ECON 480-3
LECTURE 17: THE BOOTSTRAP

Ivan A. Canay
Northwestern University

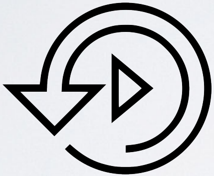
Past & Future

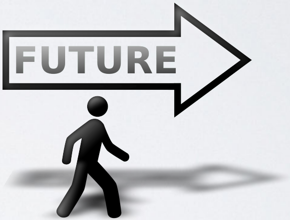
Part III so far

- HC Standard Errors
- HAC Standard Errors
- CCE Standard Errors

TODAY

- Confidence Sets and Pivots
- Bootstrap: Algorithm
- Bootstrap: Sample Mean
- Discussion





Confidence Sets

- **Data**: let $X_i, i = 1, \ldots, n$ be an i.i.d. sample of observations with distribution $P \in \mathbf{P}$.
- The **family** $\mathbf{P}$ may be a parametric, nonparametric, or semiparametric family of distributions.
- We are interested in making inferences about some **parameter** $\theta(P) \in \Theta = \{\theta(P) : P \in \mathbf{P}\}$.
- Examples of $\theta(P)$ are the mean of P, the median of P, of a regression coefficient, among others.
- **Confidence set for $\theta(P)$**: a random set $C_n = C_n(X_1, \ldots, X_n)$ such that
$$P\{\theta(P) \in C_n\} \approx 1 - \alpha \ ,$$
at least for n sufficiently large.
- The typical way of constructing such sets is based off of approximating the distribution of a **root**,
$$R_n = R_n(X_1, \ldots, X_n, \theta(P)) \ .$$
- **Root**: any real-valued function depending on both the data and the parameter of interest, $\theta(P)$.

Sampling distribution

- **Idea**: if the distribution of the root were known, then one could straightforwardly construct a confidence set for $\theta(P)$.
- Let $J_n(P)$ denote the sampling distribution of R_n and define the corresponding cdf as,

$$J_n(x, P) = P\{R_n \leqslant x\} \, .$$

- **Note**: the distribution of R_n depends on both the sample size, n, and the distribution of the data, P.
- Knowing $J_n(x, P)$ we may choose a constant c such that

$$P\{R_n \leqslant c\} \approx 1 - \alpha \, .$$

- The set $C_n = \{\theta \in \Theta : R_n(X_1, \ldots, X_n, \theta) \leqslant c\}$ is a confidence set in the sense described above.
- We may also choose c_1 and c_2 so that

$$P\{c_1 \leqslant R_n \leqslant c_2\} \approx 1 - \alpha \, ,$$

and construct the desired confidence set as

$$C_n = \{\theta \in \Theta : c_1 \leqslant R_n(X_1, \ldots, X_n, \theta) \leqslant c_2\} \, .$$

Pivots

- In some **rare** instances, $J_n(x, P)$ does **not depend** on P.
- In these instances, the root is said to be **pivotal** or a **pivot**.
- **Example**: if $\theta(P)$ is the mean of P and $\mathbf{P} = \{N(\theta, 1) : \theta \in \mathbf{R}\}$, then the root

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P))$$

  is a pivot because $R_n \sim N(0, 1)$.
- In this case, we may construct confidence sets C_n with **finite-sample validity**; that is,

$$P\{\theta(P) \in C_n\} = 1 - \alpha$$

  for all n and $P \in \mathbf{P}$.

Asymptotic Pivots

- Sometimes, the root may not be pivotal in the sense described above, but it may be **asymptotically pivotal** or an **asymptotic pivot**.
- **Asymptotic pivot**: $J_n(x, P)$ converges in distribution to a limit distribution $J(x, P)$ that does not depend on P.
- **Example**: $\theta(P)$ is the mean of P and $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance, then

$$R_n = \frac{\sqrt{n}(\bar{X}_n - \theta(P))}{\hat{\sigma}_n}$$

  is asymptotically pivotal because it converges in distribution to $J(x, P) = \Phi(x)$.
- We may then construct confidence sets that are asymptotically valid in the sense that

$$\lim_{n \to \infty} P\{\theta(P) \in C_n\} = 1 - \alpha$$

  for all $P \in \mathbf{P}$.

