Lecture 5: Bootstrap

Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

University of Washington

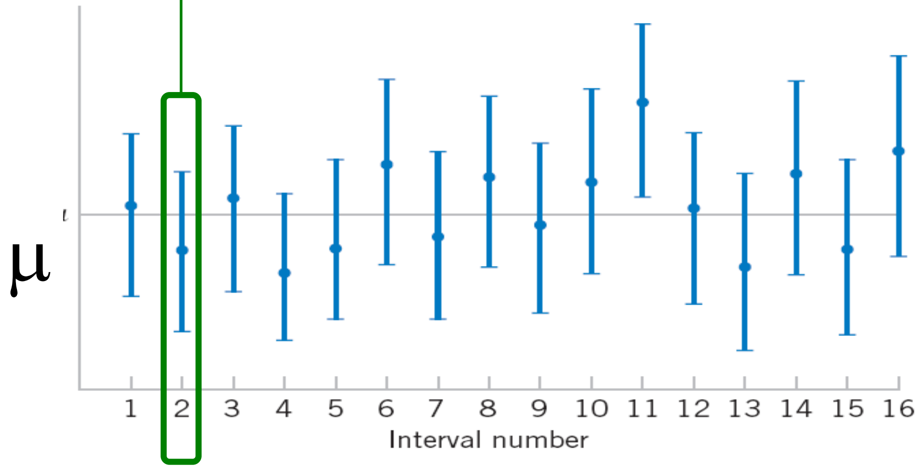
Review the rationale of hypothesis testing and confidence interval

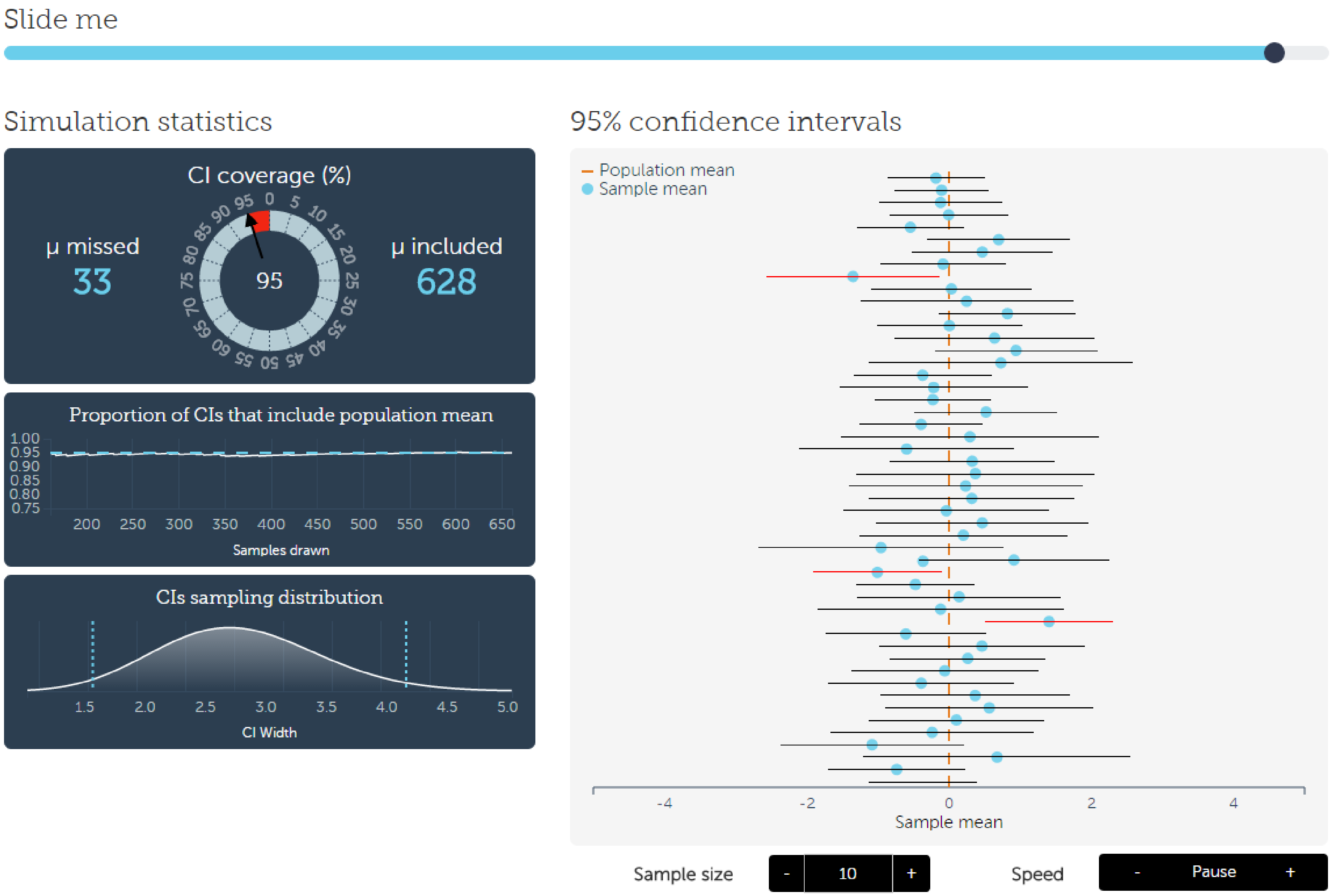
21 19 17 19 19 25 24 20 23 18 $\bar{x}_2 = 20.5$

