Bootstrap et Rééchantillonnage

CHAMPAGNE Clara
DONNOT Benjamin
PLUNTZ Matthieu

Table des matières

Introduction

La loi de Pareto a pour fonction de répartition :

$$F(x) = 1 - \left(\frac{c}{x}\right)^{\beta} \text{ pour } x > c, \, c \text{ connu}$$

La distribution ne possède de moments que pour les ordres inférieurs à β. Elle appartient au domaine d'attraction de Fréchet (distributions à queues lourdes). Il s'agit également d'une distribution fortement asymétrique. Si $\beta > 3$, on a

$$k_3(\beta) = \frac{2(1+\beta)}{\beta - 3}\sqrt{\frac{\beta - 2}{\beta}}$$

Dans ce travail, on se propose d'étudier les méthodes d'estimation des quantiles de la loi de Pareto. Les caractéristiques de cette loi, ainsi que celles des quantiles rendent cette estimation complexe. Il s'agit donc d'étudier dans quelle mesure le bootstrap constitue une solution à ces problèmes, et quel est le type de bootstrap le mieux adapté.

Cette loi est très utilisée dans les domaines actuariels. En effet, elle permet de modéliser des phénomènes ayant des impacts importants (phénomènes liés aux "queues lourdes"). Et, une des principale mesure de risque repose sur la "Value at Risk", qui n'est ni plus ni moins qu'un fractile.

On étudiera successivement deux estimateurs des quantiles : l'estimateur non paramétrique du quantile empirique puis deux estimateurs paramétriques, l'estimateur de Hill et l'estimateur du maximum de vraisemblance, reposant sur la forme paramétrique de la loi de Pareto.

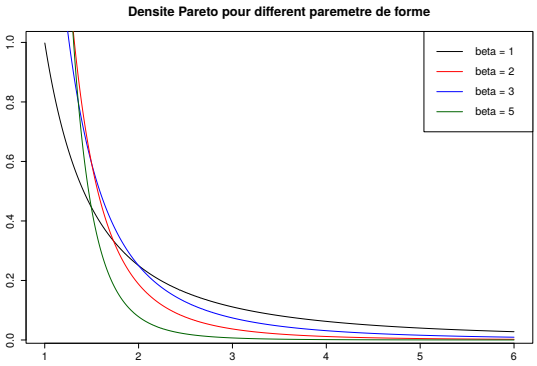


FIGURE 1 – Densité de la loi de Pareto pour $c = 1$ et différents paramètres de forme.

Comme nous pouvons le voir sur le graphique précédent, la forme de la fonction de densité de la loi de Pareto varie beaucoup selon le paramètre. Dans la suite, et sauf mention du contraire, nous avons décidé de prendre $c = 1$ et $\beta = 2$.

I Approche non paramétrique : l'estimateur du quantile empirique

1 Théorie et propriétés asymptotiques

Dans cette section, on étudie l'estimateur du quantile empirique d'ordre q, soit $\hat{Q}_n(q) = \hat{F}_n^{-1}(q)$, où $\hat{F}_n$ est la fonction de répartition empirique.
La fonction d'influence du quantile d'ordre q vaut [1] :

$$T^{(1)}[x,P] = \frac{q - \mathbf{1}_{x \leq F^{-1}(q)}}{f(F^{-1}(q))}$$

Par le théorème de Von Mises, cet estimateur est convergent et asymptotiquement normal selon (oracle 1)

$$\sqrt{n}[\hat{F}_n^{-1}(q) - F^{-1}(q)] \sim N\left(0, \frac{q(1-q)}{[f(F^{-1}(q))]^2}\right)$$

On peut ainsi construire un intervalle de confiance asymptotique autour de la valeur de $\hat{Q}_n(q)$, soit $[\hat{Q}_n(q) \pm \frac{a_\alpha \sqrt{q(1-q)}}{f(\hat{Q}_n(q))\sqrt{n}}]$, où a_α est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$.

Sans connaître la densité f, la variance asymptotique est inconnue. Un intervalle peut être calculé empiriquement pour l'échantillon $(X_1,...,X_n)$. Si on considère l'échantillon ordonné $(X_{(1)},...,X_{(n)})$, l'intervalle $[X_{nq-\sqrt{n}a_\alpha\sqrt{q(1-q)}}, X_{nq+\sqrt{n}a_\alpha\sqrt{q(1-q)}}]$, constitue un intervalle de confiance de niveau asymptotique α. Il nécessite toutefois que le nombre d'observation n soit suffisamment grand pour être correctement défini (la démonstration est fournie en annexe).

La vitesse de convergence, pour un quantile q fixé, est donc de l'ordre de $O(\frac{1}{\sqrt{n}})$. Toutefois, plus l'ordre du quantile est élevé, plus la variance asymptotique est grande : le numérateur est plus élevé et le dénominateur plus faible (car on se situe dans la queue de distribution, dans laquelle la densité f est plus faible). Par ailleurs, le coefficient de *skewness* de cette loi très asymétrique (lorsqu'il existe !) perturbe fortement les conclusions de l'approche par l'asymptotique.

Le calcul des quantiles d'une telle loi peut être utilisé comme un indicateur de risque. En particulier, on peut vouloir tester si le quantile de la loi est supérieur à une valeur fixée R, dans le cadre d'un test de niveau α.

$$H_0 : Q(q) \leq R$$
$$H_1 : Q(q) > R$$

Le test de Wald unilatéral correspondant a pour région critique :

$$\frac{n(\hat{Q}_n(q) - R)^2}{\frac{q(1-q)}{[f(F^{-1}(q))]^2}} \geq \chi^2_{1-\alpha}$$

Ce test de niveau α est consistant mais on rencontre toujours le problème de la variance asymptotique, qui nécessite de connaître la densité de la loi pour être calculée.

