Estimation des quantiles d'une loi de Pareto

Bootstrap et rééchantillonnage

Clara CHAMPAGNE, Benjamin DONNOT, Matthieu PLUNTZ

ENSAE Paristech - 3A

4 mai 2015

Structure

Loi de Pareto

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).

Loi de Pareto

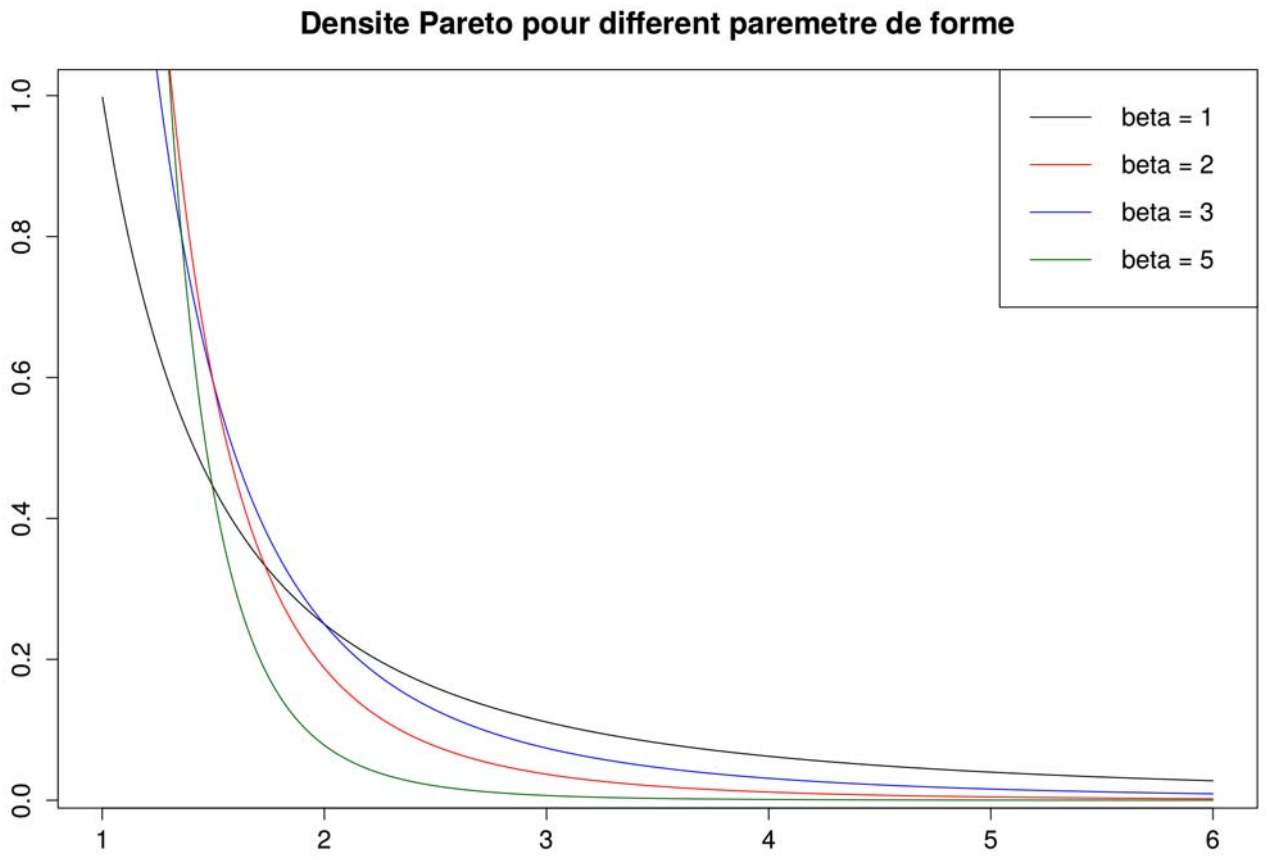

Figure: Densité de la loi de Pareto pour $c = 1$ et différents paramètres de forme (β).

