ENSAE ParisTech

Projet de Bootstrap et ré-échantillonnage

Application des méthodes de ré-échantillonage pour l'estimation du seuil et de l'indice de queue

Imen Said, Matteo Amestoy, Moussab Djerrab

Mai 2015

Table des matières

1 Introduction

Dans ce projet, nous nous intéressons au problème de l'estimation dans un cadre purement iid . Notre objectif est d'estimer des quantités relatives à des observations indépendantes tirées selon la loi Pareto. Cette loi est caractérisée par deux paramètres à savoir le seuil c et l'indice β et sa fonction de répartition est de la forme :

$$\begin{array}{ll} F(x) = 1 - (\frac{c}{x})^{\beta} & \text{pour } x > c \\ F(x) = 0 & \text{pour } x < c \end{array}$$

Cette loi peut être considérée comme la théorisation du *principe de Pareto* qui permet d'identifier les phénomènes rares qui peuvent avoir des effets très importants. En effet, en travaillant sur des données fiscaux, Vilferdo Pareto avait remarqué que 20% de la population possédait 80% de la richesse.
Cette répartition a été transposée dans différents domaines et en particulier en Actuariat : on a vite remarqué que ,pour les risques extrêmes, 20% des sinistres peuvent être à l'origine de 80% des pertes. Ceci a permis d'élaborer des modèles de tarification basés sur ce principe puis sur l'étude de la théorie des valeurs extrêmes. C'est dans ce cadre qu'on a introduit la loi Pareto en assurance/réassurance. Elle permet, en particulier, de modéliser la loi de franchissement de seuils : par exemple, niveau optimal à partir duquel un assureur devrait céder son risque au réassureur.Grâce à des queues de distribution épaisses (indice de queue strictement positif), et par conséquent une décroissance lente, elle est utilisée dans la modélisation des risques catastrophes par exemple.

Nous disposons d'un échantillon de variables aléatoires i.i.d $(X_i)_{i=1..n}$ qui suivent la loi Pareto(β, c), nous nous proposons d'estimer les quantités suivantes :

$$\theta = \mathbb{P}(X_1 > a) \text{ : la probabilité de dépasser un certain seuil a fixé}$$

$$\text{Et} \quad \xi = \text{l'estimateur de Hill de l'indice de queue de la distribution choisie}$$

Nous allons donc construire et comparer plusieurs estimateurs du paramètre de la loi et par conséquent des quantités d'intérêt et on explicitera les différentes méthodes de bootstrap utilisées selon les données dont on dispose et les hypothèses faites sur la distribution.

Dans la première partie, nous allons nous attarder sur les estimateurs classiques utilisés dans le cadre paramétrique pour retrouver les paramètres de la distribution.

Dans la deuxième partie, nous allons introduire les méthodes d'échantillonage et bootstrap et reprendre les méthodes présentées dans la première partie (pour l'approche paramétrique).

Enfin, bien qu'on se situe dans un cadre paramétrique, nous allons, tester des méthodes de bootstrap non paramétrique.

2 L'approche traditionnelle empirique :

Dans cette partie, nous allons expliquer et expliciter les méthodes classiques d'estimation des paramètres de la loi de distribution :
On dispose d'un échantillon $(X_i)_{i=1..n}$ de variables aléatoires indépendantes et identiquement distribués selon une loi Pareto(β,c) dont on ne connaît pas le paramètre β. On va donc appliquer sur tout l'échantillon les méthodes d'estimation classiques, aucune technique d'échantillonage ne sera utilisée. dans cette partie. Le but étant de pouvoir comparer les deux approches et "mesurer" l'efficacité et l'apport des méthodes de bootstrap dans l'efficacité des estimateurs.

Remarque : *Dans tout le projet, nous supposons que le seuil c soit connu et* ***fixé à c=1****, et que seuls les paramètres β et θsoient inconnus. Ces derniers vont caractériser par conséquent la distribution cherchée.*

2.1 Loi Pareto(β, c)

Nous disposons d'un échantillon d'observations indépendantes qui suivent la loi Pareto de paramètres(β, c), c'est-à-dire,$\forall i \in [1..n]$ et pour un réel c strictement positif donné on a :

$$\begin{array}{ll} \mathbb{P}(X_i < x) = F(x) = 1 - (\frac{c}{x})^{\beta} & \text{pour } x > c \\ \mathbb{P}(X_i < x) = F(x) = 0 & \text{pour } x < c \end{array}$$

Pour se familiariser un peu avec la loi, nous avons simulé pour différentes valeurs du paramètre β et de taille d'échantillon n les densités de la loi Pareto :