2 Bootstrap

L'approche par l'asymptotique présente de nombreux problèmes. D'une part, la variance asymptotique est inconnue. D'autre part, les approximations peuvent être fallacieuses, surtout pour la loi de Pareto, asymétrique et à queue lourde, et en particulier pour les quantiles d'ordre élevé.

Il s'agit désormais d'étudier différentes techniques de bootstrap dans ce contexte, afin de déterminer l'approche la plus adéquate pour pallier ces problèmes. On suppose disponible un échantillon de n observations $(X_1, ..., X_n)$.

ii Bootstrap naïf

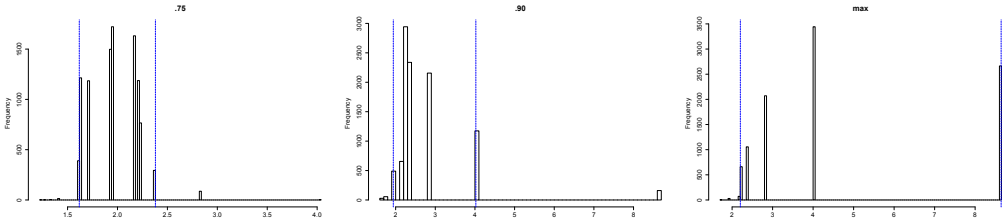
On considère tout d'abord l'approche classique par bootstrap : on réalise B tirages avec remise parmi $(X_1, ..., X_n)$ d'un échantillon de n observations $(X_1^*, ..., X_n^*)$.

Comme la statistique du quantile est Fréchet différentiable, le théorème de Von Mises s'extrapole au cas du bootstrap, et on obtient :

$$\sqrt{n}\left[\hat{F_n^*}^{-1}(q) - \hat{F_n^{-1}}(q)\right] \sim N\left(0\ ,\ \frac{q(1-q)}{[f(F^{-1}(q))]^2}\right)$$

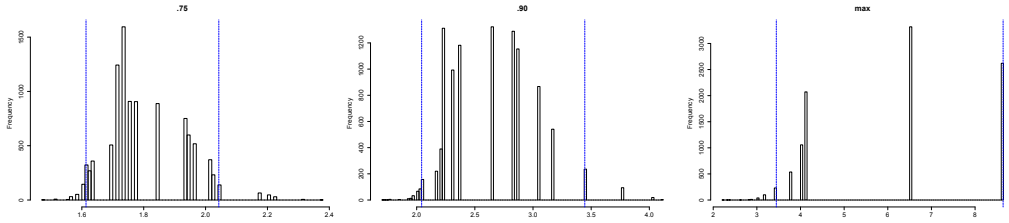
Le bootstrap est donc convergent. En revanche, il n'est pas valide au second ordre.
D'une part, la condition de Cramer n'est pas vérifiée car la fonction de répartition empirique est par définition à support discret. On n'a donc pas $\overline{lim}_{t\to+\infty}|\mathbb{E}_{P_n}e^{itX}| < 1$.
De plus, la fonction d'influence de la loi empirique n'existe pas, son dénominateur étant impossible à connaître dans le cas discret.

L'approche par bootstrap naïf est donc convergente, mais elle est pire que l'asymptotique car elle converge seulement à une vitesse en $O(n^{-1/4})$.



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 29/30

FIGURE 2 – Histogramme de la distribution du Bootstrap "naïf" pour n = 30. En bleu, l'intervalle de confiance à 95%.



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 99/100

FIGURE 3 – Histogramme de la distribution du Bootstrap "naïf" pour n = 100. En bleu, l'intervalle de confiance à 95%.

On constate graphiquement que l'estimation par bootstrap dans cette situation est bien mauvaise. Le tirage n'étant réalisé que parmi les valeurs de l'échantillon (qu'il y en ait 30 ou 100), la distribution bootstrap est concentrée sur un petit nombre de valeurs discrètes. Les estimations sont d'autant moins bonnes que l'on s'intéresse à un quantile élevé. Dans le cas du quantile maximal (1-1/n), on assiste à un échec du bootstrap qui, même au premier ordre, n'est pas convergent. En effet, l'estimation de ce quantile s'apparente à l'estimation du maximum, dont la statistique empirique ne possède pas de propriétés d'équicontinuité quelle que soit la métrique retenue. Les estimations classiques par bootstrap ne sont donc pas valides dans cette situation.

ii Bootstrap lissé

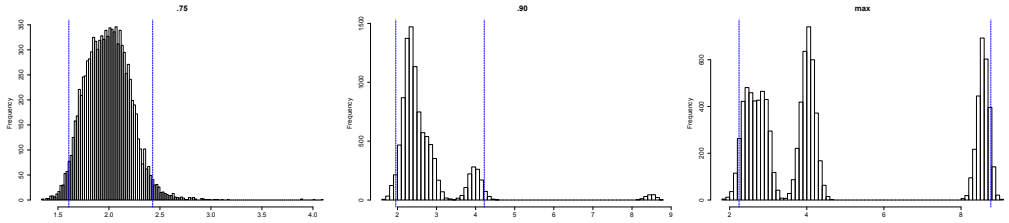
Afin de pallier ce défaut, on a recours au bootstrap lissé. On utilise pour cela un noyau K_{h_n}. On simule un B échantillons $(\varepsilon_1^i,...,\varepsilon_n^i)$ $(i = 1...B)$ selon K_{h_n}. Ensuite, l'échantillon boostrap $(X_1^*,...,X_n^*)$ numéro i est tiré avec remise dans $(X_1+\varepsilon_1^i,...,X_n+\varepsilon_n^i)$. Chaque échantillon bootstrap est donc tiré avec remise dans l'ensemble des observations perturbées.

Par exemple, dans le cas du noyau gaussien, on a $(N_1,...,N_n) \sim N(0,1)$, $(\tilde{X}_1^*,...,\tilde{X}_n^*)$ tirés indépendamment dans $\hat{F}_n$, et $X_i^* = \tilde{X}_i^* h_n N_i$, $(i = 1...n)$. On répète B fois cette opération.

