Search

Fork Sign in

👋 Welcome. This is live code! Click the left margin to view or edit.

Kris Sankaran

2

Published Sep 30, 2019

# The Bootstrap

IFT6758, Fall 2019

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."

## Example: Sample Mean

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 $\sigma^2$. Then,

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

$$= \frac{1}{n^2} n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

## Estimated Standard Error

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

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

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

the raw data.

## Estimated Standard Error

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

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

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

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

# Abstraction

This is a more abstract way of describing this process.

1. Define a statistic $\hat{\theta}\left(x_1, \dots, x_n\right)$ 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 $\log \mu$ or $\frac{\alpha}{\beta}$ in some model
- You ran some iterative algorithm to compute $\hat{\theta}$
  - Robust regression
  - 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 $\alpha$ 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

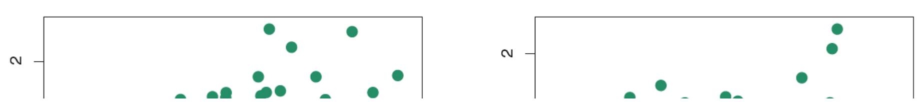
- $\alpha$ is unknown in practice, so we estimate it,

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

- 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?



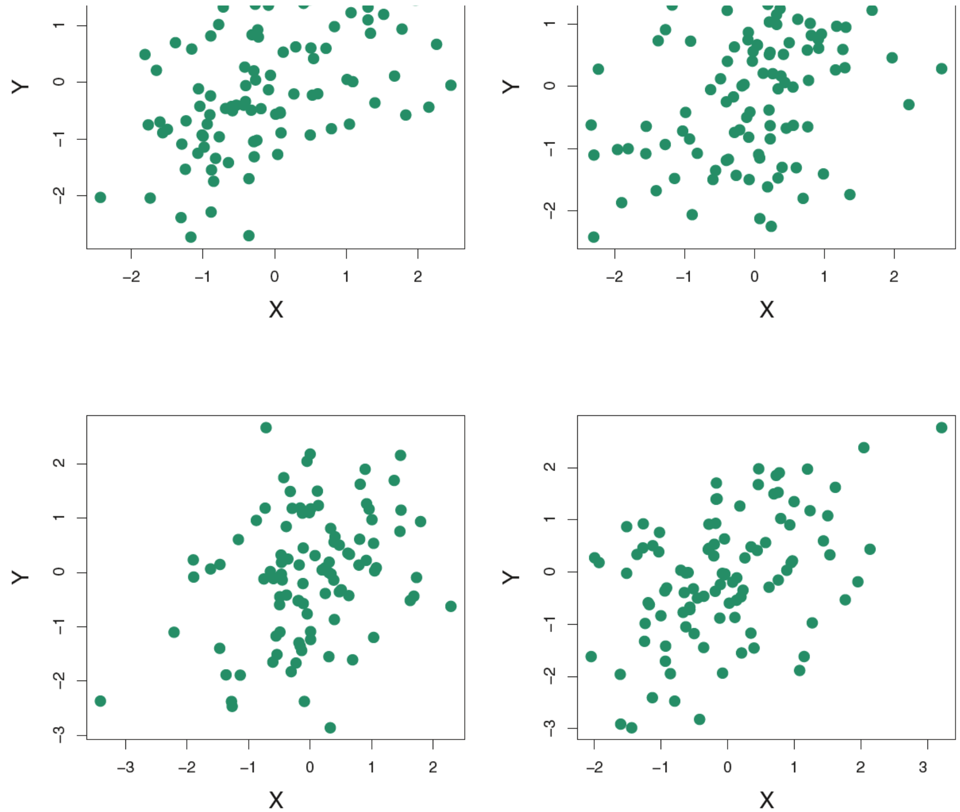



**FIGURE 5.9.** *Each panel displays* 100 *simulated returns for investments* $X$ *and* $Y$*. From left to right and top to bottom, the resulting estimates for* $\alpha$ *are* 0.576, 0.532, 0.657, *and* 0.651.

# Thought Experiment

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

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

which (unsurprisingly) is very close to the underlying 0.6.

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

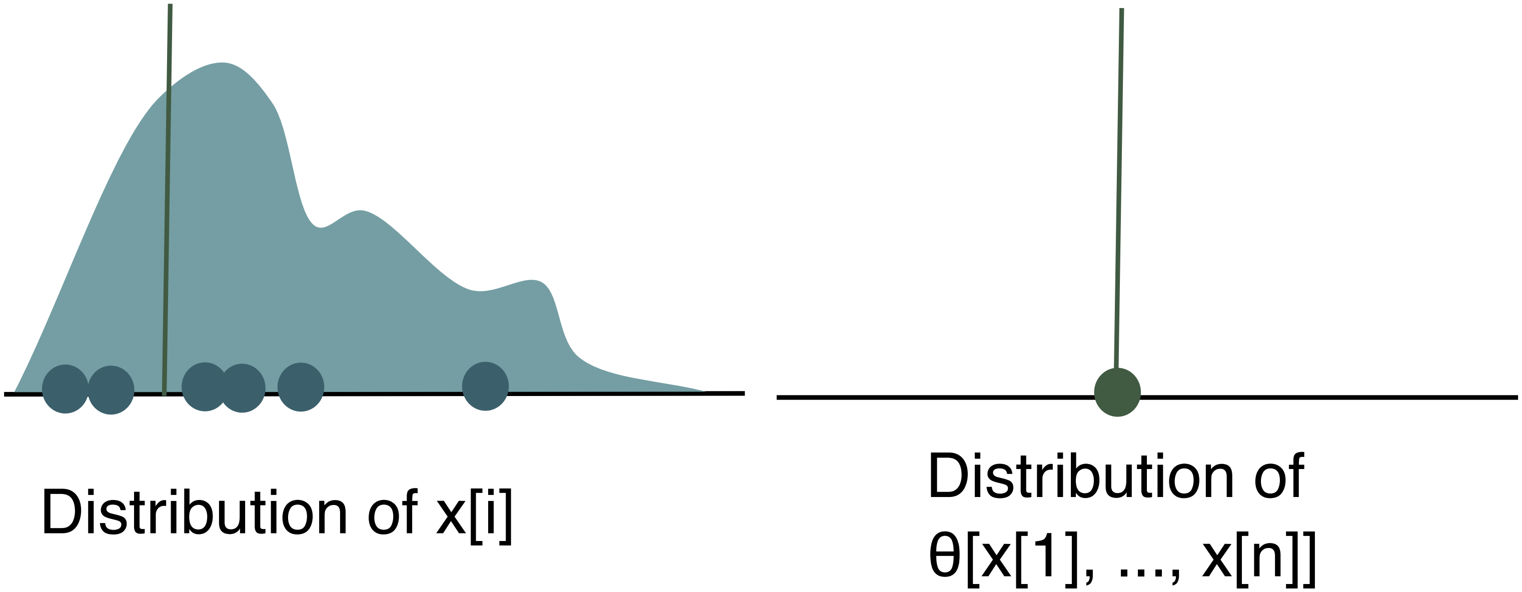
$$\widehat{\mathbf{Var}}\left[\hat{\alpha}\right] = \frac{1}{999} \sum_{r=1}^{1000} \left(\hat{\alpha}_r - \bar{\alpha}\right)^2 \approx 0.083^2,$$

