Jhelum Chakravorty

Published Sep 29
Fork of The Bootstrap by Kris Sankaran

The Bootstrap

IFT6758, Fall 2020

Reading: ISLR 5.2 - 5.3

Optional reading: CASI Chapter 1

Computation for Inference

- Increased computation has revolutionized our ability to fit functions to data
- It has also profoundly influenced inference
- The bootstrap is one important reflection of this

Algorithms vs. Inference

- Algorithms help you construct useful reductions of data (for knowledge or decision making).
- Inference helps you gauge how reliable your reductions are

"It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also be used to assess it's accuracy."

Q: Given the data, can we have an idea about how 'confident' we are about a statistic?

Example: Sample Mean

Good representation of sample population

There is a closed-form formula for the standard error of the sample mean. Suppose $x_i \overset{i.i.d.}{\sim} F$, a distribution with variance σ^2. Then,

$$\begin{aligned}\mathrm{Var}\left[\bar{x}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[x_i\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2} n\sigma^2 \\
&\boxed{= \frac{\sigma^2}{n}}
\end{aligned}
$$

Estimated Standard Error

how much variance would be in the sample mean

Therefore, if we had a way of estimating σ^2, then we can estimate the standard error of the mean by plugging this estimate in.

$$
\mathrm{Var}\left[\bar{x}\right] \approx \frac{\hat{\sigma}^2}{n}
$$

This would give us a way of understanding to what degree we can trust our estimate of the mean, computed entirely from the raw data.

Estimated Standard Error

Of course, if $x_i \overset{i.i.d.}{\sim} F$, then a reasonable estimate for σ^2 is

$$
\boxed{\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}
$$

So in summary, to estimate the variance of the mean, we

1. Compute an estimate of σ^2
2. Plug that the true expression for $\mathrm{Var}\left[\bar{x}\right]$

Abstraction: estimate the variance of the statistic from the data

This is a more abstract way of describing this process.

1. Define a statistic $\hat{\theta}\left(x_1, \ldots, x_n\right)$ of the data
2. Do math to get an expression for $\mathrm{Var}\left[\hat{\theta}\right]$
3. Replace unknowns in the expression with estimates, and argue that the result is $\approx \mathrm{Var}\left[\hat{\theta}\right]$.
   - If you want to be precise, you could call this new estimator $\widehat{\mathrm{Var}}\left[\hat{\theta}\right]$

Goals

In a lot of cases, step (2) is intractable.

- $\hat{\theta}$ is a more complex function of the x_i
  - Ratio between eigenvalues in your PCA
  - Derived statistics, like $\log \mu$ in some model
- You ran some iterative algorithm to compute $\hat{\theta}$
  - Robust regression (e.g. Least Trimmed Squares)
  - Random forests

We'd like a recipe that works even then, and which doesn't have to be rederived for every single problem.

Example: Portfolio Optimization

- Two assets that you can invest in, X and Y.
- Distribute α fraction of funds to X, and the rest to Y, i.e., invest

$$\alpha X + (1 - \alpha) Y$$

- The best strategy (in the sense of minimizing variance) can be shown to be

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Example: Portfolio Optimization

- α is unknown in practice, so we estimate it,

$$\hat{\alpha}\left(x_1, \ldots, x_n, y_1, \ldots, y_n\right) = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

- Now we want to know, how variable is this estimator?

A Thought Experiment

- Suppose you had a window into parallel universes
- How much does the estimator change across different samples?

- $\sigma_X^2 = 1,\ \sigma_Y^2 = 1.25,\ \sigma_{XY} = 0.5;\ \alpha = 0.6$

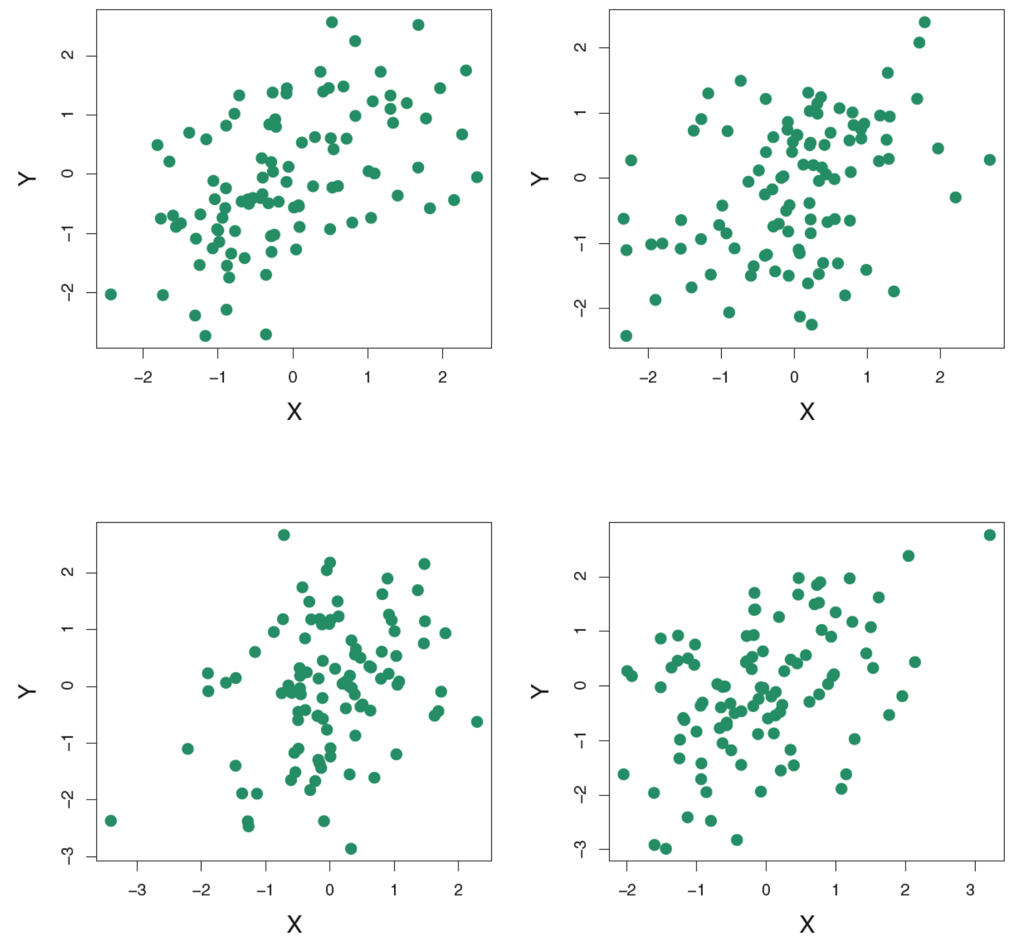


FIGURE 5.9. *Each panel displays 100 simulated returns for investments X and Y. From left to right and top to bottom, the resulting estimates for α are 0.576, 0.532, 0.657, and 0.651.*

Thought Experiment

If you simulate 1000 datasets in this way, you can get a different $\hat{\alpha}_r$, for $r = 1, 2, \ldots, 1000$.

$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996$$

which is very close to the underlying 0.6.

To get a sense of the variability across datasets, we can use

$$\widehat{\mathrm{Var}}\,[\hat{\alpha}] = \frac{1}{999} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2 \approx 0.083^2,$$