Le lissage de la fonction de répartition empirique permet l'existence de la fonction d'influence de la loi empirique. Ainsi, le développement d'Edgeworth peut se calculer et sous certaines hypothèses de régularité, on obtient la validité du bootstrap au second ordre. Dans le cas d'un noyau gaussien, on a ainsi une vitesse de convergence en $O(n^{-3/4})$ si la variance des perturbations (h_n) est bien choisies.

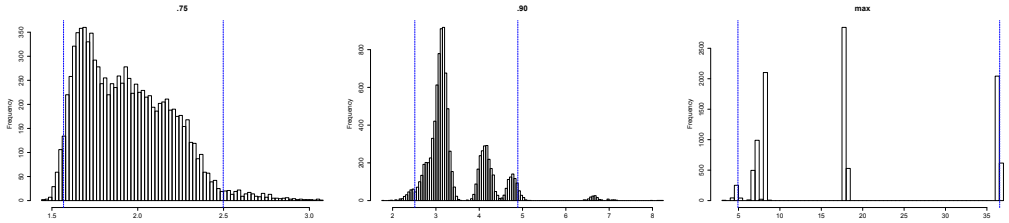
Il est nécessaire de bien choisir la fenêtre h_n : en particulier, il faut respecter les conditions $nh_n \to +\infty$ et $h_n \to 0$.

$$Pr\left[\sqrt{n}\frac{T(P_n^*)-T(P_n)}{S_n} \leq x\right] - Pr\left[\sqrt{n}\frac{T(P_n)-T(P)}{S_n} \leq x\right] = O\left(n^{-3/4}\right)$$



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 29/30

FIGURE 4 – Histogramme de la distribution du Bootstrap "lissé" pour n = 30. En bleu, l'intervalle de confiance à 95%. On utilise un noyau gaussien et on fixe $h_n = 1/\sqrt{n}$.



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 99/100

FIGURE 5 – Histogramme de la distribution du Bootstrap "lissé" pour n = 100. En bleu, l'intervalle de confiance à 95%. On utilise un noyau gaussien et on fixe $h_n = 1/\sqrt{n}$.

Le lissage améliore très sensiblement les estimations bootstrap, notamment pour les quantiles peu élevés. En revanche, l'estimation de la valeur maximale est toujours très douteuse, car en présence de valeurs très dispersées dans l'échantillon (car on se situe dans la queue de distribution), la fenêtre de lissage ne permet pas de bien estimer la densité et de couvrir l'espace d'intérêt. Le lissage ne permet donc pas de pallier l'échec du bootstrap dans le cas de l'estimation du quantile maximal. En effet, la loi de Pareto appartenant au domaine d'attraction de Fréchet, ses queues de distributions sont lourdes et le lissage gaussien sous-estime largement la probabilité d'apparition des observations élévées.

	valeur théorique	IC Oracle	IC Asympt.	IC Naif	IC Lisse
75%	2	1.94	1.94	1.97	1.99
		$[1.29, 2.58]$	$[1.62, 2.38]$	$[1.62, 2.38]$	$[1.61, 2.43]$
90%	3.16	2.42	2.42	2.68	2.71
		$[-0.37, 5.22]$	$[2.17, 8.64]$	$[1.94, 4.02]$	$[1.96, 4.22]$
max	5.48	4.18	4.18	4.69	4.68
		$[-10.87, 19.23]$	$[2.38, NA]$	$[2.2, 8.64]$	$[2.24, 8.77]$

TABLE 1 – Quantiles estimés et intervalles de confiance à 95% par les différents méthodes. Les intervalles de confiance "oracle" correspondent aux intervalles de confiance asymptotiques calculés en utilisant la connaissance de la densité f (soit une loi de Pareto de paramètre connu). Les intervalles de confiance asymptotiques correspondent à la méthode non paramétrique décrite précédemment. n = 30

	valeur théorique	IC Oracle	IC Asympt.	IC Naif	IC Lisse
75%	2	1.76	1.76	1.8	1.81
		[1.41, 2.11]	[1.62, 2.04]	[1.61, 2.04]	[1.62, 2.09]
90%	3.16	2.66	2.66	2.62	2.62
		[1.13, 4.19]	[2.17, 3.76]	[2.04, 3.45]	[2.07, 3.38]
max	10	6.57	6.57	6.05	6.09
		[−44.2, 57.33]	[4.02, *NA*]	[3.45, 8.64]	[3.46, 8.71]

TABLE 2 – Quantiles estimés et intervalles de confiance à 95% par les différents méthodes. Les intervalles de confiance "oracle" correspondent aux intervalles de confiance asymptotiques calculés en utilisant la connaissance de la densité f (soit une loi de Pareto de paramètre connu). Les intervalles de confiance asymptotiques correspondent à la méthode non paramétrique décrite précédemment. n = 100

On constate que les intervalles de confiance calculés par bootstrap sont bien plus proches des intervalles oracles que les intervalles asymptotiques non paramétriques, pour les quantiles à 75% et 90%. Dans le cadre de ces simulations, on ne décèle pas de différence notable entre l'approche par bootstrap lissé et l'approche par bootstrap naïf sur les intervalles de confiance. L'observation des histogrammes des distributions bootstrap sous-jacentes permet toutefois de souligner l'importance du lissage.
L'intervalle de confiance asymptotique oracle explose pour les grands quantiles ce qui s'explique par le fait que le comportement asymptotique ne prévaut que lorsque $n \to \infty$ à q fixé. Cela justifie une approche par bootstrap pour estimer des quantiles proches de 1 quand n n'est pas très grand.
En ce qui concerne l'estimation de la valeur maximale, surtout dans le cas du quantile à 99% (lorsque n=100), on assiste à l'échec des trois méthodes d'estimation. La valeur du quantile est largement sous-estimée, et les intervalles de confiance sont soit infinis (cas asymptotique), soit ne contiennent pas la bonne valeur du paramètre. L'estimation non paramétrique du maximum, par des techniques asymptotiques ou bootstrap est un échec même au premier ordre (non convergence vers la bonne valeur du paramètre).