## Bootstrap Idea

- 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*

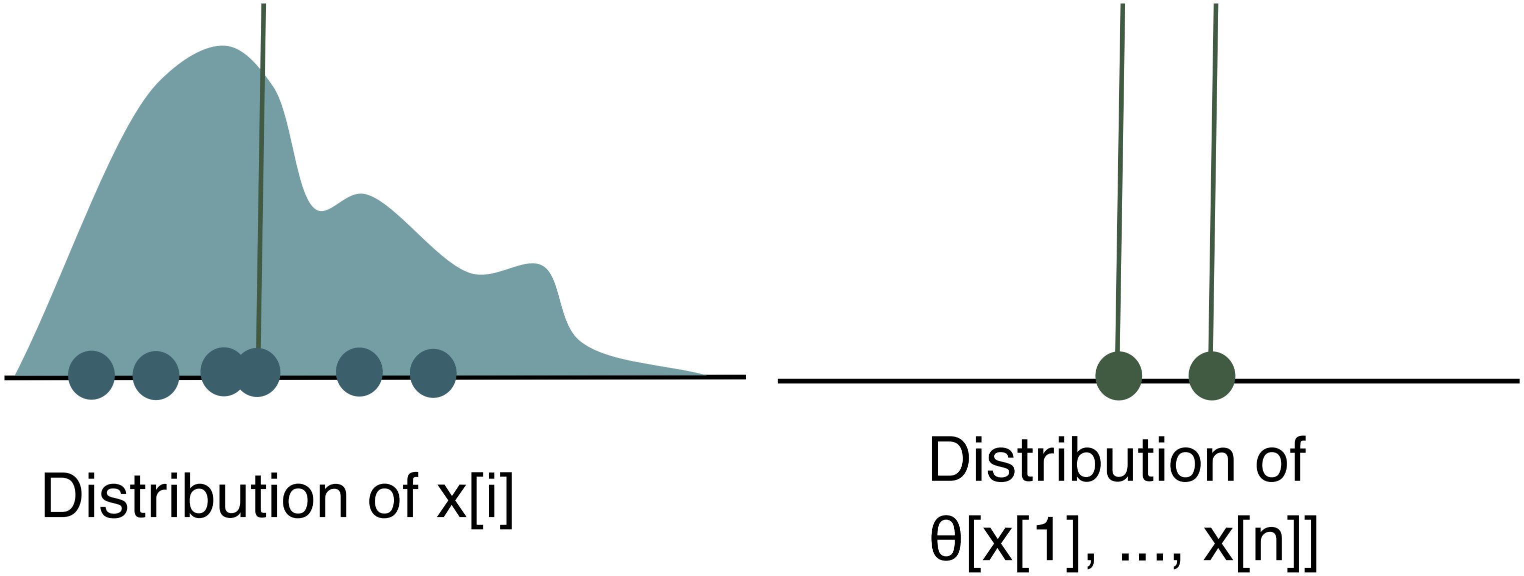
## Sampling from $F$

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



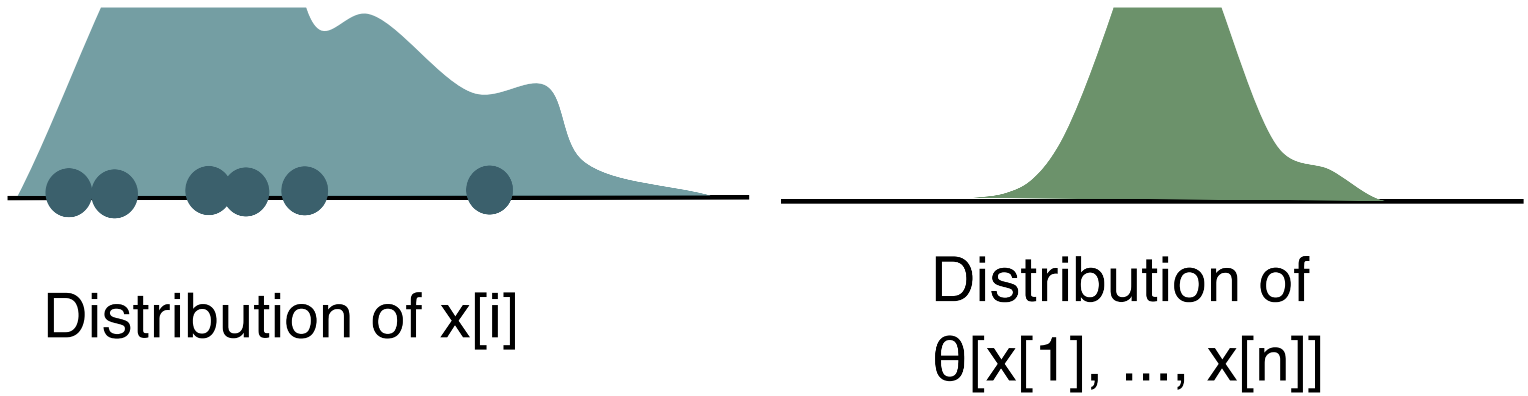

## Sampling from $F$

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



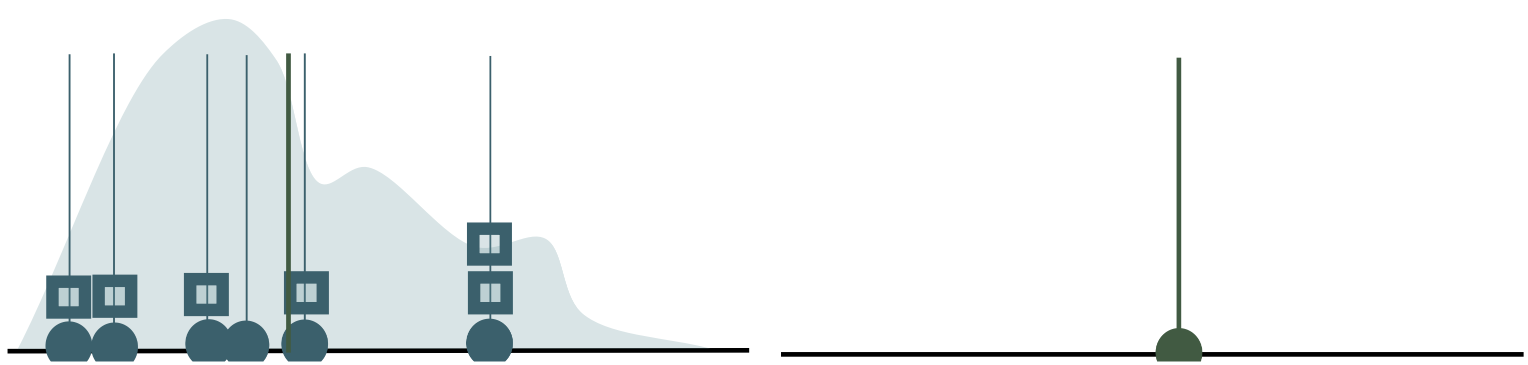

## Sampling from $F$

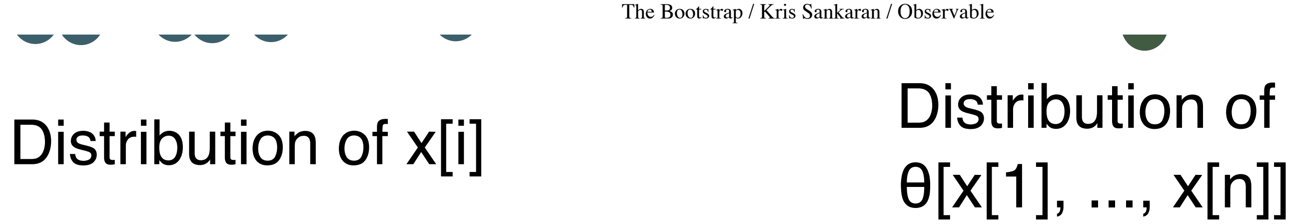
The properties of the final sampling distribution for $\hat{\theta}$ can be used to guide inference.