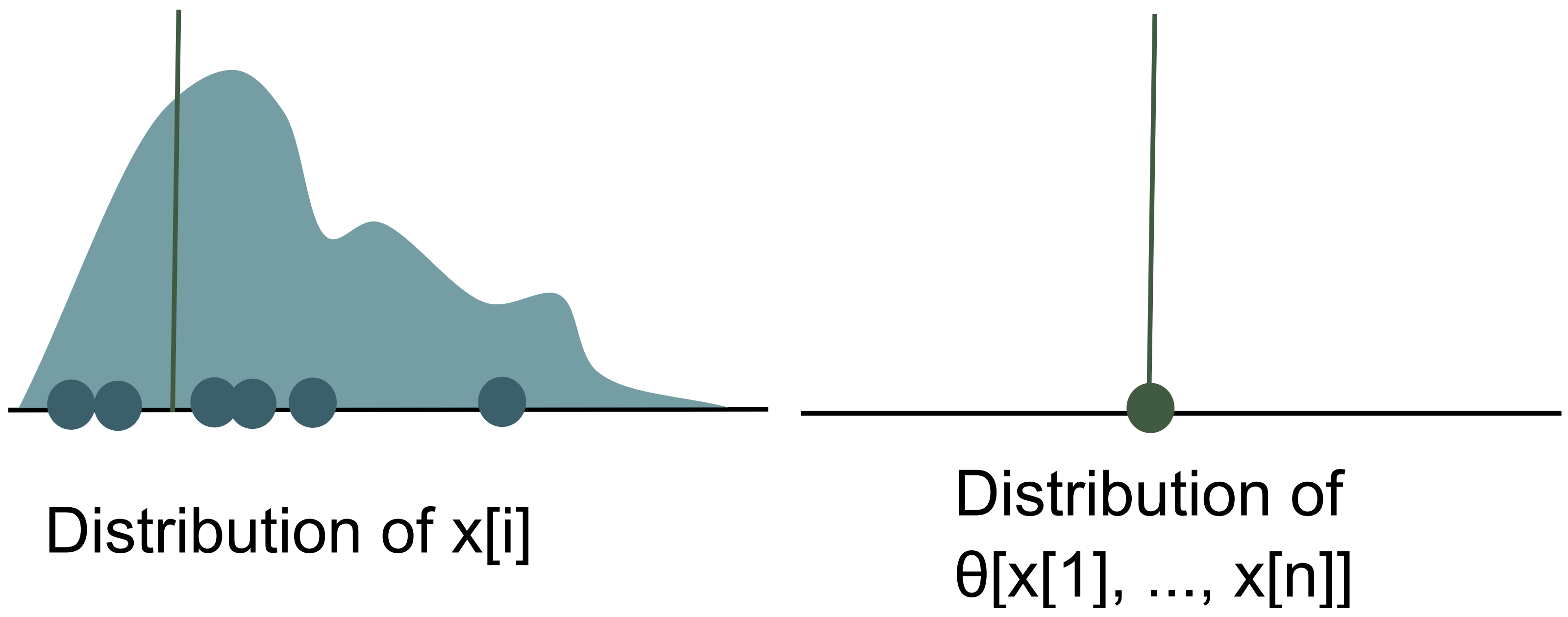
Bootstrap Idea

We cant access to the parallel universe to get the data. Hence take the samples from the samples of the population.

- **Key idea**: The population is to the sample as the sample is to the bootstrap samples.
- In our simulation, we were able to sample as many new datasets F as we wanted. In reality, we see only one.
- But we can simulate as many datasets from the empirical distribution $\hat{F}$ as we want
- It turns out that if you use $\hat{F}$ in place of F, the approach from the thought experiment *still works*

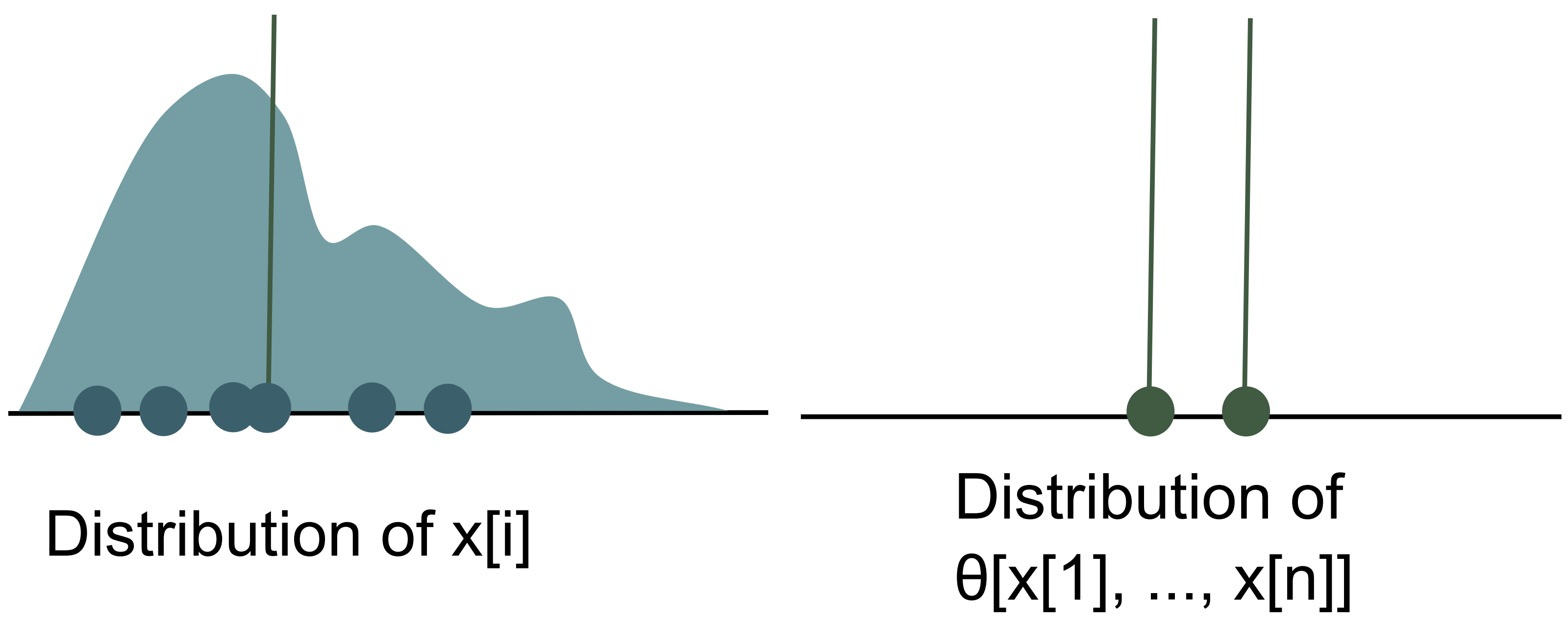
Sampling from F

In our ideal simulation world, we're able to generate many datasets from F and see how our estimator $\hat{\theta}$ changes.



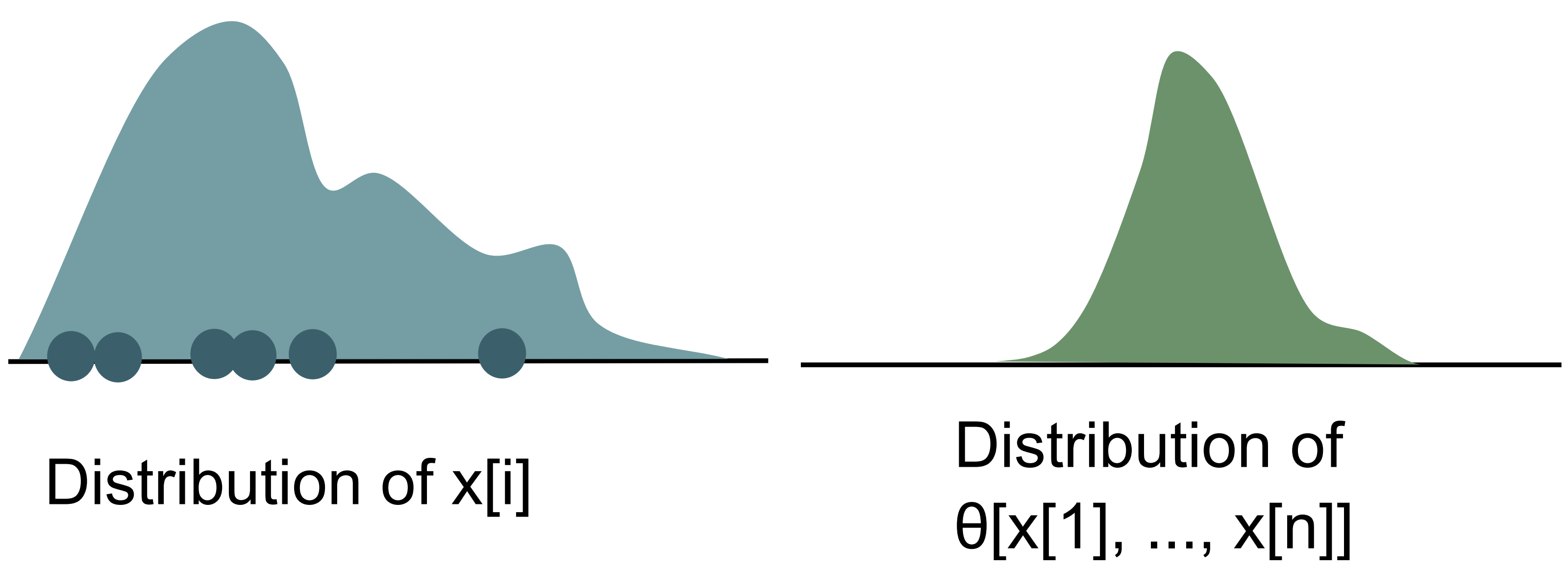
Sampling from F

In our ideal simulation world, we're able to generate many datasets from F and see how our estimator $\hat{\theta}$ changes.