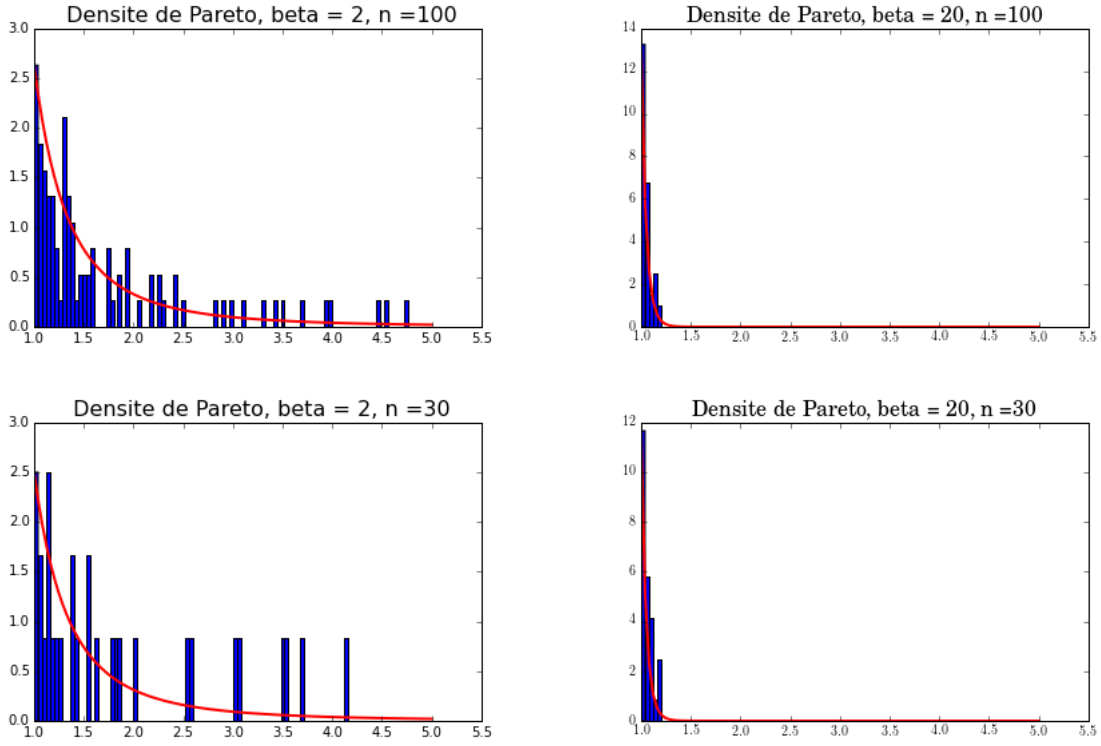

TABLE 1 – Densité de loi de Pareto

Ce graphe confirme que, la loi Pareto est une distribution à queue épaisse et appartient au domaine d'attraction de Fréchet : on remarque la décroissance plutôt lente des valeurs extrêmes. Cette décroissance est d'autant plus rapide que β augmente. On remarque que plus β augmente , plus rapide devient la décroissance. Donc, augmenter β semble être équivalent à diminuer ξ, (d'après la théorie des valeurs extrêmes , plus on augmente l'indice de queue ξ, plus lente est la décroissance). En effet, on montrera que $\xi = \frac{1}{\beta}$

2.2 Quantités à estimer

Rappelons qu'on se propose d'estimer, à partir des observations, les quantités suivantes :

- $\theta = \mathbb{P}(X_1 > a) = 1 - F(a) = (\frac{c}{a})^{\beta}$ pour un réel $a > c$ donné
- $\alpha = \beta = \frac{1}{\xi}$ avec ξ l'estimateur de Hill

Dans les deux cas, le paramètre inconnu est β . En effet, on suppose a et c fixés, estimer θ revient , par **plug-in** à estimer β, de même pour $\xi = \frac{1}{\beta}$ comme on le montre ci-dessous :

Estimateur de Hill

L'un des sujets les plus essentiels dans la théorie des valeurs extrêmes consiste à trouver des conditions équivalentes d'appartenance au domaine d'attraction de l'une des lois limites possibles (Fréchet, Gumbel ou Weibull) et qui s'expriment à partir du paramètre de forme.
La paramètre de forme ou l'indice de queue nous donne en effet un indicateur sur la vitesse de décroissance des queues). Dans ce cadre là, plusieurs estimateurs de ξ ont été construits. En 1975, Hill , a construit un estimateur semi-paramétrique qui est devenu l'un des estimateurs les plus utilisés surtout pour les distributions à queue épaisse. A partir des observations, on a :

$$\text{Pour } \xi > 0, \quad \xi_{k,n}^{Hill} = \frac{1}{k} \sum_1^k ln(X_i) - ln(X_{k+1})$$

Si $k \to \infty$, $k/n \to 0$ as $n \to \infty$, l'estimateur est **faiblement convergent**.
Il est aussi **asymptotiquement gaussien** :

$$k^{\frac{1}{2}} (\xi_{k,n}^{Hill} - \xi) \xrightarrow{L} N(0, \xi^2)$$

Dans la pratique, ce qui est délicat c'est le choix du seuil k : il y a un arbitrage biais-variance à effectuer.

L'estimateur de Hill est valide si on sait qu'a priori l'indice de queue est positif : il est donc légitime dans notre cas d'utiliser cet estimateur (la loi Pareto appartient au domaine d'attraction de Fréchet, distribution à queue épaisse).

Lien entre β et l'indice de queue ξ pour la loi Pareto (β,c)

$(X_i)_{i=1..n}$ v.a absolument continues, on peut donc définir l'inverse de la fonction de hasard pour tout x dans le support de x :

$$h(x) = \frac{1-F(x)}{f(x)} \text{ avec } f(x) \text{ la densité.}$$

Cette fonction nous donne une caractérisation simple de l'appartenance au domaine d'attraction. En effet, on admet que :

$$h'(x) \to \xi \text{ l'indice de queue.}$$

Un calcul simple nous montre que , dans le cas de la distribution Pareto définie plus haut, on a :

$$h'(x) = \frac{1}{\beta} = \xi$$

On se situe dans le cadre de **l'estimation paramétrique** : Le modèle statistique est ainsi restreint à une famille de distribution (pareto dans notre cas) dont il faut estimer les paramètres d'intérêt :

$$P_\beta = \left\{P_\beta, \beta \in \ \subset \mathbb{R}\right\} \text{ (c est fixé et égal à 1)}$$

Et on s'intéresse - comme on l'a montré- à estimer des fonctions du paramètre d'intérêt :

$$T(P_\beta) = g(\beta) \quad \text{par} \quad T(P_{\hat{\beta}}) = g(\hat{\beta})$$

$$\text{Or on a} \qquad T_1 = \theta = exp(\beta ln(\tfrac{c}{a})) = g_1(\beta)$$

$$\text{et} \qquad T_2 = \xi_{k,n}^{Hill} = \frac{1}{\beta} = g_2(\beta)$$

Avant de présenter les différents estimateurs, arrêtons nous sur le principe de la règle **plug-in** :une fois le paramètre β estimé, on utilisera cette règle pour trouver les quantités qu'on se propose d'estimer. En effet, si on est capable de construire un estimateur convergent de la distribution $\hat{P}_n$, alors la fonction $T(\hat{P}_n)$ est un estimateur naturel et convergent de $T(P)$. Bien sûr, cette règle tient sous certaines conditions : la continuité de T et la convergence faible de $\hat{P}_n$ vers P. Nous devons donc vérifier la validité de ces conditions avant d'énoncer les résultats de convergence des distributions.

2.3 Estimation par méthode des moments

Dans cette partie et la suivante , nous allons nous intéresser aux estimateurs classiques des paramètres d'intérêt : aucune méthode de bootstrap ne va être appliquée pour le moment. Ces méthodes d'estimations sont les premières mises en place dans le monde de la statistique dès que l'on se trouve dans un modèle paramétrique. L'idée principale est d'exploiter les données observées dont on dispose afin d'estimer un certain paramètre d'intérêt.
Comme nous l'avons montré dans le paragraphe précédent, les quantités à estimer sont des fonctions du paramètre inconnu β. On va donc se concentrer sur l'estimation de β.