Loi de Pareto : une loi très asymétrique

Si $\beta \leq 3$: $k_3(\beta)$ n'est pas défini.

Si $\beta > 3$, on a :

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

Ainsi $k_3(5) \simeq 4.6$ et $k_3(4) \simeq 7.1$ et $k_3(3.01) \simeq 348.7$!

Quantile empirique : comportement asymptotique

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

- dénominateur de la variance asymptotique qui explose pour les quantiles élevés
- densité de la loi inconnue dans sa version empirique
- convergence perturbée par l'asymétrie de la distribution

Quantile empirique : bootstrap naïf peu performant

Convergence à l'ordre 1 :

$$Pr\left[\sqrt{n}\frac{F_n^{*-1}(q) - F_n^{-1}(q)}{S_n} \leq x\right] - Pr\left[\sqrt{n}\frac{F_n^{-1}(q) - F^{-1}(q)}{S_n} \leq x\right] = O(n^{-1/4})$$

La condition de Cramer n'est pas vérifiée, et f_n n'existe pas : le bootstrap n'est pas valide au second ordre

Une distribution Bootstrap assez particulière

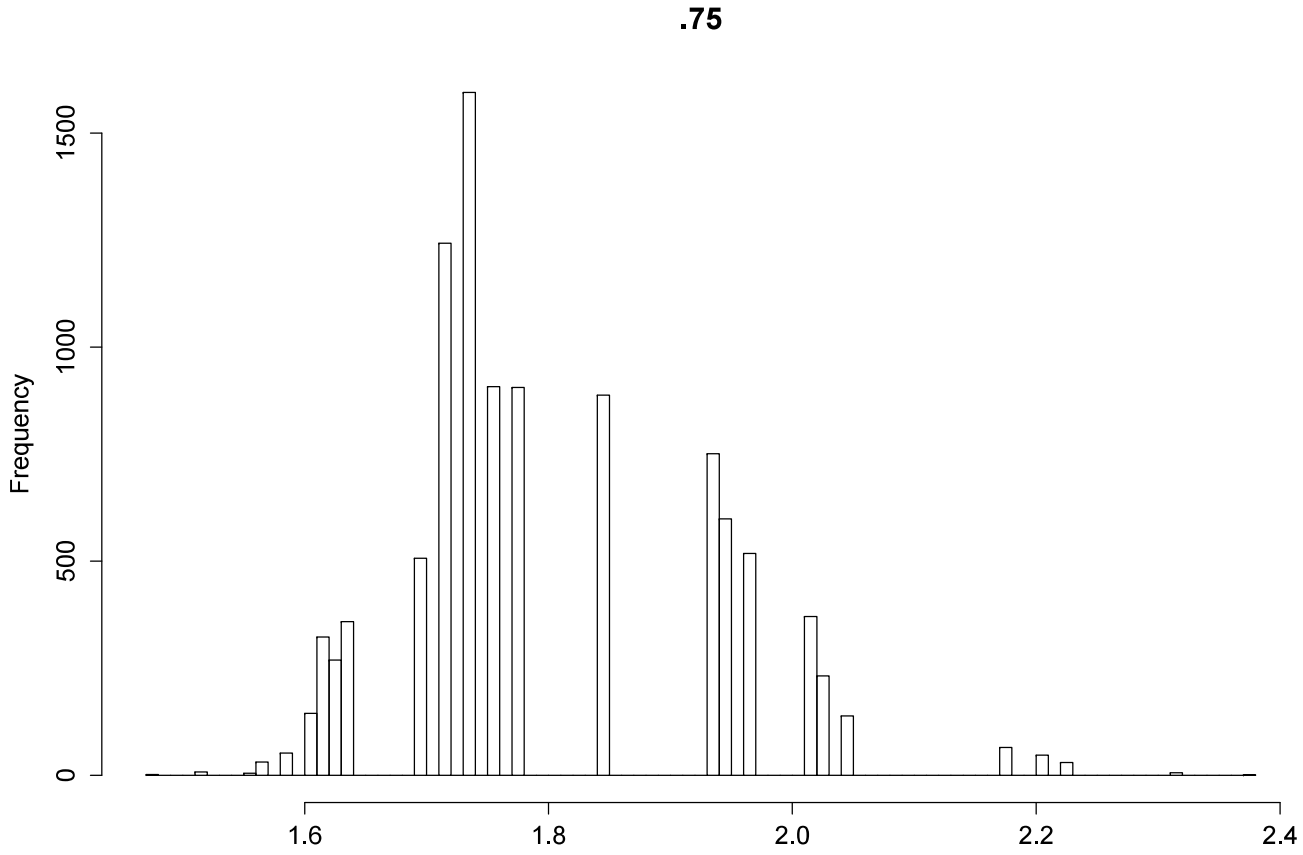

Figure: Quantile à 75%

Une distribution Bootstrap assez particulière

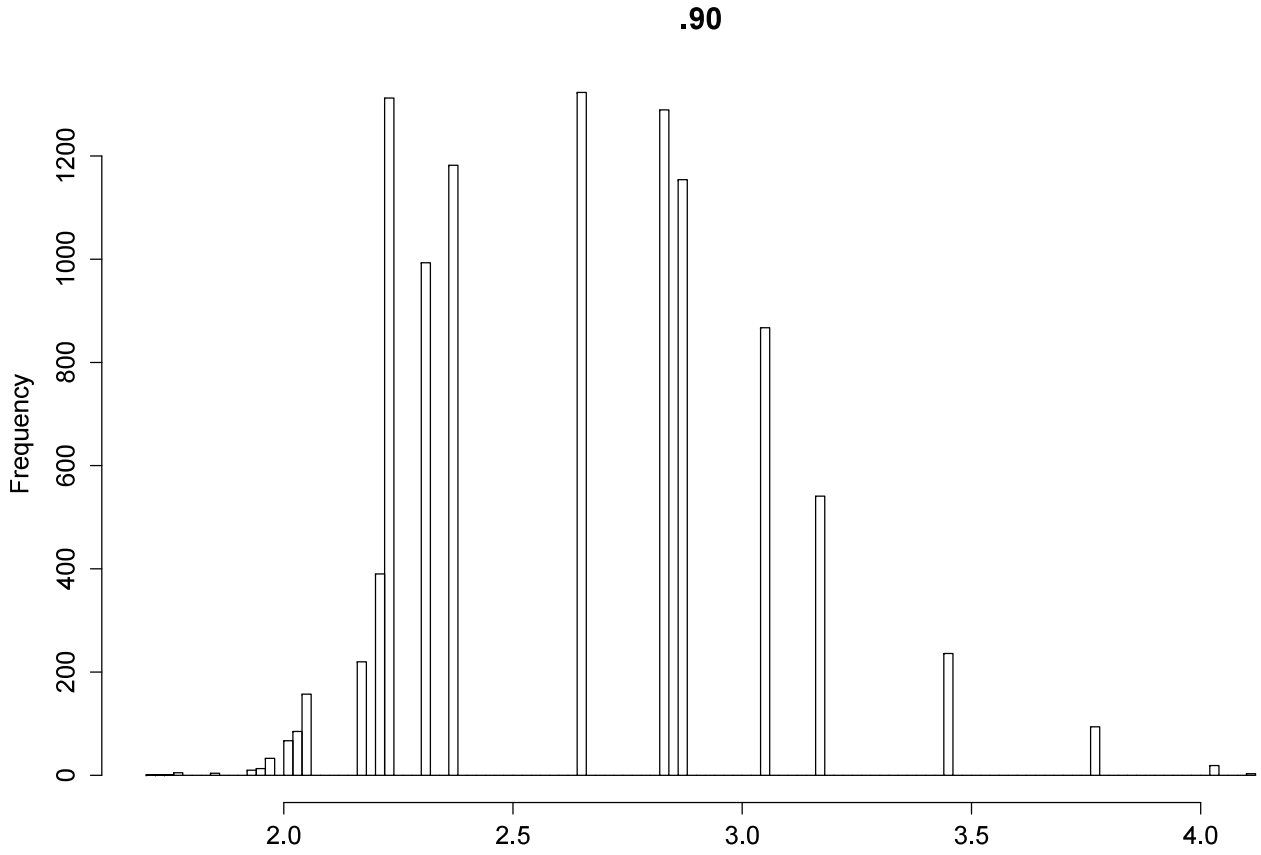