## Sampling from $\hat{F}$

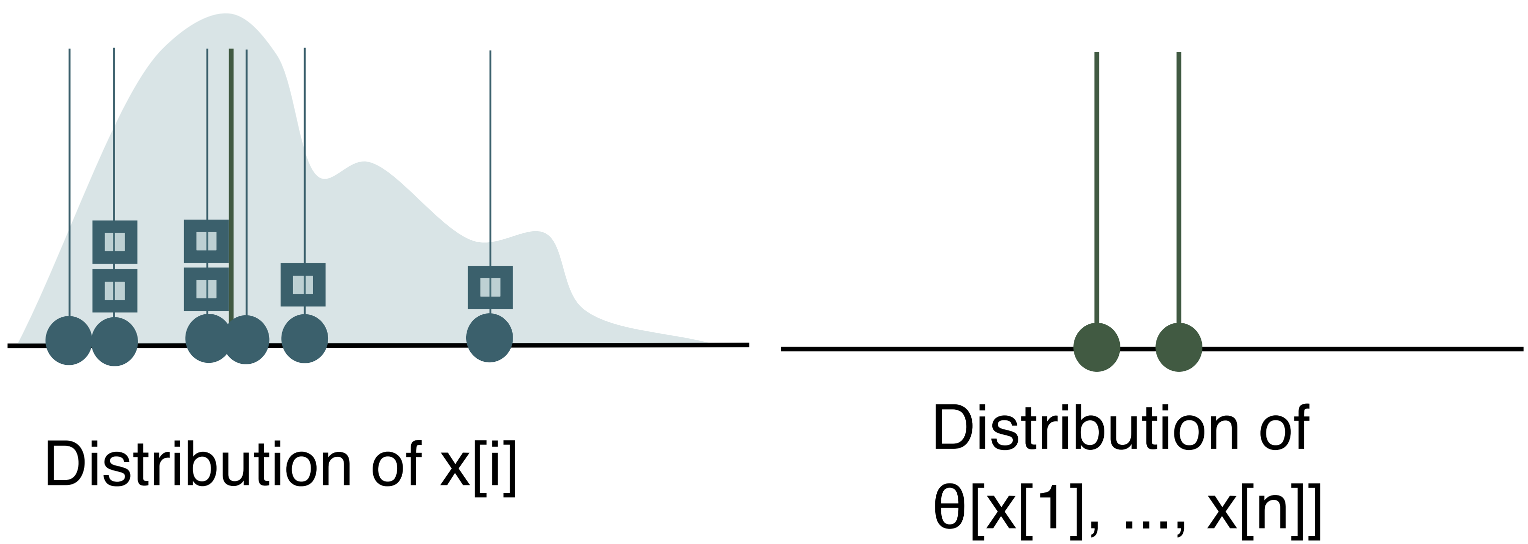
In reality, we can't just generate new datasets. We *can* draw samples from the empirical distribution, however.





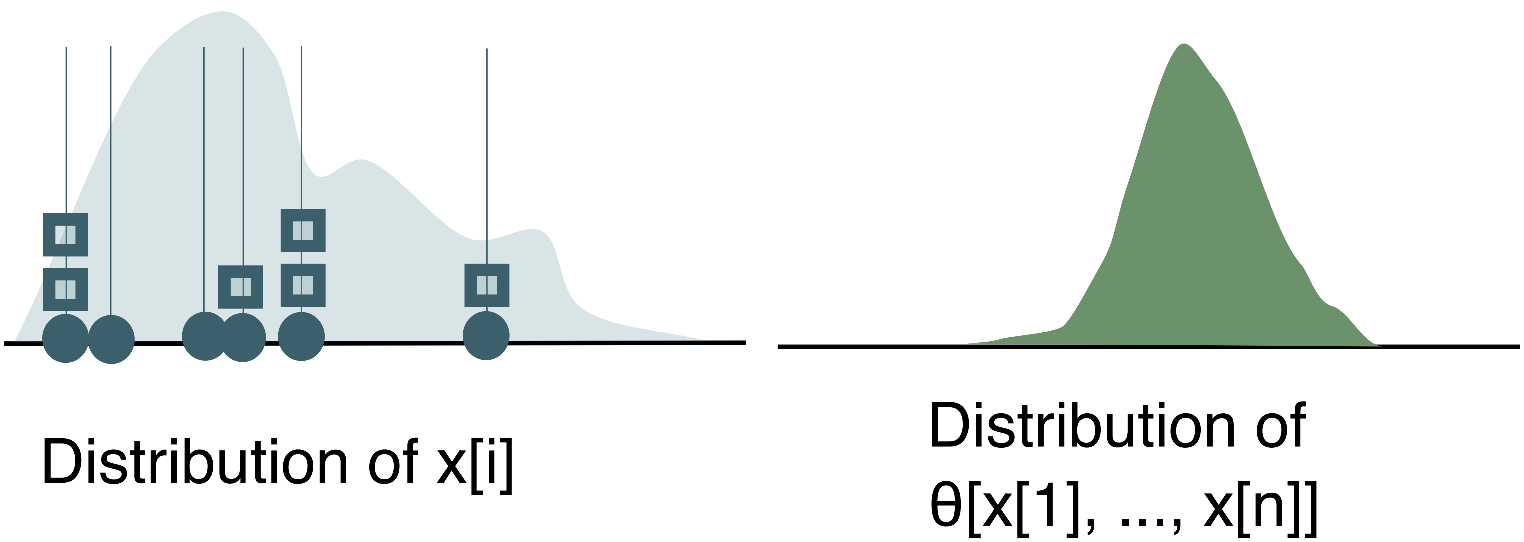

## Sampling from $\hat{F}$

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).
- The estimated variances are very similar

