Lecture 5.2: The Bootstrap

STAT 1361/2360: Statistical Learning and Data Science

University of Pittsburgh
Prof. Lucas Mentch

The Bootstrap

- Developed by Brad Efron (1979)
  - One of the most important developments in the history of statistics
  - More than 36,000 citations; among the most of any statistics paper
- You may have seen this before in one context or another; today we'll try to motivate the big idea behind the bootstrap and show how it can be applied in very general contexts

Statistical Inference Review

Confidence Intervals

- Specifically, what do we mean when we say $[a, b]$ is a 95% CI for some parameter θ?

Confidence Intervals

- Specifically, what do we mean when we say $[a, b]$ is a 95% CI for some parameter θ?

  If we were to take many samples – each of size n – and calculate a CI in exactly the same fashion, then approximately 95% of these CIs would contain the true value of the parameter θ

Confidence Intervals

- Given a sample of size n, $X_1, \ldots, X_n$, how do we get a confidence interval for the mean μ?

Confidence Intervals

- Given a sample of size n, $X_1, \ldots, X_n$, how do we get a confidence interval for the mean μ?

$$\hat{\mu} \pm t^* \frac{s}{\sqrt{n}}$$

  where $\hat{\mu}$ is the sample mean, s is the sample standard deviation, and t^* is the appropriate quantile of t-distribution (for large enough sample size, could use z^* as a normal approximation)

- To get a 95% confidence interval with reasonably large n, $t^* \approx z^* = 1.96$

Confidence Intervals

- Where does this formula come from?

Confidence Intervals

- Where does this formula come from?

Central Limit Theorem: The (sampling) distribution of the sample mean is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

$\Longrightarrow$ The most important result in all of statistics

Inference Overview

So let's review the fundamental steps in statistical inference:

1. We have some (population distribution) F with some associated parameter θ
2. We take a sample from that population and calculate an estimate $\hat{\theta}$
3. We want to use $\hat{\theta}$ to *infer* something about θ

But ... In order to make any kind of inference, we need to know something about the distribution of $\hat{\theta}$

Beyond CLTs

CIs without CLTs

Crucial Question: What if I didn't have any kind of CLT-type mathematical results that told me about the sampling distribution?

If I have a reasonably good (unbiased) estimator $\hat{\mu}$, how else could I think of getting a CI for μ?

CIs without CLTs

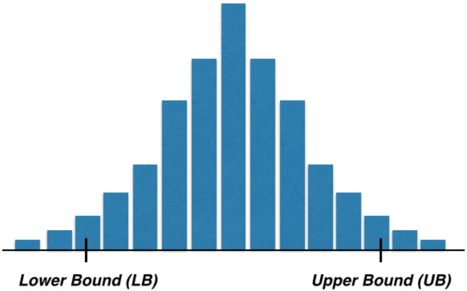
Crucial Question: What if I didn't have any kind of CLT-type mathematical results that told me about the sampling distribution?

If I have a reasonably good (unbiased) estimator $\hat{\mu}$, how else could I think of getting a CI for μ?

Hint: Think about what you could do with unlimited time and resources.

CIs without CLTs

I could just take many samples, each of size n, look at the empirical distribution (histogram) of all of the $\hat{\mu}$, and take the appropriate quantiles (0.025 and 0.975 for 95% CI)!



Histogram of values of $\hat{\mu}$ collected across many samples.

CIs without CLTs

More formally:

- LB = value such that 2.5% of all sample means calculated are less than or equal to LB
- UB = value such that 97.5% of all sample means calculated are less than or equal to UB
- Our confidence interval is then just simply $[LB, UB]$

CIs without CLTs

But wait ...

CIs without CLTs

But wait ...

As long as I have an unbiased estimator, this same idea would work for *any* parameter θ, not just the mean.

CIs without CLTs

But wait ...

As long as I have an unbiased estimator, this same idea would work for *any* parameter θ, not just the mean.

So what ... why is that important?

CIs without CLTs

But wait ...

As long as I have an unbiased estimator, this same idea would work for *any* parameter θ, not just the mean.

So what ... why is that important?

