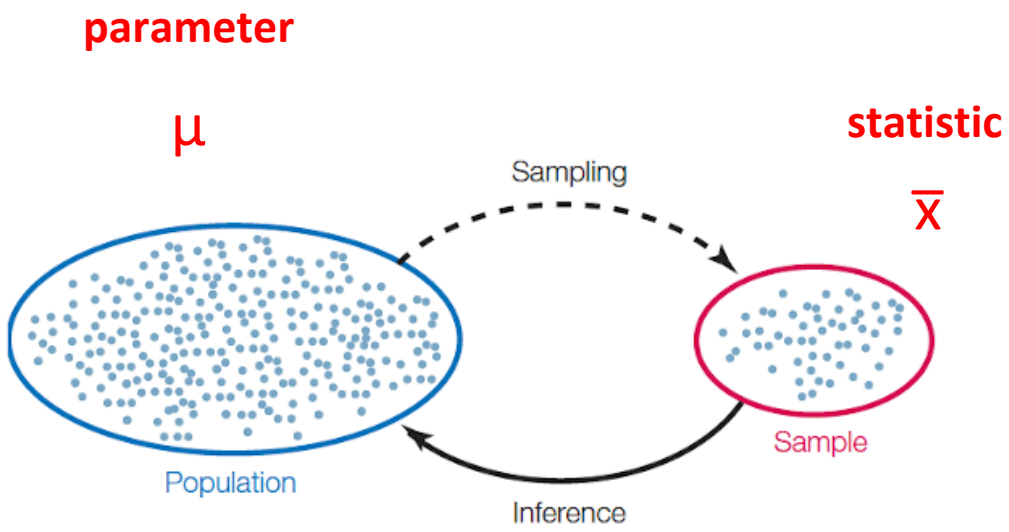
What is the value of an unknown parameter?

If you have data on the whole population:

- Just calculate the parameter value and you're done

If you only have a random sample from the population

- Use a statistic as an estimate of the parameter



Let's explore this in Jupyter!

Variability of the estimate

One sample → One estimate

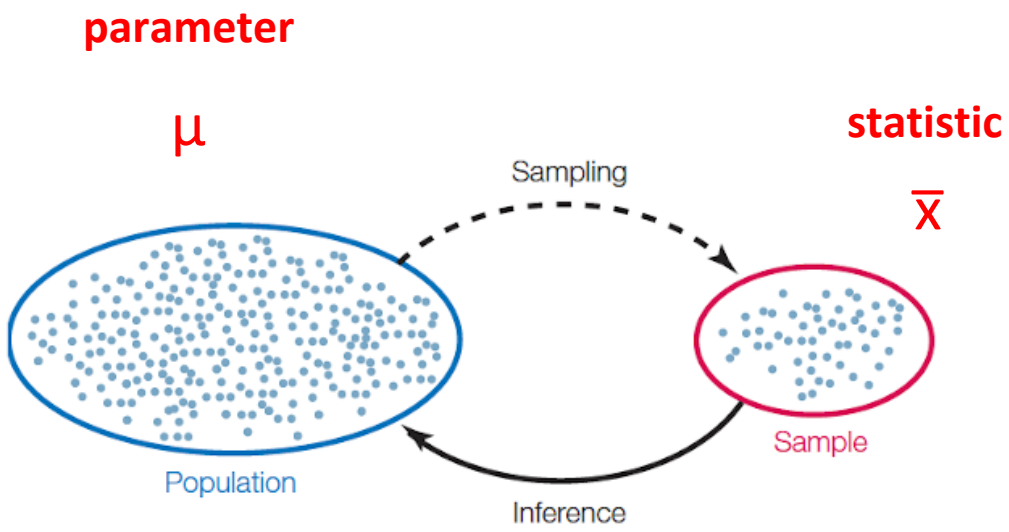
But the random sample could have come out differently

- 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?

