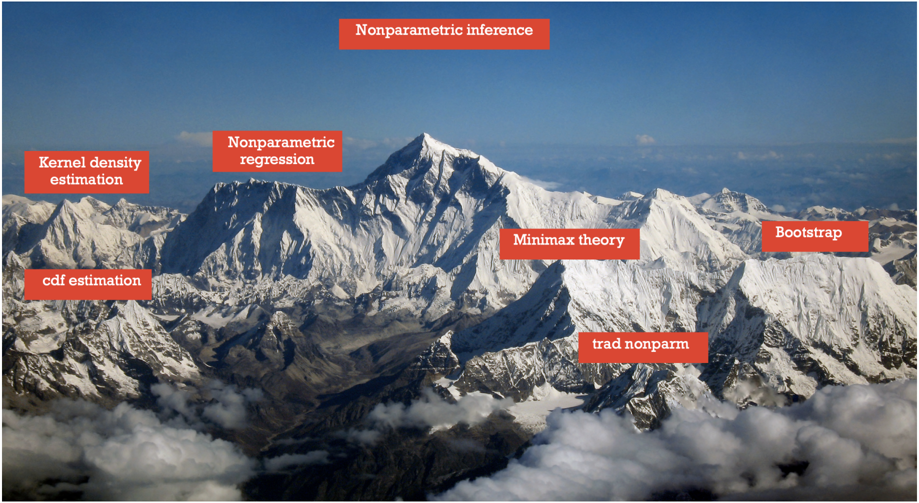
STAT 824 sp 2025 Lec 08 slides

Bootstrap for the mean

Karl Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.


Nonparametric inference
Kernel density estimation
Nonparametric regression
Minimax theory
Bootstrap
cdf estimation
trad nonparm


Percentile bootstrap
Unstudentized
Consistency
Bootstrap for regression
Bootstrap
Berry-Esseen
Statistical functionals
Studentized
Second-order correctness
Edgeworth expansion

The *bootstrap* is a method for estimating sampling distributions.

It is useful in many contexts. For now we focus on the mean of iid data:

Pivot quantities for the mean

Consider the sampling distributions of the pivots

$$Y_n = \sqrt{n}(\bar{X}_n - \mu) \quad \text{or} \quad Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \quad \text{or} \quad T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n},$$

where $X_1, \dots, X_n$ are iid with $\mathbb{E}X_1 = \mu$, $\text{Var}\, X_1 = \sigma^2$ and

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

We will call these the *unstandardized*, *standardized*, and *studentized* pivots for μ.

Application: build confidence intervals for μ based on these pivot quantities.

cdfs of pivot quantities for the mean

Define the cdfs of the pivots as

$$G_{Y_n}(x) = P\left(Y_n \leq x\right)$$
$$G_{Z_n}(x) = P\left(Z_n \leq x\right)$$
$$G_{T_n}(x) = P\left(T_n \leq x\right)$$

for all $x \in \mathbb{R}$.

The bootstrap can be used to get estimators $\hat{G}_{Y_n}$, $\hat{G}_{Z_n}$, and $\hat{G}_{T_n}$ of these cdfs.

Exercise: Give CIs for μ when G_{Y_n}, G_{Z_n}, and G_{T_n} known.

IID bootstrap for the mean

Introduce iid rvs $X_1^*, \dots, X_n^* | X_1, \dots, X_n$ with cdf $\hat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$.

Define bootstrap versions of Y_n, Z_n, and T_n as

$$Y_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n), \quad Z_n^* = \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n}, \quad \text{and} \quad T_n^* = \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{S_n^*},$$

where $\bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$, and $(S_n^*)^2 = (n-1)^{-1} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2$.

Then iid bootstrap estimators of G_{Y_n}, G_{Z_n}, and G_{T_n} are given by

$$\begin{aligned} \hat{G}_{Y_n}(x) &= P\left(Y_n^* \leq x | X_1, \dots, X_n\right) \\ \hat{G}_{Z_n}(x) &= P\left(Z_n^* \leq x | X_1, \dots, X_n\right) \\ \hat{G}_{T_n}(x) &= P\left(T_n^* \leq x | X_1, \dots, X_n\right). \end{aligned}$$

for all $x \in \mathbb{R}$.

Idea is to ask how the pivot behaves when $\hat{F}_n$ is the population cdf.