ii Tests

Dans l'approche par bootstrap, la construction d'un test est directe : il s'agit de voir si la cible du test est contenue dans l'intervalle de confiance autour de la valeur estimée. Par exemple, dans le test décrit précédemment :

$$H_0 : Q(q) \leq R$$
$$H_1 : Q(q) > R$$

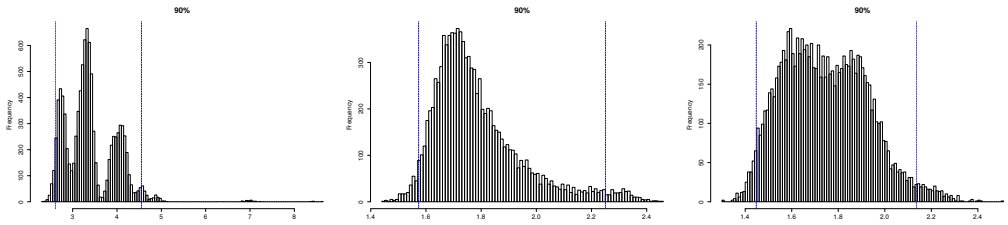
Il faut ainsi voir si la borne supérieure de l'intervalle de confiance à 90% dépasse la valeur R (dans le cadre d'un test unilatéral de niveau 5%). Dans ce cas, H_0 est rejetée.

3 Retour sur le choix du paramètre β

Dans cette sous-partie, nous nous focaliserons sur le paramètre β de la loi de Pareto (aussi appelé paramètre de forme), et tenterons de voir son influence sur la distribution du bootstrap lissé (avec un lissage normal de variance $h_n \equiv \sqrt{n}$) du quantile 90% de ces lois.

Pour $\beta = 2, 3$ puis 4, nous allons évaluer le quantile 90% d'une loi de Pareto de paramètre β. Nous ne nous intéresserons pas aux valeurs quantitatives du quantile (qui dépend bien entendu de la loi et qui sera

donc différent selon les lois considérées), mais allons regarder l'impact du paramètre de forme de la loi d'origine sur la distribution du bootstrap lissé de ce quantile.



(a) $\beta = 2$ (b) $\beta = 3$ (c) $\beta = 4$

FIGURE 6 – Histogramme de la distribution du Bootstrap "lissé" pour n = 100, pour 10 000 échantillons du bootstrap du quantile 90% d'une loi de Pareto pour différent paramètre de forme.

La figure 6 montre les résultats obtenus pour différents paramètres de forme. Dans tous les cas, les histogrammes représentent 10 000 échantillons bootstrap différents.
Comme nous pouvons le voir, dans le cas $\beta = 2$ (sous-figure 6a) , la distribution obtenue ne semble pas régulière, on distingue clairement dans le résultat les différentes normales centrées autour de l'échantillon de départ. Il aurait sans doute fallu lisser plus.
Lorsque $\beta = 3$ (6b), ce phénomène disparaît, la loi semble régulière mais n'est absolument pas symétrique. Il serait donc dangereux de considérer l'intervalle de confiance asymptotique, qui revient peu ou prou à faire un hypothèse de normalité.
Enfin lorsque $\beta = 4$ (6c), non seulement la distribution bootstrap semble régulière, mais en plus celle-ci est "quasiment" symétrique. L'hypothèse de normalité n'est dans ce cas pas à exclure, et l'asymptotique donnerait sans doute de bons résultats.

Ici, comme nous avons pu le voir, il apparaît que plus la queue est lourde (ie plus β est petit), plus l'intérêt du Bootstrap par rapport à l'asymptotique se justifie. En effet, plus la queue sera lourdre, moins la distribution bootstrap pourra être considérée comme proche d'une normale.

Ces résultats illustrent également le fait que la fenêtre de lissage doit dépendre de la forme de la distribution : plus β est petit, plus le lissage correspondant doit être élevé.

II Approche paramétrique : maximum de vraisemblance et estimateur de Hill

1 Estimation par maximum de vraisemblance

i Expression de l'estimateur

Dans cette sous-partie on utilise l'hypothèse que les données suivent une loi de Pareto de paramètre $\beta = \beta_0$ inconnu, et c connu qu'on prendra égal à 1 dans les simulations. On commence par déterminer l'estimateur du maximum de vraisemblance $\widehat{\beta}^{MV}$ de β. La vraisemblance s'écrit :

$$L(\beta|X_1) = \beta \left(\frac{X_1}{c}\right)^{-\beta-1}$$

On en déduit la log-vraisemblance :

$$l(\beta|X_1) = \log(\beta) - (\beta+1)\log\frac{X_1}{c}$$

donc pour n observations indépendantes on obtient :

$$l\left(\beta|X^{(n)}\right) = n\log(\beta) - (\beta+1)\sum_{i=1}^{n}\log\frac{X_i}{c}$$

et en dérivant :

$$\frac{\partial}{\partial\beta}l\left(\beta|X^{(n)}\right) = \frac{n}{\beta} - \sum_{i=1}^{n}\log\frac{X_i}{c}$$

d'où :

$$\frac{1}{\widehat{\beta}_{MV}} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{X_i}{c}$$

On prend maintenant comme estimateur du quantile $Q(q)$ d'ordre q, le quantile théorique d'une loi de Pareto de paramètre $\widehat{\beta}_{MV}$ qui est donné par :

$$\widehat{Q(q)}_{MV} = (1-q)^{-1/\widehat{\beta}^{MV}}$$

C'est-à-dire :

$$\widehat{Q(q)}_{MV} = (1-q)^{-E_{\hat{F}_n}\left[\log\frac{X}{c}\right]}$$

Ce qui permet de bien voir qu'à l'instar du quantile empirique, le quantile $\widehat{Q(q)}_{MV}$ est une fonctionnelle appliquée à la fonction de répartition empirique $\hat{F}_n$.