Asymptotic Approximations

- Typically, the root will be **neither** a pivot nor an asymptotic pivot.
- The distribution $J_n(x, P)$ and the limiting distribution $J(x, P)$ will typically depend on P.
- **Example**: $\theta(P)$ is the mean of P and $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance, then
$$R_n = \sqrt{n}(\bar{X}_n - \theta(P))$$
converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$.
- We can approximate this limit distribution with $\Phi(x/\hat{\sigma}_n)$, which will lead to confidence sets that are asymptotically valid in the sense described above.
- This approach depends very heavily on the limit distribution $J(x, P)$ being both **known** and **tractable**. Even if it is known, the limit distribution may be difficult to work with (e.g., it could be the supremum of some complicated stochastic process with many nuisance parameters).
- Even if it is **known** and **manageable**, the method may be poor in finite-samples because it essentially relies on a **double approximation**: 1 $J_n(x, P)$ is approximated by $J(x, P)$, then 2 $J(x, P)$ is approximated in some way by estimating the unknown parameters of the limit distribution.

The Bootstrap

- The **bootstrap** is a fourth, more general approach to approximating $J_n(x, P)$.
- The **idea** is very simple: replace the unknown P with an **estimate** $\hat{P}_n$.
- Given $\hat{P}_n$ it is possible to compute $J_n(x, \hat{P}_n)$ (either analytically or using simulation to any desired degree of accuracy).
- **Estimate:** 1 In the case of i.i.d. data, a typical choice is the **empirical distribution**.

$$\begin{array}{cccccc} \text{Data} & X_1 & X_2 & X_3 & \cdots & X_n \\ \hat{P}_n & \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{array}.$$

  2 If $P = P(\psi)$ for some finite-dimensional parameter ψ, then one may also use $\hat{P}_n = P(\hat{\psi}_n)$ for some estimate $\hat{\psi}_n$ of ψ.
- The hope is that whenever $\hat{P}_n$ is close to P, $J_n(x, \hat{P}_n)$ is close to $J_n(x, P)$.
- Essentially, this requires that $J_n(x, P)$, when viewed as a function of P, is **continuous** in an appropriate neighborhood of P.
- Often, this turns out to be true, but, unfortunately, it is not true in general.

QUESTIONS?

The case of the mean

- We will now consider the case where P is a distribution on $\mathbf{R}$ and $\theta(P)$ is the **mean** of P.

- We will consider first the **root** $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$.

- Let $\hat{P}_n$ denote the **empirical distribution** of the $\{X_i : i = 1, \ldots, n\}$.

$$\begin{array}{cccccc} \text{Data} & X_1 & X_2 & X_3 & \cdots & X_n \\ \hat{P}_n & \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{array}.$$

- **Question**: under what conditions is $J_n(x, \hat{P}_n)$ "close" to $J_n(x, P)$?

- The sequence of distributions $\hat{P}_n$ is a random sequence, so it is more convenient to answer the question first for a nonrandom sequence P_n and then extend to random sequences.

- Before presenting formal results, we discuss how to implement the bootstrap in practice for this case.

Implementation of The Bootstrap

- Most often, the bootstrap approximation $J_n(x, \hat{P}_n)$ cannot be calculated exactly.
- We can approximate this distribution to an arbitrary degree of accuracy by taking samples from $\hat{P}_n$
- Let's consider two cases for the case of the mean:

$$\mathbf{P}_{\text{np}} = \{\text{distributions with finite second moments}\} \quad \text{and} \quad \mathbf{P}_{\text{p}} = \{P : \exp(1/\theta)\} .$$

- Implementation of the bootstrap requires **4 steps**.

