

Resampling techniques for statistical modeling

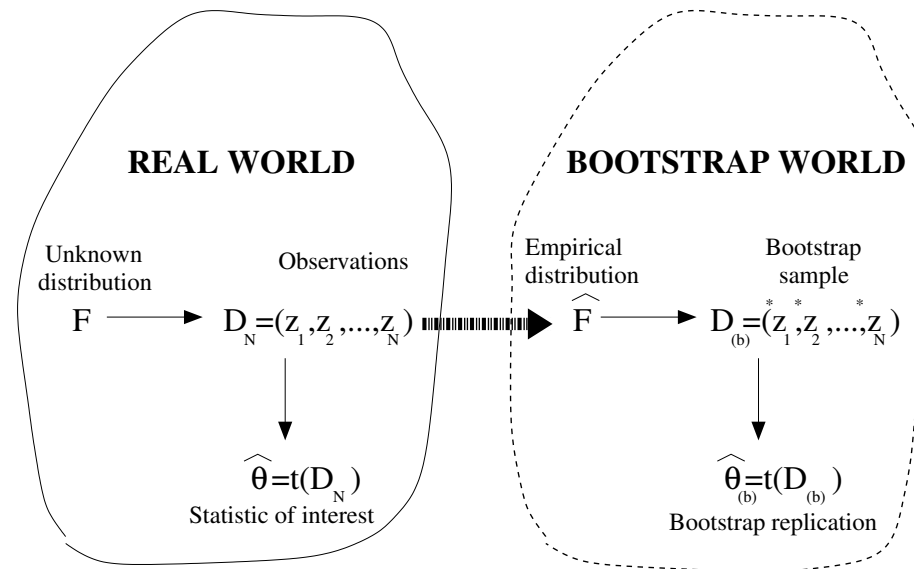
Gianluca Bontempi

Département d'Informatique
Boulevard de Triomphe - CP 212
<http://www.ulb.ac.be/di>

More complicated data structures

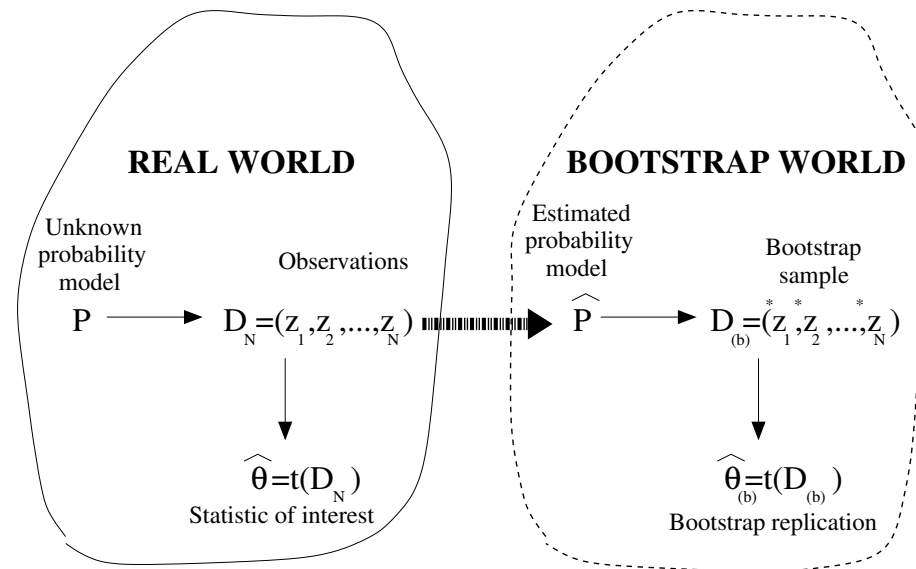
- So far we have considered the simplest possible probability model for random data: the **one-sample model** where a single unknown probability distribution F produces D_N by random sampling. Note that the individual data points z_i can themselves be quite complex (e.g. vectors) but the probability mechanism is simple.
- Many data analysis problems involve more complicated data structures (e.g. time series, regression models, censored data, stratified sampling, ...)

