TELECOM ParisTech **MDI720**
2016/2017 **François Portier, Joseph Salmon**

Devoir maison n° 2 : Bootstrap

Pour ce travail vous devez déposer un **unique** fichier **anonymisé** (votre nom ne doit apparaître nulle part y compris dans son nom lui-même) sous format `ipynb` sur le site http://peergrade.enst.fr/. Vous devez charger votre fichier, avant le dimanche 23/10/2016 23h59. La correction sera disponible sur EOLE le lundi 24 et donc les personnes qui n'auront pas déposé leur travail avant la limite obtiendront **zéro**.

Entre le lundi 24 et le vendredi 28 octobre, 23h59, vous devrez noter trois copies qui vous seront assignées anonymement, en tenant compte du barème suivant pour chaque question :

- 0 (manquant/ non compris/ non fait/ insuffisant)
- 1 (passable/partiellement satisfaisant)
- 2 (bien)

Ensuite, il faudra également remplir de la même manière les points de notation suivants :

- aspect global de présentation : qualité de rédaction, d'orthographe, d'aspect de présentation, graphes, titres, etc. (Question 19).
- aspect global du code : indentation, Style PEP8, lisibilité du code, commentaires adaptés (Question 20).
- Point particulier : absence de bug sur votre machine (Question 21)

Des commentaires adaptés pourront être ajoutés question par question si vous en sentez le besoin ou l'utilité pour aider la personne notée à s'améliorer. Enfin, veillez à rester polis et courtois dans vos retours.

Les personnes qui n'auront pas rentré leurs notes avant la limite obtiendront également **zéro**.

Rappel : aucun travail par mail accepté !

Exercice 1. (The sample mean in $\mathbb{R}^2$)

1) Generate $n = 500$ samples from a Beta distribution with parameter $(\alpha, \beta) = (2, 5)$. Display the histogram of this sample with 25 bins[1]. Then, generate $n = 500$ independent random vectors X_i, $i = 1, \dots n$, in $\mathbb{R}^2$, where all coordinates are drawn independently from a Beta distribution with parameter $(\alpha, \beta) = (2, 5)$. Compute the mean vector (in $\mathbb{R}^2$).
2) Compute $B = 500$ bootstrap estimators of the mean. On the same plot, represent the observed data, the mean and the 500 bootstrap estimators of the mean.
3) Give bootstrap estimates of the bias and the variance of the mean estimator.
4) Give jackknife estimates of the bias and the variance of the mean estimator. Verify the formula

$$\hat{\sigma}^2_{\text{jack}} = \frac{1}{n}\hat{\text{cov}}_n(X),$$

where $\hat{\text{cov}}_n(X)$ is the unbiased estimate of the variance.

1. You can initialized your random seed for reproducibility reasons.

5) Give the true variance of the estimator of the mean. Compare the bootstrap, the jackknife and the asymptotic approximation through the ℓ_1-norm.

Exercice 2. (The correlation coefficient)

6) Generate $n = 300$ independent random vectors X_i with 2 dependent components satisfying

$$X_{i2} = X_{i1} + U_i$$

where $X_{i1} \sim \mathcal{U}[0,1]$ and $U_i \sim \mathcal{U}[-.1,.1]$ are independent (and $\mathcal{U}$ stands for uniform distribution).

7) Give the (theoretical) correlation coefficient and an estimated correlation coefficient of X_{i1} and X_{i2}.

8) Compute a 5% basic bootstrap confidence interval for the correlation coefficient, for a number $B = 500$ of bootstrap replicas.

9) Compute a 5% percentile bootstrap confidence interval for the correlation coefficient, for a number $B = 500$ of bootstrap replicas.

10) Compute a 5% asymptotic confidence interval for the correlation coefficient (hint : one should propose a reasonable way to estimate the asymptotic variance).

11) By means of simulation (use for instance $M = 2000$ repetition of the experiment with a Monte-Carlo approach), evaluate the coverage probability associated to each method : basic and percentile.

12) Draw the coverage probability for $n = 30, 50, 100, 200$.

Exercice 3. (Linear regression)

As in DM1, we work on the database `auto-mpg` which could be downloaded from the link https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data-original and from which we drop the lines containing `na` values. We also let aside the discrete variables `origin` et `car name`. We assume the following model

$$\mathbf{y} = \boldsymbol{\theta_0} + \boldsymbol{X\theta} + \boldsymbol{\varepsilon}, \tag{1}$$

where

— $\mathbf{y} = (y_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ is a random column vector $n \times 1$,

— $X = (X_{i,j})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$ is a matrix $n \times p$ with random entries,

— $\theta = (\theta_i)_{1 \leqslant i \leqslant p} \in \mathbb{R}^p$ is a column vector $p \times 1$, $\theta_0 \in \mathbb{R}$ is a constant,

— $\boldsymbol{\varepsilon} = (\varepsilon_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ is a random column vector $n \times 1$ with mean 0 independent of X.

13) Compute the least-square estimator $\hat{\theta}$ of θ. Compute the estimated residuals vector and draw its density.

14) Give two 5% confidence intervals for the coefficients of θ. One based on the asymptotic approximation. Another based on the hypothesis that $\boldsymbol{\varepsilon}$ is Gaussian.

15) Implement the bootstrap method based on the residuals. On the same graph, draw the response and the bootstrap responses versus the estimated response.

16) Compute basic bootstrap and percentile confidence intervals.

17) For each method, create a figure representing the confidence intervals and the estimators of $\theta = (\theta_i)_{1 \leqslant i \leqslant p} \in \mathbb{R}^p$ (not the intercept).

18) (forward variable selection) Consider the following p models, for $k = 1, \ldots, p$, $y = \theta_0 + \theta_k X_k + \epsilon$, where X_k stands for the k-th variable in X. By computing the statistics $|\hat{\theta}_k|/\hat{\sigma}_k$, where $\hat{\sigma}_k$ is the standard error estimate of $\hat{\theta}_k$, argue which variable is the most significant.