Bootstrap Algorithm

1. Conditional on the data $(X_1, \ldots, X_n)$, draw B **samples** of **size** n from $\hat{P}_n$. Denote the jth sample by

$$(X^*_{1,j}, \ldots, X^*_{n,j}) \quad \text{for} \quad j = 1, \ldots, \mathbf{B} .$$

Non-parametric: When $\mathbf{P} = \mathbf{P}_{\text{np}}$, $\hat{P}_n$ is typically the empirical distribution and this involves resampling the original observations with replacement (each observation has probability $1/n$).

Parametric: When $\mathbf{P} = \mathbf{P}_{\text{p}}$, $\hat{P}_n = \exp(1/\hat{\theta}_n)$ where $\hat{\theta}_n$ is, for example, the MLE of θ: $\hat{\theta}_n = \bar{X}_n$.

NOTE!: Step 1, how you take your samples, determines the **type** of bootstrap (parametric, non-parametric, Wild-bootstrap, clustered samples, and many others).

Bootstrap Algorithm

2. For each bootstrap sample j, compute **the root**, i.e.,

$$R^*_{j,n} = R_n(X^*_{1,j}, \ldots, X^*_{n,j}, \hat{\theta}_n) \quad \text{for} \quad j = 1, \ldots, \mathbf{B} \, .$$

Note: $\theta(\hat{P}_n) = \hat{\theta}_n$, so in the bootstrap distribution $\theta(P)$ becomes $\hat{\theta}_n$. In the case of the mean,

$$R^*_{j,n} = \sqrt{n}(\bar{X}^*_{j,n} - \bar{X}_n) \, .$$

3. Compute the **empirical cdf** of $(R^*_{1,n}, \ldots, R^*_{B,n})$ as

$$L_n(x) = \frac{1}{B} \sum_{j=1}^{B} I\{R^*_{j,n} \leqslant x\} \, .$$

By the Glivenko-Cantelli Theorem, $\sup_{x \in \mathbf{R}} |L_n(x) - J(x, \hat{P}_n)| \to 0$ as $B \to \infty$. Thus, the Glivenko-Cantelli Theorem implies we can achieve an arbitrary degree of accuracy by choosing B sufficiently big. In practice, researcher use $B = 1,000$ or above, given enough computational power.

Bootstrap Algorithm

4. Compute the desired function of $L_n(x)$, for example, a quantile,

$$L_n^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : L_n(x) \geqslant 1-\alpha\} ,$$

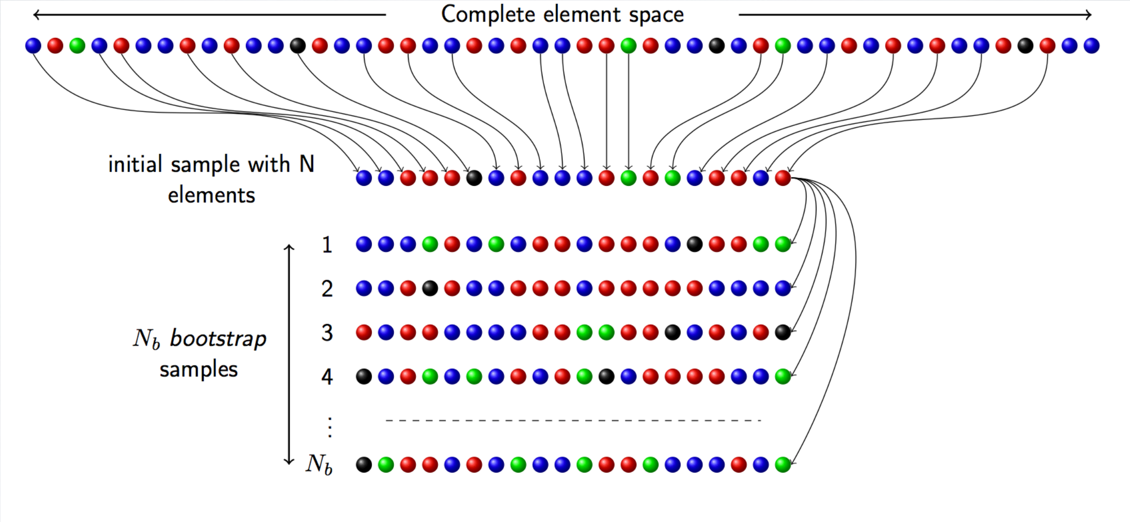
for a given significance level α.