Because we don't have results like the CLT for very many statistics!

e.g. Suppose $\hat{\theta} = \min(X_i)^{\max(X_i)}$

CIs without CLTs

Assume we have some population distribution F with some associated parameter θ. To get a 95% CI for θ

1. Take B samples, each of size n, from the population
2. From those samples, calculate B estimates of θ: $\hat{\theta}_1, \ldots, \hat{\theta}_B$
3. To get a 95% CI for θ, take the 0.025 and 0.975 quantiles of the distribution of the $\hat{\theta}$'s calculated in the previous step: $[\hat{\theta}_{0.025}, \hat{\theta}_{0.975}]$

   (Note: If you want a single point estimate, you could take $\hat{\theta} = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i$)

Problems

Obvious Problem:

Problems

Obvious Problem: We almost certainly don't have the resources to just continue taking samples. If we did, statistics wouldn't be much of a field.

Problems

Obvious Problem: We almost certainly don't have the resources to just continue taking samples. If we did, statistics wouldn't be much of a field.

Less Obvious Problem:

Problems

Obvious Problem: We almost certainly don't have the resources to just continue taking samples. If we did, statistics wouldn't be much of a field.

Less Obvious Problem: Even if we did have those resources, that would be equivalent to taking one giant sample of size $B \cdot n$ and we'd rather work with a single sample that size anyway

Bootstrapping

Sampling without Sampling

So we don't have the resources to just keep sampling as much as we like from the population, but is there something else that sort of looks like the population that we could keep taking samples from?

Sampling without Sampling

So we don't have the resources to just keep sampling as much as we like from the population, but is there something else that sort of looks like the population that we could keep taking samples from?

Yes – the original sample itself!

Bootstrap CI Procedure:

Assume we have some population distribution F with some associated parameter θ. To get a 95% CI for θ *with bootstrapping*

1. Take original sample of size n, from the population, $\{X_1, \ldots, X_n\}$, and calculate original estimate $\hat{\theta}_0$
2. Take a *bootstrap sample* of the original data, $b_1^* = \{X_1^*, \ldots, X_n^*\}$, and use this to calculate $\hat{\theta}_1^*$
3. Repeat Step 2 a total of B times to calculate B bootstrap samples and thus B estimates of θ: $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$
4. To get a 95% CI for θ, take the 0.025 and 0.975 quantiles of the distribution of the $\hat{\theta}^*$'s calculated in the previous step: $[\hat{\theta}^*_{0.025}, \hat{\theta}^*_{0.975}]$

Bootstrap Samples

So what is a *bootstrap* sample?

A bootstrap sample is simply a sample of size n from the original data $X_1, ..., X_n$ taken with replacement

E.g. Suppose the original dataset (n=5) is

$$\{3.3, 6.4, 2.5, 6.6, 4.8\}$$

Examples of a bootstrap samples would be:

$$b_1 = \{6.4, 6.4, 6.6, 3.3, 6.4\}$$
$$b_2 = \{2.5, 3.3, 2.5, 4.8, 6.6\}$$

Notes on Bootstrap Samples

- Bootstrap samples are the *same size* as the original sample
- Samples are taken *with replacement*. Otherwise ... what would we get?
- Bootstrap samples will almost always contain duplicates; that's arguably the entire point (see above bullet point)

Some terminology:

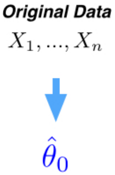
- "Bootstrap" samples *without replacement* are simply permutations (random re-orderings)
- Samples of size $k < n$ taken without replacement are usually called (proper) subsamples