i Comportement asymptotique

Remarquons d'abord que par propriété de la loi de Pareto, la variable aléatoire $\log\frac{X_1}{c}$ qui intervient dans l'estimateur suit une loi exponentielle de paramètre β_0. En effet :

$$\begin{aligned}\forall\, t \geq 0,\ P(\log\frac{X_1}{c} \leq t) &= F(c\, e^t)\\ &= 1 - e^{-\beta_0 t}\end{aligned}$$

On a donc :

$$\mathrm{E}\left(\log\frac{X_1}{c}\right) = \frac{1}{\beta_0}$$
$$\mathrm{Var}\left(\log\frac{X_1}{c}\right) = \frac{1}{\beta_0^2}$$

Par conséquent, d'après le théorème de la limite centrale, l'estimateur du paramètre a la loi asymptotique suivante :

$$\sqrt{n}\left(\frac{1}{\widehat{\beta}_{MV}} - \frac{1}{\beta_0}\right) \sim N\left(0, \frac{1}{\beta_0^2}\right)$$

Par delta-méthode, on en déduit la variance asymptotique de l'estimateur du quantile :

$$\frac{\partial\, \widehat{Q(q)}_{MV}}{\partial(\frac{1}{\widehat{\beta}_{MV}})} = -\log(1-q)\ (1-q)^{-1/\widehat{\beta}^{MV}}$$

donc, asymptotiquement (oracle 2) :

$$\sqrt{n}\left[\widehat{Q(q)}_{MV} - Q(q)\right] \sim N\left(0, \frac{\log^2(1-q)\ (1-q)^{-2/\beta_0}}{\beta_0^2}\right)$$

Cette variance asymptotique est inférieure à celle de l'estimateur non paramétrique (i) pour toutes valeurs de β_0 et q. Dans la suite on affichera comme intervalles de confiance "oracle" ceux qui correspondent au cas non paramétrique et qui constituent une sorte de borne supérieure de la variance asymptotique des estimateurs (on ne peut que faire mieux si l'on dispose de plus d'hypothèses paramétriques).

2 Bootstrap

ii Bootstrap naïf de l'estimateur du quantile par maximum de vraisemblance

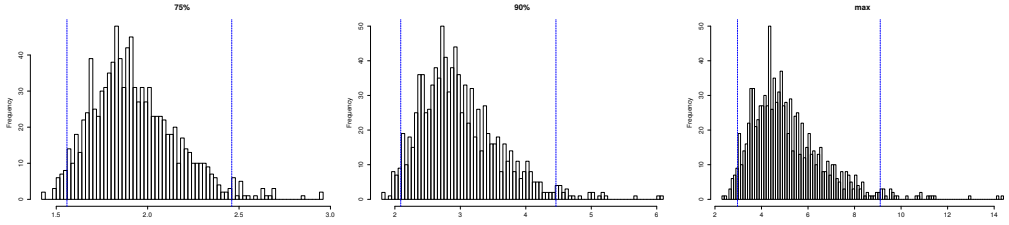
De même que pour l'estimateur non-paramétrique du quantile (voir partie 1), on peut appliquer la méthode standard de bootstrap, ou bootstrap naïf à l'estimateur par maximum de vraisemblance $\widehat{Q(q)}_{MV}$ décrit ci-dessus. On réalise B tirages avec remise parmi $(X_1,...,X_n)$ d'un échantillon de n observations : $\forall b = 1,..,B,\ X^{*b} = (X^*_{1^{(b)}},...,X^*_{n^{(b)}})$, et pour chacun de ces échantillons on calcule :

$$\widehat{Q(q)}_b = \widehat{Q(q)}_{MV}(X^{*b})$$

L'utilisation de $\widehat{Q(q)}_{MV}$ au lieu du quantile empirique permet de résoudre certains inconvénients qui empêchaient la validité au second ordre du bootstrap. En effet la fonction d'influence est la suivante :

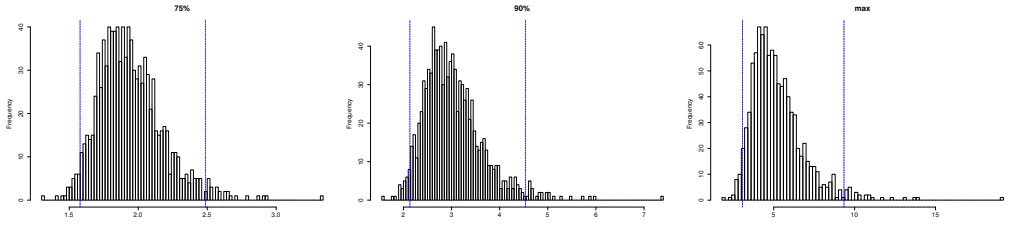
$$\begin{aligned}
\widehat{Q(q)}^{(1)}_{MV}(P,x) &= \frac{\partial}{\partial t}_{|0} \widehat{Q(q)}_{MV}((1-t)P + t\delta_x) \\
&= \frac{\partial}{\partial t}_{|0} (1-q)^{(1-t)E_P[\log\frac{X}{c}] + t\log\frac{x}{c}} \\
&= \widehat{Q(q)}_{MV}(P)\left(\log\frac{x}{c} - E_P\left[\log\frac{X}{c}\right]\right)\log(1-q)
\end{aligned}$$

Cette fonction est définie pour toute loi P à support dans $\mathbb{R}_+$, et en particulier en $P = \hat{F}_n$, alors que la fonction d'influence du quantile n'existe que lorsque P possède une densité par rapport à la mesure de Lebesgue (1), ce qui exclut les lois $\hat{F}_n$.
En plus des intervalles de confiance non-asymptotiques estimés par bootstrap, on peut estimer des intervalles de confiance asymptotiques à partir de la variance asymptotique de l'estimateur en remplaçant β_0 par $\widehat{\beta}_{MV}(X))$ dans l'équation (i).