Bootstrap in one-sample problems



- In the real world, the unknown probability distribution F gives the data D_N and from D_N we calculate the statistic $\hat{\theta} = t(D_N)$.
- In the *bootstrap world* \hat{F} generates $D_{(b)}$ by random sampling giving $\hat{\theta}_{(b)} = t(D_{(b)})$.
- The crucial step is how we construct from D_N an estimate \hat{F} of the unknown population.

Beyond one-sample problems



Two are the practical problems to be solved to extend the bootstrap approach to a generic probability setting

1. Estimate the entire probability mechanism P underlying the observations.
2. Simulate bootstrap data from \hat{P} .

Bootstrap and regression

- Regression models are among the most useful and most used of statistical methods.
- Here we will consider bootstrap for assessing the accuracy of **parameter estimators**. Later, we will consider bootstrap for assessing the prediction accuracy of a generic regression model (also nonparametric).
- Consider a regression problem where $D_N = \{\langle x_i, y_i \rangle : i = 1, \dots, N\}$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}$.
- The goal of regression is to estimate the quantity $h(x) = E[\mathbf{y}|x]$.
- The key assumption of the linear model is that

$$h(x) = x^T \beta$$

where x stands for the $[p \times 1]$ vector $x = [1, x_1, x_2, \dots, x_n]^T$ and where $p = n + 1$ is the total number of model parameters.

Bootstrap and regression (II)

- The parameter vector β is unknown. The goal of linear regression is to infer the estimate $\hat{\beta}$ on the basis of D_N .
- If the linear assumption holds, the sampling distribution of $\hat{\beta}$ is known.
- But what if the probabilistic model is not linear?
- There are two basic approaches to bootstrapping in the regression problem:
 1. First fit the model and apply the bootstrap to the residuals.
This requires that the residuals be independent and identically distributed.
 2. Apply the bootstrap procedure to the input/output dataset.

Bootstrapping regression residuals

- Suppose the least-squares estimate $\hat{\beta}$ is available.
- The N residuals ϵ_i are given by

$$\epsilon_i = y_i - x_i^T \hat{\beta}, \quad i = 1, \dots, N$$

- For $b = 1, \dots, B$, sample with replacement the set $E = \{\epsilon_1, \dots, \epsilon_N\}$ then compute the bootstrap responses by

$$y_i^* = x_i^T \hat{\beta} + \epsilon_i^*, \quad i = 1, \dots, N$$

- Once the input/output bootstrap set

$$D_{(b)} = \{\langle x_i, y_i^* \rangle\}$$

is built, we can derive the bootstrap replicates $\hat{\beta}^{(b)}$.

- The analysis of the bootstrap distribution allows the assessment of the estimator $\hat{\beta}$.

Bootstrap and time series

- Suppose that the dataset comes from a time series, a data structure for which nearby values of the time parameter indicate closely related values of the measured quantity z .
- Suppose that the dataset have been generate by a first order autoregressive scheme

$$z_t = \beta z_{t-1} + \epsilon_t \quad t = 1, \dots, N$$

where β is an unknown parameter and the disturbances ϵ_t are assumed to be a random sample from an unknown distribution F with zero expectation.

- The estimate $\hat{\theta} = \hat{\beta}$ can be obtained by the least-squares formulas. We want to assess the accuracy of $\hat{\beta}$ with a bootstrap technique.
- We have to perform the two steps $D_N \rightarrow \hat{P}$ and $\hat{P} \rightarrow D_{(b)}$

Bootstrap and time series (II)

- The probability model P has two unknowns: β and F . Once we estimate β by least-squares the empirical \hat{F} can be obtained by using the relation

$$\epsilon_t = z_t - \hat{\beta}z_{t-1}$$

- We run the above formula N times and we obtain a dataset of N approximate disturbances from which to infer \hat{F} .
- The bootstrap sample $D_{(b)}$ is then obtained by generating N samples $\{\epsilon_1^*, \dots, \epsilon_N^*\}$ and running recursively

$$z_t^* = \hat{\beta}z_{t-1}^* + \epsilon_t^* \quad t = 1, \dots, N$$

- Once $D_{(b)}$ is given the bootstrap replication $\hat{\theta}_{(b)}$ can be obtained by least-squares on $D_{(b)}$.
- The distribution of $\hat{\theta}_{(b)}$ can be used also to infer some properties of $\hat{\beta}$, e.g. if it significantly different from zero.