Rappelons les premiers moments de la loi Pareto étudiée :

$X \sim Pareto(\beta, c)$, on a donc :

$$E(X) = \frac{\beta c}{\beta - 1} \text{ pour } \beta > 1$$

$$V(X) = \frac{c^2 \beta}{(\beta-1)^2(\beta-2)} \text{ pour } \beta > 2$$

Remarque : *La Loi Pareto admet un moment d'ordre r si et seulement si* $\beta > r$

A partir d'un estimateur convergent de l'espérance, on peut donc estimer β :

$$\hat{\beta} = \frac{\overline{X}_n}{\overline{X}_n - c} \text{ avec } \overline{X}_n = \frac{1}{n}\sum_1^n X_i \overset{LGN}{\to} E(X)$$

De plus, on peut avoir les propriétés asymptotiques et de convergence de son estimateur en appliquant la **la delta méthode** :

La delta méthode

Cette méthode peut être considérée comme une extension de la loi Centrale Limite : en effet elle nous donne un résultat sur la convergence asymptotique, en loi , de la transformée d'une variable aléatoire.

$(X_i)_{i=1..n}$ une suite de v.a d'espérance α et de variance σ^2

$$\text{Si : } \quad \sqrt{n}(\overline{X}_n - \alpha) \overset{L}{\to} N(0, \sigma^2),$$

$$\text{Alors : } \forall g \text{ dérivable avec } g'(\alpha) \neq 0, \sqrt{n}(g(X_n) - g(\alpha)) \overset{L}{\to} N(0, \sigma^2 g(\alpha)^2)$$

L'avantage de la Delta-méthode c'est qu'elle nous donne directement l'intervalle de confiance asymptotique autour de $g(X_n)$, soit :

$$\left[g(X_n) \pm \frac{t_\alpha \sigma g'(\theta)}{\sqrt{n}} \right]$$

La delta méthode nous donne donc la convergence de la loi de l'estimateur avec une vitesse de l'ordre de $\bigcirc(n^{-1/2})$. Dans notre cas, α est l'espérance de la loi Pareto simulée, σ son écart type et $g(x) = \frac{x}{x-c}$ prise au point $x = \overline{X}_n$.

2.3.1 Estimateur de β par la delta méthode

Ci-dessous les graphes résumant les résultats de convergence de β et de sa distribution. :

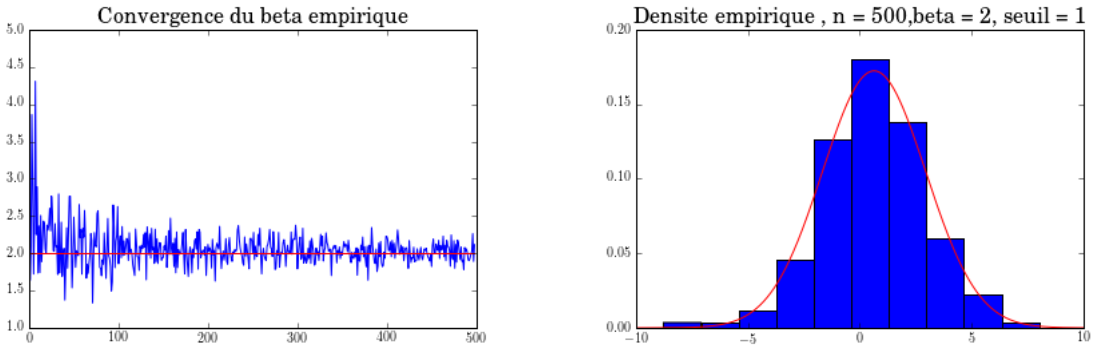

TABLE 2 – Estimation du β_n pour n = 1,..500 et densité de l'erreur pour n=500

Le premier graphe (à gauche) , nous montre que β_n converge vers la vraie valeur $\beta = 2$, la variance de l'estimateur diminue rapidement (à partir de n autour de 150 observations , on commence à obtenir la convergence).
Le deuxième graphe, est l'histogramme de la distribution de $\sqrt{n}(\hat{\beta}_n - \beta)$ pour n assez grand (n=500). On voit bien qu'il s'agit d'une distribution Gaussienne.

De la même manière, nous avons testé la convergence des deux quantités à estimer en utilisant la delta méthode. Rappelons que les vraies valeurs de θ et ξ sont :

$$\theta = (\frac{c}{a})^{\beta} = (\frac{1}{0.5})^2 = 4$$

$$\xi = \frac{1}{\beta} = \frac{1}{2} = 0.5$$

2.3.2 Estimateur de θ et de Hill ξ par la delta-méthode

— **Estimation de θ**
Nous voulons estimer , dans un premier temps la quantité $(\frac{c}{a})^{\beta}$ donc $g(X_n) = hof(\overline{X}_n)$ avec $f(\overline{X}_n) = \frac{\overline{X}_n}{\overline{X}_n - c}$ et $h(t) = (\frac{c}{a})^t$. On a $g(\overline{X}_n) = hof(\overline{X}_n)$ est dérivable et de dérivée non nulle au point E(X) : on peut donc appliquer la Delta méthode.

On a $\theta = hof(\overline{X}_n)$ avec $h(x) = exp(xln(\frac{c}{a}))$ et $f(x) = \frac{x}{x-c}$. D'où

$$(hof)'(x) = \frac{-c}{(x-c)^2} ln(\frac{c}{a}) exp(f(x) ln(\frac{c}{a}))$$

L'intervalle de confiance d'ordre 1-α est donc de la forme :

$$[hof(X_n) \pm t_\alpha \sigma((hof)'(X_n))]$$

avec t_α le quantile d'ordre $1 - \alpha$ de la loi Normale.

