Lecture 2: Bootstrap 101

Raul Riva

FGV EPGE

October, 2025

Magic 101

Magic 101

- First, we will do some magic and then explain why it works;

Magic 101

- First, we will do some magic and then explain why it works;
- Assume $X_i \sim N(\theta, 1)$;
- Let's say you have a sample of size n from this distribution, $X_1, \dots, X_n$;
- One natural estimator of θ is the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$;

Magic 101

- First, we will do some magic and then explain why it works;
- Assume $X_i \sim N(\theta, 1)$;
- Let's say you have a sample of size n from this distribution, $X_1, \dots, X_n$;
- One natural estimator of θ is the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- You would like to find a confidence interval for θ
- You know that $R_n \equiv \sqrt{n} \cdot (\bar{X}_n - \theta) \sim N(0, 1)$
- But let's say you don't know this finite-sample result...
- How can we use R_n to construct a confidence interval for θ?

I will propose a way!

- In this case, it would be easy to numerically compute the distribution of R_n because we **know** where the data comes from: resample!
- In practice you cannot ask "for more data"... you have *to pull yourself up by your bootstraps*!
- The only thing you can use are the numbers you got $(x_1, ..., x_n)$.

A Simple Strategy

You cannot resample from the original distribution… but can resample from the empirical distribution…

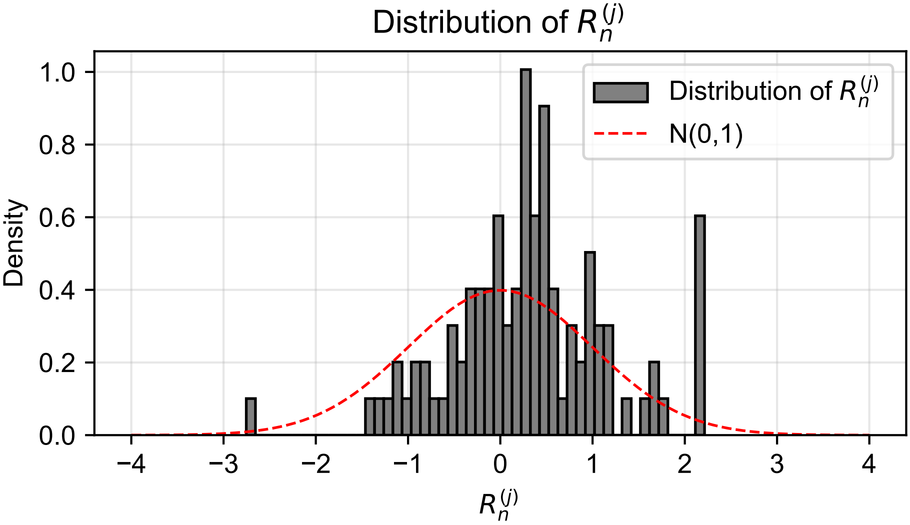
A Simple Strategy

You cannot resample from the original distribution... but can resample from the empirical distribution...

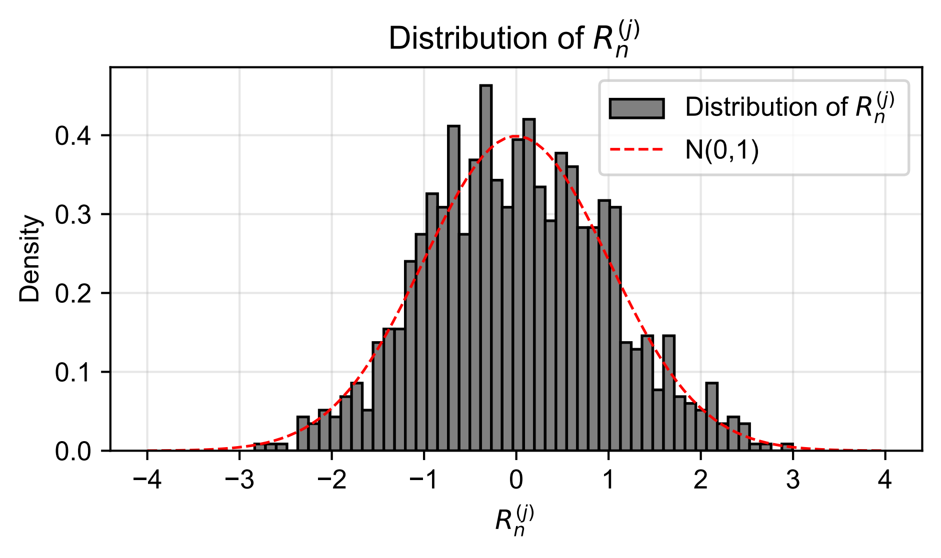
Ok, let's do this:

