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].

2) Generate $n = 500$ independent random vectors $(X_{i1}, X_{i2}) \in \mathbb{R}^2$ where each coordinate is independent of the other and both having a Beta distribution with parameter $(\alpha, \beta) = (2, 5)$. Compute the mean vector $\hat{\mu}$ (in $\mathbb{R}^2$). In the next we apply some *bootstrap* techniques to estimate the variance and the bias of $\hat{\mu}$.

3) Compute $B = 500$ *bootstrap* estimators of the mean $\hat{\mu}_1^*, \ldots, \hat{\mu}_B^*$ (where each $\hat{\mu}_b^* \in \mathbb{R}^2$). On the same plot, represent the observed data, the estimated mean $\hat{\mu}$ and the 500 *bootstrap* estimators of the mean $\hat{\mu}_b^*$. Note that since $\hat{\mu} \in \mathbb{R}^2$ the variance is a matrix (2×2).

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

4) Give *bootstrap* estimates of the bias and the variance of the mean estimator $\hat{\mu}$. The formulas given in the course in $\mathbb{R}$ become in $\mathbb{R}^2$,

$$\widehat{\text{Bias}}_{\text{boot}} = B^{-1} \sum_{b=1}^{B} \hat{\mu}_b^* - \hat{\mu},$$
$$\widehat{\text{Var}}_{\text{boot}} = B^{-1} \sum_{b=1}^{B} (\hat{\mu}_b^* - \overline{\mu}^*)(\hat{\mu}_b^* - \overline{\mu}^*)^\top,$$

where $\overline{\mu}^* = B^{-1} \sum_{b=1}^{B} \hat{\mu}_b^*$. Estimating the variance in our case is indeed estimating a symmetric matrix that contains the variance of each coordinate and the covariance between each coordinate.

Give now the *jackknife* estimates of the bias and the variance of the mean estimator. The formulas given in the course in $\mathbb{R}$ become in $\mathbb{R}^2$,

$$\widehat{\text{Bias}}_{\text{jack}} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\mu}_{-i} - \hat{\mu}),$$
$$\widehat{\text{Var}}_{\text{jack}} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{-i} - \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{-i} \right) \left(\hat{\mu}_{-i} - \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{-i} \right)^\top,$$

where $\hat{\mu}_{-i} = \frac{1}{n-1} \sum_{j \neq i}^{n} X_j$. Verify empirically the formula

$$\widehat{\text{Var}}_{\text{jack}} = \frac{1}{n} \widehat{\text{var}_n}(X),$$

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

5) Give the true variance of the estimator of the mean. As we know the true underlying distribution of the data (because we have generated the random variables X_{i1}, X_{i2}), we can compute the true variance. Indeed, we can show that

$$\text{Var}\,(\hat{\mu}) = \frac{1}{n} \text{var} \begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix} = \frac{1}{n} \text{var} \begin{pmatrix} \text{var}(X_{11}) & \text{cov}(X_{11}, X_{12}) \\ \text{cov}(X_{11}, X_{12}) & \text{var}(X_{12}) \end{pmatrix}.$$

Each quantity in the previous matrix can be explicitly computed using the moments of Beta distributions. Use the true asymptotic variance to compare the performance of the *bootstrap* and the *jackknife*. We can for instance use the distance

$$\left\| \begin{pmatrix} \hat{v}_{11} & \hat{v}_{12} \\ \hat{v}_{12} & \hat{v}_{22} \end{pmatrix} - \begin{pmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{pmatrix} \right\| = |\hat{v}_{11} - v_{11}| + |\hat{v}_{12} - v_{12}| + |\hat{v}_{22} - v_{22}|.$$

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) Compute the true correlation coefficient c_0 and an estimated correlation coefficient $\hat{c}$ 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. Using the fact that $n^{1/2}(\hat{c} - c_0) \approx \mathcal{N}(0, \sigma^2)$, the asymptotic confidence interval is

$$\left[\hat{c} - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}\,,\ \hat{c} - \frac{\sigma}{\sqrt{n}} q_{\alpha/2}\right],$$

where q_α is the α-quantile of the standard normal distribution. Since the asymptotic variance σ^2 is hard to derive here we shall estimate this variance using the Jackknife technique.

11) By means of simulation (use for instance $M = 2000$ repetitions of the experiment with a Monte-Carlo approach), evaluate the coverage probability associated to each method : basic (in question 8), percentile (in question 9) and asymptotic-*jackknife* (in question 10). If $\hat{I}$ is a confidence interval that has been computed using one of the previous technique. The coverage probability associated to $\hat{I}$ is

$$\mathbb{P}(c_0 \in \hat{I}).$$

The Monte-Carlo approach consists in generating $\hat{I}_1, \dots, \hat{I}_M$ (each time with a new data generated) and then computing

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{c_0 \in \hat{I}_m\}}.$$

This indicates the precision of the employed technique. The closer to 95% the better.

12) Draw the evaluated coverage probabilities for different sample sizes $n = 30, 50, 100, 300$.

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` and `car name`. The variable to explain is `mpg`. Each observation of `mpg` is denoted by y_i. The explicative variables are `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`. Each such 6×1 column vector of observations is denoted by X_i. We assume the following model

$$y_i = \theta_0 + X_i^\top \theta + \epsilon_i,$$

where

— $\theta = (\theta_j)_{1 \leqslant j \leqslant 6} \in \mathbb{R}^6$ is a 6×1 column vector, $\theta_0 \in \mathbb{R}$ is a constant,

— each ε_i has mean 0 and is independent of X_i.

13) Compute the least-square estimator $(\hat{\theta}_0, \hat{\theta})$ of (θ_0, θ). Compute the estimated residuals vector and draw its density.

14) Give a 5% confidence interval for the coefficient θ_1, associated to the variable `cylinders`, based on the hypothesis that each ε_i is Gaussian.

15) Implement the *bootstrap* method based on the residuals (this method is specified in the beamer of the class, here $\hat{g}(X_i) = \hat{\theta}_0 + X_i^\top \hat{\theta}$). On the same graph, draw the response y_i and the *bootstrap* responses y_i^* versus the estimated response $X_i^\top \hat{\theta}$.

16) Compute a basic *bootstrap* confidence interval for θ_1.

17) For each method, create a figure representing the confidence intervals of every coefficient $\theta_1, \dots, \theta_6$ (except the intercept θ_0) and the estimators $\hat{\theta}_1, \dots, \hat{\theta}_6$. The x-axis should be reserved for the indexes $j = 1, \dots, 6$. We shall notice that many confidence intervals contains the value 0 meaning that some are not significant.