Figure: Quantile à 90%

Une distribution Bootstrap assez particulière

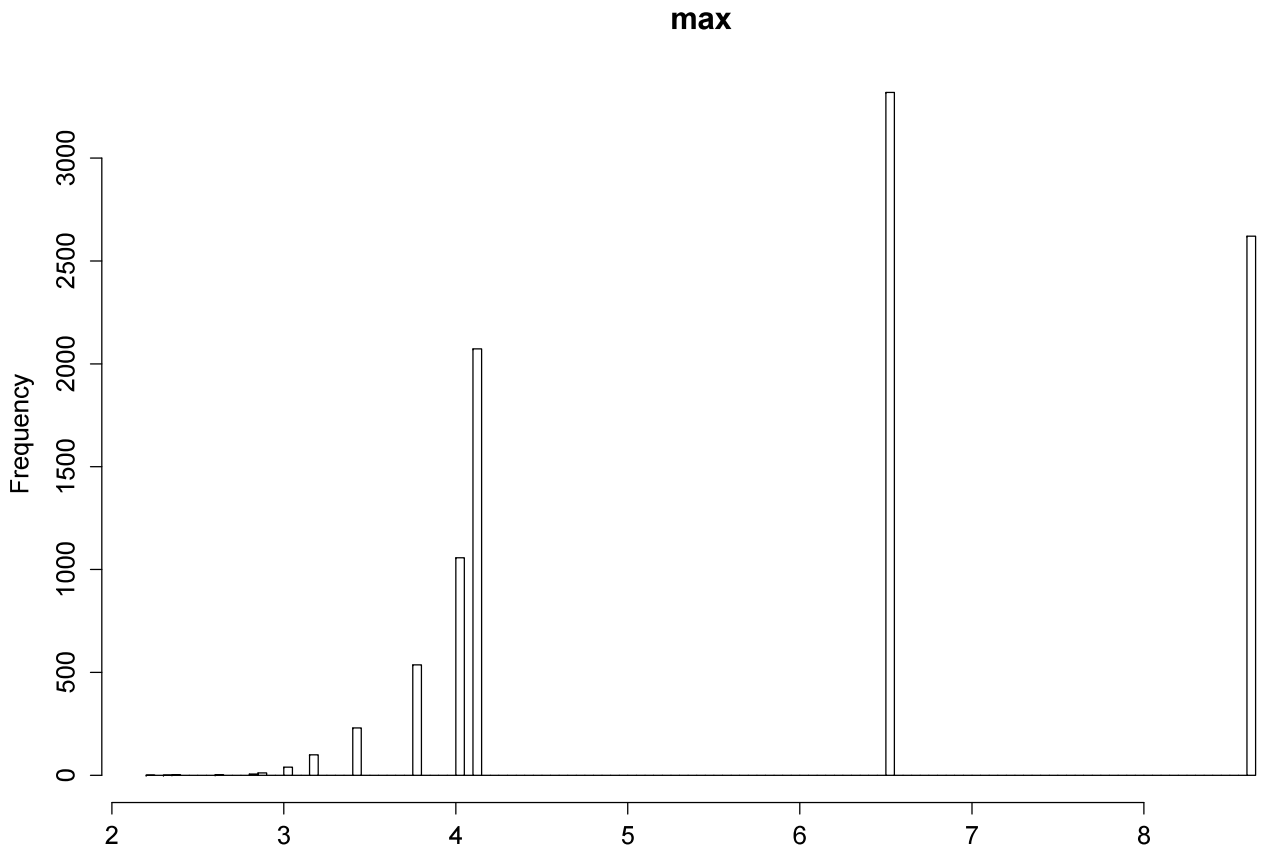

Figure: Quantile à 99/100

Quantile empirique : amélioration par bootstrap lissé

Lissage gaussien :
$(N_1, ..., N_n) \sim \mathcal{N}(0,1)$, $(\tilde{X}_1^*, ..., \tilde{X}_n^*)$ 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.

$$Pr\left[\sqrt{n}\frac{F_n^{*-1}(q) - F_n^{-1}(q)}{S_n} \leq x\right] - Pr\left[\sqrt{n}\frac{F_n^{-1}(q) - F^{-1}(q)}{S_n} \leq x\right] = O(n^{-3/4})$$

Si h_n bien choisi

Distribution "Smooth Bootstrap"