1. Draw a sample of size n from the empirical distribution of the original sample $(x_1, \ldots, x_n)$. This is just a random sample with replacement from the original sample. Call this new sample $(x_1^{(j)}, \ldots, x_n^{(j)})$;
2. Compute the sample mean of this new sample and call it $\bar{X}_n^{(j)} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(j)}$;
3. Compute $R_n^{(j)} = \sqrt{n} \cdot (\bar{X}_n^{(j)} - \bar{X}_n)$;
4. Repeat steps 1-3 B times to get $R_n^{(1)}, \ldots, R_n^{(B)}$;
5. Plot a histogram of the $R_n^{(j)}$'s;

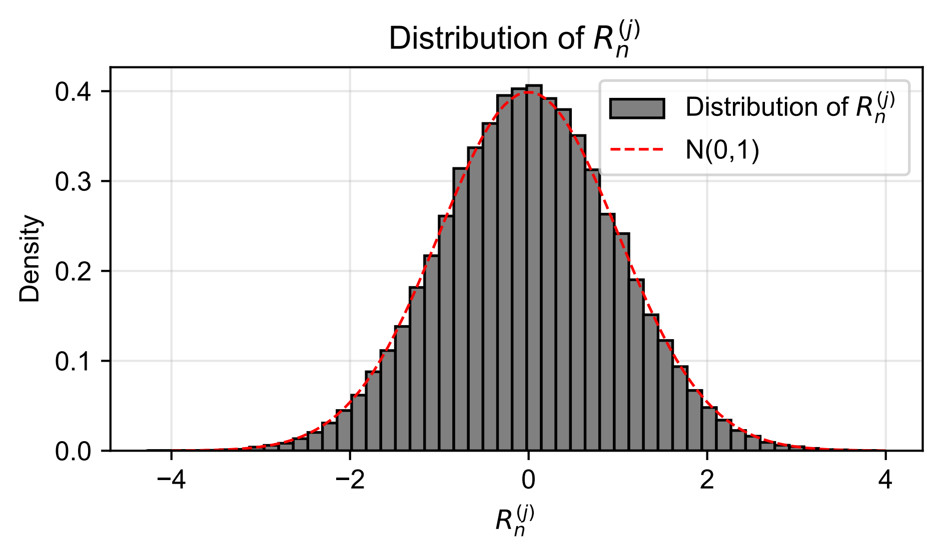
Low B


Distribution of $R_n^{(j)}$
Density
1.0
0.8
0.6
0.4
0.2
0.0
−4
−3
−2
−1
0
1
2
3
4
$R_n^{(j)}$
Distribution of $R_n^{(j)}$
N(0,1)

Medium B


Distribution of $R_n^{(j)}$
Distribution of $R_n^{(j)}$
N(0,1)
Density
0.0
0.1
0.2
0.3
0.4
−4
−3
−2
−1
0
1
2
3
4
$R_n^{(j)}$

High B


Distribution of $R_n^{(j)}$
Distribution of $R_n^{(j)}$
N(0,1)
Density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
$R_n^{(j)}$

What kind of dark magic is this? Do they teach this at Hogwarts?

- Not magic at all: just the bootstrap at work!
- As B increases, the distribution of $R_n^{(j)}$ converges to the distribution of R_n;
- Notice that we are keeping n fixed throughout the process;
- We were able to approximate the **finite-sample distribution** of this statistic;
- This lecture is a bird's eye view of the bootstrap and why it works (and why it doesn't);

What is the bootstrap?

What is the bootstrap?

- Assume X_i comes from some distribution P, and you have and i.i.d. sample $X_1, \dots, X_n$;
- Very often, we want to construct confidence intervals for a parameter $\theta(P)$;
- That is a set $C_n = C_n(X_1, \dots, X_n)$ such that $P(\theta(P) \in C_n) \approx 1 - \alpha$;

What is the bootstrap?

- Assume X_i comes from some distribution P, and you have and i.i.d. sample $X_1, \ldots, X_n$;
- Very often, we want to construct confidence intervals for a parameter $\theta(P)$;
- That is a set $C_n = C_n(X_1, ..., X_n)$ such that $P(\theta(P) \in C_n) \approx 1 - \alpha$;
- Tipically, we rely on some statistic that is a function of the data and this parameter, $R_n(X_1, \ldots, X_n; \theta(P)) \implies$ we call this a *root*.
- Obviously, the distribution of this root might depend on the distribution P;

What is the bootstrap?