Bootstrap notation

Let P_*, $\mathbb{E}_*$ and Var_* be operators such that

$$\begin{aligned} P_*(\,\cdot\,) &= P(\,\cdot\,|X_1, \dots, X_n) \\ \mathbb{E}_*(\,\cdot\,) &= \mathbb{E}(\,\cdot\,|X_1, \dots, X_n) \\ \text{Var}_*(\,\cdot\,) &= \text{Var}(\,\cdot\,|X_1, \dots, X_n), \end{aligned}$$

representing conditional probability, expectation, and variance, given $X_1, \dots, X_n$.

So P_*, $\mathbb{E}_*$ and Var_* treat $X_1^*, \dots, X_n^*$ as random and $X_1, \dots, X_n$ as fixed.

Exercise: Show that

1. $\mathbb{E}_*[X_1^*] = \bar{X}_n$
2. $\text{Var}_*[X_1^*] = \hat{\sigma}_n^2 = (n-1)S_n^2/n$.

Discuss: How to build CIs after obtaining $\hat{G}_{n,\text{U}}$, $\hat{G}_n$, and $\hat{G}_{n,\text{S}}$.

Discuss: The bootstrap estimator $\hat{G}_{Y_n}$ of G_{Y_n} is given by

$$\hat{G}_{Y_n}(x) = P_*(Y_n^* \leq x)$$

for all $x \in \mathbb{R}$. Can we compute this?

Instead of computing $\hat{G}_{Y_n}$, $\hat{G}_{Z_n}$, and $\hat{G}_{T_n}$ *exactly*, we use MC approximation.

Monte Carlo approximation to bootstrap estimators $\hat{G}_{Y_n}$, $\hat{G}_{Z_n}$, and $\hat{G}_{T_n}$

For $b = 1, \dots, B$ for large B (≥ 500, say):

1. Draw $X_1^{*(b)}, \dots X_n^{*(b)}$ with replacement from $X_1, \dots, X_n$.
2. Compute

$$\begin{aligned} & Y_n^{*(b)} = \sqrt{n}(\bar{X}_n^{*(b)} - \bar{X}_n) \\ \text{or} \quad & Z_n^{*(b)} = \sqrt{n}(\bar{X}_n^{*(b)} - \bar{X}_n)/\hat{\sigma}_n \\ \text{or} \quad & T_n^{*(b)} = \sqrt{n}(\bar{X}_n^{*(b)} - \bar{X}_n)/S_n^{*(b)}, \end{aligned}$$

where $\bar{X}_n^{*(b)} = n^{-1}\sum_{i=1}^n X_i^{*(b)}$, $(S_n^{*(b)})^2 = (n-1)^{-1}\sum_{i=1}^n(X_i^{*(b)} - \bar{X}_n^{*(b)})^2$.

We now retrieve (MC-approximated) quantiles of $\hat{G}_{Y_n}$, $\hat{G}_{Z_n}$, and $\hat{G}_{T_n}$ as

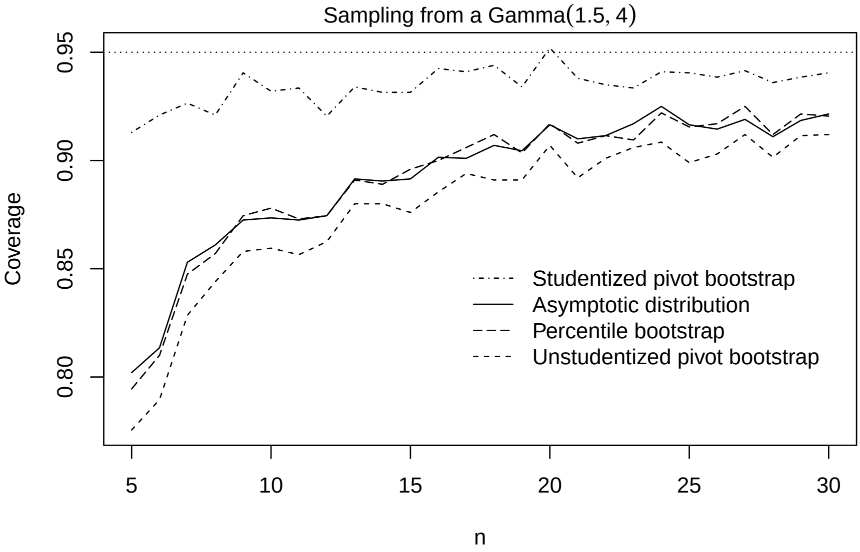
$$\hat{G}_{Y_n}^{-1}(u) = Y_n^{*(\lceil uB \rceil)} \quad \text{or} \quad \hat{G}_{Z_n}^{-1}(u) = Z_n^{*(\lceil uB \rceil)} \quad \text{or} \quad \hat{G}_{T_n}^{-1}(u) = T_n^{*(\lceil uB \rceil)}$$

after sorting the realizations of each pivot in ascending order.

