CSCI 3022

intro to data science with probability & statistics

Lecture 21
April 4, 2018

The Bootstrap wrapup
Intro to Regression (maybe)

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- HW5 due **this Friday**.

OH today 11-1

Fr 8-~~10~~ 9:45

Previously on CSCI 3022

- **Definition**: a bootstrapped resample is a set of *n* draws from the original dataset (drawn IID from *X*), sampled *with replacement*.
- **Proposition**: a suitable estimate of the 95% confidence interval for the mean of the distribution *X* is given by [*a*,*b*], where *a* and *b* are the 2.5 percentile and 97.5 percentile of the means of a large number of bootstrapped resamples.
- **In plain English**: resample your original data many times. Compute the mean for each resample. Compute the 2.5 and 97.5 percentiles of those means.

Bootstrap: why we like it

- 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 exciting?

Bootstrap: why we like it

- 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 exciting?

We can bootstrap CIs for things other than the mean!

- Median. ✓
- Standard Deviation. ✓
- Other statistical measures that we don’t have a theory for.

Bootstrap for the median

90%

- Let's write **pseudocode** for how we would bootstrap a $\overset{\vee}{}$ CI for the median:

1. Resample. Create M resampled datasets (with replacement)

2. For each resampled dataset, compute the median

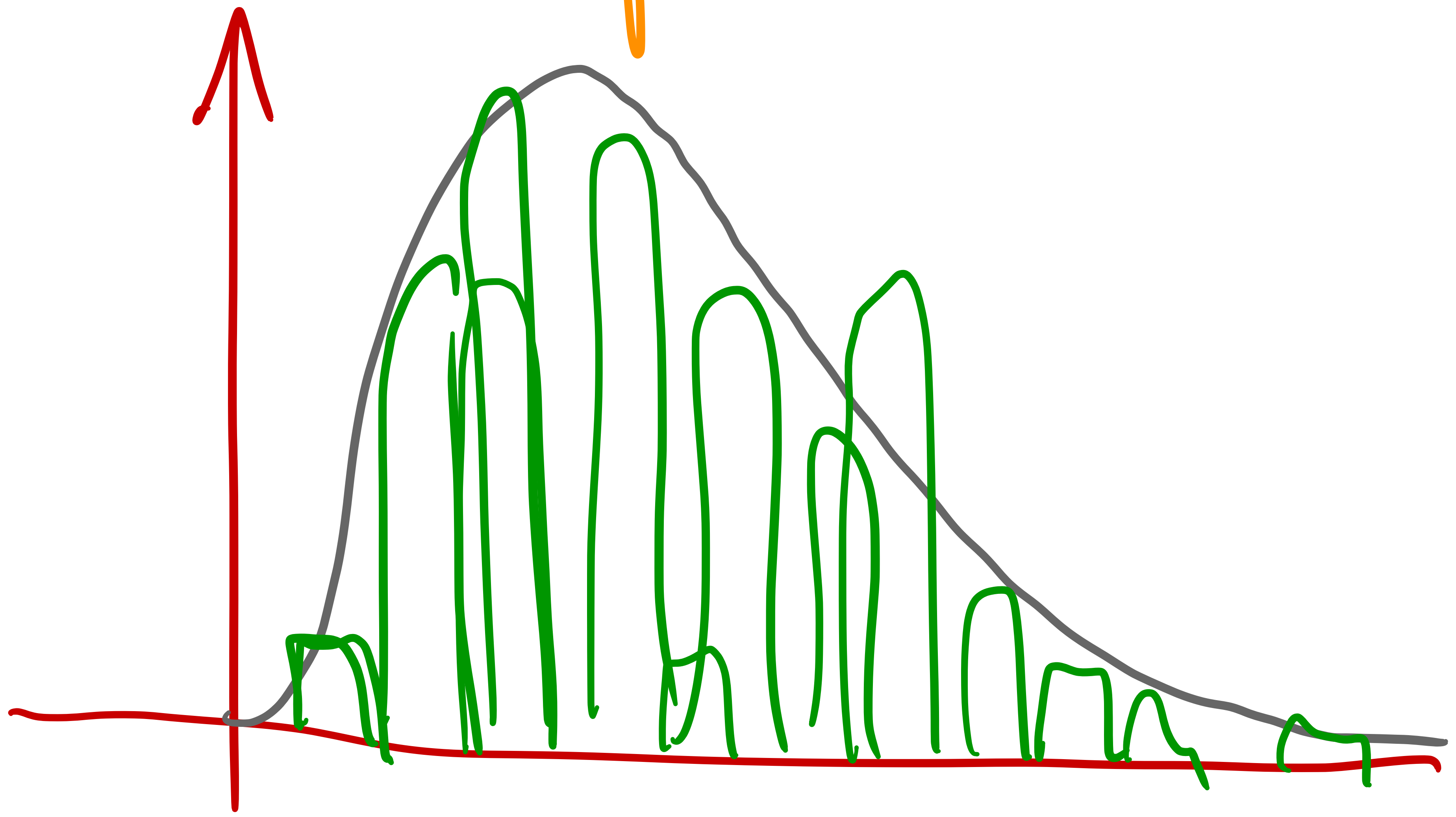
3. Take that distr. of medians, and compute 5th percentile and 95th percentile

CI: $\left[\frac{100-\alpha}{2}\%, 100 - \frac{100-\alpha}{2}\%\right]$

Bootstrap for the variance

- Let's write **pseudocode** for how we would bootstrap a CI for the variance:

No. See prev. slide.



The Non-Parametric Bootstrap

- In the literature—your book, the Wikipedia, etc—you may read about a "non-parametric bootstrap." What is this?

The Non-Parametric Bootstrap

- In the literature—your book, the Wikipedia, etc—you may read about a "non-parametric bootstrap." What is this?
- Let's decode this word, "non-parametric"
- **Definition**: *parametric statistics* assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters.
- Can you name some **examples** of distributions with parameters?

Urchin Eating Pois(λ) $N(\mu,\sigma^2)$ Bin(n,p)

- Can you name a *non*-parametric distribution we've talked about in class?

Let X be a r.v. such that $P(X=-1)=0.2$, $P(X=0)=0.5$, $P(X=19)=0.3$

The Parametric Bootstrap

Replace with your preferred distr. and statistic

- 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.

- **Definition**: the parametric bootstrap estimates a CI for a desired property in two steps: (1) repeatedly estimate the parameter(s) of the known distribution, and then (2) compute a CI for the desired property by sampling from the known known distribution using the parameters that you inferred.

1. Create M bootstrapped datasets.

2. Assume that the data came from Pois(λ), and estimate λ for each of the M datasets.

3. Use those parameters, and for each one, compute the median.

4. Compute the CI from those medians.

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.

- **Definition**: the parametric bootstrap estimates a CI for a desired property in two steps: (1) repeatedly estimate the parameter(s) of the known distribution, and then (2) compute a CI for the desired property by sampling from the known known distribution using the parameters that you inferred.

- **Why**? The parametric bootstrap can be shown to do a better job than the non-parametric bootstrap in various scenarios.

- Why not use the parametric bootstrap all the time?

  1. Might not be getting data from a parametric distrib.
  2. Might not know what the parametric distrib. is!

Let’s notebook it up!