- Let's define $J_n(x, P) \equiv P(R_n \leq x) \implies$ the (finite-sample) distribution of the root;

What is the bootstrap?

- Let's define $J_n(x, P) \equiv P(R_n \leq x) \implies$ the (finite-sample) distribution of the root;
- In some cases, it does not depend on P;
- In the "magical example", we had $J_n(x, P) = \Phi(x)$, the CDF of the standard normal;
- In that case, it would be easy to come up with a confidence interval;

What is the bootstrap?

- Let's define $J_n(x, P) \equiv P(R_n \leq x) \implies$ the (finite-sample) distribution of the root;
- In some cases, it does not depend on P;
- In the "magical example", we had $J_n(x, P) = \Phi(x)$, the CDF of the standard normal;
- In that case, it would be easy to come up with a confidence interval;
- Even if X_i were not Gaussian, we would still have $J_n(x, P) \to \Phi\left(x/\sigma(P)\right)$ as $n \to \infty$ by the standard CLT;
- Then we could create a confidence set that would be asymptotically valid, at least;

What is the bootstrap?

But these two cases are more the exception than the rule...

1. Usually, $J_n(x, P)$ depends on P in an unknown way;
2. Even if you get a CLT, what is the quality of the approximation? When is n "large enough"?

What is the bootstrap?

But these two cases are more the exception than the rule...

1. Usually, $J_n(x, P)$ depends on P in an unknown way;
2. Even if you get a CLT, what is the quality of the approximation? When is n "large enough"?

- What if you get a CLT, but the limiting distribution is super complicated?
- Also very common: the asymptotic distribution might depend on parameters that hard to estimate...

What is the bootstrap?

But these two cases are more the exception than the rule...

1. Usually, $J_n(x, P)$ depends on P in an unknown way;
2. Even if you get a CLT, what is the quality of the approximation? When is n "large enough"?

- What if you get a CLT, but the limiting distribution is super complicated?
- Also very common: the asymptotic distribution might depend on parameters that hard to estimate...

What we really want is $J_n(x, P)$!

What is the bootstrap?

The basic idea is the following:

- We don't know P but we know $\hat{P}_n$, the empirical distribution of the sample $(X_1, \dots, X_n)$;

$$\hat{P}_n(u) \equiv \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq u\}}$$

What is the bootstrap?

The basic idea is the following:

- We don't know P but we know $\hat{P}_n$, the empirical distribution of the sample $(X_1, \dots, X_n)$;

$$\hat{P}_n(u) \equiv \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq u\}}$$

- If P_n is a good approximation of P, we might have a chance to approximate $J_n(x, P)$ using $J_n(x, \hat{P}_n)$
- Intuitively, this approximation is only good if
  - $\hat{P}_n$ is a good approximation of P;
  - $J_n(x, P)$ has some "continuity" with respect to P;

What is the bootstrap?

The general algorithm for the (non-parametric) bootstrap:

Definition (Non-parametric Bootstrap)

1. Draw a sample of size n **with replacement** using the empirical distribution $\implies$ it's ok if some observations are repeated!
2. Compute the statistic of interest and use moments from $\hat{P}_n$ whenever you need population moments;
3. Repeat steps 1-2 B times to get a list of realized statistics $R_n^{(1)}, \dots, R_n^{(B)}$;
4. Use the empirical distribution of the $R_n^{(j)}$'s to approximate $J_n(x, P)$;

What is the bootstrap?

The general algorithm for the (non-parametric) bootstrap:

Definition (Non-parametric Bootstrap)

1. Draw a sample of size n **with replacement** using the empirical distribution $\implies$ it's ok if some observations are repeated!
2. Compute the statistic of interest and use moments from $\hat{P}_n$ whenever you need population moments;
3. Repeat steps 1-2 B times to get a list of realized statistics $R_n^{(1)}, \ldots, R_n^{(B)}$;
4. Use the empirical distribution of the $R_n^{(j)}$'s to approximate $J_n(x, P)$;

- Use the empirical distribution of the $R_n^{(j)}$'s to construct confidence intervals, get quantiles, etc.
- Treat the distribution from the bootstrap as if it were the true distribution of R_n;

Questions?

Why does it work?

Why does it work?

- First question: why does $\widehat{P}_n$ approximate P?

Why does it work?