Note! In practice you usually skip step 3 and compute step 4 directly using a `quantile` function:

$$L_n^{-1}(1-\alpha) = \texttt{quantile}_{1-\alpha}\left(R^*_{1,n}, R^*_{2,n}, \ldots, R^*_{B,n}\right) .$$

- In Step 1, it matters the family of distributions (parameter vs nonparametric).
- In Step 2, it matters what the parameter of interest is and what the root is.
- Step 3 may be skipped if step 4 can be calculated directly as above.
- We are now ready to start with the **formal results**

The Bootstrap: illustration with marbels



QUESTIONS?

Intermediate Step: nonrandom sequence

Theorem

Let $\theta(P)$ be the mean of P and let $\mathbf{P}$ denote the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance. Consider the root

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) .$$

Let $P_n, n \geqslant 1$ be a nonrandom sequence of distributions such that:

1. P_n converges in distribution to P
2. $\theta(P_n) \to \theta(P)$
3. $\sigma^2(P_n) \to \sigma^2(P)$.

Then,

(I) $J_n(x, P_n)$ converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$.

(II) The $1-\alpha$ quantile $J_n^{-1}(1-\alpha, P_n) = \inf\{x \in \mathbf{R} : J_n(x, P_n) \geqslant 1-\alpha\}$ converges to

$$J^{-1}(1-\alpha, P) = z_{1-\alpha}\sigma(P) .$$

PROOF OF (I)

For each n, let $X_{i,n}, i = 1, \ldots, n$ be i.i.d. with distribution P_n.

WTS: $\sqrt{n}(\bar{X}_{n,n} - \theta(P_n)) \xrightarrow{d} N(0, \sigma^2(P))$.

PROOF OF (I)

1. Suppose that Y_n and Y are real valued random variables and that $Y_n \xrightarrow{d} Y$.
   If the Y_n are uniformly bounded, then $E[Y_n] \to E[Y]$.
2. Suppose that $Y_n \xrightarrow{d} Y$. Let g be a measurable map from $\mathbf{R}$ to $\mathbf{R}$. Let C be the set of points in $\mathbf{R}$ for which g is continuous. If $P\{Y \in C\} = 1$, then $g(Y_n) \xrightarrow{d} g(Y)$.

PROOF OF (I)

Claim that $\lim_{n\to\infty} E[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\}] = 0$ and apply the Lindeberg-Feller CLT.

PROOF OF PART (II)

Part (II) follows from part (i) and Lemma 2 below applied to $F_n(x) = J_n(x, P)$ and $F(x) = J(x, P)$.

LEMMA

Let $F_n, n \geqslant 1$ and F be nonrandom distribution functions on $\mathbf{R}$ such that F_n converges in distribution to F. Suppose F is continuous and strictly increasing at

$$F^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : F(x) \geqslant 1-\alpha\}\,.$$

Then,

$$F_n^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : F_n(x) \geqslant 1-\alpha\} \to F^{-1}(1-\alpha)\,.$$

Proof: see Lecture Notes.

Main Theorem

We are now ready to pass from the nonrandom sequence P_n to the random sequence $\hat{P}_n$.

Theorem

Let $\theta(P)$ be the mean of P and let $\mathbf{P}$ denote the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance. Consider the root

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) \; .$$

Then,

(I) $J_n(x, \hat{P}_n)$ converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$ a.s.

(II) $J_n^{-1}(1-\alpha, \hat{P}_n)$ converges to $J^{-1}(1-\alpha, P) = z_{1-\alpha}\sigma(P)$ a.s.

