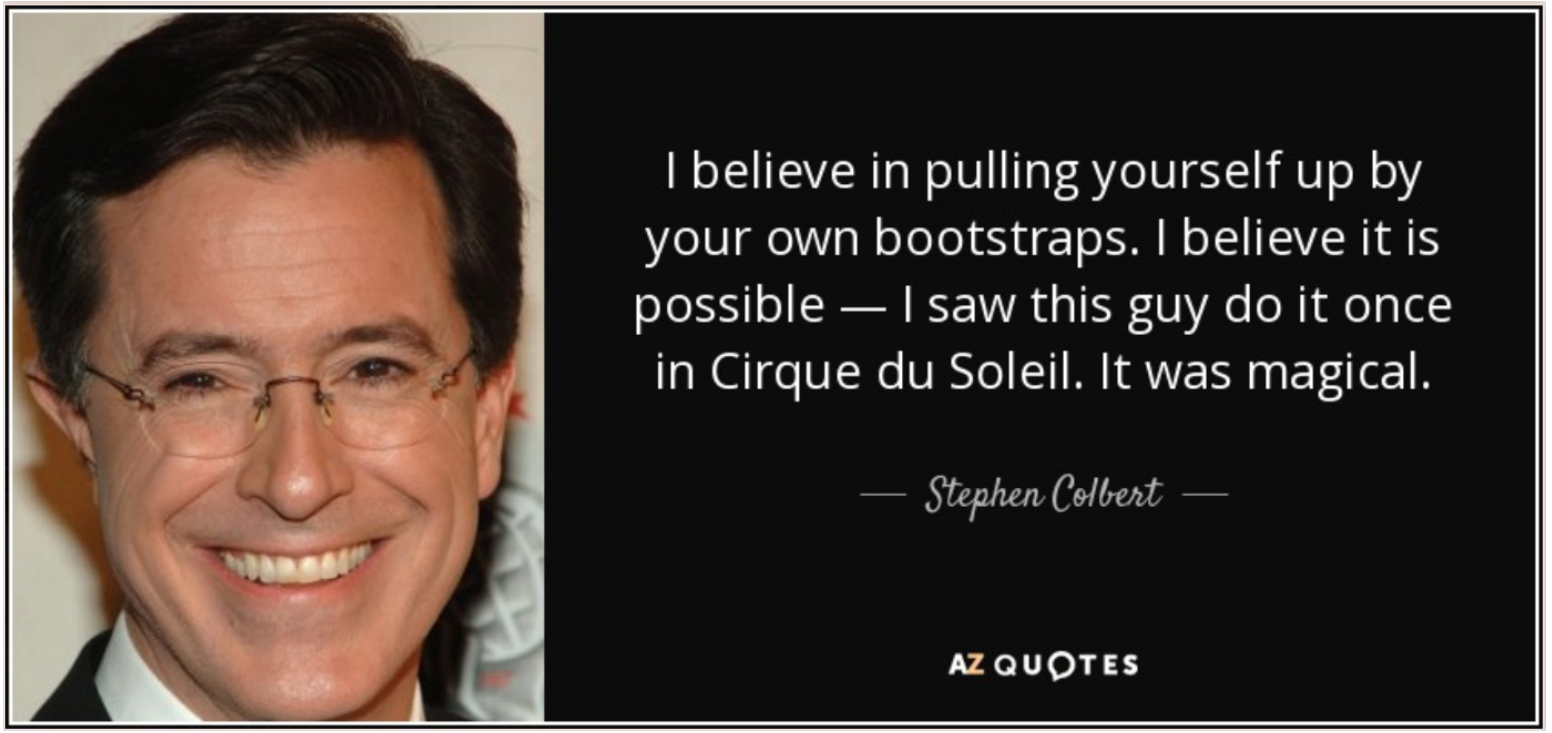
# CSCI 3022: Intro to Data Science

## Lecture 19: The Bootstrap

Rachel Cox

Department of Computer Science



# Announcements & Reminders

HW4 Due Friday at 11:59 pm

What will we learn today?

