The Bootstrap

Administrivia

- **Homework 5 posted.** Due Friday Nov 10
- **Good Milestones**:
  - Problems 1-4 **this week**
  - Problems 5 and 6 **next week**

Previously on CSCI 3022

We've looked at ways to compute confidence intervals for several different statistics:

A $100(1-\alpha)\%$ **confidence interval** for the mean μ with known sd. σ is given by

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

A $100(1-\alpha)\%$ **confidence interval** for the difference between means

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

A $100(1-\alpha)\%$ **confidence interval** for the difference between proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

What About Other Statistics?

We've seen different methods for computing CIs for means and proportions

But what about the variance? the standard deviation? the skew?

Rather than develop separate theory for each statistics, wouldn't it be nice if we had a method to compute CIs that would work for **almost all** statistics?

What About Other Statistics?

We've seen different methods for computing CIs for means and proportions

But what about the variance? the standard deviation? the skew?

Rather than develop separate theory for each statistics, wouldn't it be nice if we had a method to compute CIs that would work for **almost all** statistics...

What if We Don't Have Enough Data?

In real scenarios, data comes at a cost:

- **Money**: e.g, data from an aircraft in a wind tunnel
- **Time**: e.g, polling people in surveys is time consuming
- **Privacy Tradeoffs**: e.g., storing another person's genome in the database incurs ethical risk or cost, even when it does not cost much time or money

What if We Don't Have Enough Data?

In real scenarios, data comes at a cost:

- **Money**: e.g, data from an aircraft in a wind tunnel
- **Time**: e.g, polling people in surveys is time consuming
- **Privacy Tradeoffs**: e.g., storing another person's genome in the database incurs ethical risk or cost, even when it does not cost much time or money

Today, we'll learn a technique that enables us to tackle the not enough data problem as well as the problem of developing individual CI theory for each type of statistics

Today, we learn about **the Bootstrap!**

What are Bootstraps?

- Bootstraps are the straps that you use to pull your boots on
- To "pull yourself up by your bootstraps" is to somehow lift yourself upward by pulling on your own shoes. Obviously physically impossible.
- In statistics, however, bootstrapping means to accomplish something without aid. To accomplish what you need with what you've got.
- The statistical bootstrap is in this last sense. It allows us to really **make the most** of a small dataset without sacrificing statistical rigor or collecting more samples.

Confidence Intervals for the Mean

Recall: if we have n samples from a distribution, the Central Limit Theorem tells us that if n is sufficiently large, the confidence interval for the mean is given by

$$\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \bar{X} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}}$$

The Bootstrap is a different approach. Consider the same sample $X_1, X_2, \ldots, X_n$ as above, but instead of computing a CI analytically from the sample, instead we re-sample the sample many times and examine those.

Def: a **Bootstrapped resample** is a set of n draws from the original sample set **with replacement**.

Confidence Intervals for the Mean

Def: a **bootstrapped resample** is a set of n draws from the original sample set (drawn i.i.d. from X), sampled **with replacement**.

Example: suppose we have the data [2, 2, 4, 7, 9]

- Resample 1 might be: [2 4 7 7 9]
- Resample 2 might be: [2 2 2 2 4 7]
- Resample 3 might be: [2, 2, 4, 7, 8]

Given the example above, what does *sample with replacement* mean?