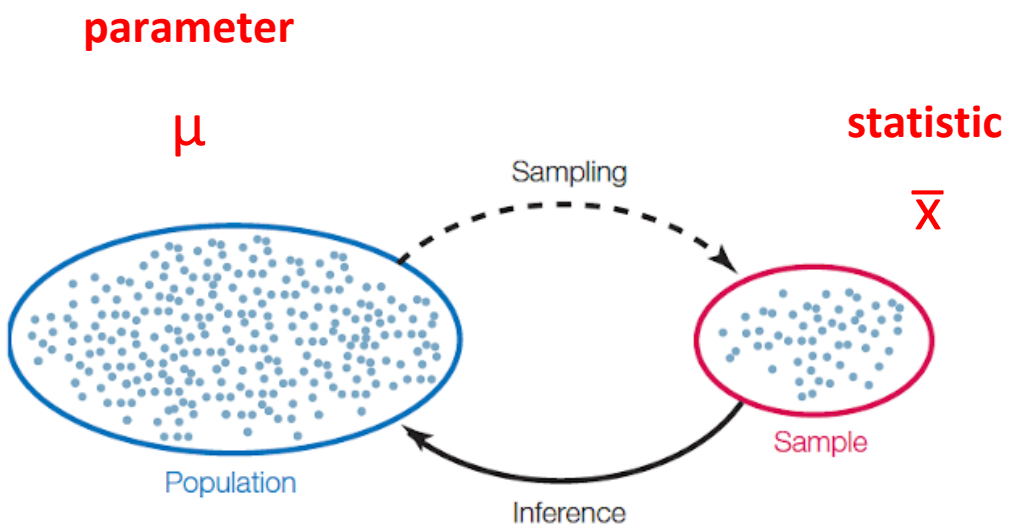
One sample → One estimate

To get many values of the estimate, we needed many random samples

Can't go back and sample again from the population:

- Too costly in terms of time and money

Stuck?



The Bootstrap

The Bootstrap

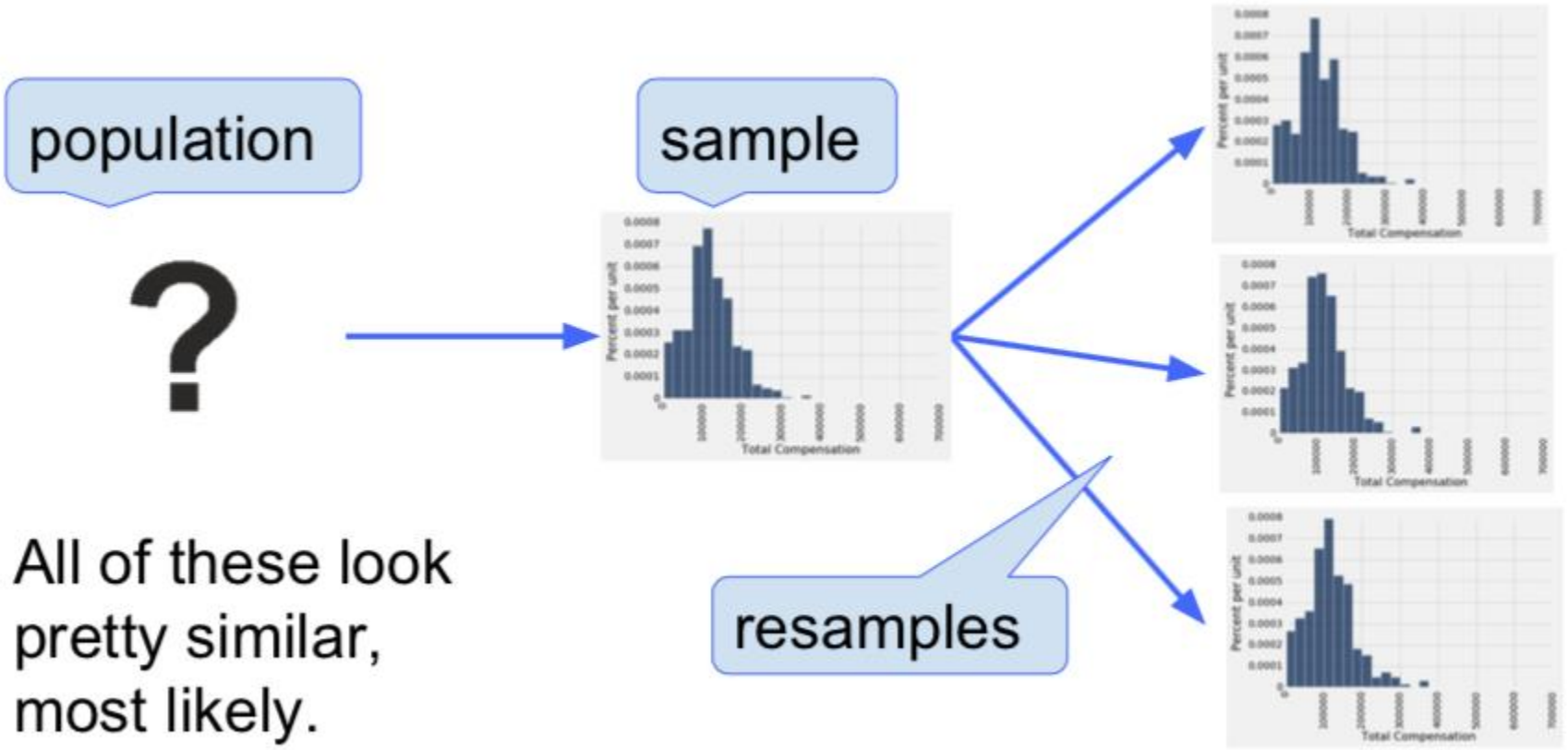
A technique for estimating confidence by simulating repeated random sampling