Statistical test and sampling

- A statistical test is based on a test statistic t which measures the discrepancy between the data and the null hypothesis.
- Given a dataset D_N the observed value of the statistic is $t(D_N)$. The level of evidence against H is measured by the p-value $\text{Prob} \{t \geq t(D_N) | H\}$.
- Sampling methods are not new to significance testing. Examples are Monte Carlo tests, randomization tests and permutation tests.

Monte Carlo tests

- The basic Monte Carlo test compares the observed statistic $t(D_N)$ to R independent values of t

$$t_{(1)}, t_{(2)}, \dots, t_{(R)}$$

which are obtained from corresponding samples independently simulated under the null hypothesis model.

- If the hypothesis H is true, the $R + 1$ samples $t(D_N), t_{(1)}, t_{(2)}, \dots, t_{(R)}$ are equally likely values of t .
- The Monte Carlo p-value is then

$$p_{\text{mc}} = \text{Prob} \{t \geq t(D_N) | H\} = \frac{1 + \#\{t_{(i)} \geq t(D_N)\}}{R + 1}$$

where $\#\{t_{(i)} \geq t(D_N)\}$ is the number of simulated values $t_{(i)}$ that exceeds $t(D_N)$

Randomization tests

- **Randomization tests** were introduced by R.A. Fisher in 1935.
- Suppose we have a *complicated* data set and believe it shows a **non random** property or pattern.
- Randomization tests make the null hypothesis of randomness and test this hypothesis against data.
- In order to test the randomness hypothesis, several random transformation of data are generated.
- Exemple: test if the pack of poker playcards are well shuffled.

A bioinformatics example

- Suppose we have a DNA sequence and we think that the number of repeated sequences (e.g. AGTAGTAGT) in the sample is greater than expected by chance. Let the number of repetition $t = 17$.
- How to test that? Let us formulate the null hypothesis that the base order is random?
- We can construct an empirical distribution under the null hypothesis by taking the original sample and scrambling $R = 1000$ times the the bases at random .
- This creates a sample with the same base frequencies as the original sample but with the order of bases assigned at random.
- Suppose that only 5 of the 1000 randomized samples has a number of repetition higher or equal than 17. The probability of seeing $t = 17$ under the null hypothesis is 0.05.

Shuffling test

Suppose we are interested in some property which is related to the **order** of data. Let the original data set $D_N = \{x_1, \dots, x_N\}$ and $t(D_N)$ some statistic which is a function of the order in the data D_N . We want to test if the value of $t(D_N)$ is due only to randomness.

- An empirical distribution is generated by scrambling (or **shuffling**) R times the N elements at random. For example the j th, $j = 1, \dots, R$ scrambled data set could be
$$D_N^{(j)} = \{x_{23}, x_4, x_{343}, \dots\}$$
- For each of the j th scrambled sets we compute a statistic $t^{(i)}$. The resulting distribution is called the **sampling distribution**.
- Suppose that the value of $t(D_N)$ is only exceeded by k of the R values of the sampling distribution.
- The probability of observing $t(D_N)$ under the null hypothesis (i.e. randomness) is only $p_t = k/R$. The null hypothesis can be accepted/rejected on the basis of p_t .

Permutation tests

- Permutation tests are a computer-intensive statistical technique that predates computers. The idea was introduced by R.A. Fisher in the 1930s, as a theoretical argument supporting the t-test.
- The basic idea is simple and free of mathematical assumption.
- It has a close connection with the bootstrap.
- Its main application concerns the two-sample problem.

Permutation tests for the two-sample problem

- Consider two r.v.s $\mathbf{x} \sim F$ and $\mathbf{y} \sim G$. Let D_N^x and D_M^y two mutually independent sets of samples.
- We wish to test the null hypothesis of no difference between F and G . If H is true, then there is no difference between the probabilistic behavior of a random dataset D_N^x and D_M^y .
- Let us compute the difference of the means

$$\hat{\theta} = \hat{\mu}_x - \hat{\mu}_y$$

The larger the value of $\hat{\theta}$, the stronger is the evidence against H .

- We are interested in computing the *significance level*

$$\text{Prob} \left\{ \hat{\theta} \geq \hat{\theta} | H \right\}$$

The smaller this value, the stronger is the evidence against H .