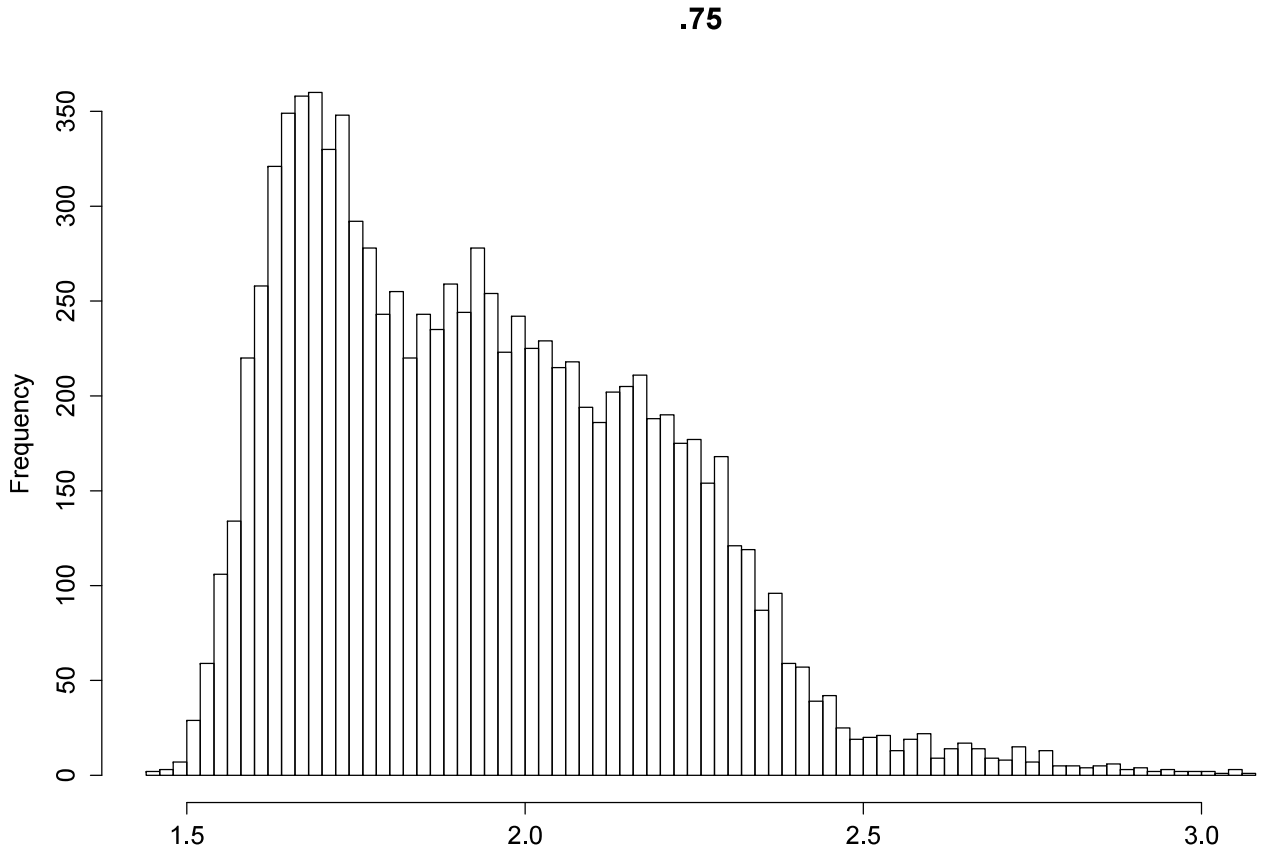

Figure: Quantile à 75%

Distribution "Smooth Bootstrap"