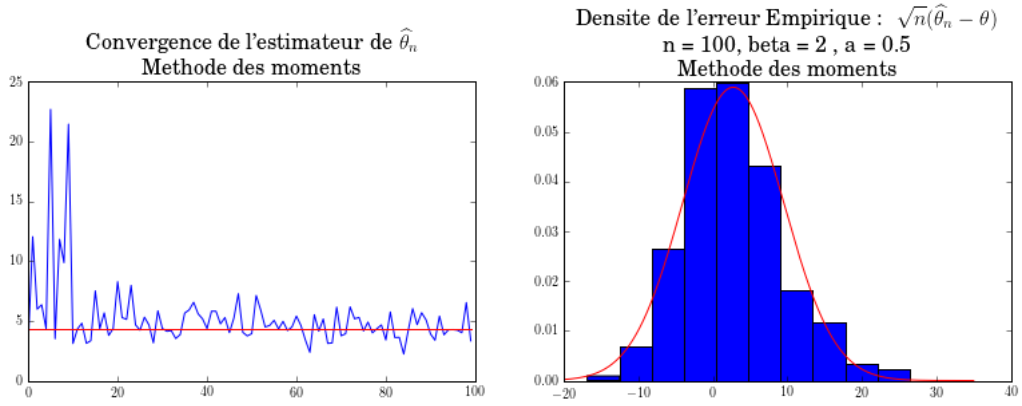

TABLE 3 – Estimation du θ_n pour n = 1,..500, a= $\frac{1}{2}$ et densité de l'erreur pour n=500

Les graphes montrent la convergence de $\hat{\theta}_n$ vers sa vraie valeur, sa vitesse de convergence est assez rapide pour qu'on puisse arrêter la taille de l'échantillon à 100 observations. On a dressé aussi l'histogramme de la distribution de l'erreur empirique, et on voir bien qu'elle converge vers une gaussienne. On s'attardera sur la valeur de la variance et les intervalles de confiance quand on introduira les méthodes d'échantillonage et bootstrap.

— **Estimateur de Hill en fonction de β**

De la même manière, nous allons appliquer la Delta méthode sur la fonction $g_1(x) = h_1 o f_1(x)$ avec $h_1(x) = \frac{1}{x}$ et $f_1(x) = f(x) = \frac{x}{x-c}$. On a donc

$$h_1 o f_1(x) = \frac{x-c}{x} = 1 - \frac{c}{x}$$

D'où

$$(h_1 o f_1)'(x) = \frac{c}{x^2}$$

De la même manière, on peut obtenir l'intervalle de confiance qui a la même forme que celui de l'estimateur de θ. Ci dessous les grahes correspondant à cet estimateur :

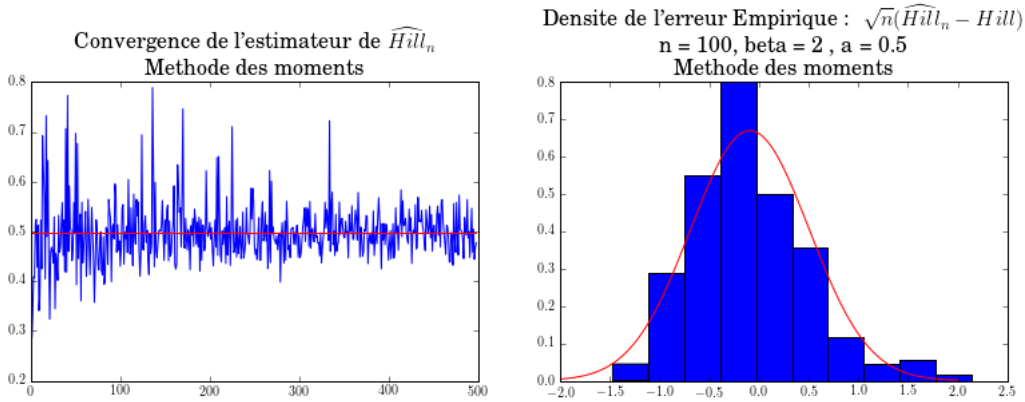

TABLE 4 – Estimation du paramètre de $Hill_n$ pour n = 1,..500, a= $\frac{1}{2}$ et densité de l'erreur pour n=500

Notons que, comme pour β, on a du augmenter beaucoup plus le nombre d'observations (n=500) pour avoir un résultat de convergence visible. L'estimateur de Hill étant étroitement corrélé à β (son inverse), cela nous semble assez cohérent.

On a généré l'histogramme pour l'erreur empirique et on voit la convergence de cette densité vers une loi Normale.

2.4 Estimation par la méthode de maximum de vraisemblance

L'une des méthodes classiques de l'estimation c'est, trouver le paramètre qui maximise la likelihood : on maximise la probabilité que nos échantillons soit distribués selon la loi Pareto($\hat{\beta}^{MLE}$,c). On va utiliser dans cette partie la règle du plug-in pour retrouver les quantités à estimer. Avant de présenter les résultats de convergence, nous présentons les principales étapes du calcul de ces estimateurs notamment en développant le calcul de la likelihood. On commence par rappeler la formule pour la fonction d'une loi de pareto $\mathcal{P}(\beta, c)$:

$$F(x) = 1 - (\frac{c}{x})^{\beta}$$

Donc , la densité de la loi Pareto étudiée est de la forme :

$$f(x) = F'(x) = \frac{c^{\beta}\beta}{x^{\beta+1}}$$

La likelihood est donc de la forme :

$$L = \prod_{i=1}^{n} f_{\beta}(x_i) = \beta^n c^{n\beta} \prod_{i=1}^{n} x_i^{-(\beta+1)}$$

Maximiser la likelihood , revient à maximiser son logarithme, or on a :

$$l = ln(L) = nln(\beta) + n\beta ln(c) - \sum_{i=1}^{n}(\beta+1)ln(x_i)$$

$$l = nln(\beta) + n\beta ln(c) - \beta\sum_{i=1}^{n} ln(x_i) - \sum_{i=1}^{n} ln(x_i)$$

En utilisant la condition de premier ordre nous donnant $\hat{\beta}^{MLE}$ vérifie la relation suivante :

$$\frac{\partial l}{\partial \beta} = 0 \Leftrightarrow \frac{n}{\beta} + nln(c) - \sum_{i=1}^{n} ln(x_i) = 0$$

i.e,

$$\hat{\beta}^{MLE} = \frac{n}{\sum_{i=1}^{n} ln(\frac{X_i}{c})}$$

D'où

$$\hat{\theta}^{MLE} = (\frac{c}{a})^{\hat{\beta}^{MLE}}$$

Et

$$\hat{\xi}^{MLE} = \frac{1}{\hat{\beta}^{MLE}}$$

Ci dessous nous présentons les résultats , sous forme de graphe, pour voir les propriétés de convergence et de distribution normale des erreurs asymptotiques. Nous réalisons les même graphes pour les deux quantités d'intérêts :