- Permutation test is a clever way of calculating the significance level for the **general** null hypothesis $F = G$. This means that we make no assumption about the form or the parameters of the two distributions.
- If the null hypothesis is true, any of the measurements x_i , $i = 1, \dots, N$ and y_j , $j = 1, \dots, M$ could have come equally well from either of the treatments.
- So we combine all the $M + N$ observations from both groups together and then we take a sample of size M without replacement to represent the first group; the remaining N observations constitute the second group.
- We compute the difference between group of means and we repeat the procedure a number R of times.

- We check the position of $\hat{\theta}$ in the distribution of differences.
- If $\hat{\theta}$ is outside the middle $(1 - \alpha)\%$ of the distribution, the two-sided permutation test **rejects** the null hypothesis at an α level.
- The permutation algorithm is quite similar to the bootstrap nonparametric algorithm. The main difference is that sampling is carried out without replacement rather than with replacement.

Example: permutation test

- Consider a small experiment where 7 out of 16 mice were randomly selected to receive a new medical treatment, intended to prolong survival times after a surgery.
- The table shows the survival time (in days) of the two groups

Treatment (x) :	94	197	16	38	99	141	23		
Control (y):	52	104	146	10	50	31	40	27	46

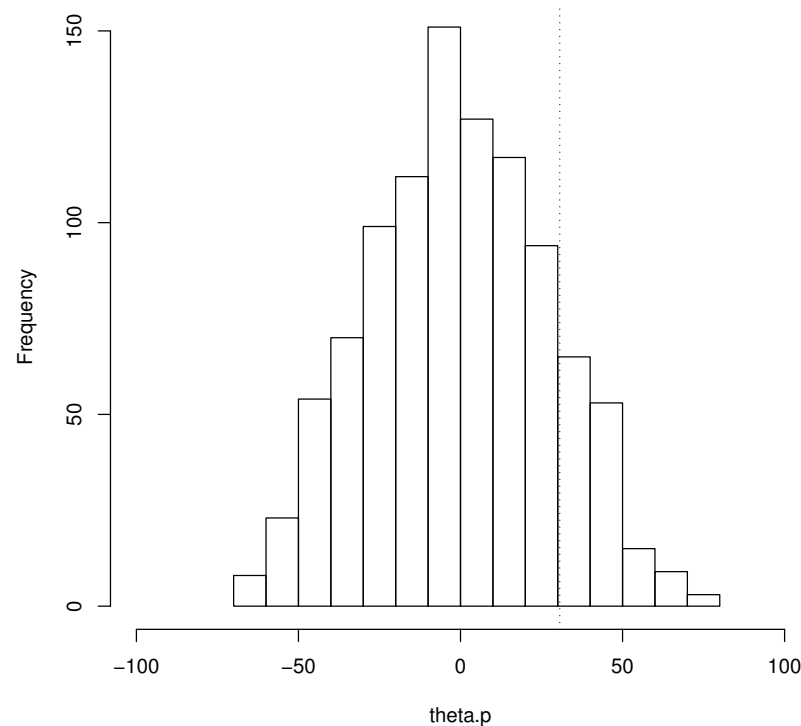
- We want to answer the question: “ Did the treatment prolong survival?” The null hypothesis is that the two treatment are not different.
- The difference of the two means is

$$\hat{\theta} = \hat{\mu}_x - \hat{\mu}_y = 30.63$$

- By performing the permutation test ($B = 100$ repetitions) we have the following distribution

Example: permutation test (II)

R file `perm.R`.



The significance level is 0.14. The data does not present enough evidence against H .

The hypothesis is **not rejected**.

Permutation lemma

The data of the above example may be ranked in the form

Group	y	x	x	y	y	x	y	y
Rank	1	2	3	4	5	6	7	8
Value	10	16	23	27	31	38	40	46
Group	y	y	x	x	y	x	y	x
Rank	9	10	11	12	13	14	15	16
Value	50	52	94	99	104	141	146	197

where the bottom line gives the ranked values and the first line gives the group.

The vector of values and the vector of groups convey the same information as $\{D_N^x, D_M^y\}$.

