ECON 480-3
LECTURE 18: SUBSAMPLING AND RANDOMIZATION TESTS

Ivan A. Canay
Northwestern University

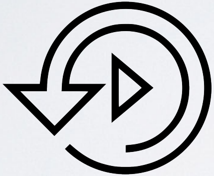
Past & Future

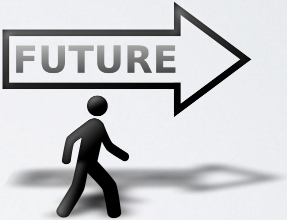
Part III so far

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

LAST CLASS!

- Subsampling
- Subsampling vs Bootstrap
- Randomization Tests
- Example: Permutation tests





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**:

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

  where root stands for a functional depending on both, the data and $\theta(P)$.

- Let $J_n(P)$ denote the **sampling distribution** of R_n and define the corresponding cdf as,

$$J_n(x, P) = P\{R_n \leqslant x\} \ . \quad (1)$$

- **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 (1994), 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 $\hat{\theta}_n$ provided

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

  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.

QUESTIONS?

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*

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

III) *Let*

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

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

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

Implementing Subsampling

Except for the first step, implementing the bootstrap and subsampling requires the **same algorithm**.

Nonparametrics Bootstrap

1. Conditional on the data $(X_1, \ldots, X_n)$, draw B **samples** of **size** n from the original observations with replacement (each observation has probability $1/n$). Denote the jth sample by

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

Subsampling

1. Conditional on the data $(X_1, \ldots, X_n)$, draw N_n **samples** of **size** b from the original observations without replacement. Denote the jth sample by

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

In practice, N_n is too large to actually compute $L_n(x)$, so what one would do is randomly sample B of the N_n possible data sets of size b and just use B in place of N_n when computing $L_n(x)$.

Comments

- **Bootstrap**: there are examples where $J_n(x,P) \to J(x,P)$, but the continuity on P fails (e.g., the extreme order statistic).

- Subsampling would have **no problems** handling the extreme order statistic.

- Typically, when both the bootstrap and subsampling are valid, the bootstrap works better in the sense of **higher-order asymptotics** but subsampling is more **generally valid**.

- There is a variant of the bootstrap known as the **m-out-of-n bootstrap**.
  - Instead of using $J_n(x,\hat{P}_n)$ to approximate $J_n(x,P)$, one uses $J_m(x,\hat{P}_n)$ where m is much smaller than n.
  - Assuming $m^2/n \to 0$, then all the conclusions of the theorem remain valid with $J_m(x,\hat{P}_n)$ in place of $L_n(x)$.
  - This follows because if $m^2/n \to 0$, then (i) $m/n \to 0$ and (ii) with probability tending to 1, the approximation to $J_m(x,\hat{P}_n)$ is the same as the approximation to $L_n(x)$ because the probability of drawing all distinct observations tends to 1 (see formal proof in class notes).

QUESTIONS?

Randomization Tests: motivation

Example (Sign changes)

- Let $X = (X_1, \ldots, X_{10}) \sim P$ be an i.i.d. sample of size 10 where each X_i takes values in $\mathbf{R}$, has a finite mean $\theta \in \mathbf{R}$, and has a distribution that is **symmetric about θ**.
- Let $\mathbf{P}$ be the collection of all distributions P satisfying these conditions.
- Consider testing

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \neq 0 \, .$$

- $n = 10$ so using an asymptotic approximation does not seem fruitful. At the same time, this is more general than the normal location model so exploiting normality is not possible.
- Suppose we decided to use the absolute value of $\bar{X}_{10}$ to test the above hypothesis: $T(X) = |\bar{X}_{10}|$.
- **Question**: how do we compute a critical value that delivers a valid test? It turns out we can do this by exploiting **symmetry**.

Randomization Tests: motivation

Example (Sign changes)

- Let ϵ_i take on either the value 1 or -1 for $i = 1, \ldots, 10$.
- Note that the distribution of $X = (X_1, \ldots, X_{10})$ is **symmetric about 0** under the null hypothesis.
- Now consider a transformation $g = (\epsilon_1, \ldots, \epsilon_{10})$ of $\mathbf{R}^{10}$ that defines the following mapping

$$(X_1, \ldots, X_{10}) \mapsto gX = (\epsilon_1 X_1, \ldots, \epsilon_{10} X_{10}) \; .$$

- Let $\mathbf{G}$ be the $M = 2^{10}$ collection of such transformations.
  $\Rightarrow$ the random variable X and gX have the **same distribution** under the null hypothesis.
- What this means is that we can get "new samples" from P by simply applying g to X. We can get a total of $M = 1,024$ samples and use these samples to simulate the distribution of $T(X)$.
- This approach leads to a test that is **valid in finite samples** as the next section shows.

Randomization Tests: definition

- **Data**: $X \sim P$ taking values in a sample space $\mathcal{X}$. Note! P is now the distribution of the entire sample.
- Want to test the null hypothesis $H_0 : P \in \mathbf{P}_0$, where $\mathbf{P}_0 \subset \mathbf{P}$.
- Let $\mathbf{G}$ be a finite group of transformations $g : \mathcal{X} \mapsto \mathcal{X}$.
- The following assumption allows for a general test construction.

Definition (Randomization Hypothesis)

Under the null hypothesis, the distribution of X is **invariant under the transformations** in $\mathbf{G}$; that is, for every $g \in \mathbf{G}$, gX and X have the **same distribution** whenever $X \sim P \in \mathbf{P}_0$.

- Note: We do not require the alternative hypothesis parameter space to remain invariant under g in $\mathbf{G}$. Only the space $\mathbf{P}_0$ is assumed invariant.
- Note: a Group always include the identity transformation.

THE TEST

- Let $T(X)$ be any **test statistic** for testing H_0. Let $|\mathbf{G}| = M$. Given $X = x$, let

$$T^{(1)}(x) \leqslant T^{(2)}(x) \leqslant \cdots \leqslant T^{(k)}(x) \leqslant \cdots \leqslant T^{(M)}(x)$$

be ordered values of $T(gX)$ as g varies in $\mathbf{G}$.

- For a nominal level $\alpha \in (0,1)$, let k be defined as

$$k = \lceil (1-\alpha)M \rceil$$

where $\lceil M\alpha \rceil$ denotes the smallest integer greater than or equal to $M\alpha$. Let

$$M^+(x) = \sum_{j=1}^{M} I\left\{T^{(j)}(x) > T^{(k)}(x)\right\} \quad \text{and} \quad M^0(x) = \sum_{j=1}^{M} I\left\{T^{(j)}(x) = T^{(k)}(x)\right\}.$$

- Now set

$$a(x) = \frac{M\alpha - M^+(x)}{M^0(x)} \qquad \text{and} \qquad \phi(x) = \begin{cases} 1 & T(x) > T^{(k)}(x) \\ a(x) & T(x) = T^{(k)}(x) \\ 0 & T(x) < T^{(k)}(x) \end{cases}. \tag{2}$$

- **Note**: $M^+(x) \leqslant M - k \leqslant M\alpha$ and $M^+(x) + M^0(x) \geqslant M - k + 1 > M\alpha$ imply $a(x) \in [0,1)$.

Comments

- Under the randomization hypothesis, Hoeffding (1952) shows that:
  1. this construction results in a test of **exact level α**,
  2. this is true for **any choice** of test statistic $T(X)$.
- This is possibly a **randomized test** if $\lfloor M\alpha \rfloor$ is not an integer and there are ties in the ordered values.
- Randomized tests are useful for theoretical purposes but not so useful for empirical practice.
- In practice, one may prefer not to randomized, and so the slightly conservative but not randomized test that rejects when $T(X) > T^{(k)}$ is level α:

$$\phi^{\mathrm{nr}}(x) = I\{T(x) > T^{(k)}(x)\}\,.$$

Theorem

Theorem

Suppose that X has distribution P in $\mathcal{X}$ and the problem is to test the null hypothesis $P \in \mathbf{P}_0$.
Let $\mathbf{G}$ be a finite group of transformations of $\mathcal{X}$ onto itself.
Suppose the **randomization hypothesis** *holds. Given a test statistic $T(X)$, let ϕ be the randomization test as described above.*
Then, $\phi(X)$ is a similar α level test, i.e.,

$$E_P[\phi(X)] = \alpha, \text{ for all } P \in \mathbf{P}_0. \quad (3)$$

Remark

The randomization test not only is of level α for all n, but also "similar", meaning that $E_P[\phi(X)]$ is never below α for any $P \in \mathbf{P}_0$.

Proof

$$M^{+}(x) = \sum_{j=1}^{M} I\{T^{(j)}(x) > T^{(k)}(x)\} \quad \text{and} \quad M^{0}(x) = \sum_{j=1}^{M} I\{T^{(j)}(x) = T^{(k)}(x)\}.$$

$$a(x) = \frac{M\alpha - M^{+}(x)}{M^{0}(x)} \quad \text{and} \quad \phi(x) = \begin{cases} 1 & T(x) > T^{(k)}(x) \\ a(x) & T(x) = T^{(k)}(x) \\ 0 & T(x) < T^{(k)}(x) \end{cases}.$$

QUESTIONS?

Special case: Permutation Tests

Economics: popular application of randomization tests are the so-called **permutation tests**.

Example (Two sample problem)

- Suppose that $Y_1, \ldots, Y_m$ are i.i.d. observations from a distribution P_Y and, independently, $Z_1, \ldots, Z_n$ are i.i.d. observations from a distribution P_Z.
- We have two samples that are not paired, i.e., Z_1 and Y_1 do not correspond to the same "unit".
- Here X is given by
$$X = (Y_1, \ldots, Y_m, Z_1, \ldots, Z_n) \, .$$
- Consider testing
$$H_0 : P_Y = P_Z \text{ vs } H_1 : P_Y \neq P_Z \, .$$
- **Group of transformations**: Let $N = m + n$ and for $x = (x_1, \ldots, x_N) \in \mathbf{R}^N$, let $gx \in \mathbf{R}^N$ be
$$(x_1, \ldots, x_N) \mapsto gx = (x_{\pi(1)}, \ldots, x_{\pi(N)}) \, , \tag{4}$$
where $(\pi(1), \ldots, \pi(N))$ is a **permutation** of $\{1, \ldots, N\}$. Let $\mathbf{G}$ be the collection of all such g, so that $M = N!$. It follows that whenever $P_Y = P_Z$, X and gX have the **same distribution**.

Comments

- In essence, each transformation g produces a **new data set** gx
- The first m elements are used as the Y sample and the remaining n as the Z sample to recompute the test statistic.
- If a test statistic is chosen that is **invariant** under permutations within each of the Y and Z samples, like
$$\bar{Y}_m - \bar{Z}_n \ ,$$
it is enough to consider the $\binom{N}{m}$ transformed data sets obtained by taking m observations from all N as the Y observations and the remaining n as the Z observations
- This is equivalent to using a subgroup $\mathbf{G}'$ of $\mathbf{G}$.
- **Note**: The randomization hypothesis here holds when $P_Y = P_Z$.

EXAMPLE (TREATMENT EFFECTS)

- **Data**: random sample $\{(Y_1, D_1), \ldots, (Y_n, D_n)\}$ from a randomized controlled trial where

$$Y = Y(1)D + (1 - D)Y(0)$$

is the observed outcome and $D \in \{0, 1\}$ is the exogenous treatment assignment.

- Suppose that we are interested in testing the **hypothesis** that the distribution Q_0 of $Y(0)$ is the same as the distribution Q_1 of $Y(1)$:

$$H_0 : Q_0 = Q_1 \text{ vs. } H_1 : Q_0 \neq Q_1 \ . \tag{5}$$

- Under the null hypothesis in (5), it follows that the distribution of

$$\{(Y_1, D_1), \ldots, (Y_n, D_n)\} \quad \text{and} \quad \{(Y_1, D_{\pi(1)}), \ldots, (Y_n, D_{\pi(n)})\}$$

are the same for any **permutation** $(\pi(1), \ldots, \pi(n))$ of $\{1, \ldots, n\}$.

- A permutation test that permutes individual from "treatment" to "control" (or from "control" to "treatment") delivers a test that is **valid in finite samples**.

Permutation Tests and Treatment Effects

- Researchers: often interested in hypotheses about the average treatment effect (ATE):

$$H_0 : E[Y(1)] = E[Y(0)] \text{ v.s. } H_1 : E[Y(1)] \neq E[Y(0)] \ . \tag{6}$$

- One may still consider the permutation test that permutes the vector of **treatment assignments**.
- Unfortunately, such an approach **does not lead to a valid test** and may over-reject in finite samples.
- These test may be **asymptotically valid** though, after carefully choosing the **test statistic**.
- The distinction between the null hypothesis in (5) and that in (6) and their implications on the properties of permutation tests are **often ignored in applied research**.
- Randomization test are **often dismissed** in applied research due to the belief that the randomization hypothesis is too strong to hold in a real empirical application. **However**:
  - Randomization tests may be asymptotically valid even when P is not symmetric. See Bugni, Canay, and Shaikh (2018) for an example in the context of randomized controlled experiments.
  - Recent developments on "approximate" randomization tests show that they may be particularly useful in regression models with a fixed (and small) number of clusters, see Canay, Romano, Shaikh (2017).

THANK YOU FOR NOT FORCING ME TO TALK TO A BLACK SCREEN EVERY WEEK!