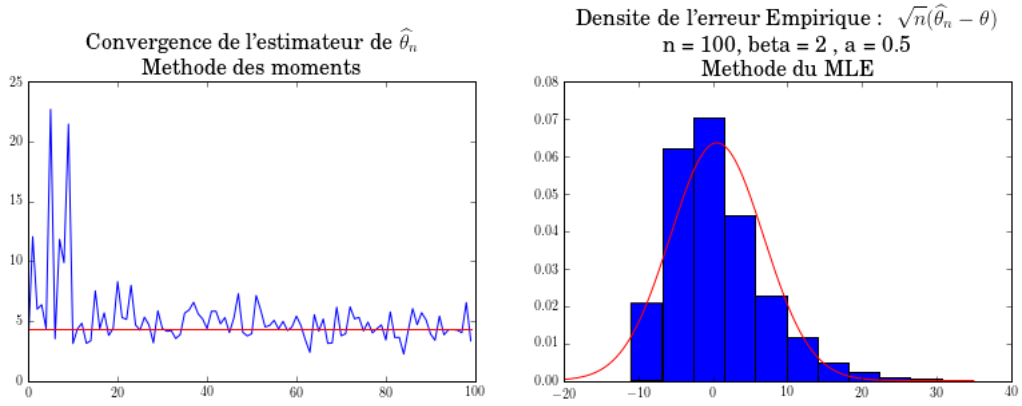

TABLE 5 – Estimation du θ_n pour n = 1,..500, a= $\frac{1}{2}$ et densité de l'erreur pour n=500

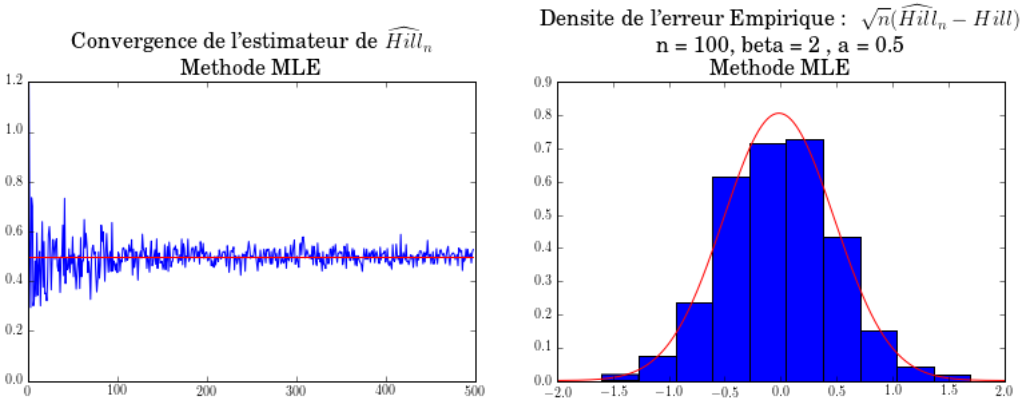

TABLE 6 – Estimation du paramètre de $Hill_n$ pour n = 1,..500, a= $\frac{1}{2}$ et densité de l'erreur pour n=500

L'estimateur de maximum de vraisemblance nous donne a priori une convergence plus rapide que celui trouvé par la méthode des moments. Ceci est d'autant plus visible dans le cas de l'estimateur de Hill. Nous allons maintenant appliquer ces méthodes non sur l'échantillon tout entier mais sur des sous échantillons en introduisant les techniques de bootstrap et de ré-échantillonage.

3 L'approche Bootstrap

L'intérêt de l'approche par bootstrap est double. Premièrement les estimateurs bootstrap nécessitent un nombre moins important de données ce qui se révèle très pratique dans la réalité car souvent on ne possède pas un nombre satisfaisant d'observations pour avoir une estimation correcte des paramètres d'intérêts. Deuxièmement les méthode bootstrap non-paramétriques permettent d'estimer les fonctions de répartitions (et donc les distributions) des lois considérées sans émettre d'hypothèses sur la famille à laquelle appartient cette distributions. Cette dernière approche permet d'évaluer des modèles beaucoup plus généraux avec un faible a priori sur le comportement probabiliste des variables.

3.1 Cadre théorique et définition

Revenons au principe général de bootstrap : supposons qu'on dispose d'un échantillon de variables $(X_i)_{i=1..n}$ indépendantes et identiquement distribués selon une distribution F qui n'est pas connue (ou qui l'est à moitié , comme dans notre cas). On cherche à estimer une fonction de cette distribution notée $\theta = T(F)$, par exemple la moyenne, ou l'indice de queue. Les méthodes de bootstrap nous permettent d'obtenir $\hat{\theta} = T(\hat{F})$. On peut distinguer deux grandes catégories de bootstrap :

- **bootstrap paramétrique** : On suppose que la distribution à estimer appartient à une famille connue de distribution (dans notre cas on suppose que les variables suivent une loi Pareto) et on cherche donc à estimer les paramètres caractéristiques de cette loi (ici, uniquement le β puisqu'on considère le seuil c connu).
- **bootstrap non paramétrique** : Dans ce cas de figure, on n'émet aucune hypothèse quant à la famille de distribution de laquelle on a tiré notre échantillon. Et par conséquent, on estime la statistique en utilisant la fonction de distribution empirique.

Dans cette partie , on tentera de focaliser notre attention sur la vitesse de convergence des estimateurs bootstrap. Ce qui nous intéresse c'est la distribution de T_n et ses propriétés asymptotiques et aussi de l'erreur asymptotique. On va donc regarder le comportement des quantités suivantes, pour une distribution P donnée et une fonction T :

$k_m(x, P) = \mathbb{P}_P((T_m(Y_1...Y_m)) \leq x)$, distribution de T_m sous P_m pour m la taille du sous-échantillon considéré.

On a la verstion centrée :

$$K_m(x, P) = \mathbb{P}_P(\tau_m(T_m(Y_1...Y_m) - T(P)) \leq x)$$

avec τ_m est un taux de convergence (typiquement $m^{-1/2}$) et $Y_1...Y_m$ une copie indépendante de l'échantillon $X_1...X_m$.

On définit de même la version réduite :

$$L_m(x, P) = \mathbb{P}_P(\tau_m S_m^{-1}(T_m(Y_1...Y_m) - T(P)) \leq x)$$

avec S_m^{-1} est une constante de normalisation (par exemple l'écart type dans le TCL et qui permet de donner la convergence vers la loi centrée réduite).