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\text{for } \alpha = 0.05: 20.5 - (1.96)\frac{2.5}{\sqrt{10}} \le \mu \le 20.5 + (1.96)\frac{2.5}{\sqrt{10}}$$

$$18.95 \le \mu \le 22.05$$





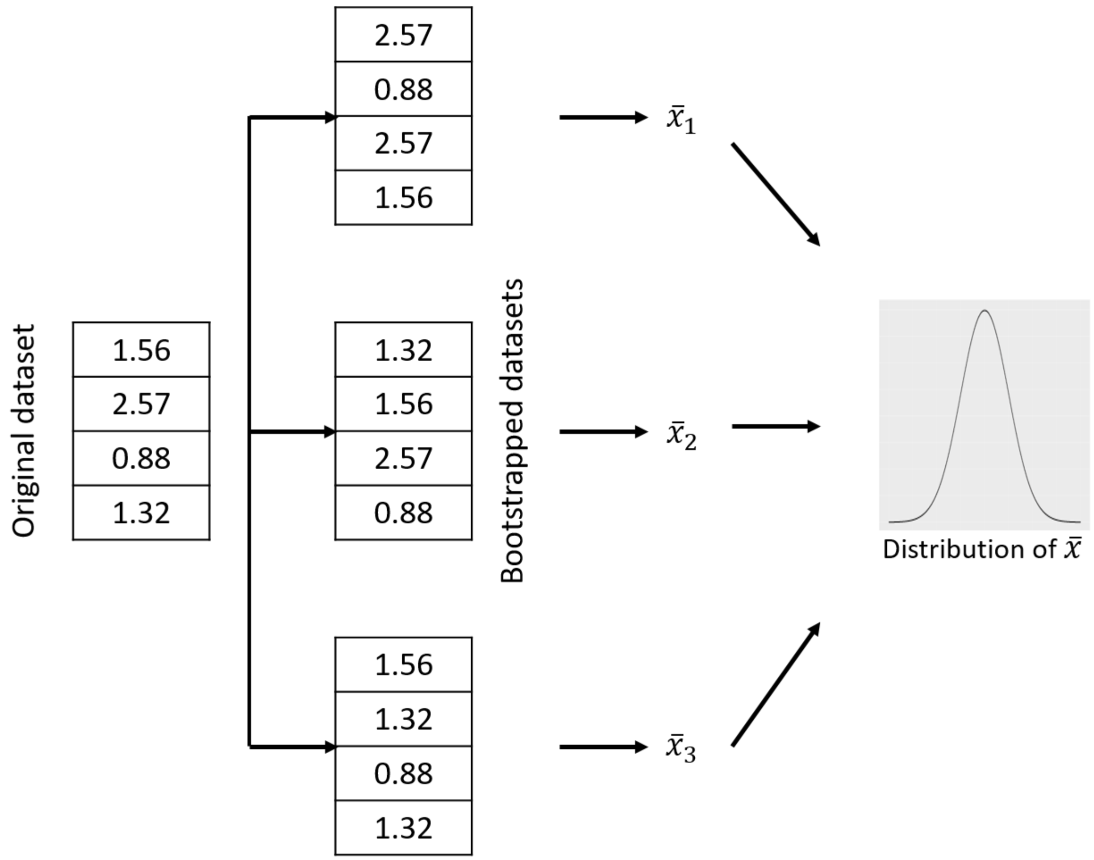
http://rpsychologist.com/d3/CI/

Without analytical tractability?