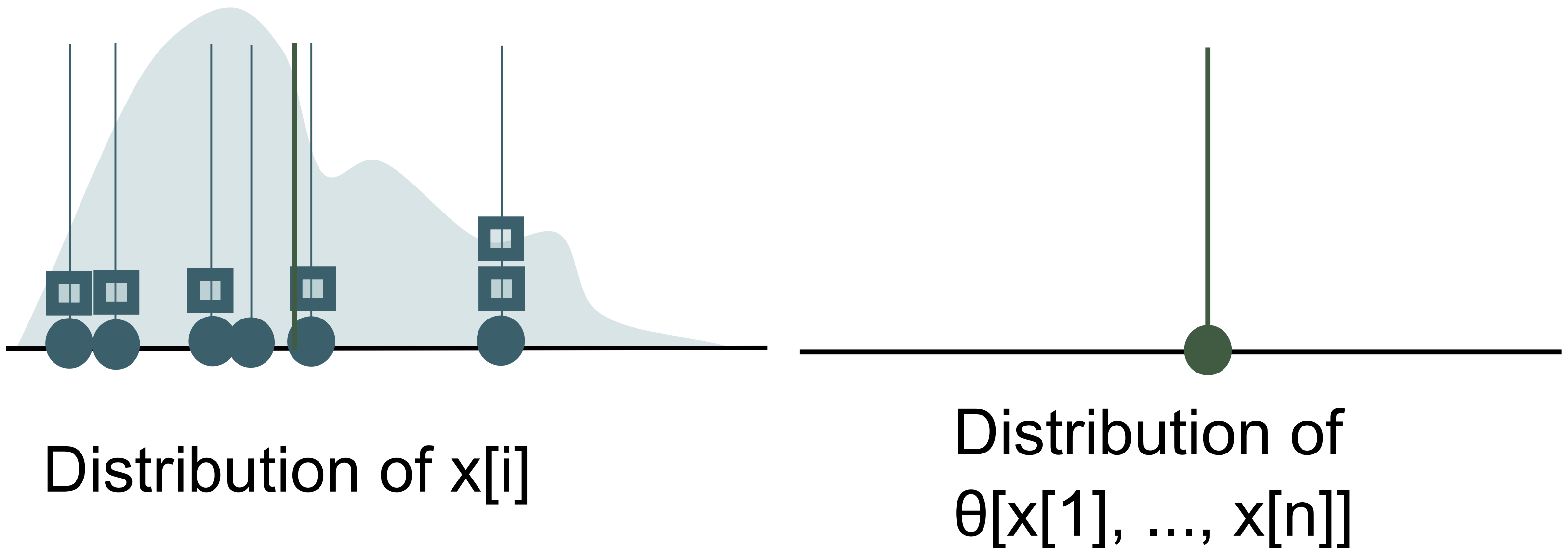
Sampling from F

The properties of the final sampling distribution for $\hat{\theta}$ can be used to guide inference.



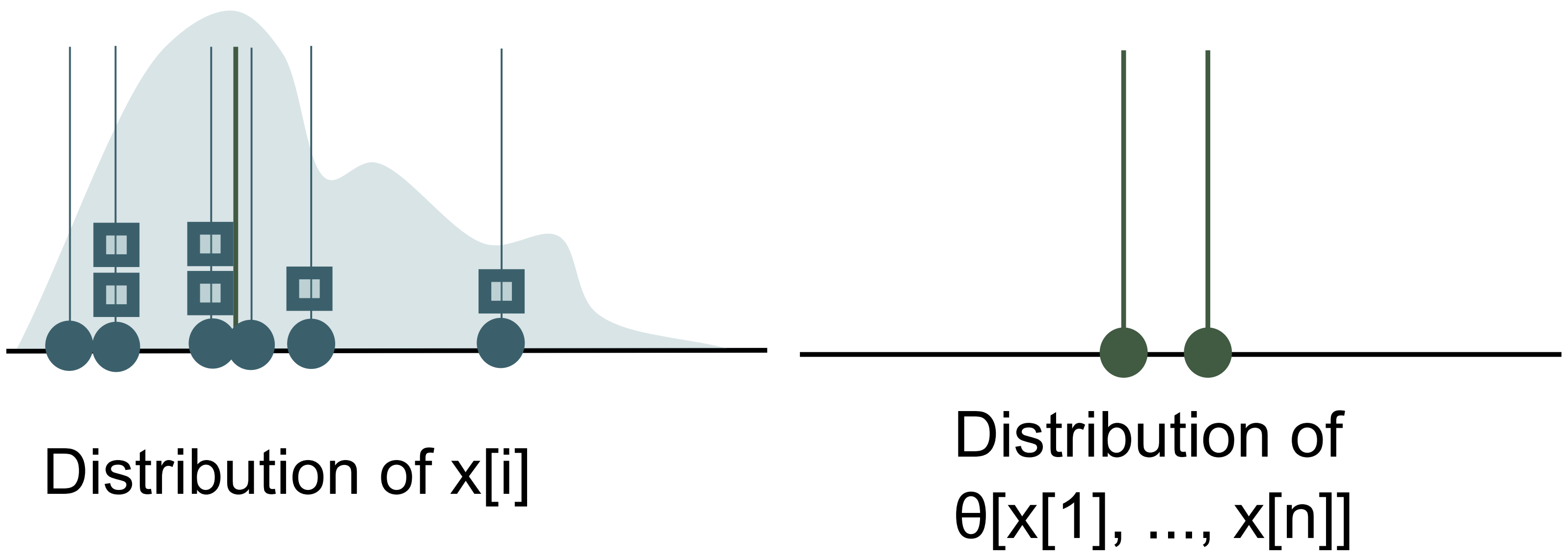
Sampling from $\hat{F}$

In reality, we can't just generate new datasets. We *can* draw samples from the empirical distribution, however.



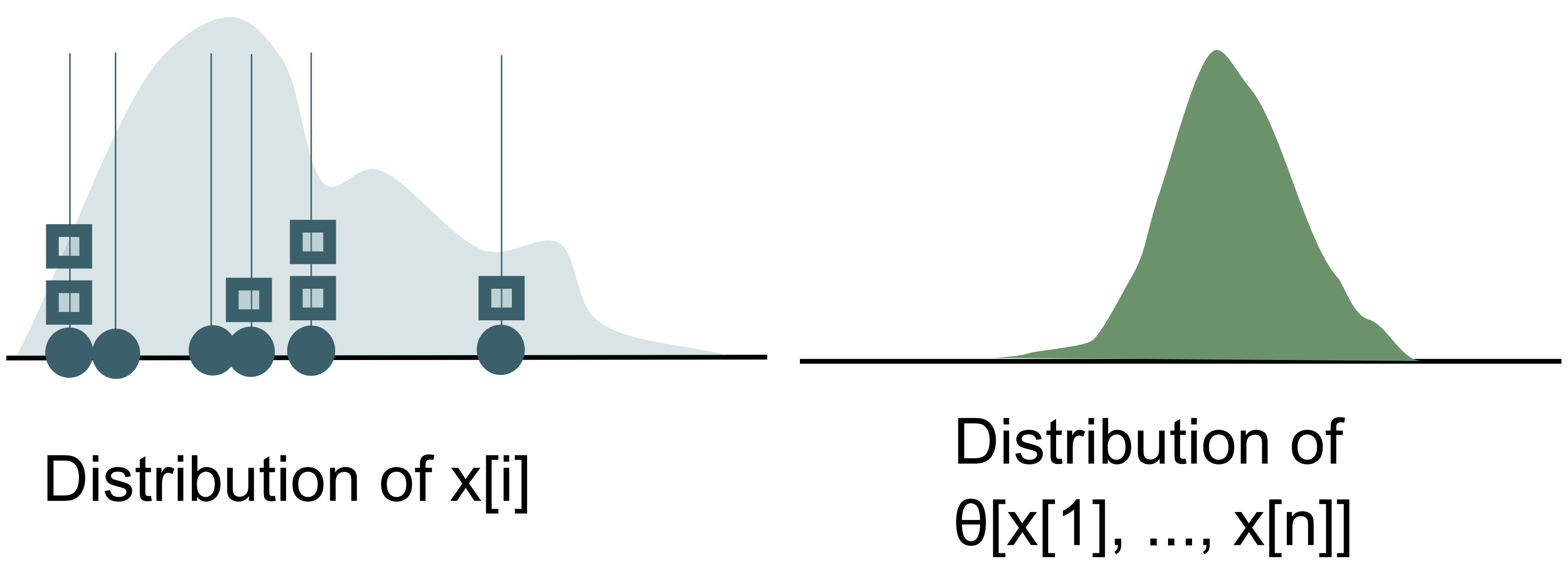
Sampling from $\hat{F}$

In reality, we can't just generate new datasets. We *can* draw samples from the empirical distribution, however.



