ECON 481-3
LECTURE 12: UNIFORMITY OF SUBSAMPLING

Ivan A. Canay
Northwestern University

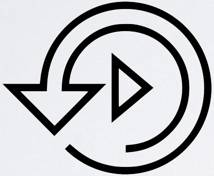
Past & Future

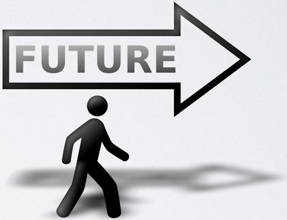
LAST CLASS

- Dist. with compact support
- Dist. with $2 + \delta$ moments
- Uniformity of the t-test
- Power of the t-test

TODAY

- Review of Subsampling
- Uniformity issues with Subsampling
- Parameter at the Boundary
- Asymptotic Size of Subsampling





Intro to Subsampling

- **Data**: $\{X_i, i = 1, \ldots, n\}$ is an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$.
- **Parameter of interest**: some real-valued $\theta(P)$
- **Estimator**: $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$.
- **Root**:

$$T_n = \sqrt{n}(\hat{\theta}_n - \theta(P)) ,$$

  where root stands for a functional depending on both, the data and $\theta(P)$.
- Let $J_n(P)$ denote the **sampling distribution** of T_n and define the corresponding cdf as,

$$J_n(x, P) = P\{T_n \leqslant x\} . \qquad (1)$$

- **Goal**: to estimate $J_n(x, P)$ so we can make inferences about $\theta(P)$. For example, we would like to estimate quantiles of $J_n(x, P)$, so we can construct confidence sets for $\theta(P)$. Unfortunately, **we do not know** P, and, as a result, we do not know $J_n(x, P)$.

Main Requirement

- **The bootstrap**: solved this problem simply by replacing the unknown P with an estimate $\hat{P}_n$.

- In the case of i.i.d. data, a typical choice of $\hat{P}_n$ is the empirical distribution of the $X_i, i = 1, \ldots, n$.

- **Condition**: for this approach to work, we essentially required that $J_n(x, P)$ when viewed as a function of P was **continuous** in a certain neighborhood of P.

- An alternative to the bootstrap known as **subsampling**, originally due to Politis and Romano (2004), does not impose this requirement but rather the following much weaker condition.

Assumption

There exists a limiting law $J(P)$ such that $J_n(P)$ converges weakly to $J(P)$ as $n \to \infty$.

Intuition

- Suppose for the time being that $\theta(P)$ **is known**.
- Suppose $X_i, i = 1, \ldots, m$ is an i.i.d. sequence of random variables with distribution P with $m = nk$ for some **very big** k (so we have many samples of size n).
- We could then estimate $J_n(x, P)$ by looking at the empirical distribution of

$$\sqrt{n}\left(\hat{\theta}_n(X_{n(j-1)+1}, \ldots, X_{nj}) - \theta(P)\right), \quad j = 1, \ldots, k\,.$$

- This is an i.i.d. sequence of k rvs with distribution $J_n(x, P)$. By the Glivenko-Cantelli theorem, we know that the empirical distribution is a good estimate of $J_n(x, P)$, at least for large k.
- **Improvement**: we can do better by using all possible sets of data of size n from the m observations,

$$\sqrt{n}\left(\hat{\theta}_{n,j} - \theta(P)\right), \quad j = 1, \ldots, \binom{m}{n}\,,$$

where $\hat{\theta}_{n,j}$ is the estimate of $\theta(P)$ using the jth set of data of size n from the original m observations.

Reality

- In practice $m = n$, so, even if we knew $\theta(P)$, this idea won't work.
- **Key idea!** replace n with some smaller number b that is much smaller than n.
- We would then expect

$$\sqrt{b}\left(\hat{\theta}_{b,j} - \theta(P)\right), \quad j = 1, \ldots, \binom{n}{b},$$

where $\hat{\theta}_{b,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size b from the original n observations, to be a good estimate of $J_b(x, P)$, at least if $\binom{n}{b}$ is large.