- ❑ The bootstrap
- ❑ Computing confidence intervals when we don't have enough samples for the CLT

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 18 & Section 23.3*

# Review

A $100 \cdot (1-\alpha)\%$ confidence interval for the mean $\mu$ when the value of $\sigma$ is known is given by:

$$\left[\bar{x} - z\alpha_{/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z\alpha_{/2}\frac{\sigma}{\sqrt{n}}\right]$$

A $100 \cdot (1-\alpha)\%$ confidence interval for the difference between means is given by:

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

A $100 \cdot (1-\alpha)\%$ confidence interval for the difference between proportions is given by:

$$(\widehat{p_1} - \widehat{p_2}) \pm z\alpha_{/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

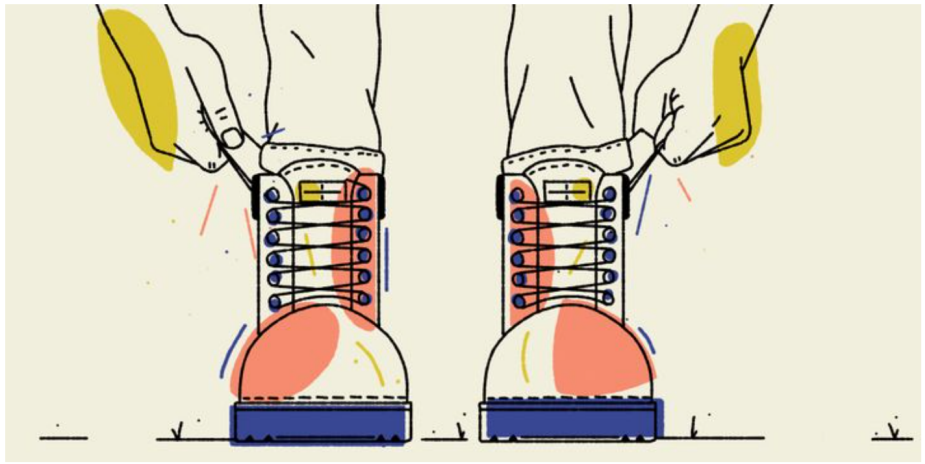
# Confidence Intervals for Other Statistics

We've seen methods for computing CIs for means, proportions, and variances. What about the median? Or the skew?

Data often comes at a cost.

- Money – e.g. data collected by aircraft
- Time – polling people in surveys is time consuming
- Privacy trade-offs – storing a person's genome in the database incurs ethical risk/cost

**The Bootstrap Principle**: A technique that enables us to tackle the "not-enough-data" problem AND the "I-want-other-statistics" problem.



ISABELLA CARAPELLA/HUFFPOST

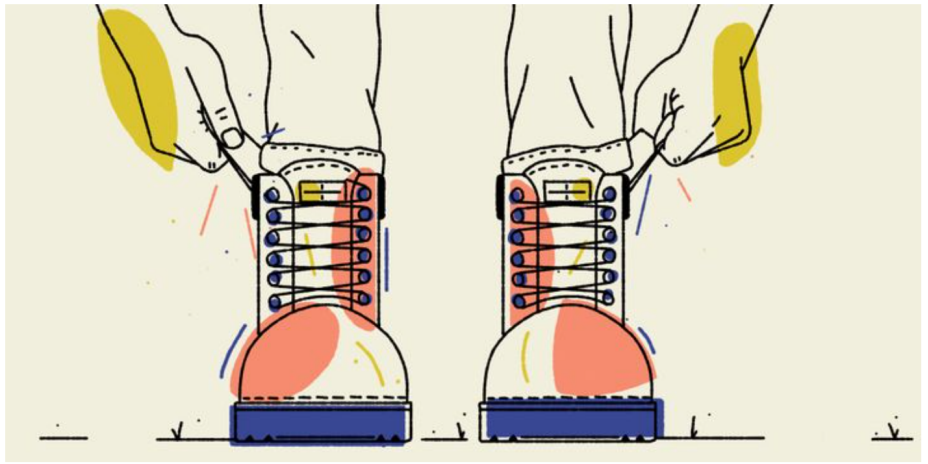
# The Bootstrap Principle

**The Bootstrap Principle**: A technique that enables us to tackle the "not-enough-data" problem AND the "I-want-other-statistics" problem.