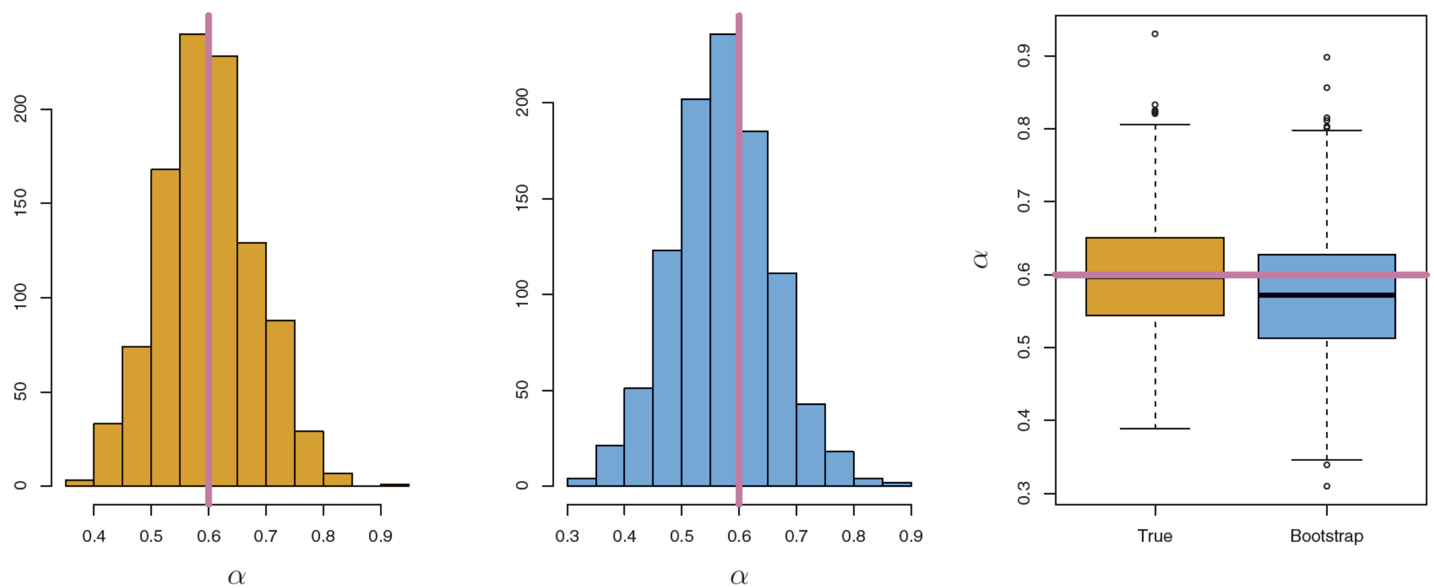



**FIGURE 5.10.** Left: *A histogram of the estimates of $\alpha$ obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the*

*estimates of $\alpha$ obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of $\alpha$ displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of $\alpha$.*

## The Bootstrap

- 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,

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

$$b=1$$

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

## Plug-in Principle

The original estimator is constructed according to

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

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

## Plug-In Principle

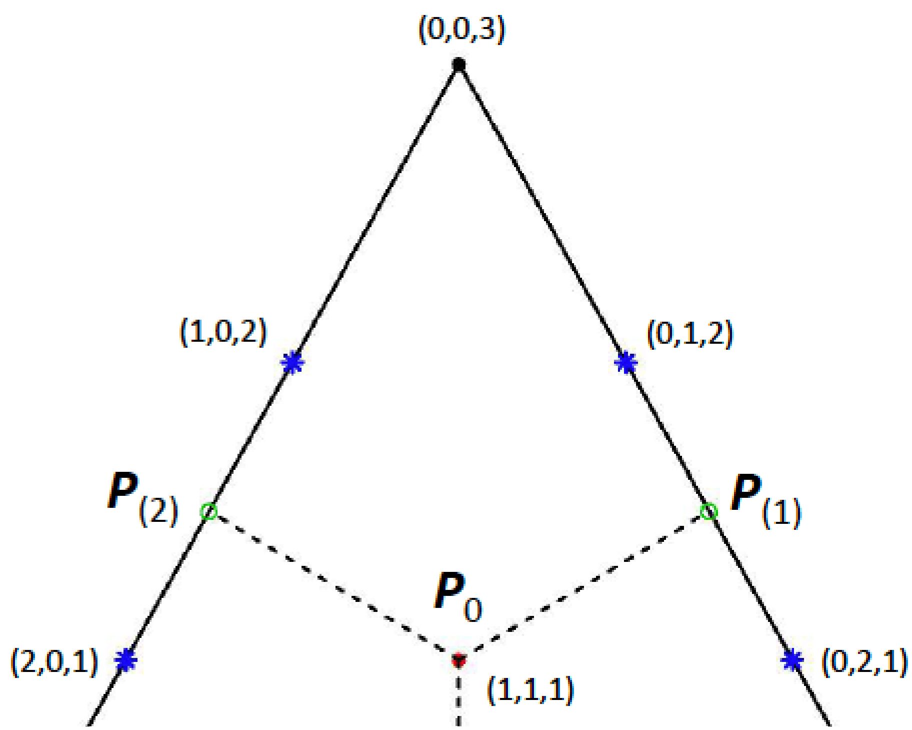
If we plug in $\hat{F}$ for $F$, we get

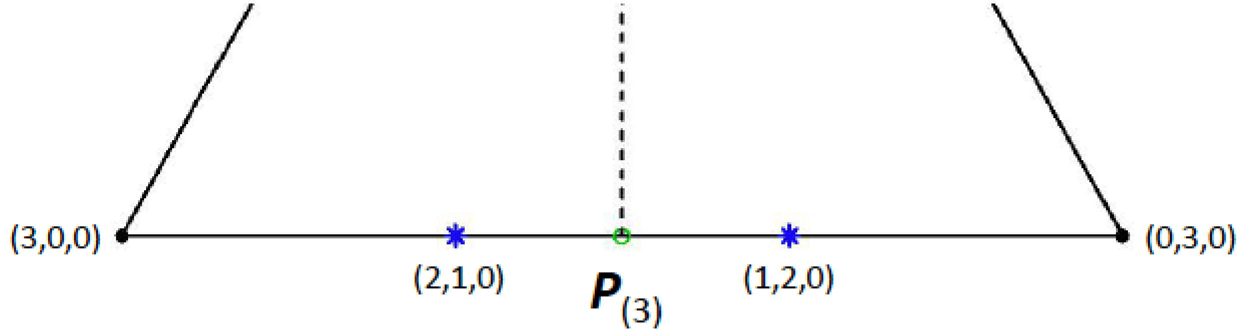
$$\hat{F} \xrightarrow{resample} x_1^b, \dots, x_n^b \xrightarrow{estimate} \hat{\theta}\left(x_1^b, \dots, x_n^b\right) := \hat{\theta}^b$$

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

## The Resampling Perspective

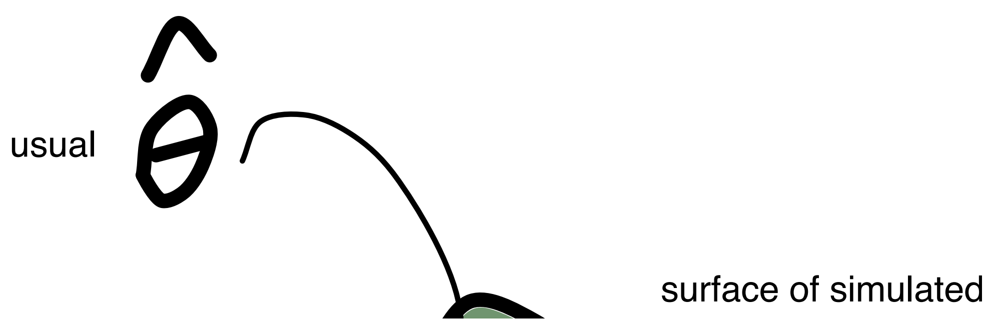
- Draws from $\hat{F}$: up or downweight original samples.
  - Some $x_i$ may not be included, other might be included multiple times
  - View this as points on the simplex (space of weights that sum to $n$)