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 29/30 : pseudo-maximum empirique

FIGURE 7 – Histogramme de la distribution du quantile estimé par EMV pour n = 30, réalisé par bootstrap naïf



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 99/100 : pseudo-maximum empirique

FIGURE 8 – Histogramme de la distribution du quantile estimé par EMV pour n = 100, réalisé par bootstrap naïf

Les distributions bootstrap sont beaucoup plus proches d'une distribution normale que celles obtenues par l'approche non paramétrique. La prise en compte de la connaissance sur la loi améliore les estimations.

	Quantile théorique	IC Oracle	IC Asympt.	IC bootstrap naïf
75%	2	1.76	1.76	1.84
		[1.41, 2.11]	[1.69, 1.83]	[1.66, 2.05]
90%	3.16	2.66	2.66	2.76
		[1.13, 4.19]	[2.47, 2.84]	[2.32, 3.3]
max	10	6.57	6.57	7.7
		[−44.2, 57.33]	[4.88, 8.25]	[5.39, 10.92]

TABLE 3 – Quantile et intervalles de confiance construits à partir de l'estimateur du maximum de vraisemblance. Les intervalles de confiance bootstrap sont calculés par bootstrap naïf selon la méthode décrite. Les intervalles de confiance asymptotiques correspondent à l'intervalle de confiance oracle dans lequel on utilise le paramètre β estimé par maximum de vraisemblance. n = 100

Dans cette situation, les estimations obtenues par bootstrap sont largement plus performantes que les estimations asymptotiques, pour les quantiles à 75% et 90% tout comme pour le quantile à $(n-1)/n$ ou quantile maximal. Les intervalles de confiance asymptotiques ne contiennent en aucun cas la vraie valeur alors que ceux du bootstrap la contiennent à chaque fois.

ii Bootstrap paramétrique pour le quantile empirique

Dans cette sous-partie on considère à nouveau l'estimateur non-paramétrique du quantile (ou quantile empirique) sur lequel portait la partie 1. On va présenter une nouvelle alternative au bootstrap naïf pour cet estimateur. L'hypothèse que les données suivent une loi de Pareto de paramètre β inconnu permet de réaliser la procédure suivante, dite de *bootstrap paramétrique* :

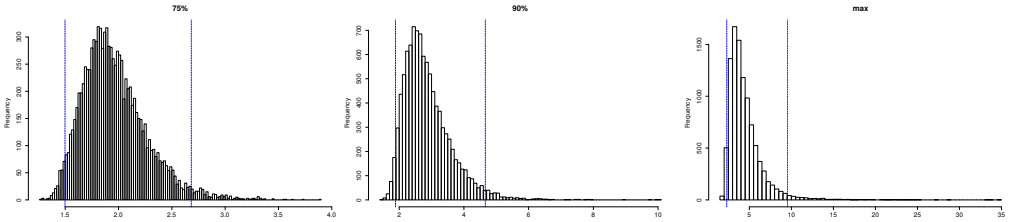
1. On calcule l'estimateur $\widehat{\beta}_{MV}(X)$
2. On simule B échantillons de taille n et de même loi :

$$\forall b = 1,..,B,\ X^{*b} = (X^*_{1^b},...,X^*_{n^b}) \sim Pareto(\widehat{\beta}_{MV}(X))$$

3. On calcule le quantile empirique $\widehat{Q(q)}_b$ de chacun de ces échantillons. On obtient ainsi un échantillon de taille b d'estimateurs du quantile empirique.

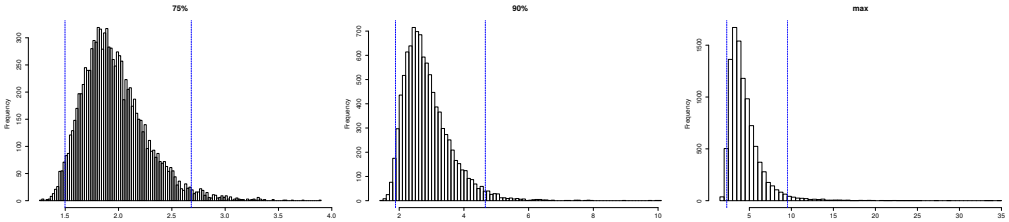
Le bootstrap paramétrique est une façon de lisser la loi F_n selon laquelle les échantillons de bootstrap sont tirés, puisqu'il s'agit ici d'une loi de Pareto. Contrairement à la fonction de répartition empirique utilisée dans un bootstrap naïf, cette loi est absolument continue par rapport à la mesure de Lebesgue donc la fonction d'influence de l'opérateur quantile (1) y est bien définie au premier et au second ordre. Ainsi, les termes du développement d'Edgeworth de la distribution empirique et de celui de la distribution bootstrap convergent vers une même valeur, ce qui rend possible la validité du bootstrap au second ordre

En revanche cette méthode est peu robuste puisque pour qu'elle impose que X suive véritablement une loi de Pareto. Dans le cas contraire F_n ne tend pas vers F et le bootstrap ne converge pas.
Paralèllement au bootstrap, la variance asymptotique du quantile empirique, dont on avait une expression théorique (1) lors de l'estimation non-paramétrique, peut désormais être estimée en remplaçant β_0 par $\widehat{\beta}_{MV}(X))$ dans cette formule, on estime ainsi de nouveaux intervalles de confiance asymptotiques du quantile empirique.



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 29/30 : maximum empirique

FIGURE 9 – Histogramme de la distribution du quantile empirique pour n = 30, obtenu par bootstrap paramétrique



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 99/100 : maximum empirique

FIGURE 10 – Histogramme de la distribution du quantile empirique pour n = 100, obtenu par bootstrap paramétrique

Dans cette situation également, les distributions bootstrap obtenues sont à nouveau plus proches d'une distribution normale que les estimations non paramétriques. Toutefois, les distributions sont fortement asymétriques à mesure que l'on s'intéresse à des quantiles plus élevés. La tendance est donc à la sous-estimation des quantiles élevés.