Bootstrapping in Pictures

Original Data

$X_1, ..., X_n$

Bootstrapping in Pictures

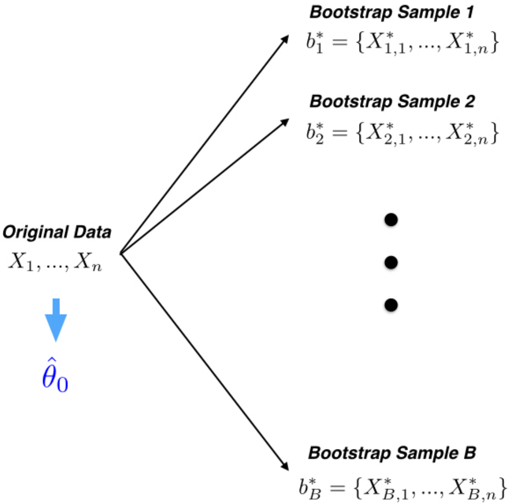


Original Data

$X_1, ..., X_n$

$\hat{\theta}_0$

Bootstrapping in Pictures



Bootstrap Sample 1

$b_1^* = \{X_{1,1}^*, ..., X_{1,n}^*\}$

Bootstrap Sample 2

$b_2^* = \{X_{2,1}^*, ..., X_{2,n}^*\}$

Original Data

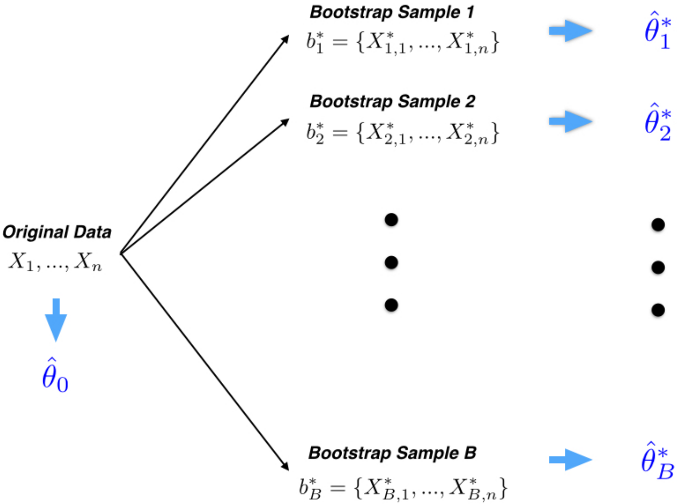
$X_1, ..., X_n$

$\hat{\theta}_0$

Bootstrap Sample B

$b_B^* = \{X_{B,1}^*, ..., X_{B,n}^*\}$

Bootstrapping in Pictures



Original Data

$X_1, ..., X_n$

$\hat{\theta}_0$

Bootstrap Sample 1

$b_1^* = \{X_{1,1}^*, ..., X_{1,n}^*\}$ → $\hat{\theta}_1^*$

Bootstrap Sample 2

$b_2^* = \{X_{2,1}^*, ..., X_{2,n}^*\}$ → $\hat{\theta}_2^*$

⋮

Bootstrap Sample B

$b_B^* = \{X_{B,1}^*, ..., X_{B,n}^*\}$ → $\hat{\theta}_B^*$

Other uses for the bootstrap

The bootstrap confidence intervals we discussed a few slides earlier are called *bootstrap percentile CIs*. What else can we do with bootstrapping?

1. **Basic Hypothesis testing:**

$$\mathrm{H}_0 : \theta = c$$

for some constant c.

$\implies$ Construct bootstrap percentile confidence interval. If $c \in [LB, UB] = [\hat{\theta}^*_{0.025}, \hat{\theta}^*_{0.975}]$ we fail to reject; otherwise we reject.

Other uses for the bootstrap

2. More stable parameter estimation:

Given bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, define a new estimate

$$\tilde{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

- This results in a more stable estimate of θ – the variance of the sampling distribution of $\tilde{\theta}^*$ is smaller than that of the original $\hat{\theta}_0$. (Think if we got a new original dataset and repeated the entire procedure)
- In general, we refer to $\tilde{\theta}^*$ as the *bagged* estimate of θ (**b**ootstrap **agg**regat**ed**), or sometimes just the *bootstrap* estimate of θ

Other uses for the bootstrap

3. Bias Correction:

For any parameter θ, the bias of an estimator $\hat{\theta}$ is defined as

$$\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

If we take our original estimate $\hat{\theta}_0$ as a substitute for θ and our bagged estimate $\tilde{\theta}^*$ as a substitute for $\mathbb{E}(\hat{\theta})$, a *bias corrected* (BC) estimate for θ is given by

$$\begin{aligned}\hat{\theta}^*_{BC} &= \hat{\theta}_0 - \widehat{\text{bias}}(\hat{\theta}) \\ &= \hat{\theta}_0 - \left(\tilde{\theta}^* - \hat{\theta}_0\right) \\ &= 2\hat{\theta}_0 - \tilde{\theta}^*\end{aligned}$$