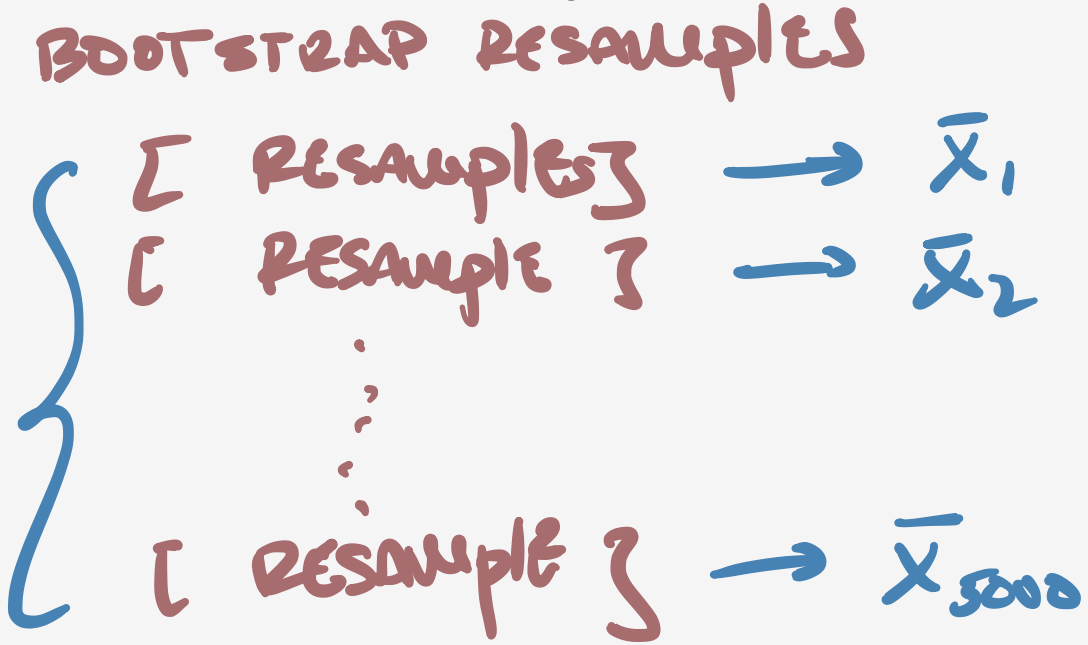
Rule of Thumb: The bootstrapped resample should contain the same number of observations of the original sample.

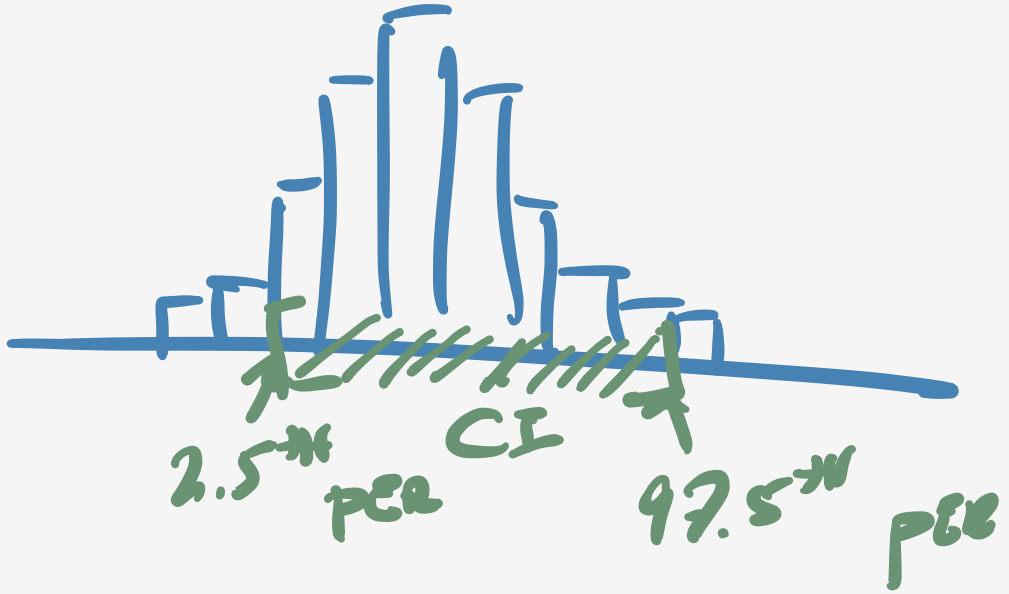
Why?

Confidence Intervals for the Mean

Def: a **bootstrapped resample** is a set of n draws from the original sample set (drawn i.i.d. from X), sampled **with replacement**.