Permutation lemma (II)

- Let v be the combined and ranked vector of values and g the vector designing the group. The vector g contains N symbols of type x and M symbols of type y .
- There are $\binom{N+M}{N}$ possible g vectors corresponding to all the possible ways of partitioning a vector of $N + M$ elements into two subsets of size N and M .
- Permutation tests depend on the following result

Lemma 1. *Under $H : F = G$, the vector g has probability $1 / \binom{N+M}{N}$ of equaling any one of its possible values.*

- In other words, all the permutations of D_N^x and D_M^y are equally likely if $F = G$.

Parametric bootstrap tests

- In some cases the distribution F of t under the null hypothesis is not perfectly known.
- A possible approach is to fit a model \hat{F}_H to the null hypothesis and use it to compute the p-value.
- For example suppose that the hypothesis is $H : \psi = \psi_H$ and that the distribution of t under H is $F(z, \psi_H, \theta)$.
- The parametric bootstrap approach consists in estimating θ by $\hat{\theta}_H$ and using the distribution $\hat{F}_H = F(z, \psi_H, \hat{\theta}_H)$.
- Then we calculate the p-value

$$p = \text{Prob} \left\{ t \geq t(D_N) \mid \hat{F}_H \right\}$$

Parametric bootstrap tests (II)

- If the above quantity cannot be computed exactly, or if there is no satisfactory approximation (normal or otherwise) we proceed by simulation. The significance probability is then approximated by

$$p_{\text{bs}} = \frac{1 + \#\{t_{(i)} \geq t(D_N)\}}{R + 1}$$

Nonparametric bootstrap tests

- The **nonparametric bootstrap test** differs from the parametric test only in the first part.
- It assumes no knowledge is available about F and consequently it does not require the fitting of a probabilistic model \hat{F} .
- As usual the bootstrap samples are generated by resampling D_N with replacement.

Bootstrap and two sample problems

- Consider two r.v.s $\mathbf{x} \sim F$ and $\mathbf{y} \sim G$. Let D_N^x and D_M^y two mutually independent sets of samples.
- Let the combined sample be $D_N = \{D_N^x, D_M^y\}$
- Let the null hypothesis be $H : F = G$.
- Bootstrap samples $D_{(b)}$ are obtained by sampling D_N with replacement. We call the first N samples by x_i^* and the remaining M observations by y_i^*

$$D_{(b)} = \{x_1^*, \dots, x_N^*, y_1^*, \dots, y_M^*\}$$

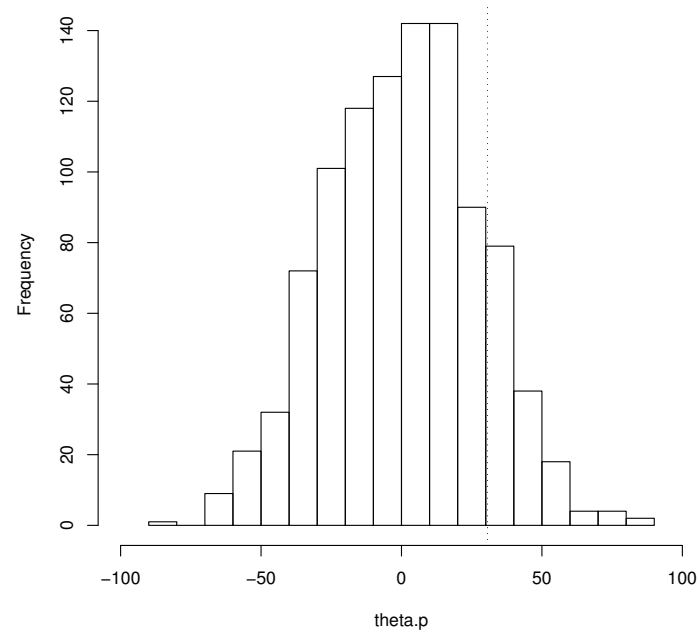
- If the statistic of interest is the difference of the means

$$\hat{\theta} = \hat{\mu}_x - \hat{\mu}_y$$

the bootstrap distribution of $\hat{\theta}$ is given by the B values $\hat{\theta}_{(b)}$.

Example: two-sample bootstrap test

R file `boot2s.R`.