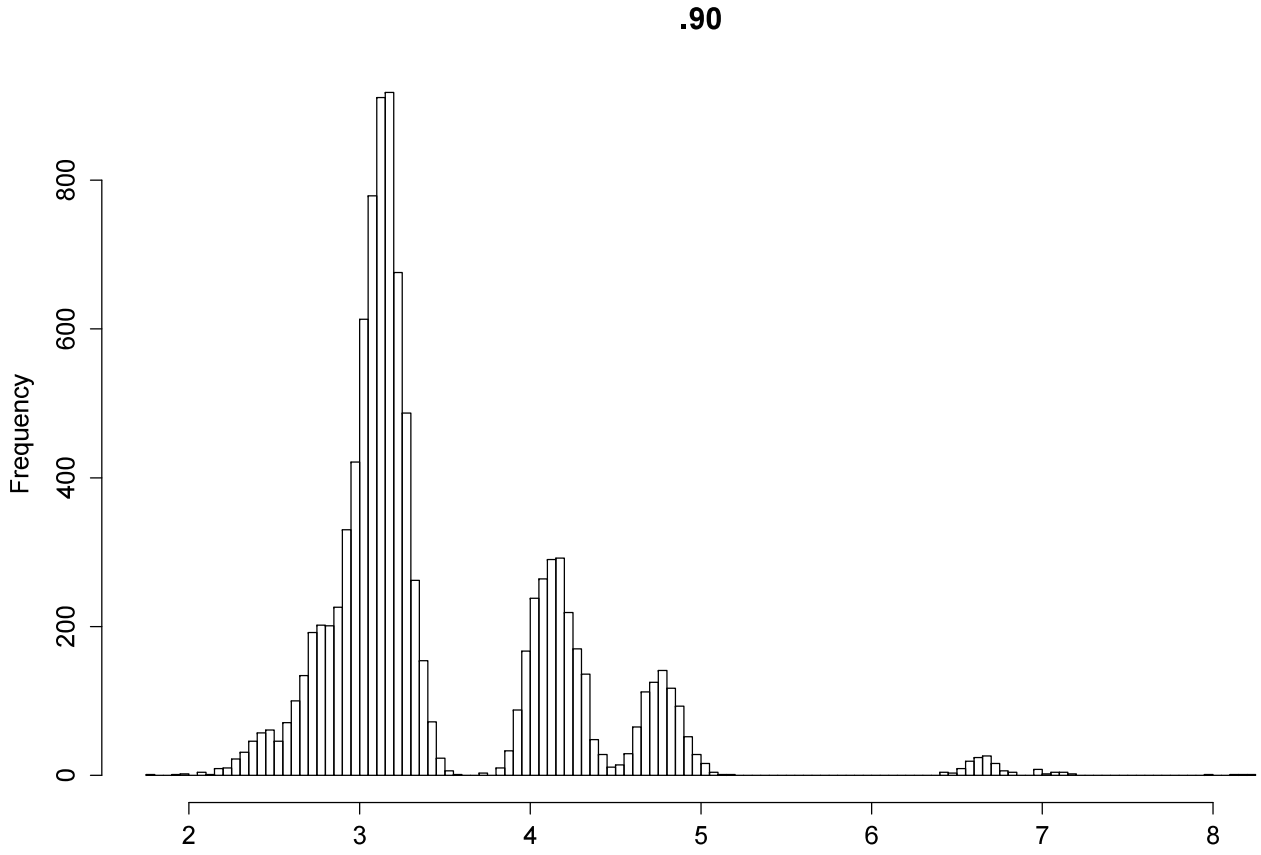

Figure: Quantile à 90%

Distribution "Smooth Bootstrap"