Exercise: Compare via simulation the performance of these intervals:

1. $(\bar{X}_n - z_{\alpha/2} S_n/\sqrt{n},\ \bar{X}_n - z_{\alpha/2} S_n/\sqrt{n})$. Asymptotic.
2. $(2\bar{X}_n - \bar{X}_n^{*(\lceil(1-\alpha/2)B\rceil)}, 2\bar{X}_n - \bar{X}_n^{*(\lceil(\alpha/2)B\rceil)})$. The Y_n-based interval.
3. $(\bar{X}_n - \hat{G}_{T_n}^{-1}(1-\alpha/2)\hat{S}_n/\sqrt{n},\ \bar{X}_n - \hat{G}_{T_n}^{-1}(\alpha/2) S_n/\sqrt{n})$. The T_n-based.
4. $(\bar{X}_n^{*(\lceil(\alpha/2)B\rceil)}, \bar{X}_n^{*(\lceil(1-\alpha/2)B\rceil)})$. Called the *percentile interval*.

for small n when the population distribution is non-Normal.


Sampling from a Gamma(1.5, 4)
Coverage
0.80
0.85
0.90
0.95
5
10
15
20
25
30
n
Studentized pivot bootstrap
Asymptotic distribution
Percentile bootstrap
Unstudentized pivot bootstrap

We now present a result on the consistency of the bootstrap.

Bootstrap "works" for the mean

Let $X_1, \dots, X_n$ be iid with $\mathbb{E}X_1 = \mu$, $\text{Var}\, X_1 = \sigma^2 \in (0, \infty)$, $\mathbb{E}|X_1|^3 < \infty$. Then

$$\sup_{x \in \mathbb{R}} \Big| P_* \left(Y_n^* \leq x \right) - P \left(Y_n \leq x \right) \Big| \to 0 \text{ w.p. } 1 \text{ as } n \to \infty.$$

Consequently the coverage probability of the interval

$$(\bar{X}_n - \hat{G}_{Y_n}^{-1}(1 - \alpha/2)\sigma/\sqrt{n},\ \bar{X}_n - \hat{G}_{Y_n}^{-1}(\alpha/2)\sigma/\sqrt{n})$$

converges to $(1 - \alpha)$ w.p. 1 as $n \to \infty$.

We will prove that the bootstrap works with the following amazing theorem:

Berry–Esseen theorem

For $X_1, \dots, X_n$ iid with $\mathbb{E}X_1 = \mu$, $\text{Var}\, X_1 = \sigma^2$, and $\mathbb{E}|X_1|^3 < \infty$, we have

$$\sup_{x \in \mathbb{R}} \left| P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x \right) - \Phi(x) \right| \leq C \cdot \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3 \sqrt{n}}$$

for each $n \geq 1$ and $0 \leq C \leq \sqrt{\frac{2}{\pi}} \left(\frac{5}{2} + \frac{12}{\pi} \right) < 5.05$. See pg 361 of [1].

Exercise: Prove the bootstrap works with B–E and results on next slide.

Special case of Marcinkiewz–Zygmund SLLN

Let $Y_1, \dots, Y_n$ be iid, $p \in (0,1)$. Then if $\mathbb{E}|Y_1|^p < \infty$, $n^{-1/p} \sum_{i=1}^n Y_i \to 0$ w.p. 1.

Minkowski's inequality

For any rvs $X, Y \in \mathbb{R}$, $p \in (1, \infty)$, we have $(\mathbb{E}|X+Y|^p)^{\frac{1}{p}} \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}}$.

Jensen's inequality

If $g : \mathbb{R} \to \mathbb{R}$ is a convex function, then for any rv X we have

$$g(\mathbb{E}X) \leq \mathbb{E}g(X),$$

provided $\mathbb{E}|X| < \infty$ and $\mathbb{E}|g(X)| < \infty$.

Why was the bootstrap interval based on T_n superior to that based on Y_n?

We will answer this question using Edgeworth expansions...

Krishna B Athreya and Soumendra N Lahiri.
Measure theory and probability theory.
Springer Science & Business Media, 2006.