- The significance level is 0.135. The data does not present enough evidence against H . Hence, the hypothesis is **not rejected**.
- The only difference between this bootstrap nonparametric algorithm and the permutation test is that samples are here drawn with replacement rather than without replacement.

Studentized bootstrap test

- Consider two r.v.s $x \sim F$ and $y \sim G$. Let D_N^x and D_M^y two mutually independent sets of samples.
- The null hypothesis is $H : F = G$.
- The test statistic is no more the differences of the mean but

$$t(D_N) = \frac{\hat{\mu}_x - \hat{\mu}_y}{\hat{\sigma} \sqrt{1/N + 1/M}}$$

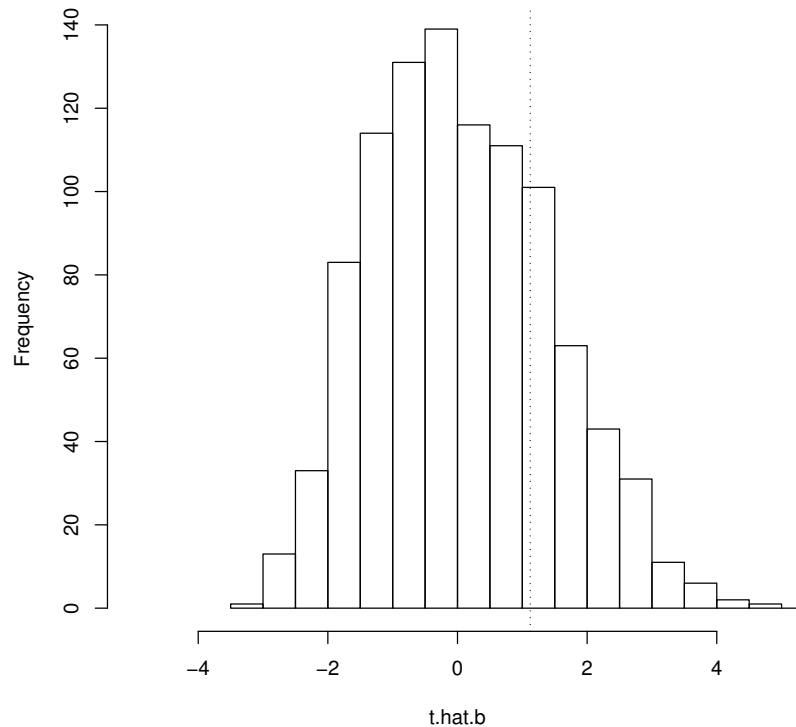
where

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{\mu}_x)^2 + \sum_{i=1}^m (y_i - \hat{\mu}_y)^2}{N + M - 2}}$$

- The bootstrap algorithm is the same as before. The only difference concerns the statistic which is adopted.

Example: studentized bootstrap test

R file `bootstu2s.R`.



- The significance level is 0.26. Hence, the hypothesis is not rejected.

Permutation test and bootstrap

- A permutation test exploits some special symmetry that exists under the null hypothesis.
- As a result of this symmetry, the significance level of a permutation test is exact (once all the permutations of the dataset are done). In a practical test, only simulation error is present.
- The bootstrap estimates the probability mechanism underlying the null hypothesis and then samples from it.
- The bootstrap is guaranteed to be accurate as the sample size goes to infinity (statistical error).
- On the other hand, bootstrap does not require the special symmetry which is needed for a permutation test and so it can be applied more generally (for example bootstrap can test also equal means and not exclusively equal distributions).

Summary procedure for bootstrap testing

In order to carry out a bootstrap hypothesis test two are the quantities to be chosen:

1. a test statistic $t(D_N)$,
2. a null distribution \hat{F} for the data under H .

Given these quantities, we generate B bootstrap values of $t(D_{(b)})$ under \hat{F} and we estimate the achieved significance level by

$$p_{\text{bs}} = \frac{\#\{t(D_{(b)}) \geq t(D_N)\}}{B}$$

General considerations

- The choice of $t(D_N)$ and \hat{F} are not obvious.
- The problem come from composite null hypothesis.
- A good choice for \hat{F} is a distribution which obeys H and is the *most reasonable* for our data.
- Bootstrap tests are useful when the alternative hypothesis is not well specified.
- In case when there is a parametric alternative hypothesis, likelihood or Bayesian methods are preferable.