Sampling from $\hat{F}$

In reality, we can't just generate new datasets. We *can* draw samples from the empirical distribution, however.



Bootstrap vs. Simulation Estimates

- The left are estimates $\hat{\alpha}_r$ when you simulate from F (impossible in practice), while the right are when you simulate from $\hat{F}$ (possible in practice).

simulate from $\hat{F}$ (possible in practice).

- The estimated variances are very similar

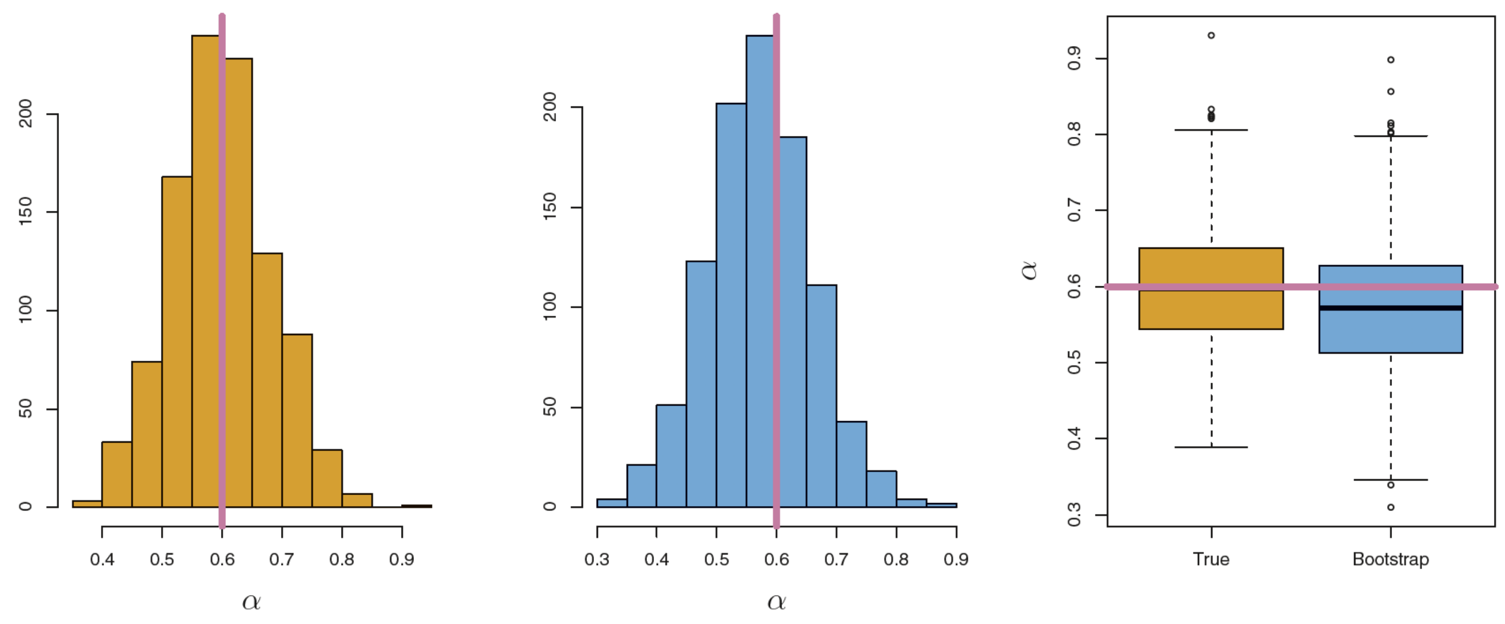


FIGURE 5.10. Left: *A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.*

The Bootstrap

Bootstrap algo, step by step

- Input: A statistic $\hat{\theta}$, number of desired simulations B
- For $b = 1, \ldots, B$,
  - Simulate $x_1^b, \ldots, x_n^b \overset{i.i.d.}{\sim} \hat{F}$
  - Compute $\hat{\theta}^b := \hat{\theta}\left(x_1^b, \ldots, x_n^b\right)$
- Estimate the variance of the original $\hat{\theta}$ by looking at the variance in the simulation output,

b is the bootstrap

$$\operatorname{Var}\left[\hat{\theta}\left(x_1, \ldots, x_n\right)\right] \approx \frac{1}{B-1} \sum_{b=1}^{B}\left(\hat{\theta}^b - \bar{\theta}\right)^2,$$

where $\bar{\theta}$ is the average of all the $\hat{\theta}^b$.

Plug-in Principle

The original estimator is constructed according to

$$F \xrightarrow{\text{sample}} x_1, \ldots, x_n \xrightarrow{\text{estimate}} \hat{\theta}\left(x_1, \ldots, x_n\right)$$

This is only done once, so you can't estimate the standard error from it alone.

Plug-In Principle

If we plug in $\hat{F}$ for F, we get

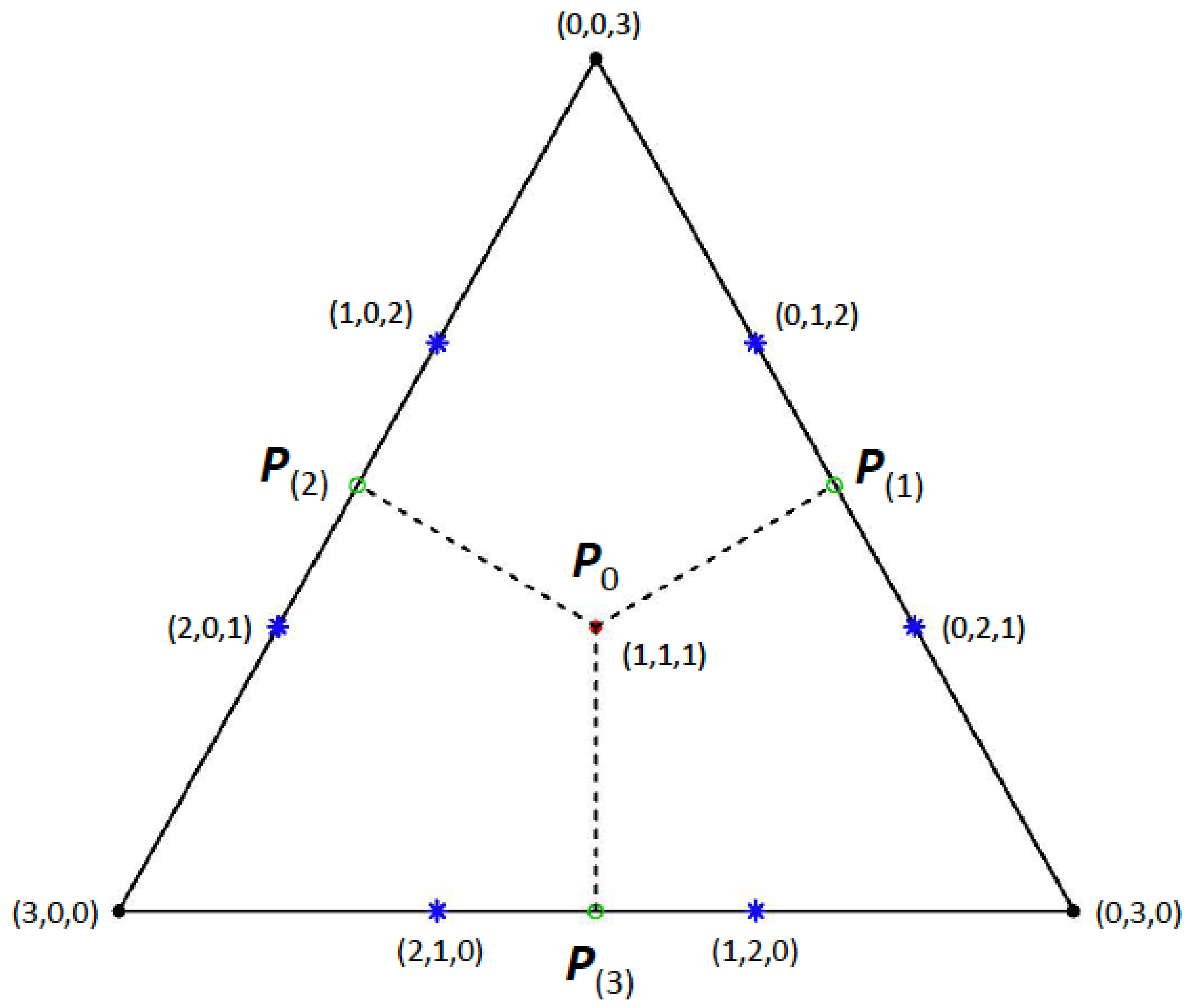
$$\hat{F} \xrightarrow{resample} x_1^b, \ldots, x_n^b \xrightarrow{estimate} \hat{\theta}\left(x_1^b, \ldots, x_n^b\right) := \hat{\theta}^b$$

This can be done as much as we want, so we can estimate the variance across $\hat{\theta}^b$.