	Quantile théorique	IC Oracle	IC Asympt.	IC Bootstrap paramétrique (MV)
75%	2	1.76 $[1.41, 2.11]$	1.76 $[1.69, 1.83]$	1.83 $[1.59, 2.14]$
90%	3.16	2.66 $[1.13, 4.19]$	2.66 $[2.47, 2.84]$	2.72 $[2.14, 3.54]$
pseudo-max	10	6.57 $[-44.2, 57.33]$	6.57 $[4.88, 8.25]$	6.72 $[3.59, 14.25]$

TABLE 4 – Estimation du quantile empirique et intervalles de confiance, obtenus par bootstrap paramétrique fondé sur l'estimateur du maximum de vraisemblance. Les intervalles de confiance asymptotiques correspondent à l'intervalle de confiance oracle dans lequel on utilise le paramètre β estimé par maximum de vraisemblance. n=100

Dans cette situation également, les estimations par bootstrap sont nettement plus performantes que les approximations par l'asymptotique. On note toutefois une tendance à la sous-estimation des quantiles plus on s'intéresse à la queue de distribution (qui résulte de l'asymétrie des distributions bootstrap commentée précédemment). Ceci peut être problématique lorsque les estimations de quantiles sont utilisées pour quantifier des risques. Les intervalles de confiance sont toutefois plus larges et mieux centrés autour de la vraie valeur que dans la méthode précédente, ce qui est plus adéquat.

3 Estimateur de Hill : plus robuste à la spécification

L'estimateur de Hill consiste à ajuster une loi de Pareto tronquée sur les $k-1$ plus grandes valeurs de l'échantillon, où $k \leq n$. Plus précisément, en notant $X_{(1)},..,X_{(k)}$ les k plus grandes valeurs de l'échantillon $X^{(n)}$, on suppose que :

$$X_{(1)},..,X_{(k-1)} \sim Pareto(\beta, c = X_{(k)})$$

Cette hypothèse est plus faible que celle qui affirme que X suit une loi de Pareto de manière globale, utilisée jusqu'ici dans cette partie. Pour n grand $k = o(n)$, on peut l'appliquer à une large classe de lois de probabilité (dites du domaine de Fréchet) dont la queue est similaire à celle d'une loi de Pareto [2]

L'estimateur de Hill est alors l'estimateur du maximum de vraisemblance de ce modèle :

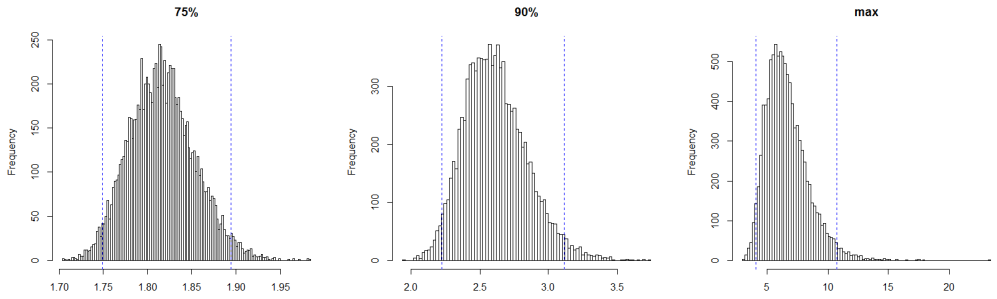
$$\frac{1}{\widehat{\beta}_{k,H}(X)} = \frac{1}{\widehat{\beta}_{MV}(X_{(1)},..,X_{(k-1)})} = \frac{1}{n}\sum_{j=1}^{k-1} \log \frac{X_{(j)}}{X_{(k)}}$$

Du paramètre estimé par la méthode de Hill, on peut déduire l'estimateur suivant des quantiles suffisament élevés, également appelé estimateur de Hill [2] :

$$\forall\, q > 1 - \frac{k}{n},\ \widehat{Q(q)}_{k,H} = X_{(k)} \left(\frac{n}{k}(1-q)\right)^{-\frac{1}{\widehat{\beta}_{k,H}}}$$

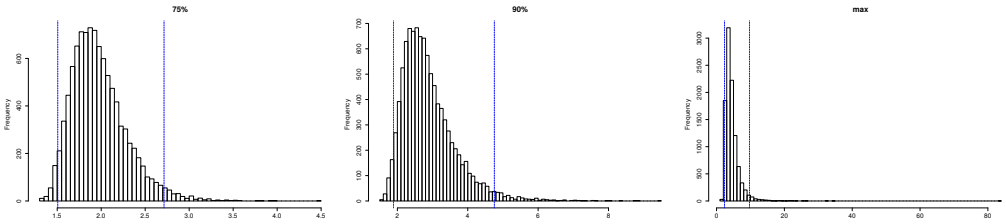
Cet estimateur est convergent si X suit une loi de Pareto dès lors que $k \to \infty$ et $n \to \infty$, mais aussi pour toute distribution du domaine de Fréchet si on a de plus $k = o(n)$. Pour satisfaire la condition $q > 1 - \frac{k}{n}$, il est nécessaire que q tende vers 1 quand $n \to \infty$, par exemple $q = 1 - \frac{1}{n}$ dont le quantile $Q(1 - \frac{1}{n})$ est estimé par le maximum d'un échantillon de taille n.

A partir de cet estimateur, on peut réécrire les deux techniques de bootstrap citées précédemment, soit le bootstrap naïf de l'estimation du paramètre de forme et le bootstrap paramétrique du quantile empirique. Les résultats sont présentés ci-dessous. Dans les simulations, on a retenu une valeur de $k = n/3$.