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



Key to resampling

From the original sample:

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

The sample (n = 10)

10, 3, 3, 3, 4, 3, 2, 6, 4, 5

→ 3, 3, 3, 5, 3, 4, 5, 2, 2, 10 $\bar{x}^* = 4$

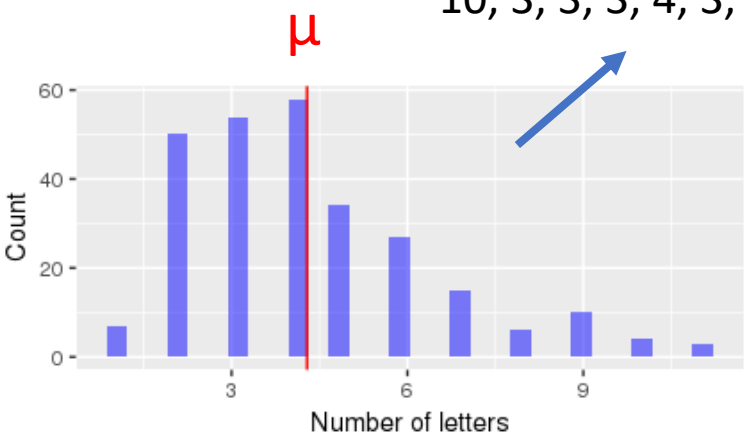
→ 3, 3, 2, 3, 6, 4, 6, 5, 3, 6 $\bar{x}^* = 4.1$

The size of the new sample has to be the same as the original one, so that we are replicating the process of drawing samples from the population

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

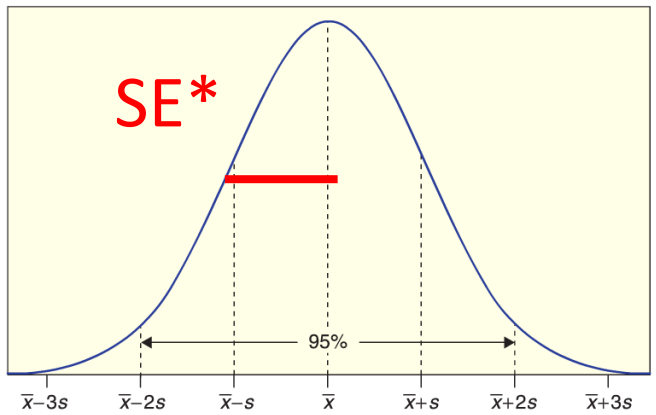
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

$\bar{x}^* = 3.9$



Bootstrap distribution!

Let's explore this in Jupyter!