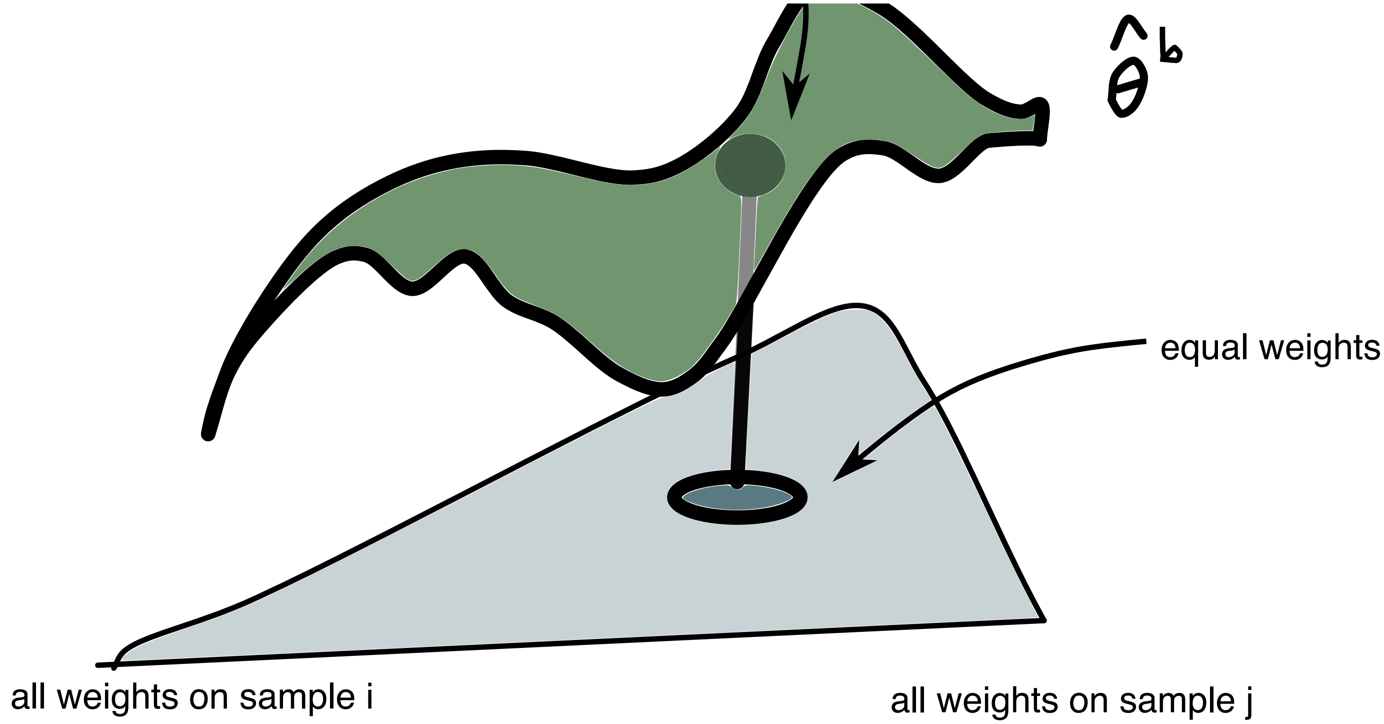





## The Resampling Perspective

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






# Bootstrap Confidence Intervals

- A lot of times, we use estimates of the variance to build confidence intervals confidence intervals

- If $\hat{\theta}$ is approximately normal, it can be shown that

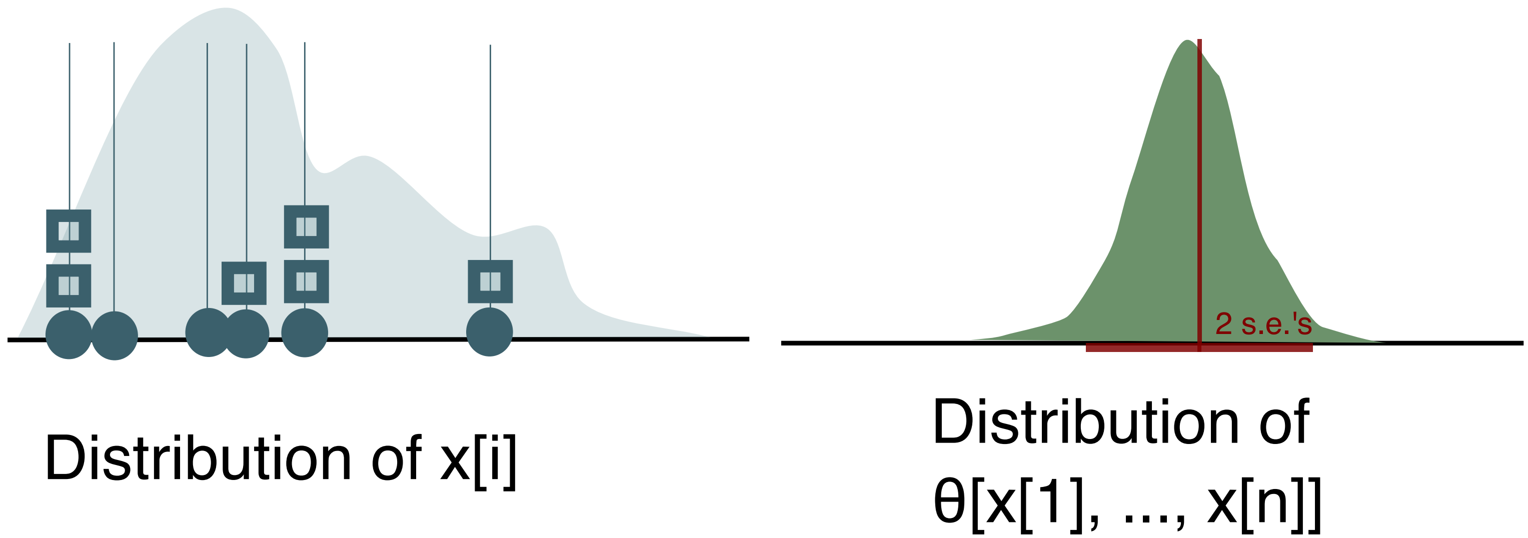
$$\hat{\theta} \pm 1.96\sqrt{\text{Var}\left(\hat{\theta}\right)}$$