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 99/100

FIGURE 11 – Histogramme de la distribution du quantile estimé par la méthode de Hill pour n = 100, réalisé par bootstrap naïf



(a) Quantile à 75% (b) Quantile à 90% (c) Quantile à 99/100

FIGURE 12 – Histogramme de la distribution du Bootstrap paramétrique par la méthode de Hill pour n = 100

	Quantile théorique	IC Asympt. (Hill)	IC Hill	
75%	2	1.76 $[1.41, 2.11]$	1.76 $[1.69, 1.82]$	1.82 $[1.75, 1.89]$
90%	3.16	2.66 $[1.13, 4.19]$	2.66 $[2.5, 2.82]$	2.61 $[2.23, 3.11]$
max	10	6.57 $[-44.2, 57.33]$	6.57 $[5.17, 7.97]$	6.65 $[4.1, 10.79]$

TABLE 5 – Quantile et intervalles de confiance construits à partir de l'estimateur de Hill. Les intervalles de confiance bootstrap sont calculés par bootstrap naïf selon la méthode décrite. Les intervalles de confiance asymptotiques correspondent à l'intervalle de confiance oracle dans lequel on utilise le paramètre β estimé par Hill. n = 100

Dans cette situation, les intervalles de confiance de l'estimateur de Hill obtenus par bootstrap sont nettement meilleurs que les intervalles asymptotiques, toutefois ils ne contiennent la vraie valeur du paramètre

que pour le quantile maximal. Le bootstrap est ici plus pertinent pour l'estimation de quantiles élevés. L'estimateur de Hill est donc un bon choix pour l'estimation de quantiles proches de 1 d'une loi à queue lourde dont on ne connaît pas la forme globale, d'autant plus que le bootstrap permet alors de décrire correctement le comportement de l'estimateur.

Conclusion

Afin d'estimer les quantiles d'une loi de Pareto, l'approche par l'asymptotique présente de nombreux défauts, liés à la forte asymétrie et aux queues de distribution de la loi. L'utilisation du bootstrap est donc nécessaire. Cette approche est d'autant plus nécessaire qu'on veut estimer des quantiles d'ordre élevé d'une loi de Pareto à queue lourde (β faible).

La méthode standard de bootstrap appliquée à l'estimateur du quantile empirique est d'une efficacité médiocre, qui a cependant pu être améliorée grâce à l'utilisation de méthodes de bootstrap lissé. Cette méthode est convergente dans la queue de distribution à l'exception du quantile extrême. Afin d'obtenir des estimations plus justes, il est très pertinent d'utiliser les connaissances sur la forme paramétrique de la loi. Ainsi, les quantiles sont plus proches de leurs valeurs théoriques, les intervalles de confiance sont plus fiables, même pour l'estimation de la valeur extrême. En pratique, si la véritable loi des observations est inconnue, mais qu'on sait que ses queues de distribution sont lourdes (l'assimilation à une loi de Pareto n'est pas certaine), il est intéressant d'utiliser une méthode semi-paramétrique fondée sur l'estimateur de Hill. De manière générale, des précautions sont toutefois à prendre dans l'estimation des quantiles les plus extrêmes, qui peuvent être sous-estimés.

Références

[1] Patrice Bertail. Bootstrap(s) in the i.i.d. case.

[2] Frédéric Planchet. Utilisation de la théorie des valeurs extrêmes dans le contexte solvabilité 2, 2012-2013.

Annexe 1 : Calcul des intervalles de confiance non paramétriques pour les quantiles

Ces calculs sont tirés du cours : `http://cermics.enpc.fr/~alfonsi/mrf-quantile.pdf`

On pose $Z_n = \sqrt{n}.\frac{\frac{1}{n}S_n - q}{\sqrt{p.(1-p)}} = \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x_q}$. Par le théorème de la limite centrale, la suite $(Z_n, n \geq 1)$ converge en loi vers la loi gaussienne centrée réduite. Pour n suffisamment grand, on a $1 \leq i_n \leq j_n \leq n$ et

$$P\left(X_{(i_n,n)} \leq x_q \leq X_{(j_n,n)}\right) = P\left(i_n \leq S_n \leq j_n\right) = P\left(\sqrt{n}\frac{\frac{1}{n}i_n - q}{\sqrt{q(1-q)}} \leq Z_n \leq \frac{\frac{1}{n}j_n - q}{\sqrt{q(1-q)}}\right)$$

Par la définition de i_n et j_n, on déduit que pour $n \geq n_0 \geq 1$, on a

$$\begin{aligned} P\left(-a_\alpha \leq Z_n \leq a_\alpha - \frac{1}{\sqrt{n_0}\sqrt{q(1-q)}}\right) & \\ \leq P\left(\sqrt{n}\frac{\frac{1}{n}i_n - p}{\sqrt{q(1-q)}} \leq Z_n \leq \frac{\frac{1}{n}j_n - q}{\sqrt{q(1-q)}}\right) & \\ \leq P\left(-a_\alpha - \frac{1}{\sqrt{n_0}\sqrt{q(1-q)}} \leq Z_n \leq a_\alpha\right) & \end{aligned}$$

On en déduit donc, en faisant tendre n puis n_0 vers l'infini, que

$$\lim_{n\to+\infty} P\left(\sqrt{n}\frac{\frac{1}{n}i_n - q}{\sqrt{q(1-q)}} \leq Z_n \leq \frac{\frac{1}{n}j_n - q}{\sqrt{q(1-q)}}\right) = P\left(-a_\alpha \leq Z \leq a_\alpha\right) = 1 - \alpha$$

On a donc obtenu

$$\lim_{n\to+\infty} P\left(X_{(i_n,n)} \leq x_q \leq X_{(j_n,n)}\right) = 1 - \alpha$$

L'intervalle aléatoire $[X_{(i_n,n)}, X_{(j_n,n)}]$ est bien défini pour n suffisamment grand et c'est un intervalle de confiance pour x_q de niveau asymptotique $1 - \alpha$.

Annexe 2 : Code R

Le code utilisé est disponible sur `https://github.com/BDonnot/Bootstrap`