Ces deux dernières quantités donnent la distribution de l'erreur de l'estimateur de bootstrap. Cependant la première converge vers une distribution normale de l'erreur alors que la seconde converge ver une loi de Student. On notera les distributions aléatoires générées par chaque échantillon : $(K_m(., \hat{P}_n))_{m\in\mathbb{N}}$: ce sont les distributions bootstrap.
Pour chacune des méthodes de bootstarp utilisées, on va s'intéresser à :

— **la consistance** de la distribution bootsrtap , i.e,

$$\mathbf{K_m}(., \hat{\mathbf{P}}_\mathbf{n}) - K(., P) \to 0$$

Ou

$$\mathbf{K_m}(., \hat{\mathbf{P}}_\mathbf{n}) - K_m(., P) \to 0$$

Ou

$$\mathbf{K_m}(., \hat{\mathbf{P}}_\mathbf{n}) - K_n(., P) \to 0$$

— **le taux de convergence** de ces estimateurs et les comportements asymptotiques des distributions.

Rappelons qu'en plus de la distribution des estimateurs, on peut évaluer l'efficacité de l'estimateur et la vitesse de convergence vers la loi Gaussienne. En effet, **les expansions d'Edgeworth** nous donnent des résultats sur les comportements asymptotiques des quantités qu'on a défini plus haut sous certaines conditions de définition des moments d'ordre 4 et 6 en particulier.

3.2 Bootstrap paramétrique

Ici, encore une fois , on suppose connue la forme de la distribution (formule de Pareto). Dans ce cadre on a vu qu'il y avait une relation entre θ et β :

$$\theta = \left(\frac{c}{a}\right)^\beta$$

$$i.e \ \frac{ln(\theta)}{ln(\frac{c}{a})} = \beta$$

Nous avons éstimé ces deux paramètres par la méthode des moments et du MLE. Ici nous allons entreprendre d'estimer les même paramètres en appliquant les techniques de ré-échantillonnage bootstrap. Les étapes sont les suivantes :

— Déterminer β par réchantillonnage bootstrap
— utilser ce β pour déterminer directement θ (plug-in)

Puisque nous aurons besoin de construire les intervalles de confiances nous rappelons les formules de la variance et de la moyenne pour la loi de pareto et , par la delta méthode, β :

$$\sigma\left(\mathcal{P}(\beta, c)\right) = \sqrt{\frac{c^2\beta}{(\beta-1)(\beta-2)}} \ , \ \beta > 3$$

$$Mean\left(\mathcal{P}(\beta, c)\right) = \frac{c\beta}{\beta - 1} \ , \ \beta \neq 1$$

pour obtenir la variance de l'estimateur du β on a :

$$\sigma(\beta) = \left(f'(Mean\left(\mathcal{P}(\beta, c)\right)\sigma\left(\mathcal{P}(\beta, c)\right)\right)^2 \ , \ \beta > 3$$

avec $f'(x) = \frac{c}{c-x}$ la dérivée de la fonction de lien entre β et la moyenne d'une pareto de même paramètre.

3.2.1 Estimation du β par technique de bootstrap

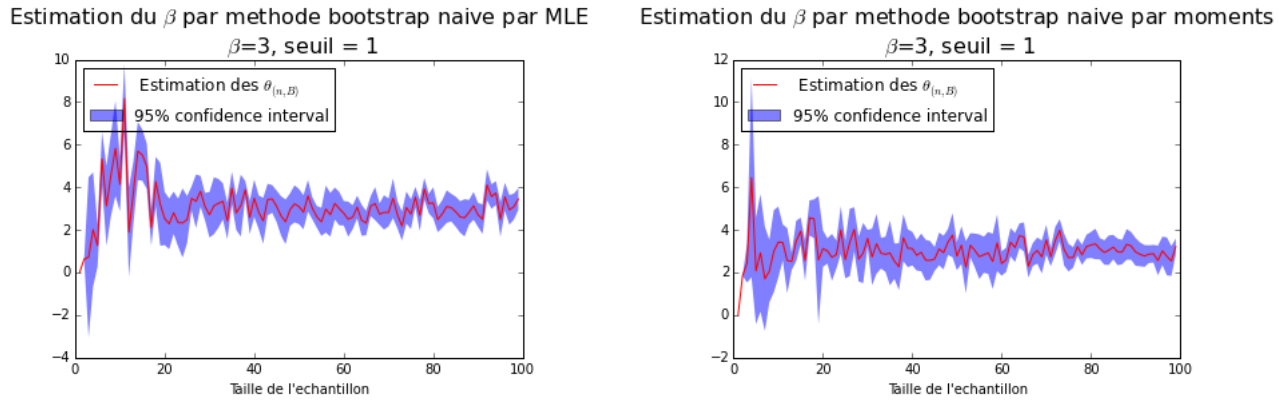

TABLE 7 – Estimation du paramètre de β par bootstrap Naive pour n = 1,..100

Notons l'efficacité de l'application de méthodes de ré-échantillonage dans la diminution de nombre d'observations nécessaires pour atteindre la convergence par rapport à la même méthode (on passe de 500 à 100).

3.2.2 Estimation du θ par méthode bootstrap

A partir de l'estimation de ce paramètre et prenant en compte la relation entre β et θ dans le cadre paramétrique , nous pouvons estimer ce dernier paramètre en fonction du dépassement du seuil que nous considérons. Nous rappelons encore la forme que doit prendre l'intervalle de confiance à travers l'écriture de la variance asymptotique :

$$h \circ f(x) = \left(\frac{c}{a}\right)^{\frac{x}{x-c}}$$

C'est la fonction qui exprime le θ en fonction de l'échantillon tiré. On a par conséquent :

$$V(\theta_n) \approx \frac{1}{n} ln \left(\frac{c}{a}\right)^2 \frac{c^2}{(c-\bar{X}_n)^4} \left(\frac{c}{a}\right)^{\frac{2\bar{X}_n}{\bar{X}_n - c}} \frac{c^2\beta}{(\beta-1)(\beta-2)}$$

De là , encore une fois , on peut construire les intervalles de confiance :

$$\theta_n \in [\theta \pm t_\alpha * \sqrt{V(\theta_n)}]$$

Le t_α est déterminé à partir de la loi normale centrée réduite (son quantile d'ordre 1-α). A noter que pour déterminer les intervalles de confiances bootstrap de θ il suffit de tirer profit des intervalles de confiance de β déjà calculés. Les tableaux suivant présentent les résultats comparatifs (Figure 1 et 2). Maintenant nous allons comparer les résultats pour les différentes méthodes et configurations, nous garderons pour tous les tests les paramètres suivants :

$$\beta = 3.5\ ,\ c = 1\ ,\ a = 2$$

n = 30 , $\theta = 0.088$	MLE	Moment
Empirique	0.11 [-0.076 ;0.253]	0.064 [-0.093 ;0.269]
Bootstrap naive	0.063 [0.025 ;0.157]	0.134 [0.06 ;0.298]