  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

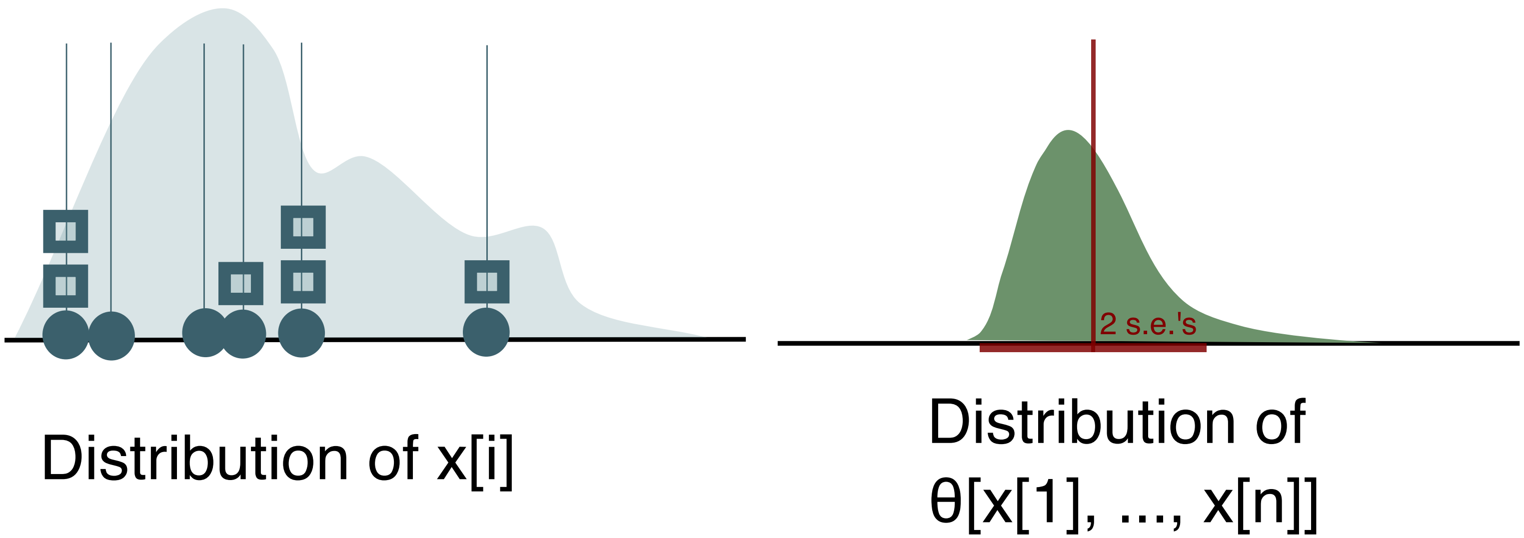
$$\hat{\theta} \pm 1.96\sqrt{\widehat{\text{Var}}\left(\hat{\theta}\right)}$$




