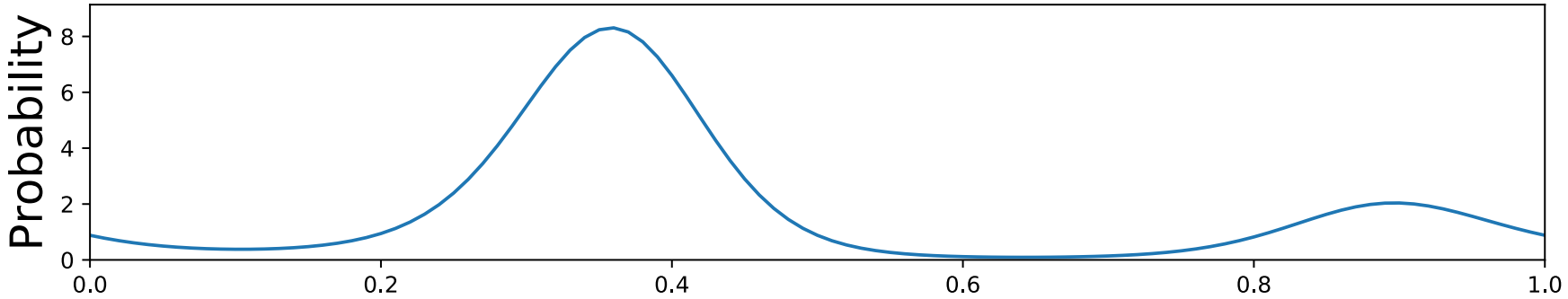
Bootstrapping

Or, reusing data for maximum benefit!

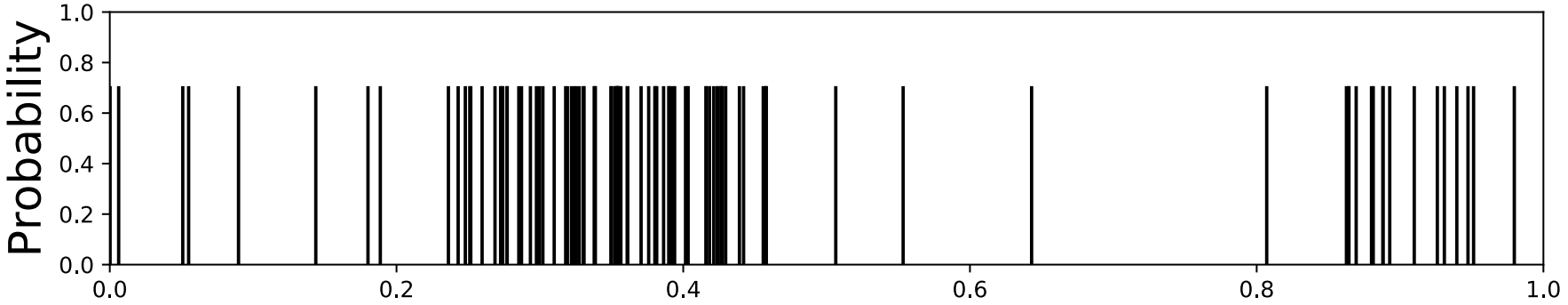
Prof. Michael Shirts
University of Colorado, Boulder
i-CoMSE MC/MD Workshop
Oklahoma State University, July 2022

The bootstrap

- I'd like to collect 1000 samples from my population. But TOO EXPENSIVE!
- What distribution am I collecting samples from?



- What distribution do I actually know about after sampling?



- How do I sample from THIS distribution????

Bootstrap Etymology

. . . to succeed or elevate yourself without any outside help

Yes, technically impossible!

In this case, we are attempting to get **out-of-sample estimates** using **in-sample data**.

It works when our sample is **representative** of our population.

Generating a bootstrap error estimate

My sample

Some function computed from my sample

F( ) = A

Pick from my sample with replacement
Compute A for each of these resamples

F( ) = A_1

F( ) = A_2

F( ) = A_3

F( ) = A_4

F( ) = A_5

F( ) = A_6

Calculate standard deviation or confidence intervals over bootstrap distribution collected

How many bootstrap samples?

- Standard errors of the mean
  - 50-200 bootstraps usually good
- 95% Confidence intervals of the mean
  - Probably need more like 1,000-2,000 bootstraps
  - Because the 95% confidence intervals depends on only a small number of outliers
- One caveat:
  - You never multiply by $\frac{1}{\sqrt{B-1}}$ bootstraps; more samples make the answer more precise. You are finding the standard deviation of the entire process, which already includes the *N* samples.

What else can I use bootstraps for?

- Other more complicated functions
  - What are the confidence intervals of $F(X) = \sin(\text{abs}(\ln X)^{3.5})$?
  - **Doesn't matter** how complicated, bootstrap still works!
- Other more complicated statistics
  - What is the median?
  - What is the value of the sample that is 10 samples lower than the median?

Limitations

- Doesn't work with correlated data
  - There's a version called the 'block bootstrap' you can use for correlated data, but easier to just select a subset of uncorrelated data
- If N is small (10 or 20), often inconsistent results
  - Especially true for confidence intervals, which are underestimated you simply don't have enough sampling in the tails
  - But, medians and means are much less dependent on N, and work well for quite small (20-50?) samples.