FIGURE 1 – B=300

n = 100 , $\theta = 0.088$	MLE	Moment
Empirique	0.107 [0.004 ;0.173]	0.078 [-0.037 ;0.214]
Bootstrap naive	0.07 [0.049 ;0.099]	0.109 [0.07 ;0.169]

FIGURE 2 – B=1000

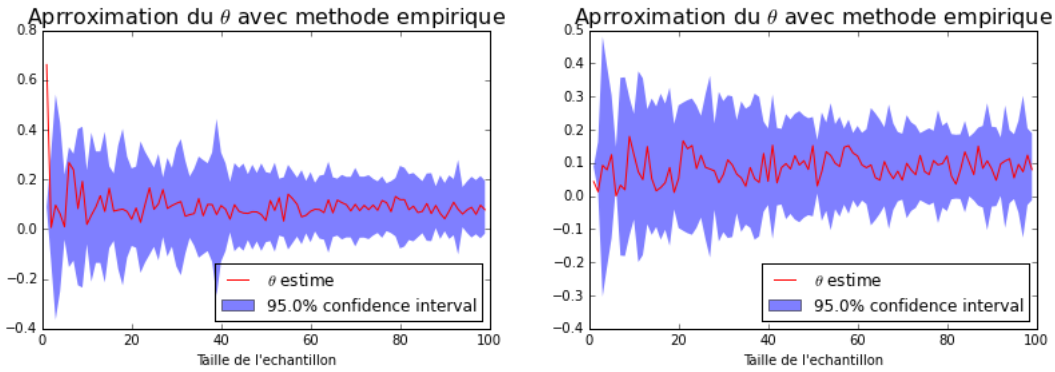

TABLE 8 – Estimation du paramètre de θ par méthode empirique pour n = 1,..100

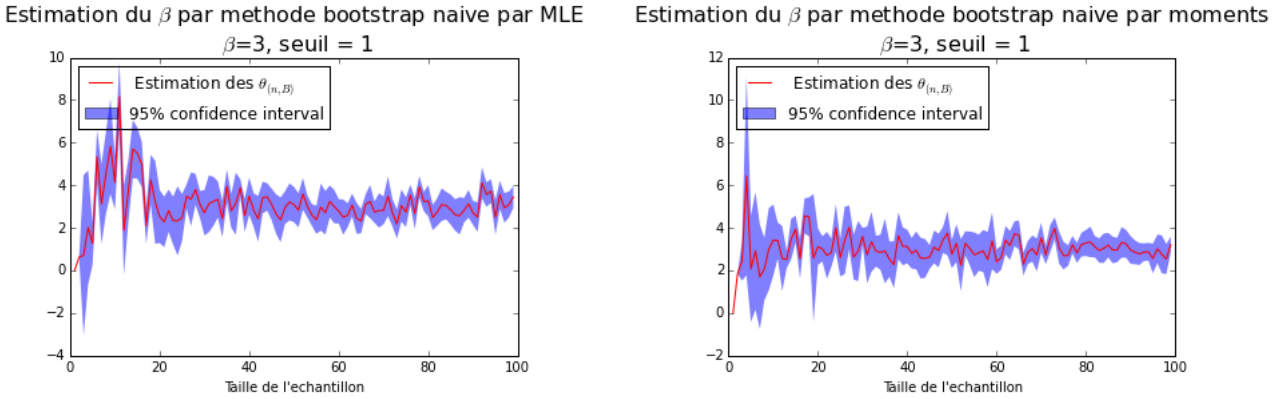

TABLE 9 – Estimation du paramètre de θ par méthode bootstrap pour n = 1,..100

De tous les tests effectués, et des méthodes envisagées , c'est le bootstrap MLE qui semble donner les meilleurs précisions pour l'estimation du θ. De plus nous avons un vrai problème pour les intervalles de confiance empiriques, la borne inférieur étant négatif. Après avoir réalisé plusieurs tests avec des paramètres différents (on a surtout fait varier le β) on comprend rapidement que cela est dû au fait que le θ est très faible dans notre exemple (lié au choix de a : le quantile à dépasser).

3.3 Bootstrap Non paramétrique

Dans cette partie, nous allons estimer les quantités d'intérêt en utilisant la fonction de répartition empirique. Ainsi on aura :

$$\theta = \mathbb{P}(X_1 > a) = 1 - \hat{F}(a)$$

3.3.1 Principe théorique de l'estimation non paramétrique par bootstrap

Effron a introduit la méthode de Bootstrap naïve : il s'agit d'approcher par simulation et méthodes de Monte Carlo la distribution de l'estimateur quand on ne connaît pas la loi de l'échantillon. On remplace donc les hypothèses probabilistes par les hypothèses de convergence des simulations. Il a proposé en plus de prendre m=n et effectuer des tirages avec remise. La méthode naîve de bootstrap suppose qu'on approche directement la distribution P par :

$$\hat{P}_{n,x} = \frac{1}{n}\sum_{1}^{n} \mathbb{P}(X_i \leq x)$$

Et on remplace par plug-in dans notre fonction pour avoir le résultat de convergence souhaité , $T(\hat{P}_n) \to T(P_n)$. Dans ce cas là, la distribution de notre distribution s'écrit de manière plus simple :

$$K_m(., P_n) = n^{-m} \sum_{i1} ... \sum_{in} \mathbb{I}_{(\tau_m(T_m - T(P)))}$$

Et la distribution noramlisée étant :

$$L_m(., P_n) = n^{-m} \sum_{i1} ... \sum_{in} \mathbb{I}_{(\tau_m S_m^{-1}(T_m - T(P)))}$$

Si on dispose de **B** échantillons , on peut utiliser **les méthodes de monte Carlo** pour l'approximation de cette distribution par la quantité suivante :

$$K_m^B(., P_n) = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}_{(\tau_m(T_m(X_j) - T(P)))}$$