## 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 $\hat{\theta}$

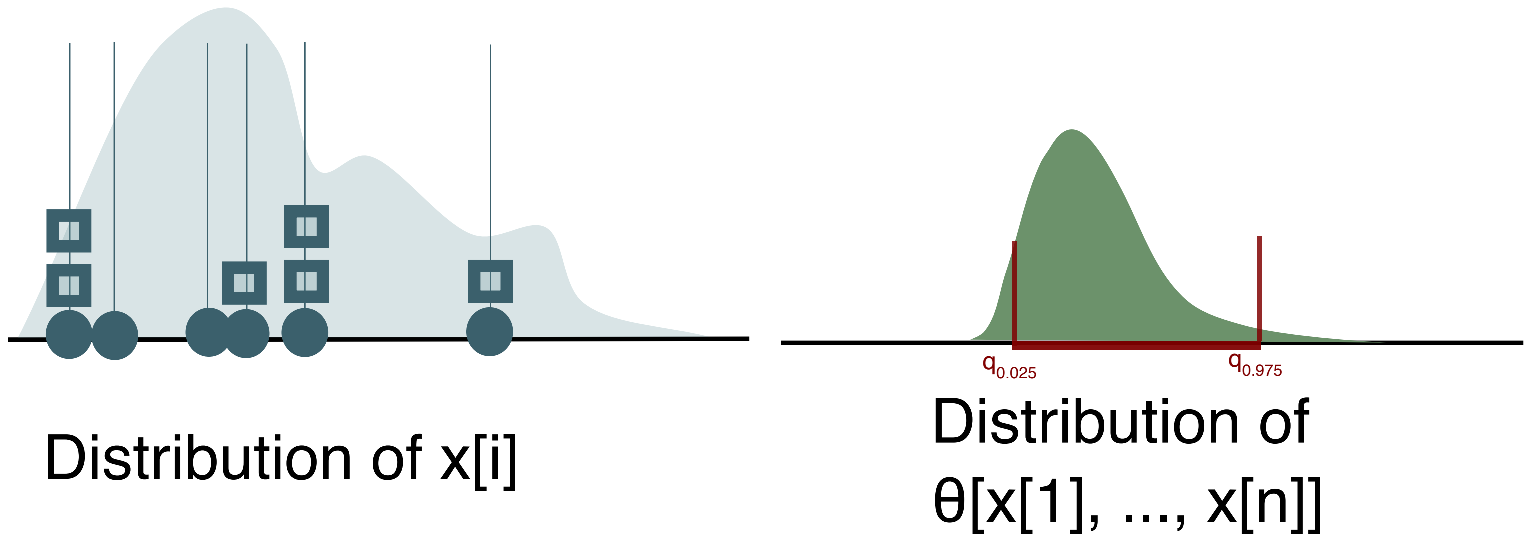
to the distribution of $\theta$




## 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

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

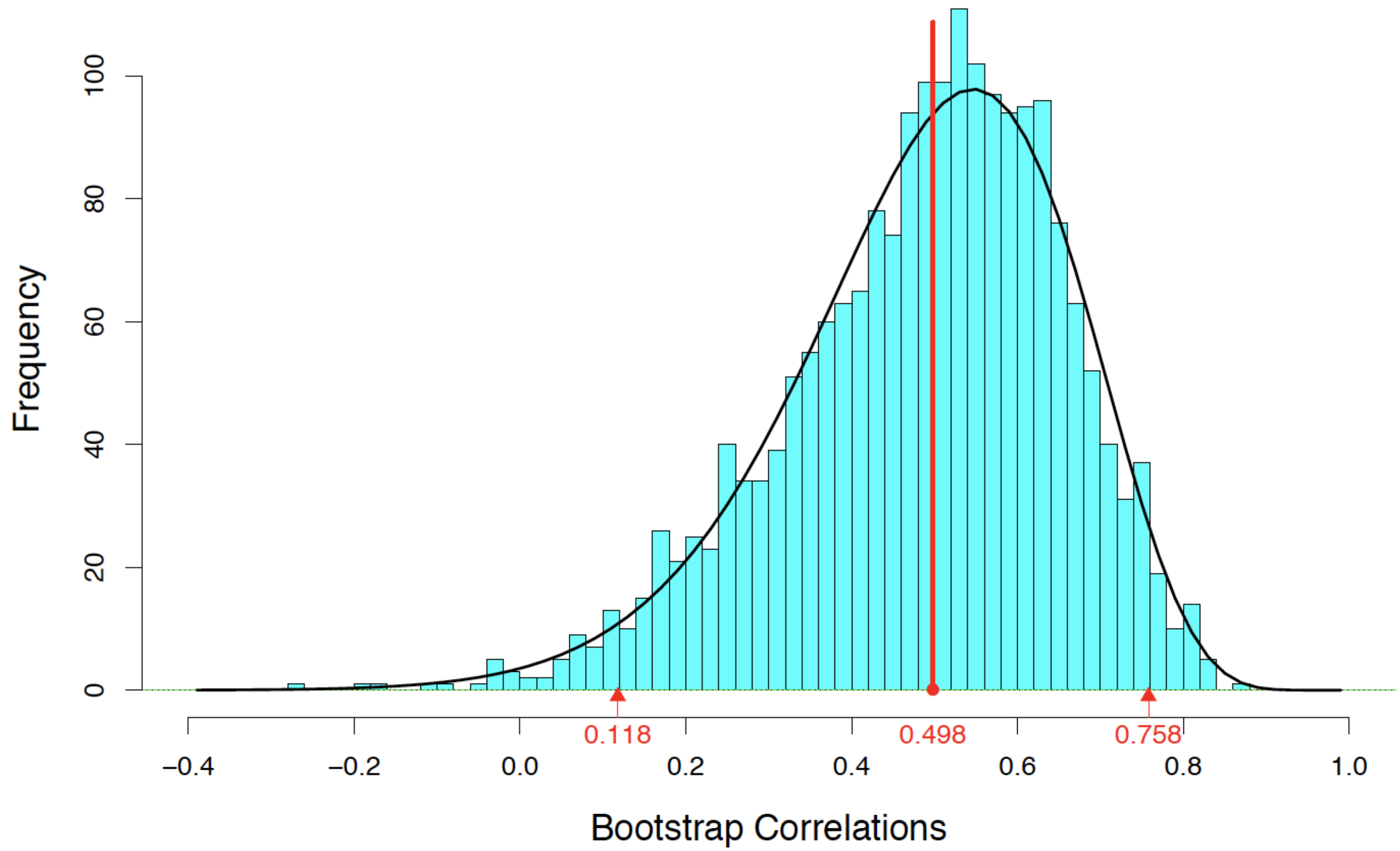
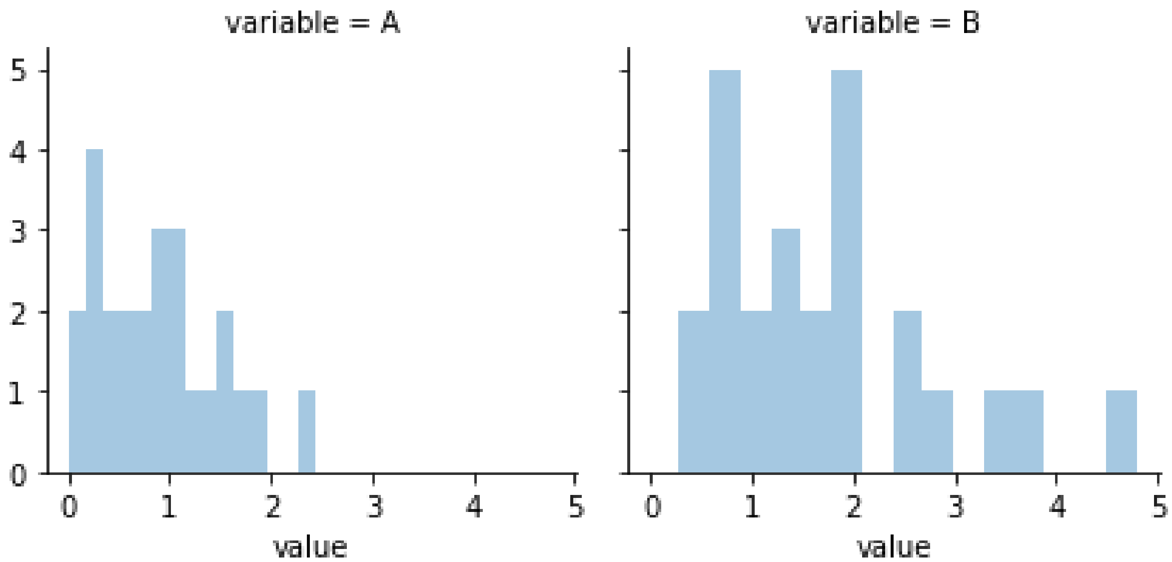



**Figure 11.3** Histogram of $B = 2000$ nonparametric bootstrap replications $\hat{\theta}^*$ for the student score sample correlation; the solid curve is the ideal parametric bootstrap distribution $f_{\hat{\theta}}(r)$ as in Figure 11.1. Observed correlation $\hat{\theta} = 0.498$. Small triangles show histogram's 0.025 and 0.975 quantiles.