Other uses for the bootstrap

4. Bias Corrected CIs:

Given that we can estimate and correct for the bias in a point estimate, we can also correct for the bias in our bootstrap percentile CIs:

$$[2\hat{\theta}_0 - UB, 2\hat{\theta}_0 - LB] = [2\hat{\theta}_0 - \hat{\theta}^*_{0.975}, 2\hat{\theta}_0 - \hat{\theta}^*_{0.025}]$$

These bias-corrected versions of bootstrap percentile CIs are generally just referred to as *standard* bootstrap CIs

Note: The above definition is *not* a typo

Bootstrap Extensions

Assessing Correlation via the Bootstrap

Let's look at one context that's a bit of an extension of the basic procedure we outlined:

Suppose we have two variables X_1 and X_2 and we want a confidence interval for the *correlation* between them

X_1	X_2
$X_{1,1}$	$X_{2,1}$
$X_{1,2}$	$X_{2,2}$
$\vdots$	$\vdots$
$X_{1,n}$	$X_{2,n}$

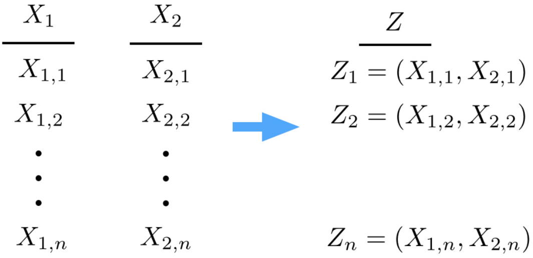
Assessing Correlation via the Bootstrap

How could I use the bootstrap to get a CI? Why can't I just take bootstrap samples of both X_1 and X_2?

Assessing Correlation via the Bootstrap

How could I use the bootstrap to get a CI? Why can't I just take bootstrap samples of both X_1 and X_2?

- X_1 and X_2 need to be resampled *together*



X_1	X_2		Z
$X_{1,1}$	$X_{2,1}$		$Z_1 = (X_{1,1}, X_{2,1})$
$X_{1,2}$	$X_{2,2}$	→	$Z_2 = (X_{1,2}, X_{2,2})$
$\vdots$	$\vdots$		
$X_{1,n}$	$X_{2,n}$		$Z_n = (X_{1,n}, X_{2,n})$

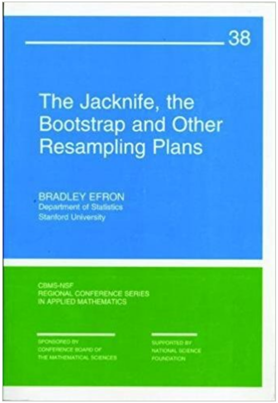
- Bootstrapping the pairs (Z_i) preserves the relationship between the variables

Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap

Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap



38

The Jacknife, the Bootstrap and Other Resampling Plans

BRADLEY EFRON
Department of Statistics
Stanford University

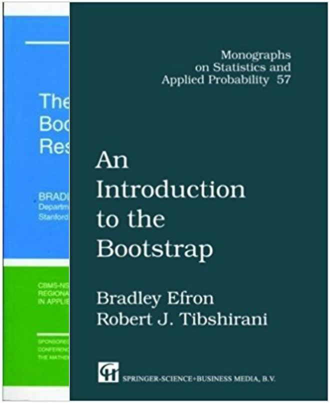
CBMS-NSF
REGIONAL CONFERENCE SERIES
IN APPLIED MATHEMATICS

SPONSORED BY
CONFERENCE BOARD OF
THE MATHEMATICAL SCIENCES

SUPPORTED BY
NATIONAL SCIENCE
FOUNDATION

Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap



Monographs
on Statistics and
Applied Probability 57

An
Introduction
to the
Bootstrap

Bradley Efron
Robert J. Tibshirani

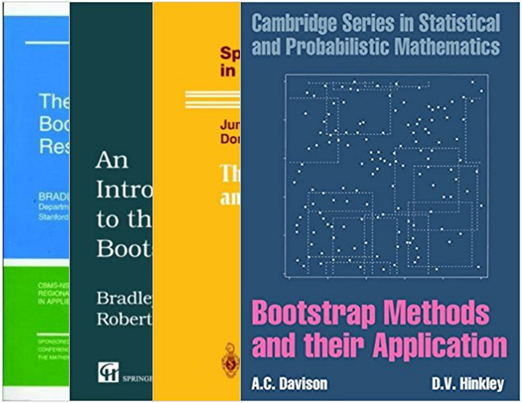
SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap

