# STAT 824 sp 2023 Lec 08 slides

# Bootstrap for the mean

Karl B. 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.

# Table of Contents

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

We wish to estimate the sampling distributions of the pivots

$$T_{n,\mathrm{U}} = \sqrt{n}(\bar{X}_n - \mu) \quad \text{or} \quad T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \quad \text{or} \quad T_{n,\mathrm{S}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n},$$

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

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

We will call these the *unstandardized*, *standardized*, and *studentized* pivots for $\mu$.

Application: build confidence intervals for $\mu$ based on these pivot quantities.

## cdfs of pivot quantities for the mean

Define the cdfs of the pivots as

$$G_{n,\mathrm{U}}(x) = P\left(\sqrt{n}(\bar{X}_n - \mu) \leq x\right)$$
$$G_n(x) = P\left(\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x\right)$$
$$G_{n,\mathrm{S}}(x) = P\left(\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}_n \leq x\right)$$

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

The bootstrap can be used to get estimators $\hat{G}_{n,\mathrm{U}}$, $\hat{G}_n$, and $\hat{G}_{n,\mathrm{S}}$ of these cdfs.

**Exercise:** Give CIs for $\mu$ when $G_{n,\mathrm{U}}$, $G_n$, and $G_{n,\mathrm{S}}$ known.

## IID bootstrap for the mean

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

The iid bootstrap estimators $\hat{G}_{n,\mathrm{U}}$, $\hat{G}_n$, and $\hat{G}_{n,\mathrm{S}}$ of $G_{n,\mathrm{U}}$, $G_n$, and $G_{n,\mathrm{S}}$ are given by

$$\hat{G}_{n,\mathrm{U}}(x) = P\left(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x | X_1, \ldots, X_n\right)$$
$$\hat{G}_n(x) = P\left(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/\hat{\sigma}_n \leq x | X_1, \ldots, X_n\right)$$
$$\hat{G}_{n,\mathrm{S}}(x) = P\left(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/\hat{\sigma}_n^* \leq x | X_1, \ldots, X_n\right)$$

for all $x \in \mathbb{R}$, with $\bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$, and $(\hat{\sigma}_n^*)^2 = n^{-1} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2$.

Idea is to ask how the pivot behaves when $F_n$ is the population cdf.

**Exercise:** Show that

1. $\mathbb{E}[X_1^* | X_1, \ldots, X_n] = \bar{X}_n$
2. $\mathrm{Var}[X_1^* | X_1, \ldots, X_n] = \hat{\sigma}_n^2$.

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

## Bootstrap notation

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

$$P_*(\,\cdot\,) = P(\,\cdot\,|X_1, \ldots, X_n)$$
$$\mathbb{E}_*(\,\cdot\,) = \mathbb{E}(\,\cdot\,|X_1, \ldots, X_n)$$
$$\text{Var}_*(\,\cdot\,) = \text{Var}(\,\cdot\,|X_1, \ldots, X_n),$$

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

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

**Discuss:** The bootstrap estimator $\hat{G}_{n,U}$ of $G_{n,U}$ is given by

$$\hat{G}_{n,U}(x) = P_*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x)$$

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

Instead of computing $\hat{G}_{n,\mathrm{U}}$, $\hat{G}_n$, and $\hat{G}_{n,\mathrm{S}}$ *exactly*, we use MC approximation.

## Monte Carlo approximation to bootstrap estimators $\hat{G}_{n,\mathrm{U}}$, $\hat{G}_n$, and $\hat{G}_{n,\mathrm{S}}$

For $b = 1, \ldots, B$ for large $B$ ($\geq 500$, say):

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

$$T_{n,U}^{*(b)} = \sqrt{n}(\bar{X}_n^{*(b)} - \bar{X}_n)$$

$$\text{or} \quad T_n^{*(b)} = \sqrt{n}(\bar{X}_n^{*(b)} - \bar{X}_n)/\hat{\sigma}_n$$

$$\text{or} \quad T_{n,S}^{*(b)} = \sqrt{n}(\bar{X}_n^{*(b)} - \bar{X}_n)/\hat{\sigma}_n^{*(b)},$$


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

We now retrieve (MC-approximated) quantiles of $\hat{G}_{n,U}$, $\hat{G}_n$, and $\hat{G}_{n,S}$ as

$$\hat{G}_{n,U}^{-1}(u) = T_{n,U}^{*(\lceil uB \rceil)} \quad \text{or} \quad \hat{G}_n^{-1}(u) = T_n^{*(\lceil uB \rceil)} \quad \text{or} \quad \hat{G}_{n,S}^{-1}(u) = T_{n,S}^{*(\lceil uB \rceil)}$$

after sorting the realizations of each pivot in ascending order.