- First question: why does $\hat{P}_n$ approximate P?
- A pointwise result is simple to get:

$$\hat{P}_n(u) \xrightarrow{p} P(u) \text{ as } n \to \infty, \forall u \in \mathbb{R}$$

because $\mathbb{E}(I_{X_i \leq u}) = P(u)$ and we can apply the LLN.

Why does it work?

- First question: why does $\hat{P}_n$ approximate P?
- A pointwise result is simple to get:

$$\hat{P}_n(u) \xrightarrow{p} P(u) \text{ as } n \to \infty, \forall u \in \mathbb{R}$$

because $\mathbb{E}(I_{X_i \leq u}) = P(u)$ and we can apply the LLN.

- But we can do so much better than that!

Theorem (Glivenko-Cantelli)

Let $X_1, \dots, X_n$ be scalar random variables with distribution P. Then, as $n \to \infty$, we have that

$$\sup_{u \in \mathbb{R}} \left| \hat{P}_n(u) - P(u) \right| \xrightarrow{p} 0.$$

Glivenko-Cantelli Withouth Math

- As n grows, the empirical distribution $\hat{P}_n$ converges uniformly to the true distribution P;
- This is: for any point in the support of P, $\hat{P}_n$ will do a great job as an approximation!
- If you have enough data, $\hat{P}_n$ is almost as good as being able to observe P directly;
- Caveat: how much is "enough data"?
- You can see the proof on Hansen's book or any Probability book - it's not hard.
- It's possible to generalize this result to vector-valued random variables, but we will not do that here.

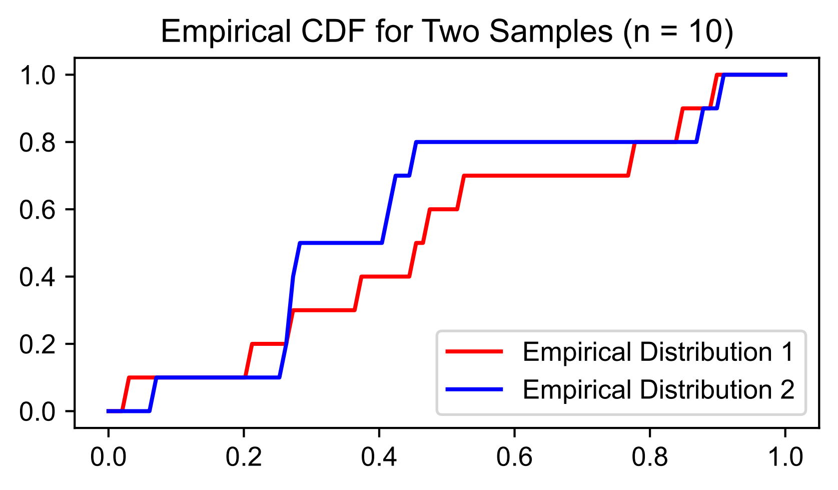
What about the continuity of $J_n(x, P)$?

- This is quite involved and each type of root might need a different treatment;
- However, we have a well-developed theory for averages t-statistic-type roots;
- We will focus on understanding these results and not on proving them;
- You should definitely check out Bruce Hansen's book on Chapter 10;

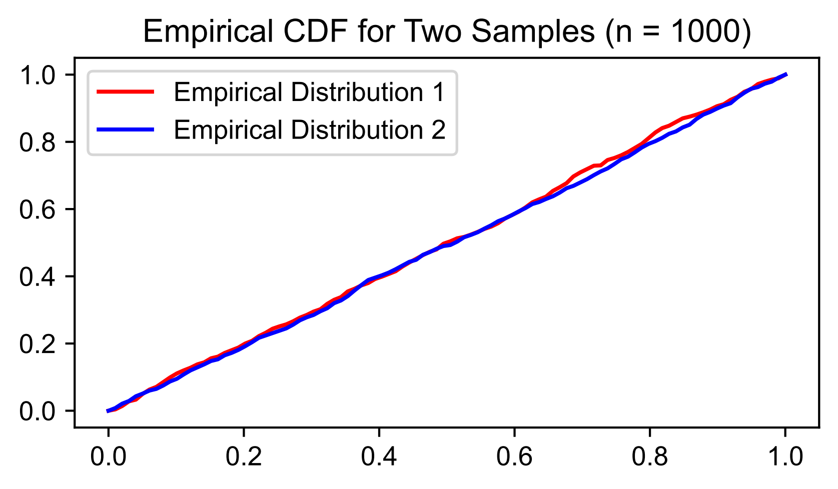
Asymptotic Theory

Asymptotic Theory

- The main complication is that $\hat{P}_n$ is a **random function**;
- For a given size n, two different samples will induce two different measures;
- To see that, let $X_i \sim U[0, 1]$ and $n = 10$ and let's compute the empirical distributions;
- Then we will increase n... what will happen?