Donc augmenter **B** revient à augmenter l'efficacité de notre estimateur. Il faut par contre faire un arbitrage entre augmenter le nombre de classes **B** (et par conséquent on diminue le nombre d'individus dans chaque échantillon) ou avoir moins de classes plus peuplées : un arbitrage biais-variance.
La convergence de cette quantité est immédiate par la LGN mais le théorème de **Berry Esséen** nous donne en plus la vitesse de la convergence de la distribution bootstrap naive, pour C une constante strictement positive :

$$\left\| K_m^B(., P_n) - K_m(., P_n) \right\|_\infty \leq CB^{-1/2}$$

Rappelons que par validité asymptotique du bootstrap naif, on entend :

$$K_n(., \hat{P}_n) - K_n(., P) \to 0$$

Ou

$$L_n(., \hat{P}_n) - L_n(., P) \to 0$$

3.3.2 Estimation des processus empiriques

Comme annoncé nous allons estimer directement la fonction de répartition du processus empirique par méthode bootstrap en vérifiant que les propriétés asymptotiques sont vérifiées. (Voir Table 10)

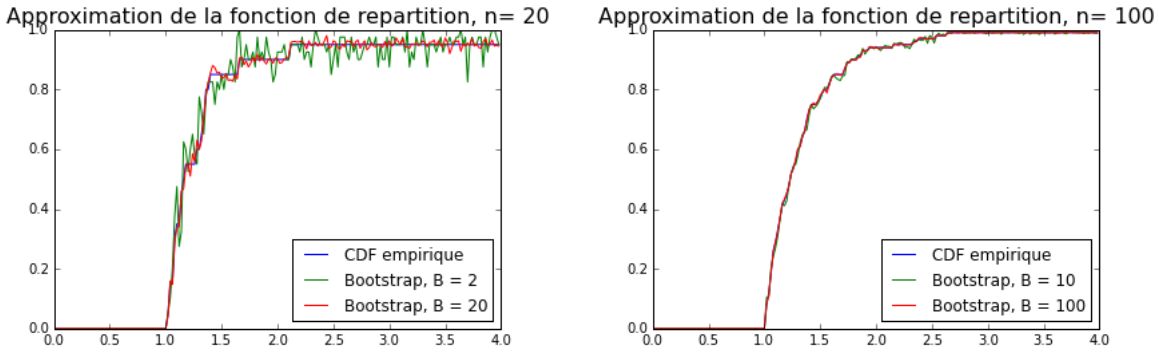

TABLE 10 – Estimation de la fonction de répartition de la loi de pareto

Dans ces premiers résultats (Table 10), nous retrouvons le principe selon lequel l'approximation bootstrap converge vers l'approximation empirique , lorsque B et n (le niveau de ré-échantillonnage et la taille de l'échantillon) converge vers l'infini. Le théorème de **Givenko-Cantelli** nous donne la convergence de la fonction de répartition empirique, par transduction on a la convergence de la fonction de répartition bootstrap vers celle théorique. Il faut maintenant voir comment l'erreur asymptotique se comporte. Pour réaliser cette tâche nous avons besoin de passer de l'estimateur $k_m(.P)$ à sa version centrée par la formule :

$$k_m(x, P) = K_m(\tau_m(x - T(P)), P)$$

Ceci nous permet de profiter de la normalité asymptotique du $K_m(., P)$ afin de pouvoir déduire des intervalles de confiance pour nos estimations asymptotiques

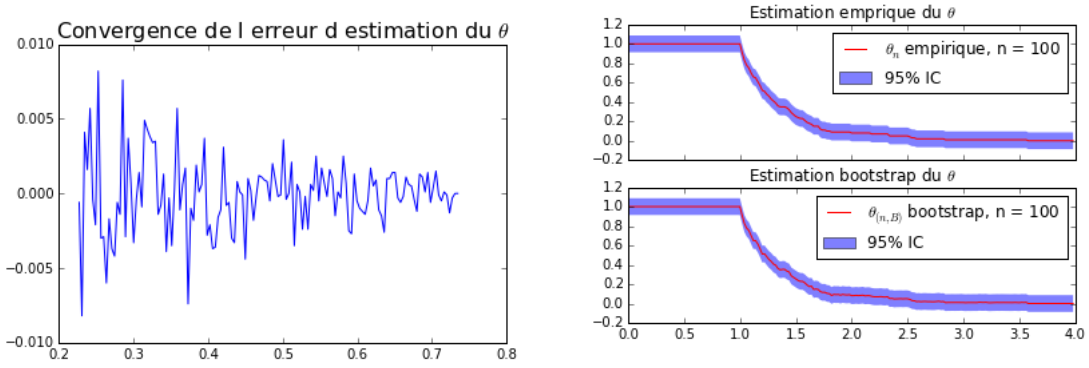

TABLE 11 – Estimation non-paramétrique de θ

On constate que l'erreur d'estimation est très proche entre la version bootstrap et celle empirique . Cela est dû à la convergence rapide des deux méthodes. Cependant dans le cadre de cette méthode les intervalles calculées sont moins satisfaisants que dans la méthode paramétrique (borne inf négatif).

4 Conclusion

Tout au long de ce travail nous avons essayé de montrer la validité pratique de la méthode de bootstrap, et aussi l'utilité d'une telle techniques : amélioration des intervalles de confiance et de l'estimateur. Si l'estimation non-paramétrique nous semble moins satisfaisante que sa version paramétrique , elle n'en reste pas moins assez robuste en comparaison de l'estimation empirique.

De plus, le rapport ne mentionne pas directement la série de tests réalisés avec différents paramètres. En effet les sensibilités des modèles liées aux paramètres a une incidence directe sur la précision des estimateurs et des intervalles de confiance. En effet plus le β théorique est élevé et plus la distribution est écrasée, ce qui a un effet inverse sur les erreurs d'estimations.

Enfin, les résultats du Bootstrap naïf ont été présentés, cependant (voir le code source du projet), d'autres méthodes de ré-échantillonnage ont été envisagées (jacknife). Nous avons fait le choix de ne pas présenter ces derniers résultats pour ne pas être redondants, dans la mesure où la précision des estimateurs est sensiblement la même.

5 Références

[1] Bootstrap(s) in the i.i.d case , Bertail
[2] Méthodes de bootstrap pour les queues de distribution, 2005, Berkane Hassiba
[3] Estimating the Parameters of a Pareto Distribution, Joseph Lee Petersen
[4] Estimation et tests en théorie des valeurs extrêmes, Gwladys Toulemonde

6 Code source

Le projet et les résultats sont issues d'une programmation en python. Le code est visible sur le compte github du groupe : **https ://github.com/MoussabDjerrab/Bootstrap/blob/master/Code/test_02052015.ipynb**