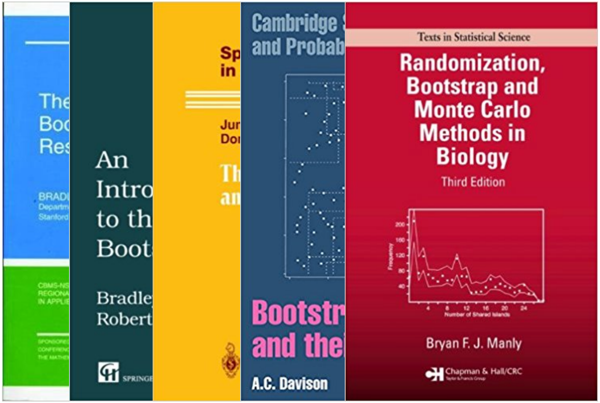
Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap


The
Boo
Res
BRADL
An
Intro
to th
Boot
Bradle
Robert
SPRINGE
Sp
in
Jur
Dor
Th
an
Cambridge Series in Statistical and Probabilistic Mathematics
Bootstrap Methods and their Application
A.C. Davison
D.V. Hinkley

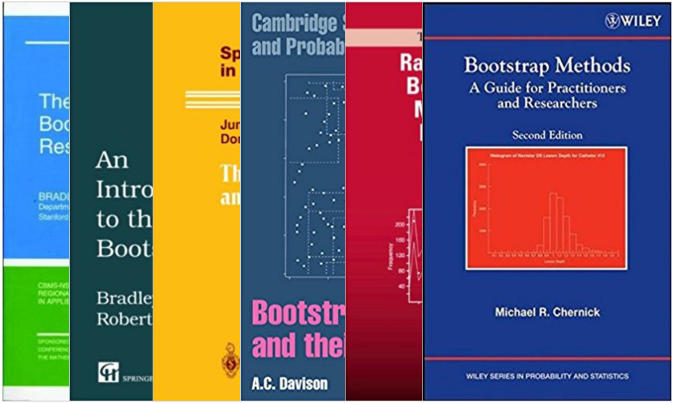
Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap



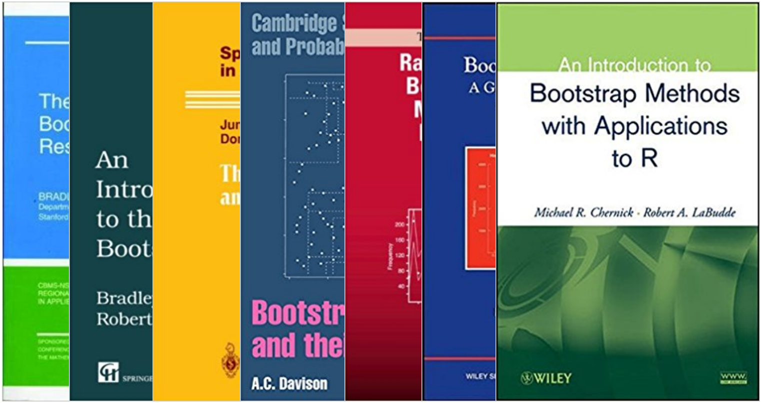
Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap



Bootstrap Extensions

We are barely scratching the surface of varieties of things that can be done with the bootstrap



Final Remarks on Bootstrapping

- Bootstrapping provides a very nice way of performing inference (CIs, hypothesis tests) that doesn't require you to know the sampling distribution of the statistic being calculated
  - Often times, such distributions are very hard to obtain and in those cases, bootstrap procedures are almost always used
- Beyond simply doing inference, bootstrapping also provides a means of getting a more stable estimate and correcting bias
- Bootstrapping works well for *most* statistics, but not all – see homework