Empirical CDF for Two Samples (n = 10)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Empirical Distribution 1
Empirical Distribution 2

Glivenko-Cantelli works, baby!



Asymptotic Theory

- We actually work with a single *realization* of the empirical distribution $\hat{P}_n$;
- I will denote this realization by P_n^*;
- Given a sample $(x_1, ..., x_n)$, this is

$$P_n^*(u) \equiv \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq u\}}$$

- Important: this is like a conditional CDF!

To get to the main theorem we will need some definitions...

Convergence in Bootstrap Probability

Definition (Convergence in Bootstrap Probability)

We say that a random vector Z_n^* **converges in bootstrap probability** to Z as $n \to \infty$, denoted $Z_n^* \xrightarrow{p^*} Z$, if for all $\epsilon > 0$

$$P_n^* \left(\|Z_n^* - Z\| > \epsilon \right) \xrightarrow{p} 0.$$

- How is this different than standard convergence in probability?
- There are two probability measures involved: who are they?

Convergence in Bootstrap Distribution

Definition (Convergence in Bootstrap Distribution)

Let Z_n^* be a sequence of random vectors with conditional distributions $G_n^*(x) = P_n^*\left[Z_n^* \leq x\right]$. We say that Z_n^* **converges in bootstrap distribution** to Z as $n \to \infty$, denoted $Z_n^* \xrightarrow{d^*} Z$, if for all x at which $G(x) = \mathbb{P}[Z \leq x]$ is continuous,

$$G_n^*(x) \xrightarrow{p} G(x) \text{ as } n \to \infty.$$

- How is this different than standard convergence in distribution?

The Main Theorem

Theorem (Asymptotic Bootstrap Theorem)

If $\{Y_i\}$ *are i.i.d. random vectors,* $\mathbb{E}\|Y\|^2 < \infty$*, and* $\Sigma = \mathrm{var}[Y] > 0$*, then as* $n \to \infty$*,*

$$\sqrt{n}\left(\overline{Y}^* - \overline{Y}\right) \xrightarrow{d^*} \mathcal{N}(0, \Sigma).$$

where $\overline{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^*$ *is the sample mean of a bootstrap sample and* $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ *is the sample mean of the original sample.*

The Main Theorem

Theorem (Asymptotic Bootstrap Theorem)

If $\{Y_i\}$ *are i.i.d. random vectors,* $\mathbb{E}\|Y\|^2 < \infty$*, and* $\Sigma = \mathrm{var}[Y] > 0$*, then as* $n \to \infty$*,*

$$\sqrt{n}\left(\overline{Y}^* - \overline{Y}\right) \xrightarrow{d^*} \mathcal{N}(0, \Sigma).$$

where $\overline{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^*$ *is the sample mean of a bootstrap sample and* $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ *is the sample mean of the original sample.*

- Notice that this is the same asymptotic distribution we would get for $\sqrt{n}(\overline{Y} - \mathbb{E}[Y_i])$;
- Why is this theorem useful?

The Main Theorem

Theorem (Asymptotic Bootstrap Theorem)

If $\{Y_i\}$ are i.i.d. random vectors, $\mathbb{E}\|Y\|^2 < \infty$, and $\Sigma = \mathrm{var}[Y] > 0$, then as $n \to \infty$,

$$\sqrt{n}\left(\overline{Y}^* - \overline{Y}\right) \xrightarrow{d^*} \mathcal{N}(0, \Sigma).$$

where $\overline{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^*$ is the sample mean of a bootstrap sample and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the sample mean of the original sample.

- Notice that this is the same asymptotic distribution we would get for $\sqrt{n}(\overline{Y} - \mathbb{E}[Y_i])$;
- Why is this theorem useful?
- The centering happens at the sample mean. Why? Any intuition?
- Importantly: the continuous mapping theorem and the Delta method are conserved!
- See details on Hansen's book;

Questions?

Why don't we just use the bootstrap everywhere?

- It might be super slow: imagine bootstrapping something that takes long to do *once*... now you have to do it B times!
- It might be less efficient than plug-in estimators;
- There are cases where it does not work at all $\implies$ you will see one example in the problem set;
- Usually, it will break if your root is not a "smooth" function;
- Example: the maximum or minimum of a sample;

The End