The more is the number of samples, the better the plug-in principle works.

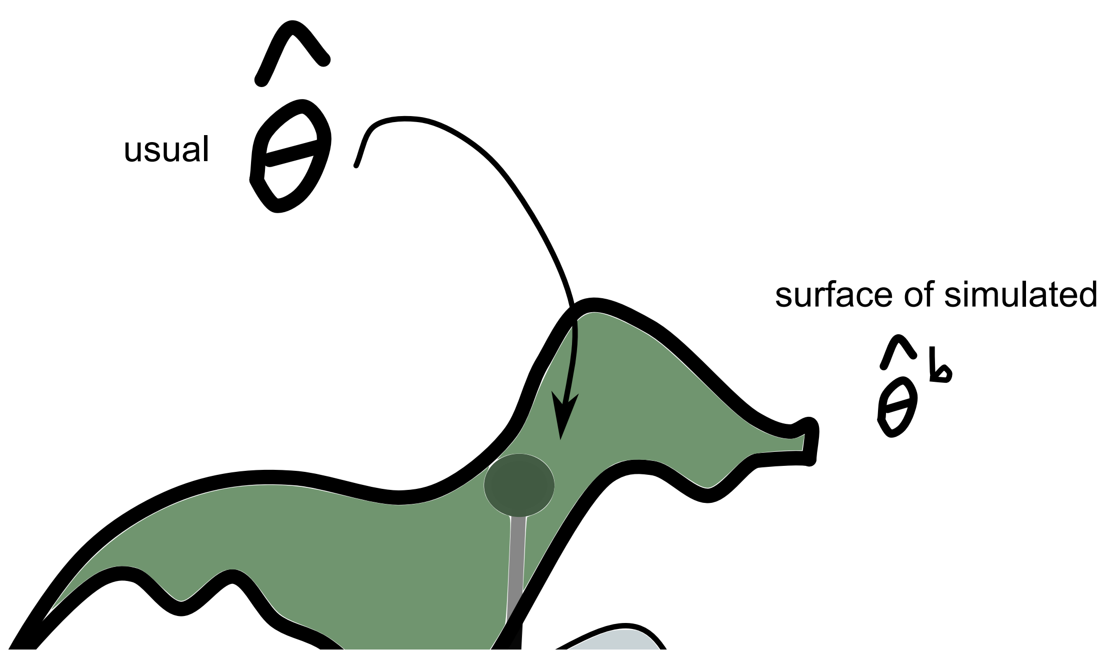
The Resampling Perspective

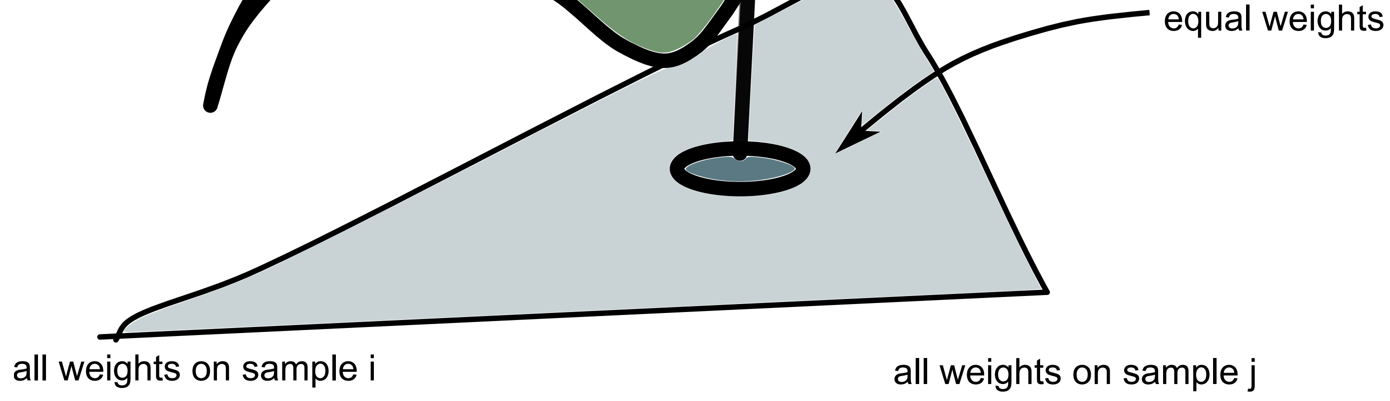
- Draws from $\hat{F}$: up or downweight original samples.
  - Some x_i may not be included, other might be included multiple times
  - View this as points on the simplex (space of weights that sum to n)



The Resampling Perspective

- The bootstrap measures the sensitivity of $\hat{\theta}$ to different weightings of the original points





Bootstrap Confidence Intervals

- Optional read: Bootstrap CI
- A lot of times, we use estimates of the variance to build confidence intervals
- Larger sample size leads to tighter confidence intervals
- If $\hat{\theta}$ is approximately normal, it can be shown that

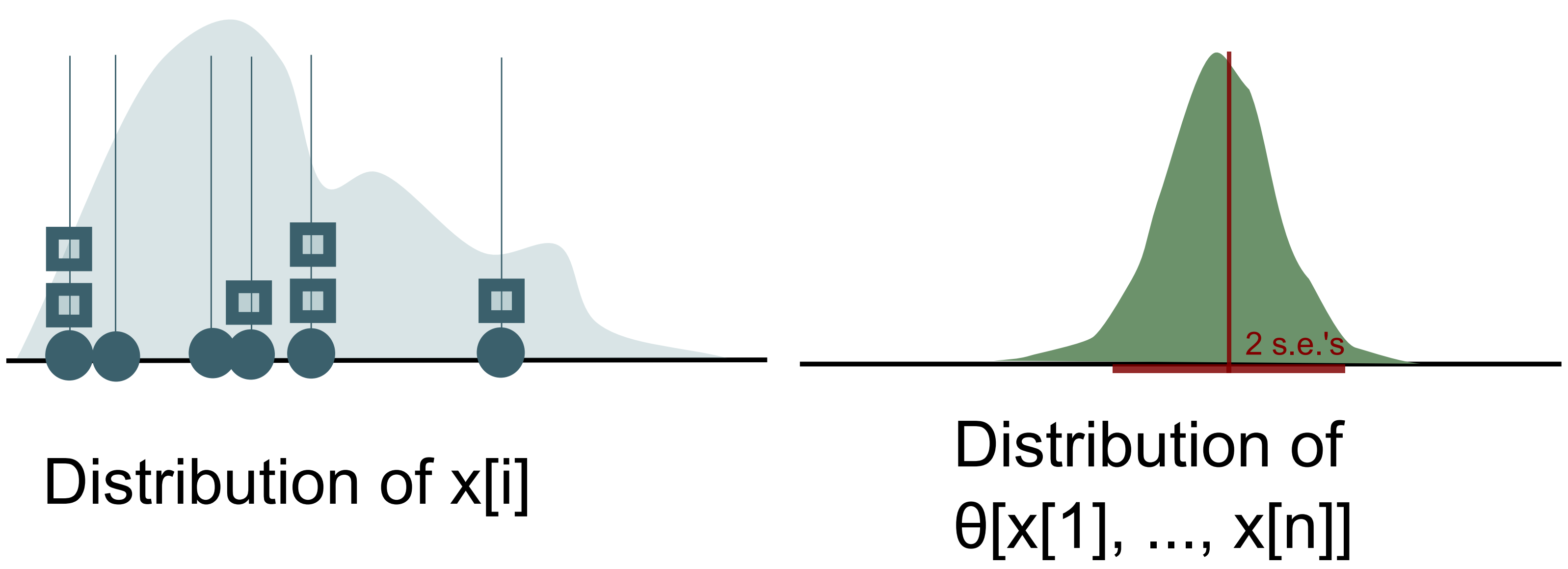
$$\hat{\theta} \pm 1.96\sqrt{\operatorname{Var}\left(\hat{\theta}\right)}$$

  is a valid confidence interval

Approach 1

Therefore, if we can use the bootstrap to estimate the variance, we can directly use it to make a confidence interval

$$\hat{\theta} \pm 1.96\sqrt{\widehat{\operatorname{Var}}\left(\hat{\theta}\right)}$$