In statistics, bootstrapping means to accomplish what you need with what you've got.

The statistical bootstrap is to make the most of a smaller data set without sacrificing statistical rigor or collecting more samples.

The bootstrap involves the following:
Consider a sample $X_1, X_2, \ldots, X_n$, but instead of computing a CI analytically from the sample, we instead **re-sample the sample** many times and examine the re-samples.



ISABELLA CARAPELLA/HUFFPOST

# The Bootstrap Principle

A **bootstrapped resample** is a set of n draws from the original sample set with replacement.

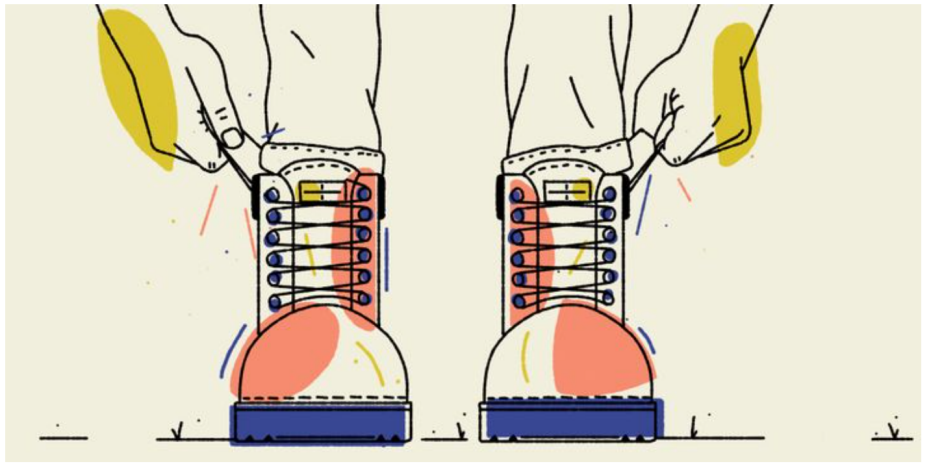
**Example**: Suppose we have the data: $X = [2, 2, 4, 7, 9]$

Resample 1 might be: [2, 4, 4, 4, 7] $\tilde{x}_1 = 4$

Resample 2 might be: [4, 4, 4, 4, 4] $\tilde{x}_2 = 4$

Resample 3 might be: [2, 4, 7, 9, 9] $\tilde{x}_3 = 7$

**General rule**: The bootstrapped resample should contain the same number of observations as the original sample.



ISABELLA CARAPELLA/HUFFPOST

# The Bootstrap Principle

A **bootstrapped resample** is a set of n draws from the original sample set with replacement.

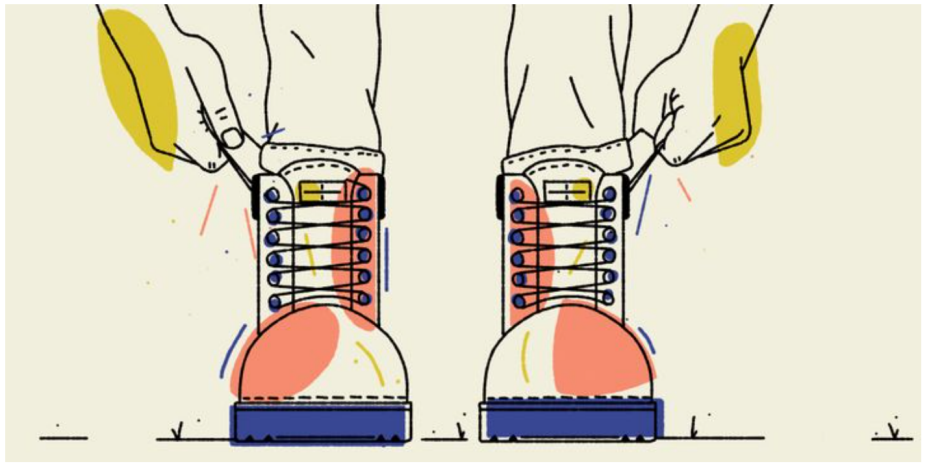
**Example**: Suppose we have the data: $X = [2, 2, 4, 7, 9]$

Resample 1 might be:

Resample 2 might be:

Resample 3 might be:

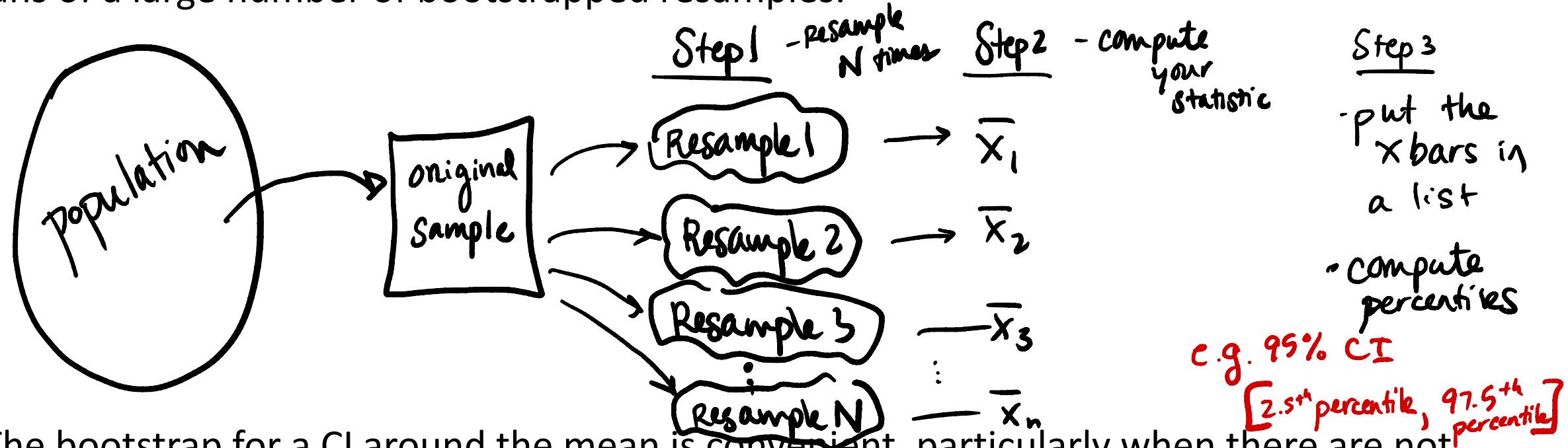
**General rule**: The bootstrapped resample should contain the same number of observations as the original sample.



ISABELLA CARAPELLA/HUFFPOST

# The Bootstrap Principle

**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 and 97.5th percentiles, respectively, of the means of a large number of bootstrapped resamples.




- The bootstrap for a CI around the mean is convenient, particularly when there are not enough samples to use the CLT.

- If we can use the CLT – we should.

# The Bootstrap Principle

We can use bootstrap CIs for things aside from the mean!

- Median
- SD
- Other statistical measures for which we don't have theory

**Example**: Let's write pseudocode for how we would bootstrap a 90% CI for the median.

```
medians = [ ]
given : original sample
for a large number of resamples:
        Resample = random.choice(original sample, with replacement)
        median_resample = np.median(resample)
        medians.append(median_Resample)

CI = np.percentile(medians, [5, 95])
Return CI
```

# The Bootstrap Principle

**Example**: Let's write pseudocode for how we would bootstrap a 90% CI for the variance.

Same thing as previous slide but we compute the variance (np.var) instead of median (np.med)

# Non–parametric bootstrap

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

- Name some examples of distributions with parameters.

$N(\mu, \sigma^2)$ , $\chi^2(\nu)$

- Name an example of a non-parametric distribution.

[1, 2, 3, 3, 3, 4, 10, 12, 100]

❖ We call the bootstrap discussed in class today so far the non-parametric bootstrap because it doesn't assume any particular parametric distribution.

# Parametric bootstrap

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

The parametric bootstrap estimates a CI for a desired property in three steps:

1) Estimate the parameter(s) of the known distribution from your sample.
2) Draw bootstrap resamples from the distribution, assuming the estimated parameter.
3) Compute a CI for the desired property from your resamples.

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

**Con**: While the method works well if the population has the distribution you assumed that it did, it does not work well if you got that wrong.

# *Next Time:*

- Regression