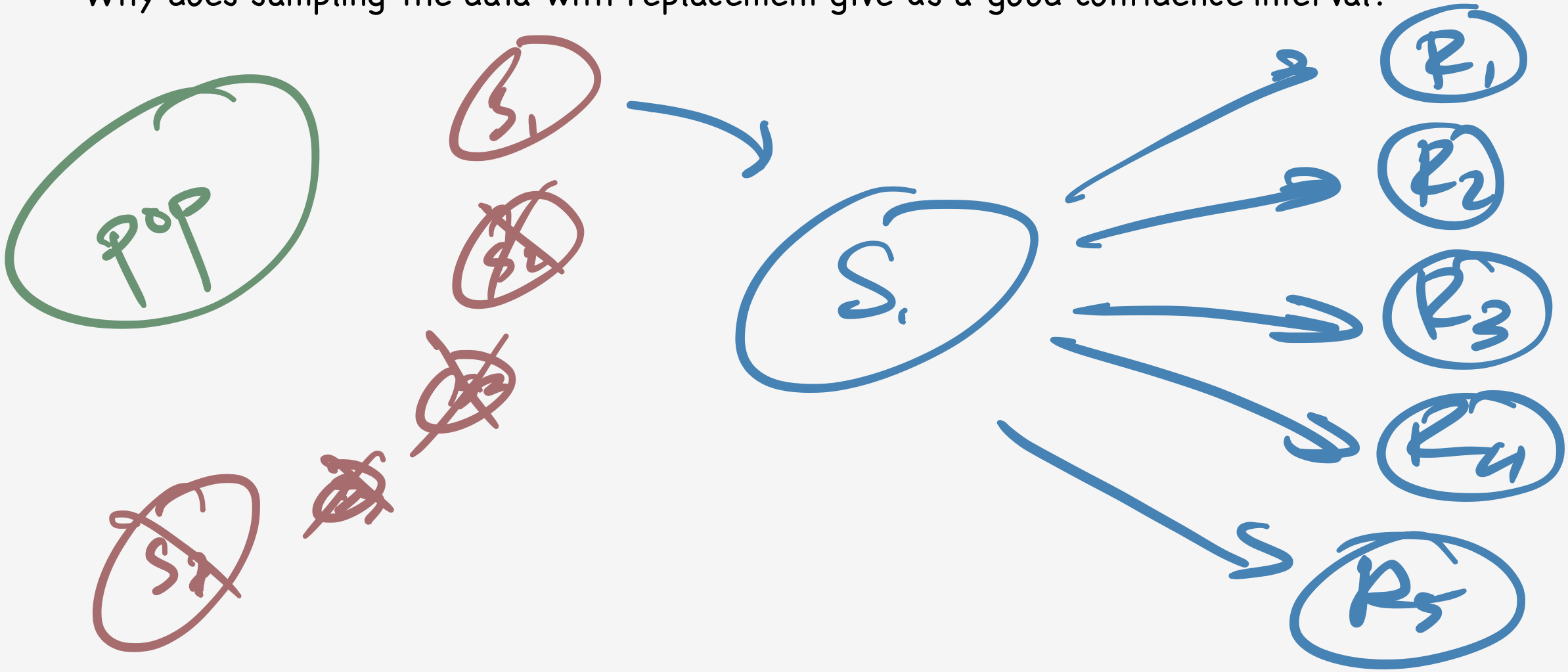
Proposition: a suitable estimate of the 95% confidence interval for the mean of the population X is given by [L, U] where L and U are the 2.5th percentile and 97.5th percentile of the means of a large number of bootstrapped resamples.





Sample with Replacement Intuition

Why does sampling the data with replacement give us a good confidence interval?



Less Talking More Hacking!

Get into groups and open the Lecture 16 in-class notebook

Goals:

- Write a function that takes in a samples, and computes a 95% confidence interval for the mean by bootstrapping the sample.
- Compare the bootstrapped CI with the traditional 95% confidence interval
- Come up with a way to test empirically whether this is working or not ...

Why we Love Bootstrap and You should Too!

- The bootstrap for a confidence interval around the mean is convenient, particularly when there are **not enough samples** to use the CLT
- Of course, if we can use the CLT, we should. So why is the Bootstrap so great?

Why we Love Bootstrap and You should Too!