Proof: it follows from the previous theorem if we show that $\hat{P}_n$ satisfies the requirements imposed on P_n a.s.:

1. $\hat{P}_n$ converges in distribution to P a.s.
2. $\theta(\hat{P}_n) \to \theta(P)$ a.s.
3. $\sigma^2(\hat{P}_n) \to \sigma^2(P)$ a.s.

Remarks

- Similar results hold for the **studentized root** where $\hat{\sigma}_n$ is consistent for $\sigma(P)$.
- Using this root leads to the so-called **Bootstrap-t**, as the root is just the t-statistic.
- A **key step** in the proof of this result is to show that $\hat{\sigma}_n$ converges in probability to $\sigma(P)$ under an appropriate sequence of distributions. We do this in 481.
- The **advantage** of working with a studentized root is that the limit distribution of R_n is **pivotal**, which affects the properties of the bootstrap approximation as discussed in the next section.
- **By Slutsky's Theorem** confidence sets of the form

$$C_n = \left\{\theta \in \mathbf{R} : R_n(X_1, \ldots, X_n, \theta) \leqslant J_n^{-1}(1-\alpha, \hat{P}_n)\right\} ,$$

  which are known as **symmetric** confidence sets, or

$$C_n = \left\{\theta \in \mathbf{R} : J_n^{-1}\left(\frac{\alpha}{2}, \hat{P}_n\right) \leqslant R_n(X_1, \ldots, X_n, \theta) \leqslant J_n^{-1}\left(1-\frac{\alpha}{2}, \hat{P}_n\right)\right\} ,$$

  which are known as **equi-tailed** confidence sets, satisfy

$$P\{\theta(P) \in C_n\} \to 1-\alpha \quad \text{for all} \quad P \in \mathbf{P} .$$

Asymptotic Refinements

- **Q**: is the bootstrap better than an asymptotically normal approximation?
- **Yes**: under certain conditions (ensuring existence of so-called Edgeworth expansions of $J_n(x, P)$) it follows that one-sided confidence sets C_n based off such an asymptotic approximation satisfy

$$\left|P\{\theta(P) \in C_n\} - (1-\alpha)\right| = O\left(\frac{1}{\sqrt{n}}\right) . \tag{1}$$

- One-sided confidence sets based off of the bootstrap and the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$ also satisfy (1), though there is some evidence to suggest that it does a bit better in the size of $O(n^{-1/2})$ term.
- On the other hand, one-sided confidence sets based off the bootstrap-t, i.e., using the root

$$R_n = \frac{\sqrt{n}(\bar{X}_n - \theta(P))}{\hat{\sigma}_n}$$

satisfy

$$\left|P\{\theta(P) \in C_n\} - (1-\alpha)\right| = O\left(\frac{1}{n}\right) . \tag{2}$$

- **Refinement**: the coverage error of the bootstrap-t interval is $O(n^{-1})$ and is of **smaller order** than that provided by the normal approximation or the bootstrap based on a nonstudentized root.

Asymptotic Refinements

- A heuristic reason why the bootstrap based on the studentized root outperforms the bootstrap based on the nonstudentized root is as follows.
  - **Nonstudentized case**: the bootstrap is estimating a distribution that has mean 0 and unknown variance $\sigma^2(P)$. The main contribution to the estimation error is the implicit estimation of $\sigma^2(P)$ by $\sigma^2(\hat{P}_n)$.
  - **Studentized case**: the studentized root has a distribution that is nearly independent of P since it is an asymptotic pivot.
- The bootstrap may also provide a refinement in two-sided tests. For example, symmetric intervals based on the absolute value of the studentized root are $O(n^{-2})$, versus the asymptotic approximation that is of order $O(n^{-1})$.
- Note that, by construction, such intervals are symmetric about $\hat{\theta}_n$.
- **Final comment**: The bootstrap may fail and work poorly in finite samples when $J(x, P)$ is **not** continuous in P. This happens in some settings relevant to economics: Auction models, Entry Game Models, Missing data settings, etc.

THE END!