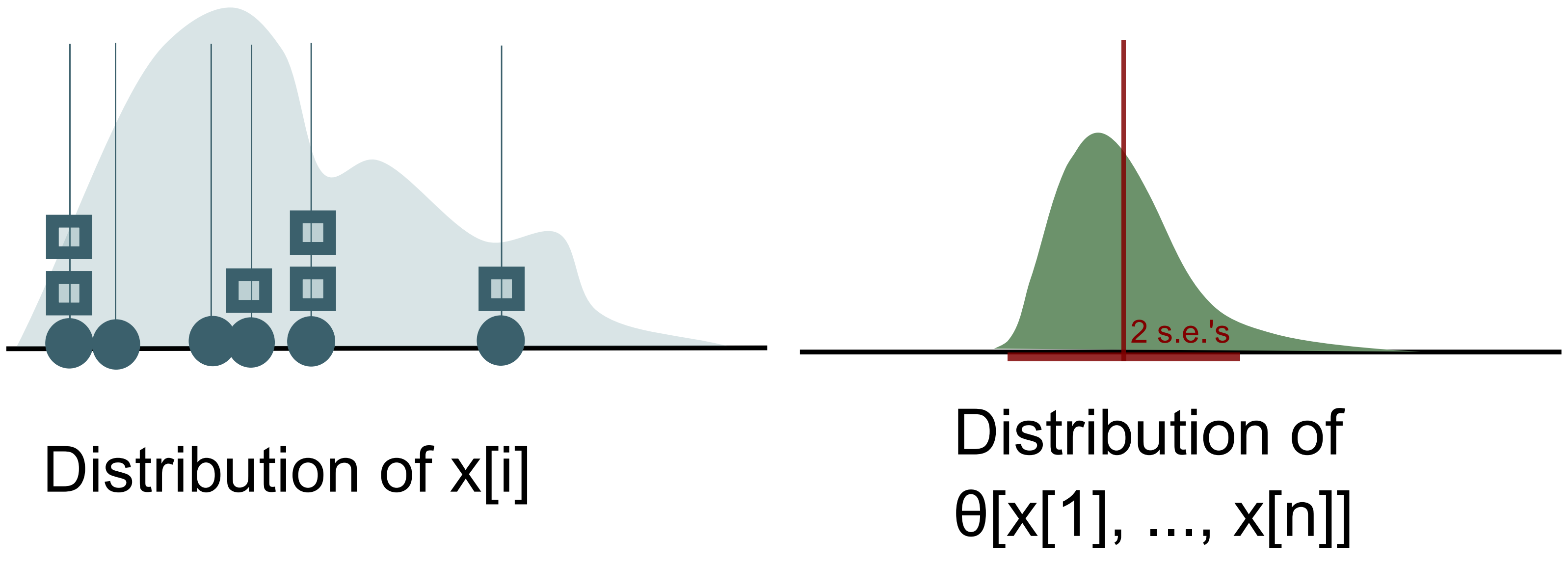


Approach 2

- What if $\hat{\theta}$ isn't approximately normal?
- In classical stats, you'd need new theory to find an alternative confidence interval
- However, the bootstrap gives us access to something close to the distribution of $\hat{\theta}$

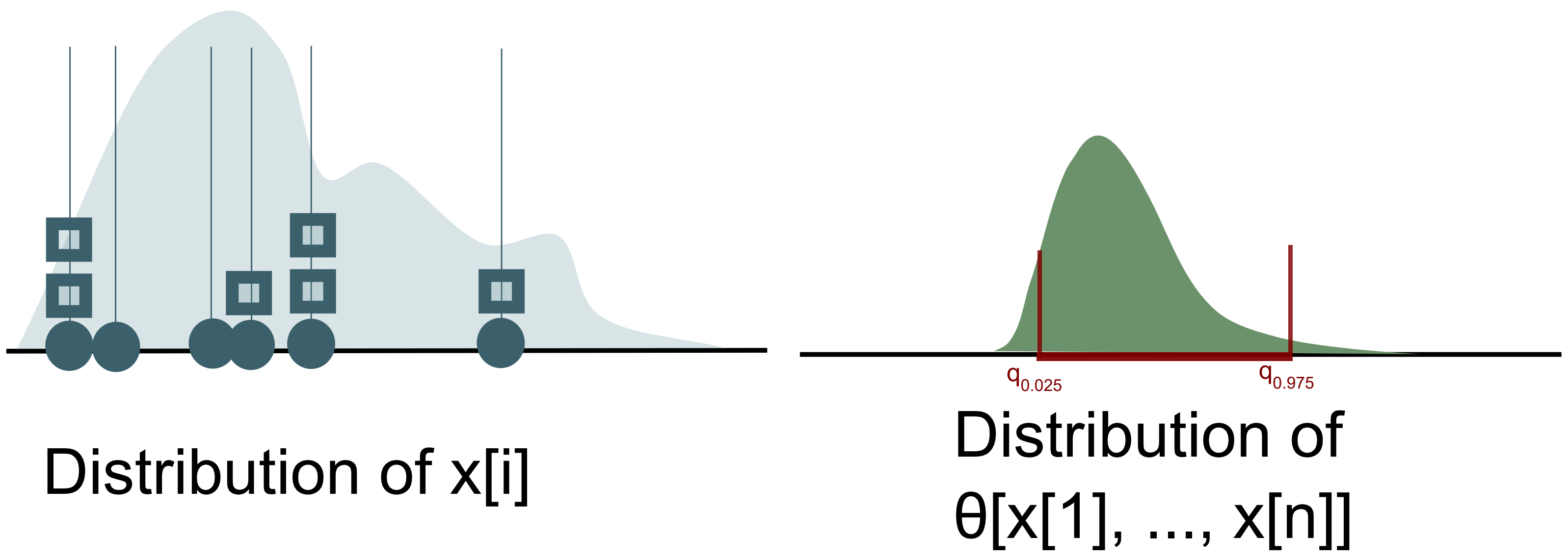
to the distribution of $\hat{\theta}$:

- Get more samples and apply (bootspraр) central limit theorem.
- Use Bias corrected and accelerated bootstrap method



Approach 2

- Main idea is to directly use quantiles of the simulated $\hat{\theta}^b$
- No longer requires normality (or even symmetry)
- However, requires more simulation samples, since quantiles are harder to estimate than variances



Approach 2

- Main idea is to directly use quantiles of the simulated $\hat{\theta}^b$
- No longer requires normality (or even symmetry)
- However, requires more simulation samples, since quantiles are harder to estimate than variances

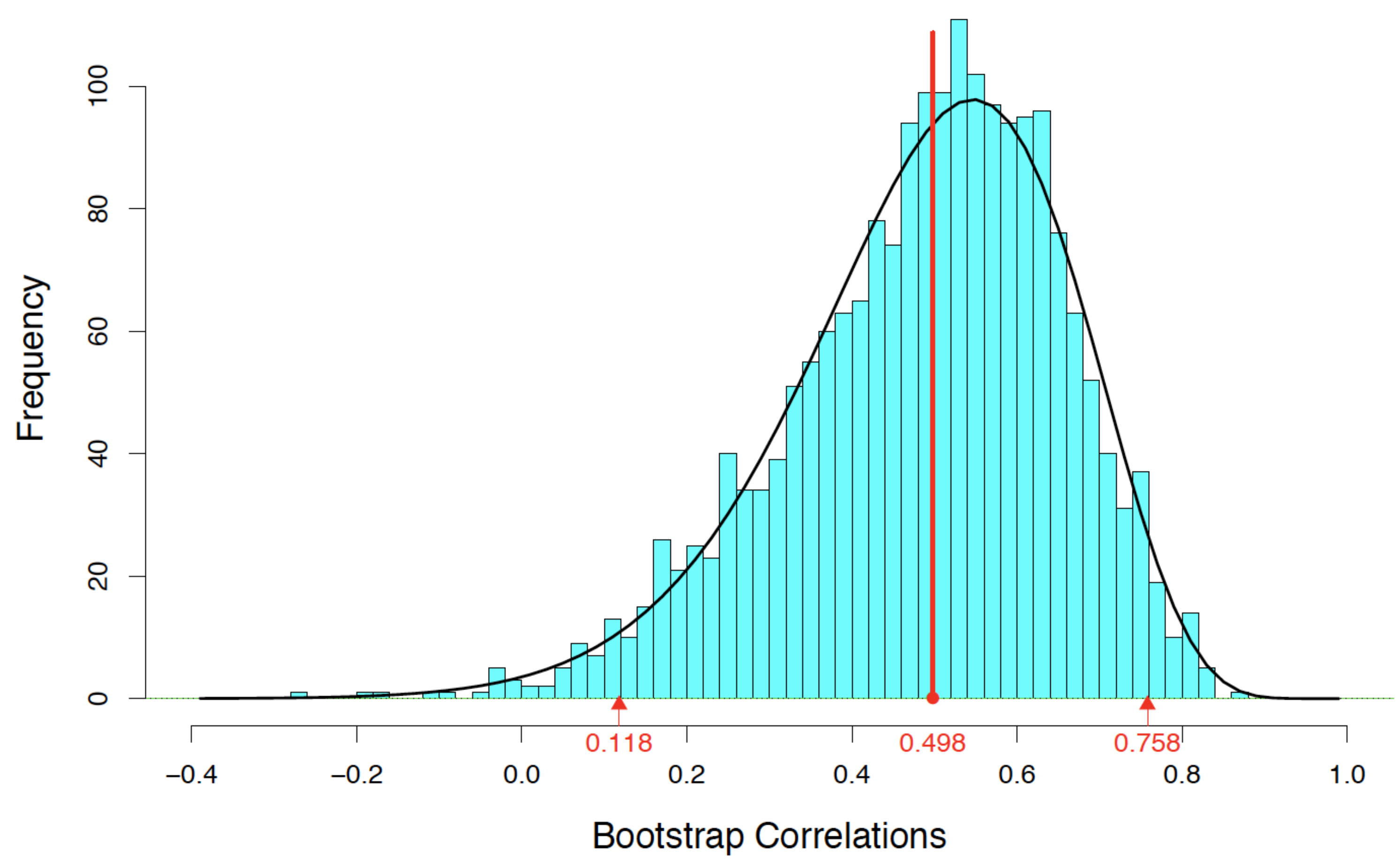


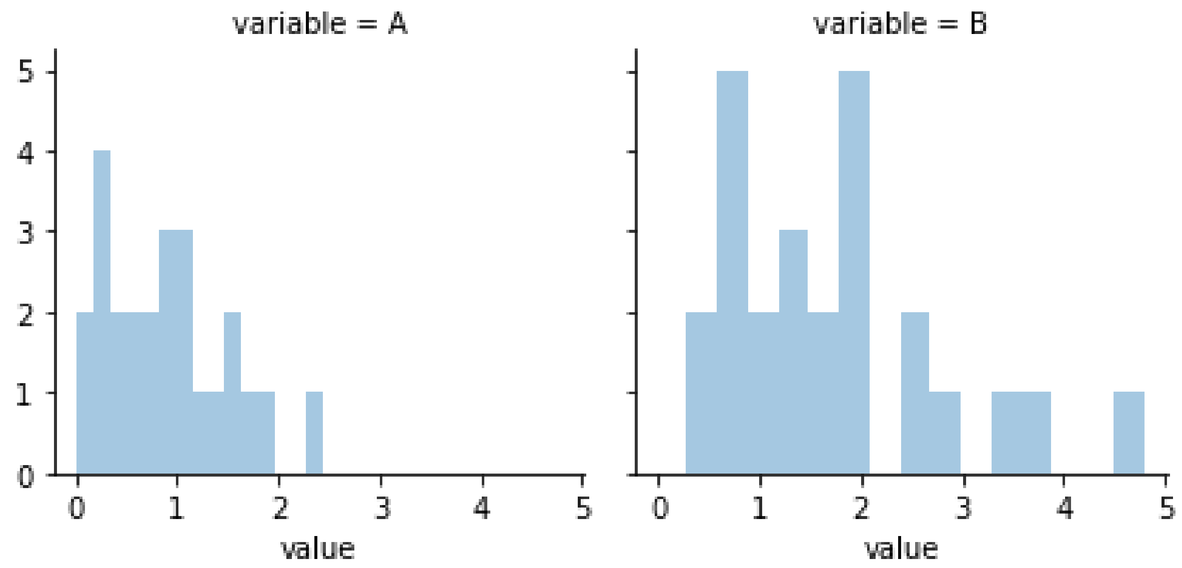
Figure 11.3 Histogram of $B = 2000$ nonparametric bootstrap replications $\hat{\theta}^*$ for the student score sample correlation; the solid curve is the ideal parametric bootstrap distribution $f_{\hat{\theta}}(r)$ as in

Figure 11.1. Observed correlation $\hat{\theta} = 0.498$. Small triangles show histogram's 0.025 and 0.975 quantiles.

- Python (e.g., to get 95% confidence interval):
  numpy.percentile(bootstrap_dist, 2.5)
  numpy.percentile(bootstrap_dist, 97.5)

Additional Examples: Difference in Means

- Suppose we want to test the difference in means between two groups.
- We can define a reference distribution using the bootstrap
- Idea is to sample repeatedly from $\hat{F}_1$ and $\hat{F}_2$, and look at the distribution in the difference in means



Additional Examples: Difference in Means

- Suppose we want to test the difference in means between two groups.
- We can define a reference distribution using the bootstrap
- Idea is to sample repeatedly from $\hat{F}_1$ and $\hat{F}_2$, and look at the distribution in the difference in means