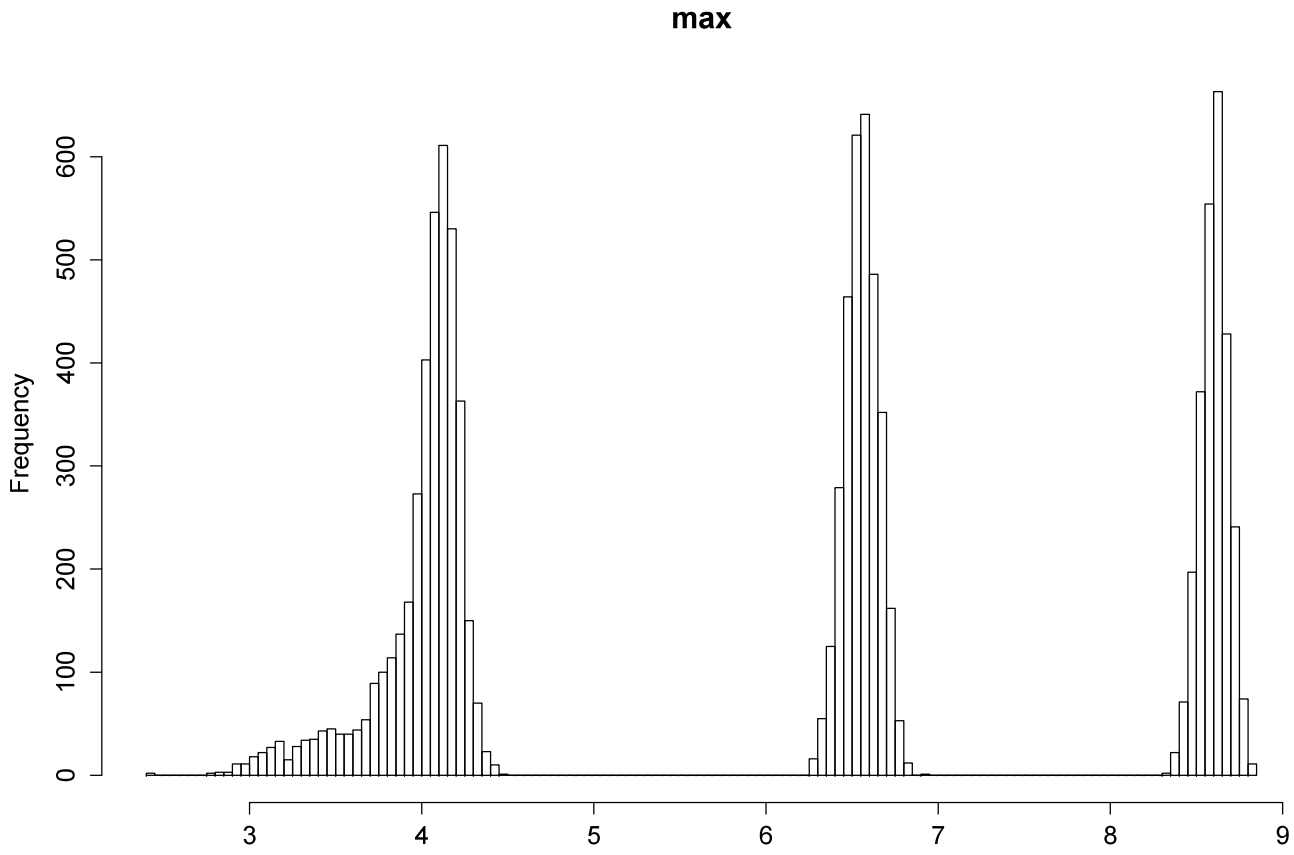

Figure: Quantile à 99/100

Quantile empirique : comparaison des approches

	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]
pseudo 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: Quantiles estimés et intervalles de confiance à 95% par les différentes méthodes (n = 100)

Retour sur le paramètre β

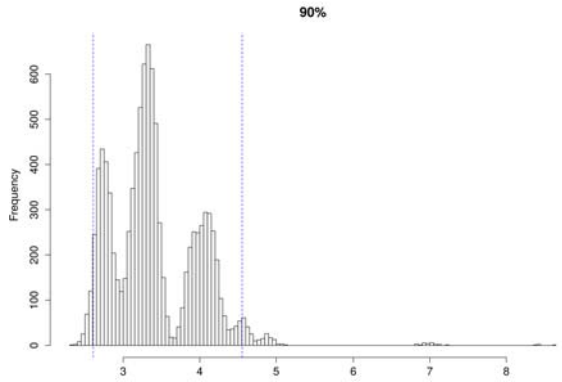