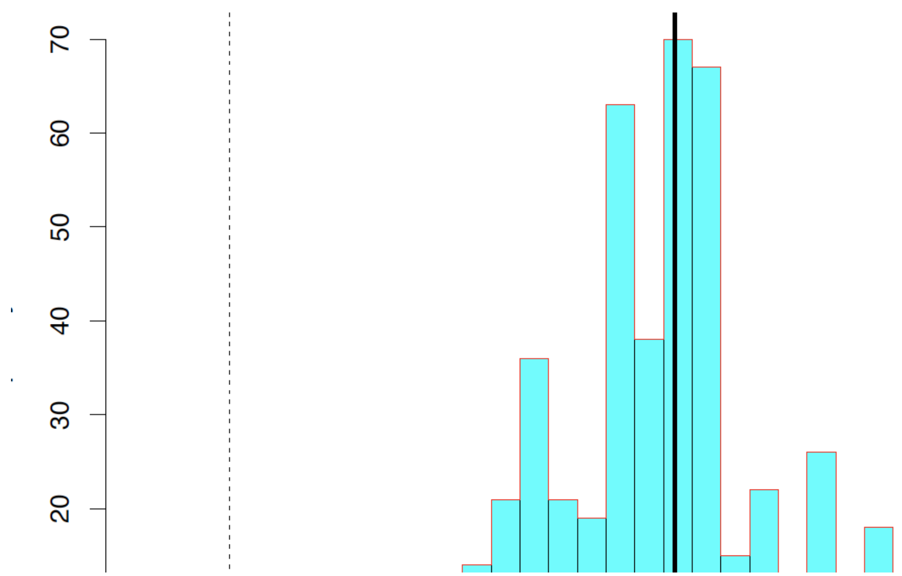
# Extra Examples: Difference in Means

- 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



## Extra 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



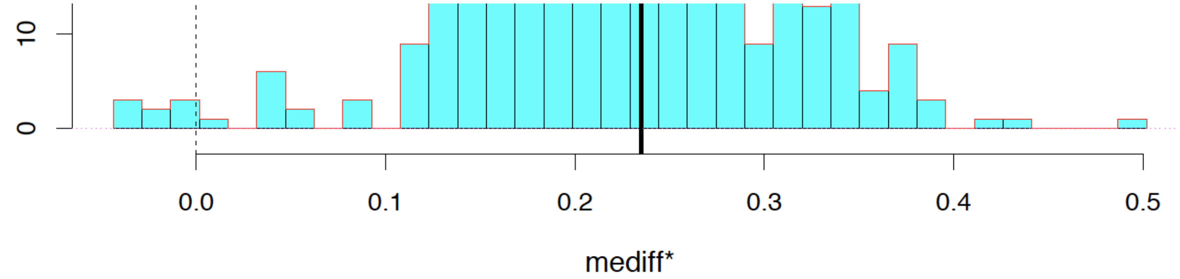
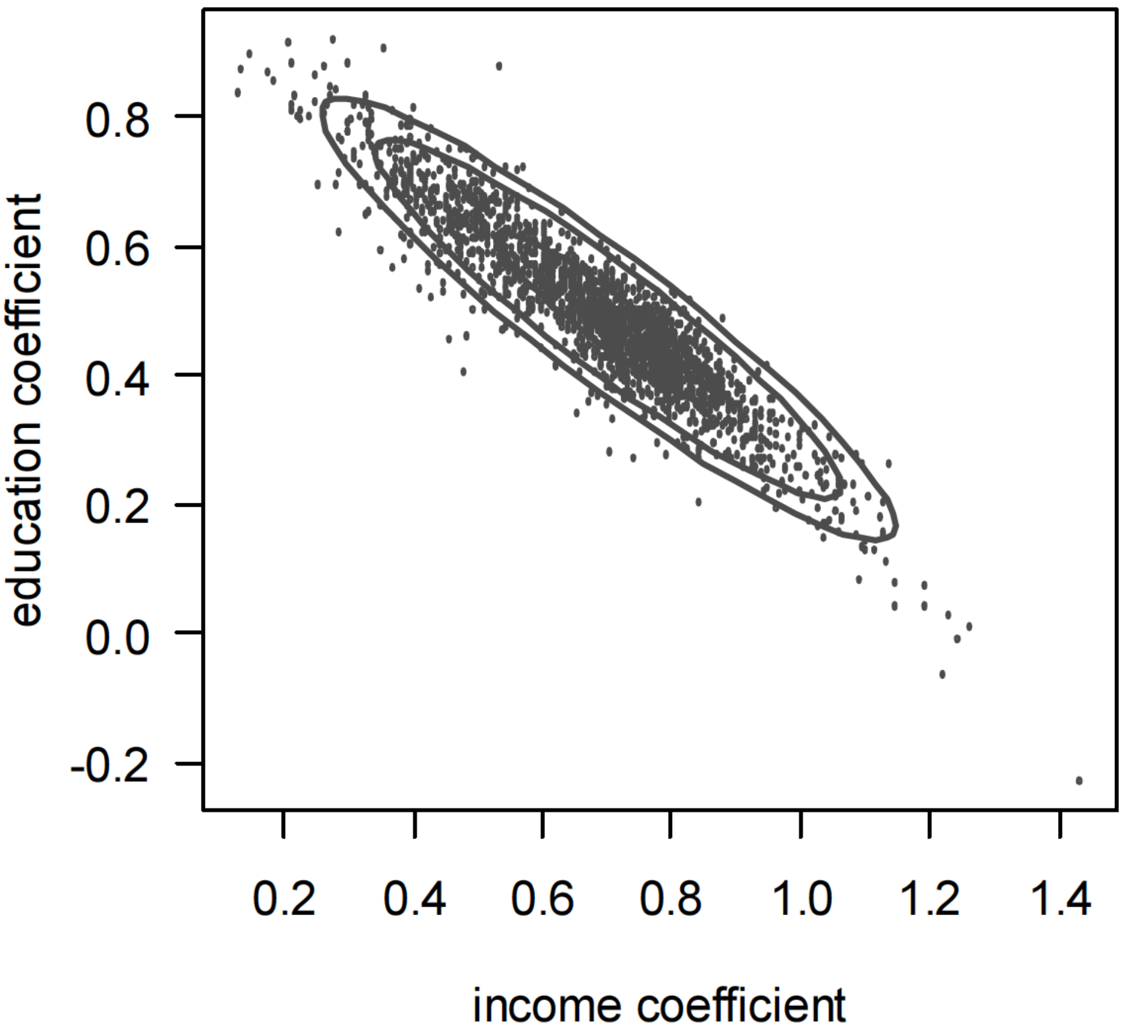



**Figure 10.4** $B = 500$ bootstrap replications for the median difference between the `AML` and `ALL` scores in Figure 1.4, giving $\widehat{\text{se}}_{\text{boot}} = 0.074$. The observed value `mediff` $= 0.235$ (vertical black line) is more than 3 standard errors above zero.

## Extra Examples: Regression

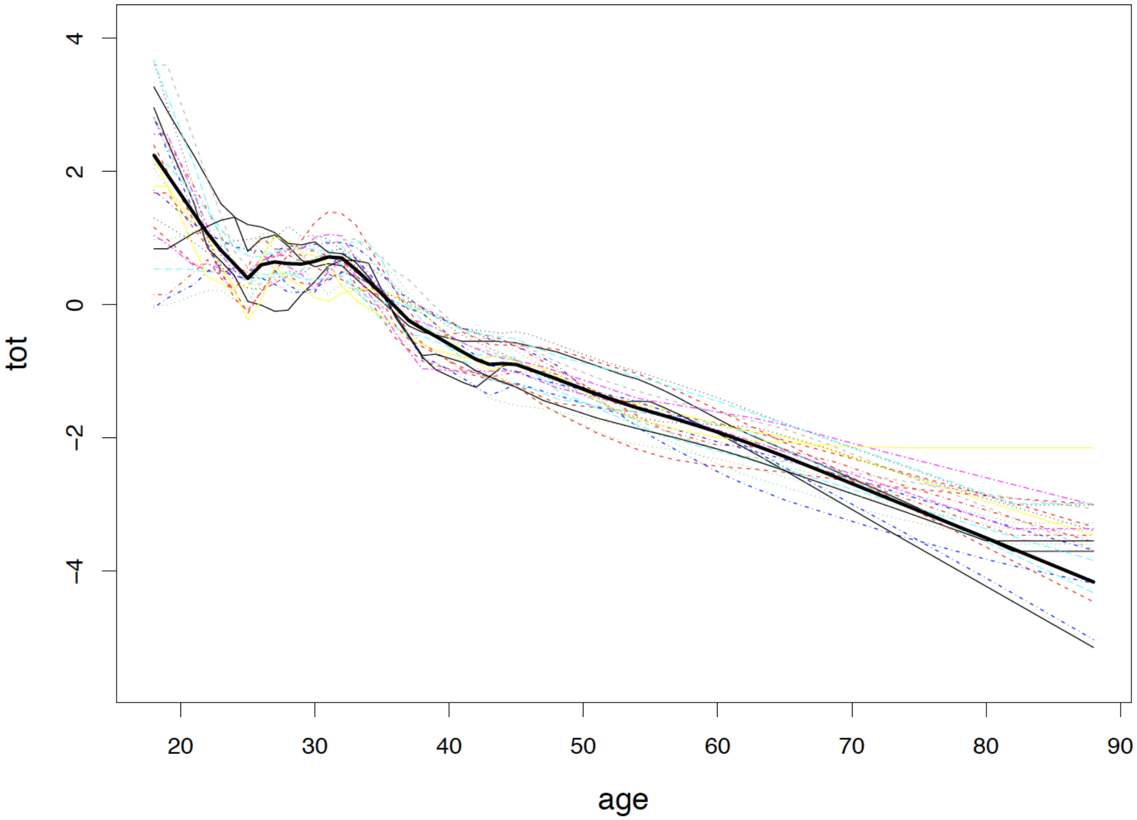
- You can bootstrap regression models as well
- Fit the regression many times, across resampled versions of the data
- Works for variants of regression with no analytical s.e. formula




Example robust regression bootstrap from Bootstrapping Regression Models

# Extra Examples: Regression

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




and even Random Forests

and even Random Forests.

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

```
<style>
```

```
mtex_block = ƒ()
```

```
mtex = ƒ()
```