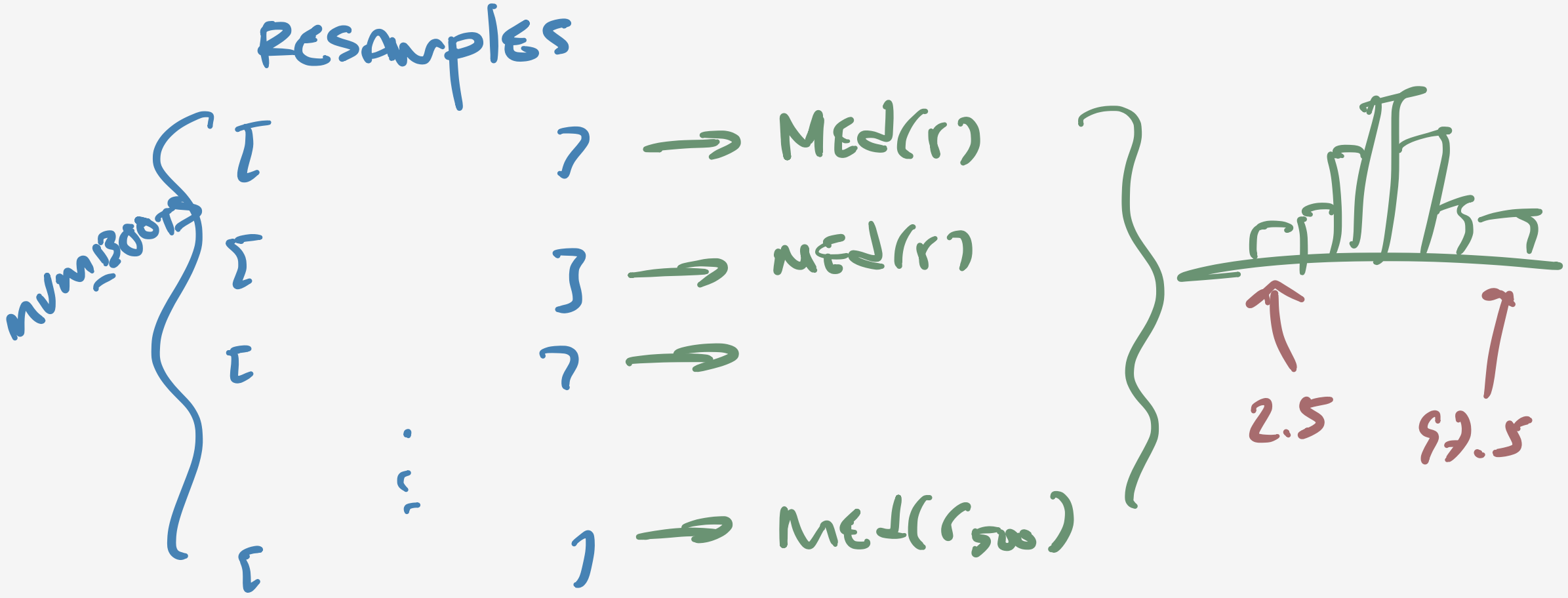
- The bootstrap for a confidence interval around the mean is convenient, particularly when there are **not enough samples** to use the CLT
- Of course, if we can use the CLT, we should. So why is the Bootstrap so great?

We can bootstrap CIs for things other than the mean!

- Median
- Standard Deviation
- Other statistical measures that we don't have theory for

Bootstrapped CIs for the Median

How could we come up with a confidence interval for the Median using Bootstrap?



The Non-Parametric Bootstrap

The literature (your book, Wiki) describe the previous methodology as a non-parametric bootstrap or empirical bootstrap. What is this?

Def: **parametric statistics** assumes that the sample data comes from a population that follows a probability distribution based on a fixed set of parameters.

Question: Can you name some **examples** of distributions with parameters?

The Parametric Bootstrap

We call the bootstrap discussed in class today the non-parametric bootstrap because it doesn't assume any parametric distribution. What you resample is what you get.

Def: the parametric bootstrap estimates a CI for a desired property in 2 steps

1. Repeatedly estimate the parameter(s) of the known distribution via bootstrap

2. Compute a CI for the desired property by sampling from the known distribution using parameters that you inferred.

EXP(λ) ESTIMATOR: RESAMPLE $X_1^*, X_2^*, \ldots, X_n^*$

ESTIMATOR $\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i^*}$

The Non-Parametric Bootstrap

We call the bootstrap discussed in class today the non-parametric bootstrap because it doesn't assume any parametric distribution. What you resample is what you get.

Def: the parametric bootstrap estimates a CI for a desired property in 2 steps

1. Repeatedly estimate the parameter(s) of the known distribution via bootstrap

2. Compute a CI for the desired property by sampling from the known distribution using parameters that you inferred.

Pro: the parametric bootstrap can be shown to do a better job than the non-parametric bootstrap in various scenarios.

Con: works great if the population has the distribution you have assumed. Not so great otherwise.

OK! Let's Go (Back) to Work!

Get in groups, get out laptop, and open the Lecture 16 In-Class Notebook again

Let's:

- Generate some bootstrapped CIs for the median and the standard deviation
- Explore the parametric bootstrap