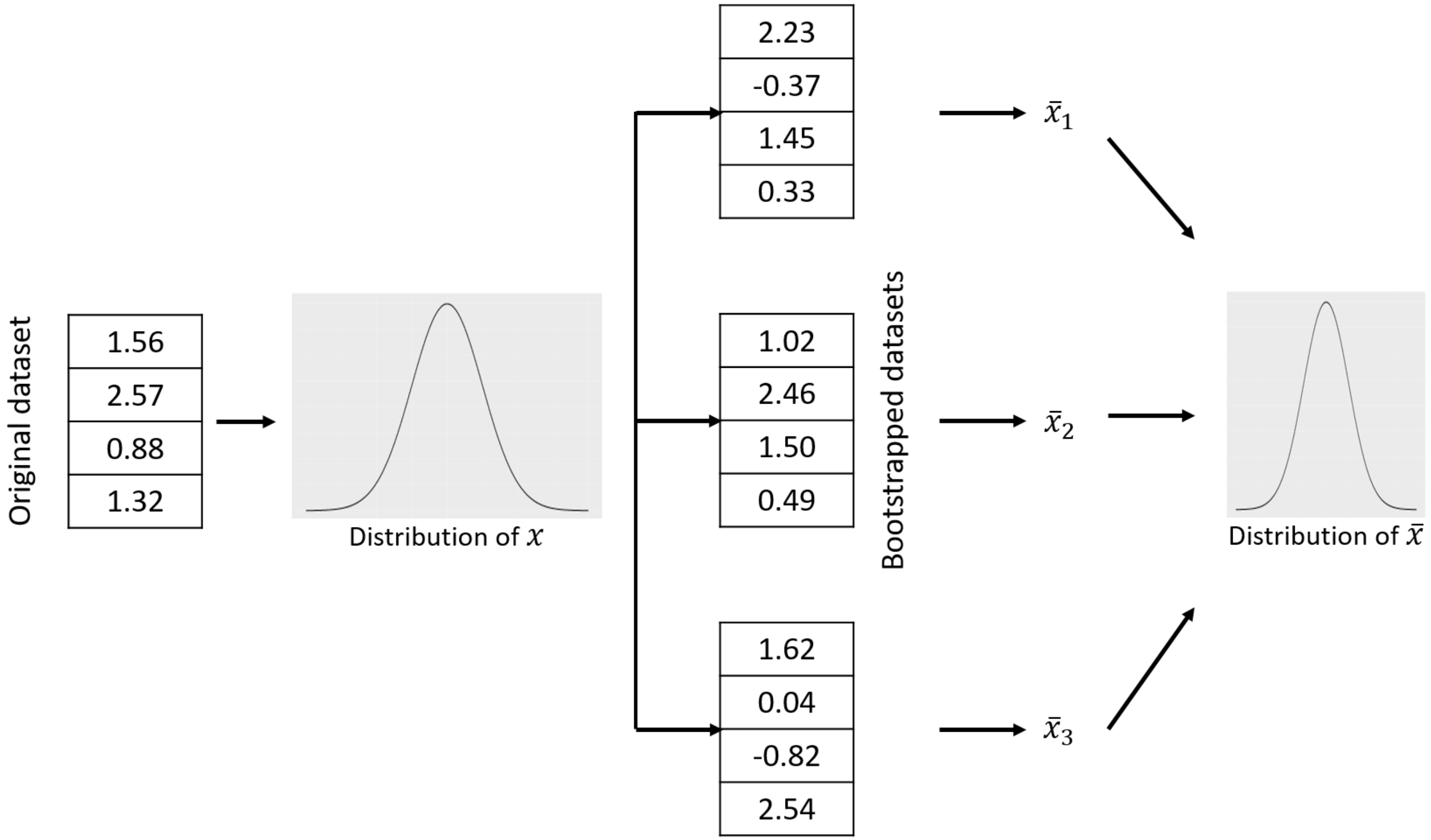
- The idea of Bootstrap to computationally mimic the sampling process

Complete dataset	X_1	X_2	X_3	X_4	X_5
Bootstrapped dataset 1	X_3	X_1	X_3	X_3	X_5
Bootstrapped dataset 2	X_5	X_5	X_3	X_1	X_2
Bootstrapped dataset 3	X_5	X_5	X_1	X_2	X_1
…					
Bootstrapped dataset K	X_4	X_4	X_4	X_4	X_1

A nonparametric Bootstrap scheme



A parametric Bootstrap scheme



Bootstrap for regression models

- Option 1: we could simply resample the data points (i.e., the $(\boldsymbol{x}, y)$ pairs) similarly as the nonparametric Bootstrap scheme. Then, for each sampled dataset, we can fit a regression model and obtain the fitted regression parameters.
- Option 2: we could simulate new samples of X using the nonparametric Bootstrap method on the samples of X only. Then, for the new samples of X, we draw samples of Y using the fitted conditional distribution model $P(Y|X)$.
- Option 3: we could fix the X, only sample for Y. In this way we implicitly assume that the uncertainty of the dataset mainly comes from Y. To sample Y, we draw samples using the fitted conditional distribution model $P(Y|X)$.

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets