Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMS091)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 14
Bootstrap and MC tests
March 2, 2017
Plan of today’s lecture

1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
The frequentist approach to statistical inference

We assume that we have at hand

- observations \( y \)
- and a (possibly parametric) model \( \mathcal{P} \) for the data.

In this setting we want to make inference about some property (estimand) \( \tau = \tau(\mathbb{P}_0) \) of the distribution \( \mathbb{P}_0 \) that generated the data. For instance,

\[
\tau(\mathbb{P}_0) = \int x f_0(x) \, dx, \quad \text{(mean)}
\]

where \( f_0 \) is the density of \( \mathbb{P}_0 \).

The estimand \( \tau \) is estimated using a statistic \( t(y) \).
Some remarks:

- It is important to always keep in mind that the estimate $t(y)$ is an observation of a random variable $t(Y)$. If the experiment was repeated, resulting in a new vector $y$ of random observations, the estimator would take another value.

- In the same way, the error $\Delta(y) = t(y) - \tau$ is a realization of the random variable $\Delta(Y) = t(Y) - \tau$.

- To assess the uncertainty of the estimator we thus need to analyze the distribution of the error $\Delta(Y)$ (error distribution).
Using the bootstrap algorithm we deal with this matter by

1. replacing $P_0$ by an data-based approximation $\hat{P}_0$ and

2. analyzing the variation of $\Delta(Y)$ by MC simulation from the approximation $\hat{P}_0$.

A generic way to obtain the approximation $\hat{P}_0$ is to use the empirical distribution.
The empirical distribution (ED)

The empirical distribution (ED) $\hat{P}_0$ associated with the data $y = (y_1, y_2, \ldots, y_n)$ gives equal weight ($1/n$) to each of the $y_i$'s (assuming that all values of $y$ are distinct).

Consequently, if $Z \sim \hat{P}_0$ is a random variable, then $Z$ takes the value $y_i$ with probability $1/n$.

The empirical distribution function (EDF) associated with the data $y$ is defined by

$$\hat{F}_n(z) = \hat{P}_0(Z \leq z)$$

$$= \frac{1}{n} \sum_{i=1}^{n} 1\{y_i \leq z\} = \text{fraction of } y_i \text{'s that are less than } z.$$
The EDF
Consequently a sample $Y^*$ of size $n$ from the empirical distribution $\hat{P}_0$ associated with the observations $y = (y_1, y_2, \ldots, y_n)$ is generated by

1. drawing indices $I_1, I_2, \ldots, I_n$ independently from the uniform distribution on the integers $\{1, 2, \ldots, n\}$, and

2. letting $Y^* = (y_{I_1}, y_{I_2}, \ldots, y_{I_n})$.

Note that this algorithm draws $n$ values from the set $\{y_1, y_2, \ldots, y_n\}$ with replacement.
The bootstrap

- Having access to data $y$, we may now replace $P_0$ by $\hat{P}_0$.
- Any quantity involving $P_0$ can now be approximated by plugging $\hat{P}_0$ into the quantity instead. In particular,

$$\tau = \tau(P_0) \approx \hat{\tau} = \tau(\hat{P}_0),$$

which, e.g., in the case of the mean, becomes

$$\tau = \int yf_0(y)\,dy \approx \frac{1}{n} \sum_{i=1}^{n} y_i.$$

- Moreover, the uncertainty of $t(y)$ can be analyzed by drawing repeatedly $Y^* \sim \hat{P}_0$ and look at the variation (histogram) of $\Delta(Y^*) = t(Y^*) - \tau \approx t(Y^*) - \hat{\tau}$.
- In the case of the empirical distribution, simulation from $\hat{P}_0$ is carried through by simply drawing, with replacement, among the values $y_1, \ldots, y_n$. 
The algorithm goes as follows.

- Construct an approximation \( \hat{P}_0 \) of \( P_0 \) from the data \( y \).
- Simulate \( B \) new data sets \( Y_b^* \), \( b \in \{1, 2, \ldots, B\} \), where each \( Y_b^* \) has the size of \( y \), from \( \hat{P}_0 \).
- Compute the values \( t(Y_b^*) \), \( b \in \{1, 2, \ldots, B\} \), of the estimator.
- By setting in turn \( \Delta_b^* = t(Y_b^*) - \hat{\tau} \), \( b \in \{1, 2, \ldots, B\} \), we obtain values being approximately distributed according to the error distribution. These can be used for uncertainty analysis.
1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
The bootstrap algorithm considered above is non-parametric in the sense that we have no assumptions on the distribution \( P_0 \) apart from the samples being i.i.d.; in particular, we do not assume that \( P_0 \) belongs to a certain parametric family.

Our approximation \( \hat{P}_0 \) of \( P_0 \) is the empirical distribution function.

The simulation step boils down to drawing from the empirical distribution, i.e. drawing from the data with replacement.
1. **Last time:** Introduction to bootstrap (Ch. 9)

2. **More on the bootstrap** (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. **MC methods for hypothesis testing** (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
Example: law schools

We have average test scores (LSAT and GPA) from 15 american law schools and want to investigate if the two scores are correlated, i.e. $\tau$ is the correlation between the two datasets.

1. Our data consists of pairs $(x, y) = ((x_1, y_1), \ldots, (x_{15}, y_{15}))$.
2. Estimate the correlation of the data using the sample correlation

$$\hat{\tau} = t(x, y) = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{n \sum_i x_i^2 - (\sum_i x_i)^2} \sqrt{n \sum_i y_i^2 - (\sum_i y_i)^2}} \approx 0.776.$$

3. Create bootstrap samples $(X, Y)^*_b$, $b \in \{1, 2, \ldots, B\}$, where each sample $(X, Y)^*_b$ is generated by drawing 15 times with replacement from the pairs $(x_i, y_i)$, $i \in \{1, \ldots, 15\}$.
4. Calculate the correlation $t((X, Y)^*_b)$ for each random sample.
Given the \((X, Y)_b^*\)'s we create variables \(\Delta_b^* = t((X, Y)_b^*) - \hat{\tau}\), \(b \in \{1, 2, \ldots, B\}\), being approximately distributed according to the error distribution.

This gives that the bias of our estimate is approximately
\[
\mathbb{E}(\Delta(X, Y)) \approx \Delta^* = -0.0057.
\]

The bias-corrected estimate is
\[
t(x, y) - \Delta^* = 0.783.
\]

A one-sided 95\%-confidence interval for the correlation is consequently
\[
I_{0.05} = (\hat{\tau} - F^{-1}_\Delta(0.95), 1)
\approx (\hat{\tau} - \Delta^*[0.95B], 1)
= (0.614, 1).
\]
1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
In the non-parametric bootstrap we had no assumptions on the distribution function apart from the observed data $y$ being i.i.d.

In the **parametric** bootstrap we assume that data comes from a distribution $P_0 = P_{\theta_0} \in \{P_{\theta}; \theta \in \Theta\}$ belonging to some parametric family.

Instead of using the ED, we find an estimate $\hat{\theta} = \hat{\theta}(y)$ of $\theta_0$ from our observations and

1. generate new bootstrapped samples $Y_b^*$, $b \in \{1, 2, \ldots B\}$, from $\hat{P}_0 = \hat{P}_{\hat{\theta}}$.

2. After this we form, as usual, bootstrap estimates $\hat{\theta}(Y_b^*)$ and errors $\Delta_b^* = \hat{\theta}(Y_b^*) - \hat{\theta}$, $b \in \{1, 2, \ldots B\}$. 
A toy example: exponential distribution

We let $y = (y_1, \ldots, y_{20})$ be i.i.d. observations of $Y_i \sim \text{Exp}(\theta_0)$, with unknown mean $\theta_0$. The MLE of $\theta_0$ is $\hat{\theta}(y) = \bar{y}$ and following plot displays $\text{Exp}(\hat{\theta}(y))$ vs. $\text{Exp}(\theta_0)$. 

![Graph](https://via.placeholder.com/150)

Estimated vs. true density function, $n=20$

- Estimated density
- True density
In Matlab:

```matlab
n = 20;
B = 500;
theta_hat = mean(y);
boot = zeros(1,B);
for b = 1:B, % bootstrap
    y_boot = exprnd(theta_hat,1,n);
    boot(b) = mean(y_boot);
end
delta = sort(boot - theta_hat); % sorting to obtain quantiles
alpha = 0.05; % CB level
L = theta_hat - delta(ceil((1 - alpha/2)*B)); % forming CB
U = theta_hat - delta(ceil(alpha*B/2));
```
A toy example: exponential distribution (cont.)
1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
Semi-parametric bootstrap

- We assume a parametric model for the data, for instance
  \[ Y_i = k x_i + m + \epsilon_i, \quad i \in \{1, 2, \ldots n\}, \]

  and a non-parametric model for the residuals \( \epsilon_i \).

- Our only assumption on the residuals is that they are i.i.d.

- Given data \( y = (y_1, \ldots, y_n) \) we want to construct estimators \( \hat{k}(y) \) and \( \hat{m}(y) \) of the parameters \( k \) and \( m \) and assess the uncertainty of the estimates.

- To do the latter, we would generate bootstrap samples \( Y^*_b \) and parameter estimates \( \hat{k}(Y^*_b) \) and \( \hat{m}(Y^*_b) \) and study the variation of e.g. \( \Delta^*_b = \hat{k}(Y^*_b) - \hat{k}(y) \).

- A confidence interval for \( k \) is then given by
  \[
  \left( \hat{k}(y) - \Delta^*_{B(1-\alpha/2)} \right, \hat{k}(y) - \Delta^*_{B\alpha/2} \right).
  \]
We proceed as follows:

- Find estimators $\hat{k} = \hat{k}(y)$ and $\hat{m} = \hat{m}(y)$ for the parameters using least squares.
- Estimate the residuals as

$$\hat{\epsilon}_i = y_i - \hat{k}x_i - \hat{m}, \quad i \in \{1, 2, \ldots, n\}.$$ 

- Now, the $\hat{\epsilon}_i$’s approximately form an i.i.d. sample from an unknown distribution. For $b = 1, 2, \ldots, B$,
  1. resample the residuals to generate a bootstrap sample $\epsilon^*_b = (\epsilon_1, \ldots \epsilon_n)^*_b$ and
  2. Use the bootstrapped residuals to generate bootstrapped observations

$$(Y^*_i)_b = \hat{k}x_i + \hat{m} + (\epsilon_i)^*_b.$$ 

- Given the bootstrapped observations, estimate the parameters to obtain $\hat{k}(Y^*_b)$ and $\hat{m}(Y^*_b)$. 

Example: linear regression

As an example,

- assume that \( Y_i = kx_i + m + \epsilon_i \), with standard Gaussian residuals.
- To test the semi-parametric bootstrap we simulate a data set with \( m = 3 \) and \( k = 4 \).
- Given data, the parameters are estimated using least squares estimation.
- For comparison, we know from the theory of linear regression that an exact confidence interval for \( k \) is given by

\[
I_\alpha = \left( \hat{k} - t_{\alpha/2}(n-2)s_b, \hat{k} + t_{\alpha/2}(n-2)s_b \right).
\]

where

\[
s_b^2 = \frac{1}{n-2} \sum_i \hat{\epsilon}_i^2 \frac{1}{\sum_i (x_i - \bar{x})^2}.
\]
Example: Simple regression

Applying this to the given data set yields

$$I_{0.05} = (3.84, 4.79).$$

For a comparison we applied semi-parametric as well as parametric bootstrap to the same data set.

- Using semi-parametric bootstrap, where we resample the estimated residuals, we obtain the interval

$$I_{0.05} = (3.85, 4.78).$$

- Instead using parametric bootstrap, where we draw new residuals from $\mathcal{N}(0, \hat{\sigma}^2)$, we obtain

$$I_{0.05} = (3.86, 4.77).$$
Summary: Different types of bootstrap

- **Non-parametric bootstrap**
  - makes no assumptions on the distribution apart from i.i.d.
  - needs more data than parametric.

- **Parametric bootstrap**
  - assumes that data comes from a parametric family of distributions.
  - needs less data to get good estimates due to stronger assumptions.
  - may however be sensitive to assumptions.

- **Semi-parametric bootstrap**
  - assumes a parametric model, coupled with non-parametric nuisance variables, often residuals.
  - is typically used for regression.
Last time: Introduction to bootstrap (Ch. 9)

More on the bootstrap (Ch. 9)
- Example: law schools
- Parametric bootstrap
- Semi-parametric bootstrap

MC methods for hypothesis testing (Ch. 9)
- Preliminaries
- MC tests
- Permutation tests
Last time: Introduction to bootstrap (Ch. 9)

More on the bootstrap (Ch. 9)
- Example: law schools
- Parametric bootstrap
- Semi-parametric bootstrap

MC methods for hypothesis testing (Ch. 9)
- Preliminaries
- MC tests
- Permutation tests
A statistical hypothesis is a statement about the distributional properties of data.

The goal of a hypothesis test is to see if data agrees with the statistical hypothesis.

Rejection of a hypothesis indicates that there is sufficient evidence in the data to make the hypothesis unlikely.

Strictly speaking, a hypothesis test does not accept a hypothesis; it fails to reject it.
Testing hypotheses

The basis of a hypothesis test consist of

- a **null hypothesis** $H_0$ that we wish to test.
- a **test statistic** $t(y)$, i.e. a function of the observed data $y$.
- a **critical region** $R$.

If the test statistic falls into the critical region, then we **reject** the null hypothesis $H_0$. 
Important concepts

**Significance**  The probability (risk) that the test incorrectly rejects the null hypothesis.

**Power**  The probability that the test correctly rejects the null hypothesis. Is a function of the true, unknown parameter.

**p-value**  The probability, given the null hypothesis, of observing a result at least as extreme as the test statistic.

**Type I error**  Incorrectly rejecting the null hypothesis.

**Type II error**  Failing to reject the null hypothesis.
Testing simple hypotheses

- A **simple hypothesis** specifies completely a single distribution for the data, e.g. $Y \sim \mathcal{N}(\theta, 1)$ with $\mathcal{H}_0 : \theta = 0$.
- We construct/define a test statistic $t(y)$ such that large values of $t(y)$ indicate evidence **against** $\mathcal{H}_0$.
- The $p$-value of the test is now $p(y) = \mathbb{P}(t(Y) \geq t(y) | \mathcal{H}_0)$.
- The rejection region is $R = \{y : p(y) \leq \alpha\}$, where $\alpha$ is the level of the test.
- Thus, to evaluate the $p$-value we need to find the distribution of $t(Y)$ under $\mathcal{H}_0$. 
1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   • Example: law schools
   • Parametric bootstrap
   • Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   • Preliminaries
   • MC tests
   • Permutation tests
MC test of a simple hypothesis

An MC-based algorithm for testing simple hypotheses is as follows:

1. Draw $N$ samples, $Y_1, \ldots, Y_N$, from the distribution specified by $H_0$.
2. Calculate the test statistic $t_i = t(Y_i)$ for each sample.
3. Estimate the $p$-value using MC integration by letting
   \[
   \hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{t_i \geq t(y)\}.
   \]
4. If $\hat{p}(y) \leq \alpha$, reject $H_0$. 

M. Wiktorsson } Monte Carlo and Empirical Methods for Stoch
1. Last time: Introduction to bootstrap (Ch. 9)

2. More on the bootstrap (Ch. 9)
   - Example: law schools
   - Parametric bootstrap
   - Semi-parametric bootstrap

3. MC methods for hypothesis testing (Ch. 9)
   - Preliminaries
   - MC tests
   - Permutation tests
Permutation tests

- The random variables of a set $Y$ is said to be exchangeable if they have the same distribution for all permutations.
- The distribution of $Y$ given the ordered sample is then the uniform distribution on the set of all permutations of $Y$.
- Conditioning on the ordered variables leads to permutation tests.
- Permutation tests can be very efficient in testing an exchangeable null-hypothesis against a non-exchangeable alternative, e.g. for testing if two samples differ in some way.
MC permutation test

An MC-based permutation test can be implemented as follows.

1. Draw $N$ permutations, $Y_1, \ldots, Y_N$, of the vector $y$.
2. Calculate the test statistic $t_i = t(Y_i)$ for each permutation.
3. Estimate the $p$-value using MC integration according to
   \[
   \hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{t_i \geq t(y)\}.
   \]
4. If $\hat{p}(y) \leq \alpha$, reject $\mathcal{H}_0$. 

M. Wiktorsson
Monte Carlo and Empirical Methods for Stoch
Example: pH data

- We have 273 historical and current pH-measurements of 149 lakes in Wisconsin and want to test if the pH-levels have increased.
- We assume that all measurements are independent and that historical measurements have a distribution $F_0$ and that new measurements have a distribution $G_0$.
- We want to test $H_0 : F_0 = G_0$ against $H_1 : F_0 \neq G_0$
Example: pH data (cont.)

- Assume that the distribution for current data can be written as $G_0(y) = F_0(y - \theta)$. That is, the mean of the current data is the mean of the historical data plus $\theta$.
- We now want to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$.
- Under $H_0$, all data are i.i.d. and thus exchangeable.
- We use the difference in the sample means as a test statistic.
- A permutation test using 10000 random permutations gives $p = 0.0185$. 