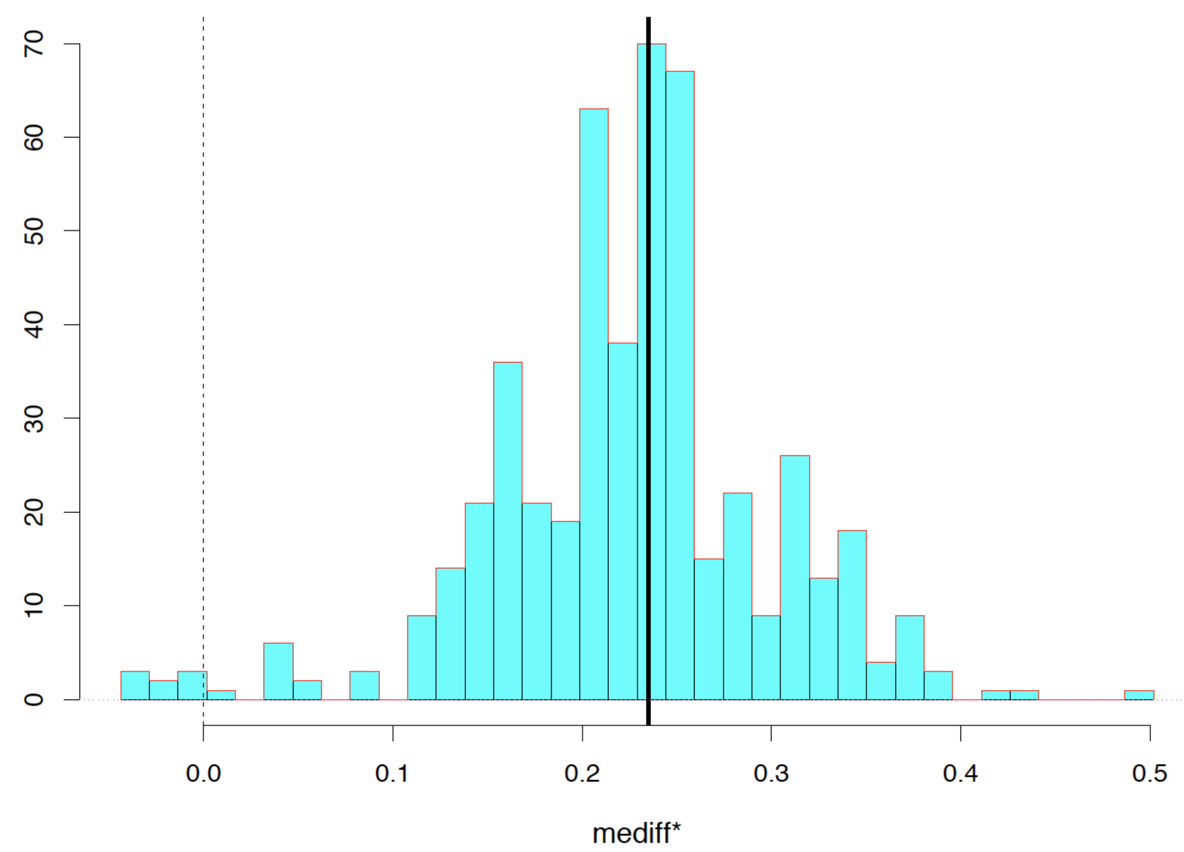
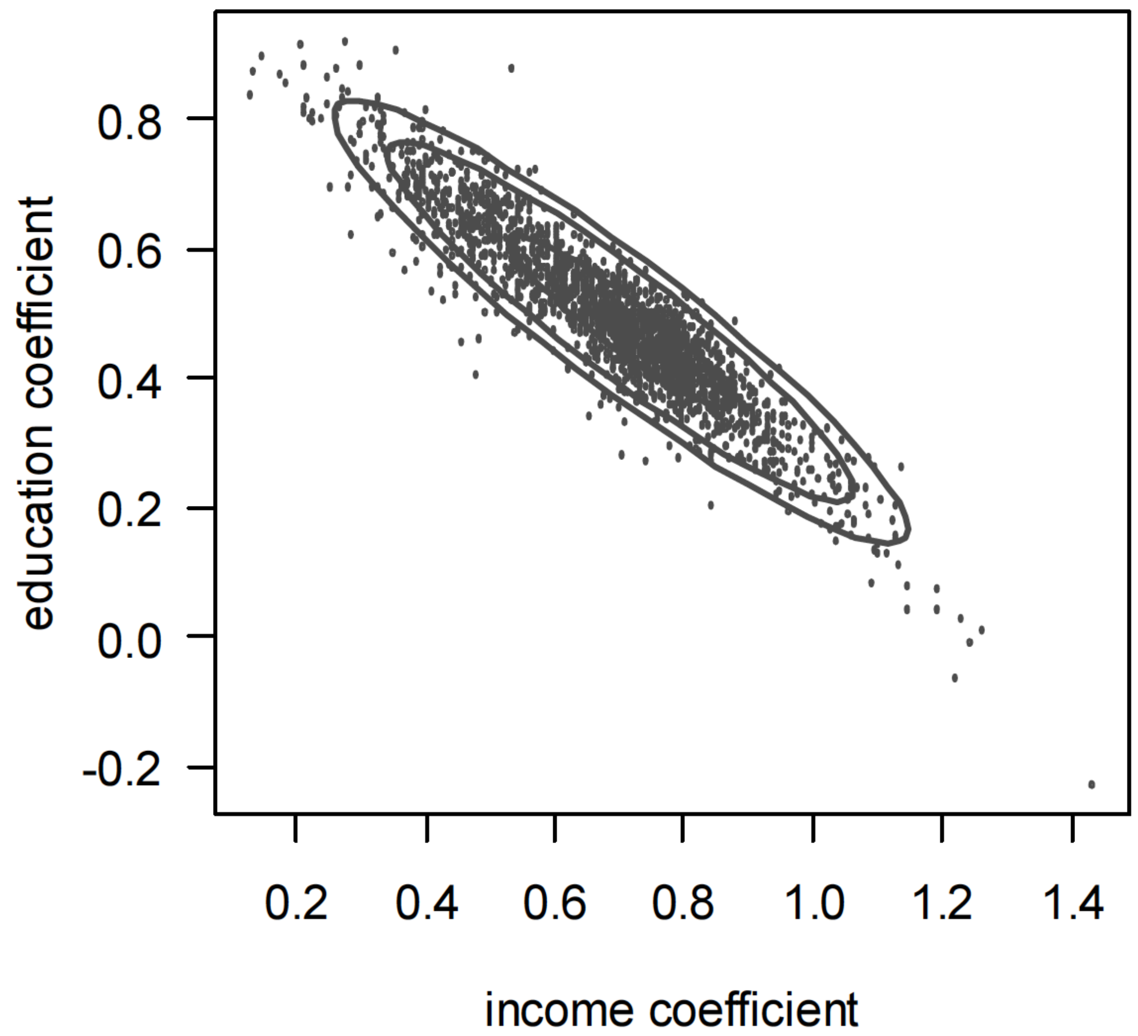


Figure 10.4 $B = 500$ bootstrap replications for the median difference between the **AML** and **ALL** scores in Figure 1.4, giving $\widehat{\text{se}}_{\text{boot}} = 0.074$. The observed value **mediff** $= 0.235$ (vertical black line) is more than 3 standard errors above zero.

Additional Examples: Regression

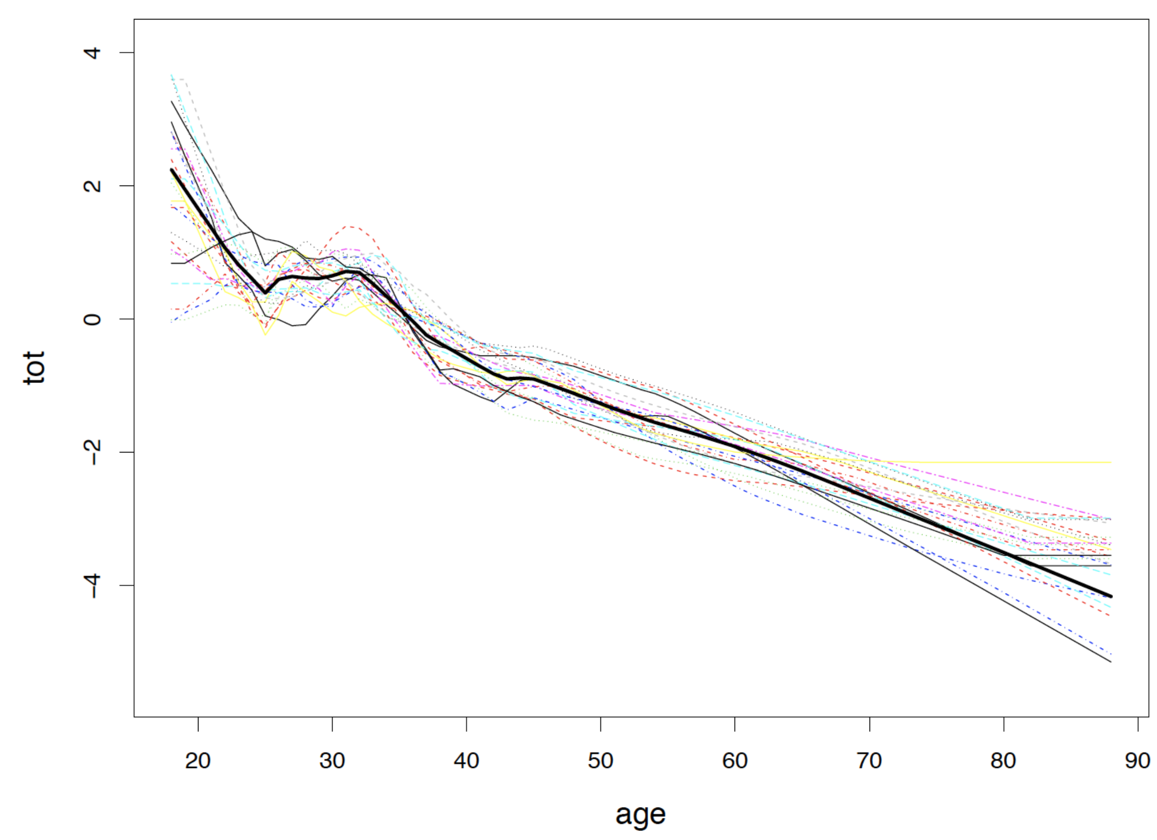
- You can bootstrap regression models as well
- Fit the regression many times, across resampled versions of the data, evaluate confidence intervals
- Works for variants of regression with no analytical s.e. formula



Ref: Bootstrapping Regression Models

Additional Examples: Regression

It can even be applied to nonparametric regression models, e.g., lowess regressions,



and even Random Forests.

- Fit models using bootstrap samples and evaluate the confidence intervals on the test statistic

```
import {slide} from @mbostock/slide
```

```
<style>
```