- **But**: we are interested in $J_n(x, P)$, not $J_b(x, P)$. We therefore need some way to force $J_n(x, P)$ and $J_b(x, P)$ to be close to one another.
- To ensure this, it suffices to assume that $J_n(x, P) \to J(x, P)$. Therefore, $J_b(x, P)$ and $J_n(x, P)$ are both close to $J(x, P)$, and thus close to one another as well, at least for large b and n.

$$|J_b(x, P) - J_n(x, P)| \leqslant |J_b(x, P) - J(x, P)| + |J_n(x, P) - J(x, P)| \, .$$

Intuition

- **Both b and $\binom{n}{b}$ need to be large**: it suffices to assume that $b \to \infty$, but $b/n \to 0$.

- This procedure is still not feasible because in practice we typically do not know $\theta(P)$. But we can replace $\theta(P)$ with $\hat{\theta}_n$ provide

$$\sqrt{b}(\hat{\theta}_n - \theta(P)) = \frac{\sqrt{b}}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta(P))$$

  is **small**, which follows from $b/n \to 0$ in this case.

- All we required was that $J_n(x, P)$ converged in distribution to a limit distribution $J(x, P)$. The bootstrap required this and that $J_n(x, P)$ was continuous in a certain sense.

- Showing continuity of $J_n(x, P)$ is very **problem specific**. On the flip side, we now have a tuning parameter: b.

Main Theorem

Theorem

Assume Assumption A. Also, let $J_n(P)$ denote the sampling distribution of $\tau_n(\hat{\theta}_n - \theta(P))$ for some normalizing sequence $\tau_n \to \infty$, $N_n = \binom{n}{b}$, and assume that $\tau_b/\tau_n \to 0$, $b \to \infty$, and $b/n \to 0$ as $n \to \infty$.

I) *If x is a continuity point of $J(\cdot, P)$, then $L_{n,b}(x) \to J(x, P)$ in probability, where*

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{j=1}^{N_n} I\{\tau_b(\hat{\theta}_{n,b,j} - \hat{\theta}_n) \leqslant x\} .$$

II) *If $J(\cdot, P)$ is continuous, then*

$$\sup_x |L_{n,b}(x) - J_n(x, P)| \to 0 \textit{ in probability} .$$

III) *Let*

$$c_{n,b}(1-\alpha) = \inf\{x : L_{n,b}(x) \geqslant 1-\alpha\} \quad \textit{and} \quad c(1-\alpha, P) = \inf\{x : J(x, P) \geqslant 1-\alpha\} .$$

If $J(\cdot, P)$ is continuous at $c(1-\alpha, P)$, then

$$P\{\tau_n(\hat{\theta}_n - \theta(P)) \leqslant c_{n,b}(1-\alpha)\} \to 1-\alpha \textit{ as } n \to \infty .$$

QUESTIONS?

Uniformity Issues with Subsampling

- Andrews and Guggenberger (2010) study the properties of subsampling in a broad class of non-regular models.
- They consider cases in which a test statistic has a discontinuity in its asymptotic distribution as a function of the true distribution that generates the observations (P).
- In such cases bootstrap procedures typically do not provide pointwise asymptotically valid inference.
- For such problems subsampling has often been advocated, but the arguments in favor have been based on “point-wise” asymptotics.
- Start with a **simple example**: parameter is at the boundary of the parameter space.

Parameter at the Boundary

Example (Parameter at the boundary)

- Suppose $X_i, i = 1, \ldots, n$ are i.i.d. with distribution $P \in \mathbf{P} = \{N(\theta(P), 1) : \theta(P) \geqslant 0\}$.

- **Maximum Likelihood Estimator**: $\hat{\theta}_n = \max\{\bar{X}_n, 0\}$.

- Consider the **root**

$$T_n = \sqrt{n}(\hat{\theta}_n - \theta(P)) = \sqrt{n}\Big(\max\{\bar{X}_n, 0\} - \theta(P) \Big)$$

$$= \max\left\{ \sqrt{n}(\bar{X}_n - \theta(P)), -\sqrt{n}\theta(P) \right\} \xrightarrow{d} \begin{cases} \max\{Z, 0\} & \text{if } \theta(P) = 0 \\ Z & \text{if } \theta(P) > 0 \end{cases}$$

  where $Z \sim N(0, 1)$.

- **Notation**: $J_0 \equiv \max\{Z, 0\}$ and $J_\infty \equiv Z$.

- Before moving to subsampling, we will show that $J_n(x, \hat{P}_n)$ (the bootstrap approximation) **does not** converge to $J(x, P)$ a.s. in this particular case.

Failure of the Bootstrap I

- For each n: let $X_{n,i}, i = 1, \ldots, n$ be i.i.d. with distribution P_n (not necessarily in $\mathbf{P}$) st

$$(i)\ P_n \text{ converges in distribution to } P, \quad (ii)\ \theta(P_n) \to \theta(P), \quad (iii)\ \sigma^2(P_n) \to \sigma^2(P)$$

- The distribution $J_n(x, P_n)$, under P_n is simply the distribution of

$$T_n = \sqrt{n}(\hat{\theta}_{n,n} - \theta(P_n)) = \sqrt{n}\Big(\max\{\bar{X}_{n,n}, 0\} - \theta(P_n)\Big) = \max\left\{\sqrt{n}(\bar{X}_{n,n} - \theta(P_n)), -\sqrt{n}\theta(P_n)\right\}.$$

- **WLOG** $\theta(P) = 0$. Let $c > 0$ and

$$(iv)\ \text{ suppose } \sqrt{n}\theta(P_n) > c \text{ for all } n.$$

For such a sequence P_n,

$$T_n \leqslant \max\{\sqrt{n}(\bar{X}_{n,n} - \theta(P_n)), -c\} \xrightarrow{d} \max\{Z, -c\},$$

under P_n, which is dominated by the distribution of $\max\{Z, 0\}$.

- To complete the argument, it suffices to show that $\hat{P}_n$ satisfies a.s. the requirements on P_n.

Failure of the Bootstrap II

- **(i)**: By the SLLN $\hat{P}_n$ converges in distribution to P a.s.
- **(ii)**: By the SLLN $\theta(\hat{P}_n) \to \theta(P)$ a.s.
- **(iii)**: By the SLLN $\sigma^2(\hat{P}_n) \to \sigma^2(P)$ a.s.
- **(iv)**: It remains to determine whether $\sqrt{n}\theta(\hat{P}_n) > c$ for all n a.s. Equivalently, we need to determine whether

$$\bar{X}_n > \frac{c}{\sqrt{n}} \text{ for all } n \text{ a.s.}$$

Unfortunately, the SLLN will not suffice for this purpose. Instead, we will need the following refinement of the SLLN known as the **Law of the Iterated Logarithm (LIL)**:

Theorem

Let $Y_i, i = 1, ..., n$ be an i.i.d. sequence of random variables with distribution P on $\mathbf{R}$. Suppose $\mu(P) = 0$ and $\sigma^2(P) = 1$. Then,

$$\limsup_{n\to\infty} \frac{\bar{Y}_n}{\sqrt{\frac{2\log\log n}{n}}} = 1 \textit{ a.s.} \quad \textit{and} \quad \liminf_{n\to\infty} \frac{\bar{Y}_n}{\sqrt{\frac{2\log\log n}{n}}} = -1 \textit{ a.s.}$$

Failure of the Bootstrap III

$$\limsup_{n\to\infty} a_n = a \iff \text{for any } \epsilon > 0 \quad a_n > a - \epsilon \text{ i.o.} \quad \text{and} \quad a_n < a + \epsilon \text{ for } n \text{ sufficiently large .}$$

- **LIL then implies**: $\bar{Y}_n > (1-\epsilon)\sqrt{\frac{2\log\log n}{n}}$ i.o. a.s.

QUESTIONS?

Subsampling: pointwise behavior

- **Subsampling Estimator**: $\hat{\theta}_{b_n,j} = \max\{\bar{X}_{b_n,j}, 0\}$, where $\bar{X}_{b_n,j}$ is the sample average of the b_n observations in the jth subsample.

- **Subsampling Root**:

$$T_{b_n,j}(P) = \sqrt{b_n}(\hat{\theta}_{b_n,j} - \theta(P)) = \sqrt{b_n}\left(\max\{\bar{X}_{b_n,j}, 0\} - \theta(P)\right) = \max\left\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta(P)), -\sqrt{b_n}\theta(P)\right\}.$$

- **Immediate**:

$$\text{If } \theta(P) = 0 \quad \Rightarrow \quad T_{b_n,j}(P) \xrightarrow{d} J_0 = \max\{Z, 0\}$$

$$\text{If } \theta(P) > 0 \quad \Rightarrow \quad T_{b_n,j}(P) \xrightarrow{d} J_\infty = Z\ .$$

- As opposed to the bootstrap, subsampling provides the right limiting behavior under standard asymptotics based on a fixed probability distribution.

- Andrews and Guggenberger show that if a sequence of test statistics has an asymptotic null distribution that is **discontinuous in a nuisance parameter** (as in the previous example), then a subsample test does not necessarily yield the desired asymptotic level.

Subsampling: uniform behavior

- **Subsampling Feature**: there are two different rates of drift such that over-rejection and under-rejection can occur. We will show this using the previous example

- Let γ_n be a localization sequence that measures how “far” or “close” we are from $\theta(P) = 0$.

- Consider a sequence of null distributions P_n such that $\theta_n = \theta(P_n) = \gamma_n$ and look at the behavior of T_n and $T_{b_n,j}$ along the sequence.

- **Complication**: the asymptotic distribution of T_n is discontinuous at $\gamma = 0$.

- **Remark**: Here θ_n and γ_n are the same parameter but they may be different in general.

Uniformity Issues with Subsampling I

- **Drift Sequence 1**: $\gamma_n = \frac{h}{\sqrt{n}}$. Study $T_n = \sqrt{n}(\hat{\theta}_n - \theta_n)$ vs $T_{b_n,j} = \sqrt{b_n}(\hat{\theta}_{b_n,j} - \theta_n)$.
- **Full sample test statistic**:
- **Sub-sample test statistic**:

Comments

- **Full-sample test statistic**: asymptotic dist. depends on a "local parameter" h.
- **Subsample test statistic** : asymptotic distribution for the case $h = 0$.
- **Claim**: $J_h(x) \geqslant J_0(x)$ for all x, where $J_h(x) = P\{\max\{Z, -h\} \leqslant x\}$.

$$J_h(x) = J_0(x) \text{ for all } x \geqslant 0 \text{ while } J_h(x) > J_0(x) \text{ for all } x \in [-h, 0) \ ,$$

- **Subsampling**: gives a good approximation in the right tail but a poor one in the left-tail.
- **Result**: an upper one-sided subsample CI for $\theta(P)$, which relies on a subsample critical value from the right tail of the subsample distribution, will perform well.

Comments

- **Upper one-sided CI**: let $c_h(1-\alpha)$ be the $1-\alpha$-quantile of J_h: $1 - J_h(c_h(1-\alpha)) = \alpha$. Then

  $$P_n\{T_n > c_0(1-\alpha)\} \to 1 - J_h(c_0(1-\alpha)) \ ,$$

  and

  $$1 - J_h(c_0(1-\alpha)) \leqslant 1 - J_0(c_0(1-\alpha)) = \alpha \ .$$

- Indeed, if we look at the **quantiles**:

  $$c_h(1-\alpha) = c_0(1-\alpha) > 0 \text{ for } \alpha < 1/2$$
  $$c_h(1-\alpha) < c_0(1-\alpha) = 0 \text{ for } \alpha > 1/2$$

- **Equal-tailed and symmetric two-sided SS CIs**: will perform poorly. Let $\bar{c}_h(1-\alpha) > 0$ be st

  $$1 - J_h(\bar{c}_h(1-\alpha)) + J_h(-\bar{c}_h(1-\alpha)) = \alpha \ .$$

  For this critical value, we have

  $$P_n\left\{|T_n| > \bar{c}_0(1-\alpha)\right\} \to 1 - J_h(\bar{c}_0(1-\alpha)) + J_h(-\bar{c}_0(1-\alpha))$$
  $$\geqslant 1 - J_0(\bar{c}_0(1-\alpha)) + J_0(-\bar{c}_0(1-\alpha)) = \alpha \ .$$

- Subsampling may lead to **over-rejection**. **Example**: for $h = 2$ the 95% quantile of the distribution of $|\max\{Z, 0\}|$ is 1.63, while the 95% quantile of the distribution of $|\max\{Z, -h\}|$ is 1.96.

Uniformity Issues with Subsampling II

- **Drift Sequence 2**: $\gamma_n = \frac{g}{\sqrt{b_n}}$. Study $T_n = \sqrt{n}(\hat{\theta}_n - \theta_n)$ vs $T_{b_n,j} = \sqrt{b_n}(\hat{\theta}_{b_n,j} - \theta_n)$
- **Full sample test statistic**:
- **Sub-sample test statistic**:

Asymptotic Size of Subsampling

- Consider the family of distributions $\mathbf{P} = \{P_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$. Here γ might be infinite dimensional.

- The **exact size** of a test that rejects $H_0 : \theta(P) = \theta_0$ when $T_n(\theta_0) > c_{1-\alpha}$ is:

$$ExSz_n = \sup_{\gamma \in \Gamma} P_{\theta_0,\gamma} \{T_n(\theta_0) > c_{1-\alpha}\} \ ,$$

  since $\mathbf{P}_0 = \{P_{\theta,\gamma} : \theta = \theta_0, \gamma \in \Gamma\}$.

- The **asymptotic size** of the test is defined by

$$AsySz = \limsup_{n \to \infty} \sup_{\gamma \in \Gamma} P_{\theta_0,\gamma} \{T_n(\theta_0) > c_{1-\alpha}\} \ .$$

- Our interest is in the exact finite-sample size of the test; we just use asymptotics to approximate this.

- Uniformity over $\gamma \in \Gamma$, which is built into the definition of $AsySz_n$, is necessary for the asymptotic size to give a good approximation to the finite-sample size.

Comments

- What we learned from the example in the previous section is in fact a general result.
- Let J_h denote the **asymptotic distribution** of T_n under a sequence γ_n (i.e., under $P_n = P_{\theta_0,\gamma_n}$) st

$$h = \lim_{n\to\infty} \sqrt{n}\gamma_n \quad \text{and} \quad g = \lim_{n\to\infty} \sqrt{b_n}\gamma_n \, ,$$

  for some $h \in H = [0,\infty]$ and $g \in H = [0,\infty]$.
- Under the **same** sequence P_n, let J_g denote the **asymptotic distribution** of $T_{b_n,j}$.
- The set of all possible pairs of localization parameters (g,h) is denoted by GH and is defined by

$$\mathbf{GH} = \Big\{(g,h) \in H \times H : g = 0 \text{ if } h < \infty \;\&\; g \in [0,\infty] \text{ if } h = \infty\Big\} \, .$$

- **NOTE**: $g \leqslant h$ for all $(g,h) \in \mathbf{GH}$. In the previous example, we got $(g,h) = (0,0)$ and $(g,h) = (\infty,\infty)$ by standard asymptotics; and $(g,h) = (0,h)$ and $(g,h) = (g,\infty)$ using drifting sequences.

General Case

Lemma

Suppose that for all $h \in H$ *and all sequences* $\{\gamma_n : n \geqslant 1\}$, $T_n \xrightarrow{d} J_h$ *under* $\{P_{\theta_0,\gamma_n} : n \geqslant 1\}$ *for some distribution* J_h. *Then,*

$$AsySz = \sup_{(g,h)\in \mathbf{GH}} \left[1 - J_h\Big(c_g(1-\alpha)\Big)\right] ,$$

provided a certain assumption in Andrews and Guggenberger (2010) holds.

- Therefore: $AsySz \leqslant \alpha$ **iff** $c_g(1-\alpha) \geqslant c_h(1-\alpha)$.

- The general results can be used to show for example that:
  1. In an instrumental variables (IVs) regression model with potentially weak IVs, all nominal level $1-\alpha$ one-sided and two-sided subsampling tests concerning the coefficient on an exogenous variable and based on the two-stage least squares (2SLS) estimator have asymptotic size equal to one;
  2. In models where (partially-identified) parameters are restricted by moment inequalities, subsampling tests and CIs based on suitable test statistics have correct asymptotic size.

Comments

- The approach given above is based on sequences of nuisance parameters and it requires verifying certain assumptions for all possible sequences of nuisance parameters.

- In particular, proving that a test controls asymptotic size typically requires to argue that it is not possible to find a non-stochastic (sub)sequence of parameters γ_n such that:

$$\lim_{n \to \infty} P_{\theta(P), \gamma_n} \{T_n > c_{1-\alpha}\} > \alpha \,.$$

- Proving the latter typically involves deriving the asymptotic distribution of the test statistic along all possible non-stochastic sequences $\gamma_n \in \Gamma$.

- A **different approach** includes the one in Romano and Shaikh (2012), who show that subsampling tests are valid whenever the family $\mathbf{P}$ satisfies,

$$\lim_{n \to \infty} \sup_{P \in \mathbf{P}} \sup_{x \in \mathbf{R}} |J_b(x, P) - J_n(x, P)| = 0 \,.$$

  The authors also provide uniform results for Bootstrap tests.

THE END