**Discuss:** Why not just use

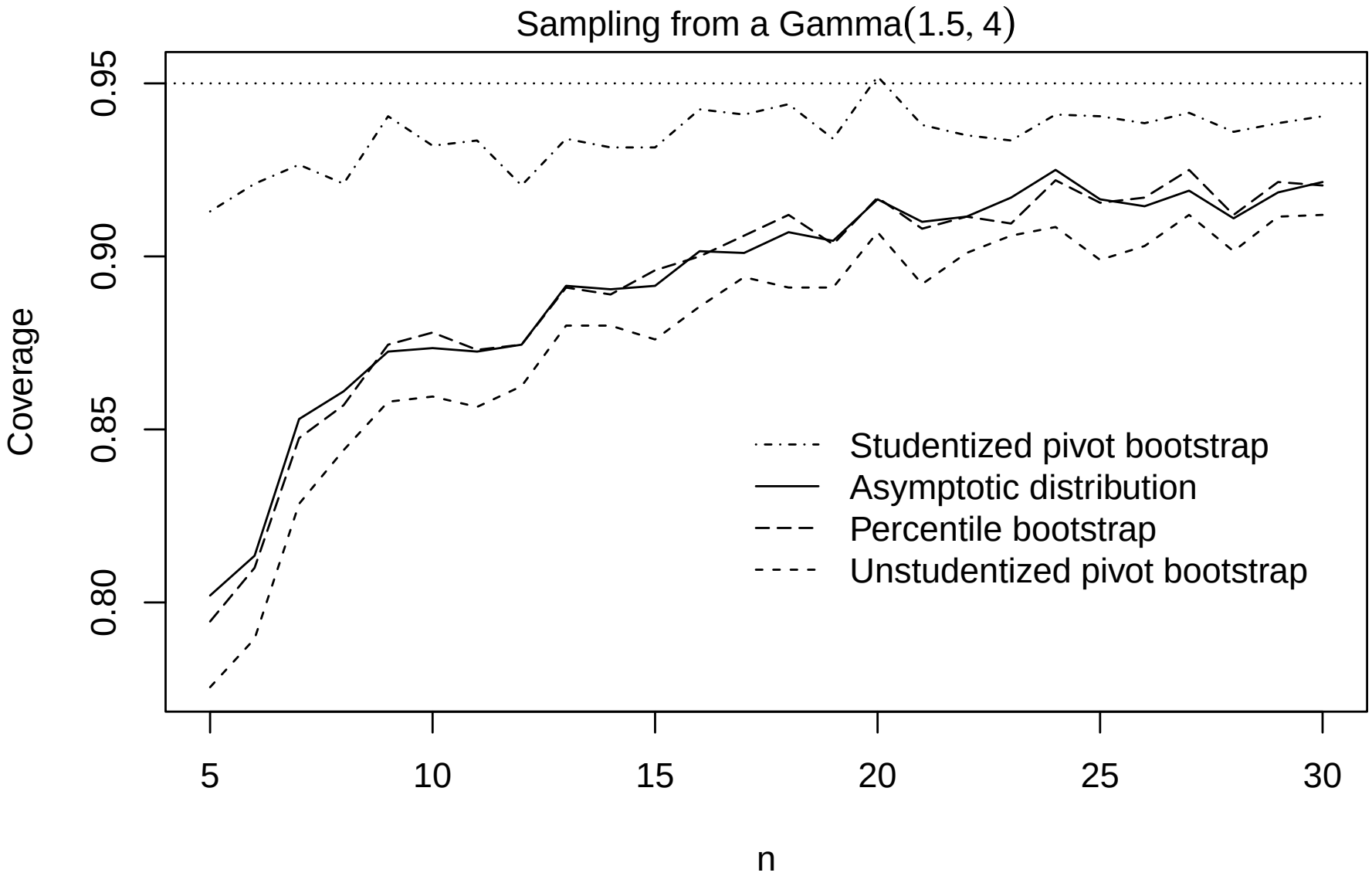
$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \to \text{Normal}(0,1) \text{ in distribution as } n \to \infty$$

and the corresponding asymptotically correct interval $\bar{X}_n \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$ ?

**Exercise:** Compare via simulation the performance of these intervals:

1. $(\bar{X}_n - \Phi^{-1}(1-\alpha/2)\hat{\sigma}_n/\sqrt{n},\ \bar{X}_n - \Phi^{-1}(\alpha/2)\hat{\sigma}_n/\sqrt{n})$. 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 $G_{n,U}$-based interval.
3. $(\bar{X}_n - \hat{G}_{n,S}^{-1}(1-\alpha/2)\hat{\sigma}_n/\sqrt{n},\ \bar{X}_n - \hat{G}_{n,S}^{-1}(\alpha/2)\hat{\sigma}_n/\sqrt{n})$. The $G_{n,S}$-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

# Table of Contents

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

## Bootstrap "works" for the mean

Let $X_1, \ldots, X_n$ be iid with $\mathbb{E}X_1 = \mu$, $\operatorname{Var} X_1 = \sigma^2 < \infty$. Then

$$\sup_{x \in \mathbb{R}} \left| P_* \left( \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{\sigma}_n} \leq x \right) - P\left( \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x \right) \right| \to 0 \text{ w.p. 1 as } n \to \infty.$$

We could also write this as $\|\hat{G}_n - G_n\|_\infty \to 0$ w.p. 1 as $n \to \infty$.

Consequently the coverage probability of the interval

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

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

**Discuss:** Sufficient to show $\|\hat{G}_n - \Phi\|_\infty \to 0$ w.p. 1 as $n \to \infty$.

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

## Berry–Esseen theorem

For $X_1, \ldots, X_n$ iid with $\mathbb{E}X_1 = \mu$, $\operatorname{Var} X_1 = \sigma^2$, 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 $Y_1, \ldots, Y_n$ 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 $\mathbb{E}|g(X)| < \infty$.

Our bootstrap result concerns only the standardized pivot $T_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$.

But these pivots are the ones we can use (since we don't know $\sigma$):

$$T_{n,\mathrm{U}} = \sqrt{n}(\bar{X}_n - \mu) \quad \text{and} \quad T_{n,\mathrm{S}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n}$$

We can also use B–E to prove that the bootstrap "works" for these pivots.

**Discuss:** Is one better than the other??

# Table of Contents

## Fisher's z transformation

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid realizations of $(X, Y)$. Let

$$\rho = \text{corr}(X, Y) \quad \text{and} \quad \hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}.$$

If $(X, Y)$ are bivariate Normal then $\sqrt{n}(\zeta(\hat{\rho}) - \zeta(\rho)) \xrightarrow{\text{D}} \text{Normal}(0, 1)$, where

$$\zeta(\rho) = \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right). \quad \text{(can show by delta method)}$$

Then an asymptotic $(1-\alpha) \times 100\%$ CI for $\rho$ is given by

$$\left[\zeta^{-1}\left(\zeta(\hat{\rho}) - \frac{z_{\alpha/2}}{\sqrt{n}}\right), \zeta^{-1}\left(\zeta(\hat{\rho}) + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\right], \quad \zeta^{-1}(z) = \frac{e^{2z} - 1}{e^{2z} + 1}.$$

**Exercise:** Compare asymp. interval with percentile boot interval in simulation:

1. When $(X, Y)$ bivariate Normal with correlation $\rho = 1/2$.
2. When $Y|X \sim \text{Normal}(X, \sigma^2)$, $X \sim \text{Expo}(\lambda)$, $\text{corr}(X, Y) = \sqrt{\lambda^2/(\lambda^2 + \sigma^2)}$.


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