Figure: $\beta = 2$

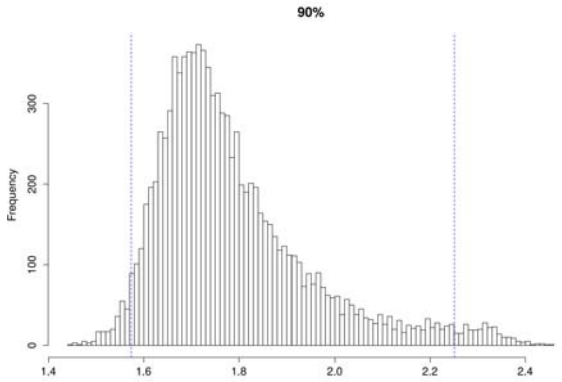

Figure: $\beta = 3$

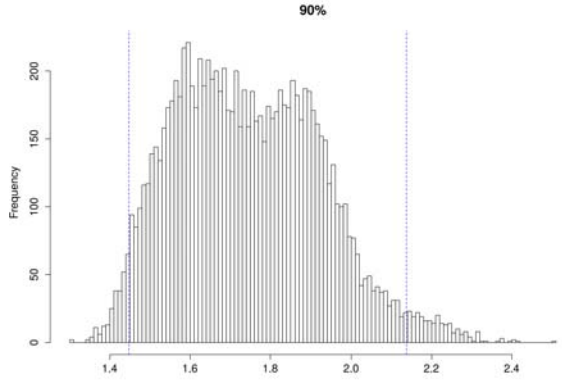

Figure: $\beta = 4$

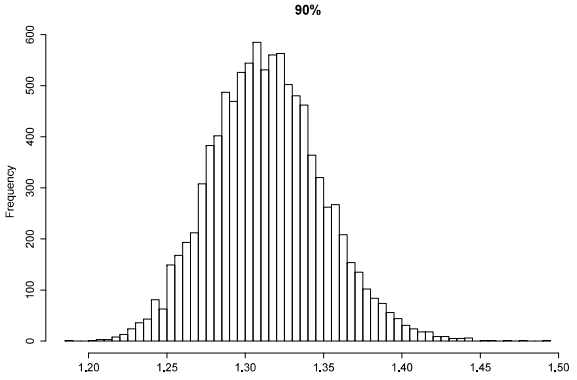

Figure: $\beta = 10$

Estimateur du maximum de vraisemblance

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

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.

Bootstrap paramétrique

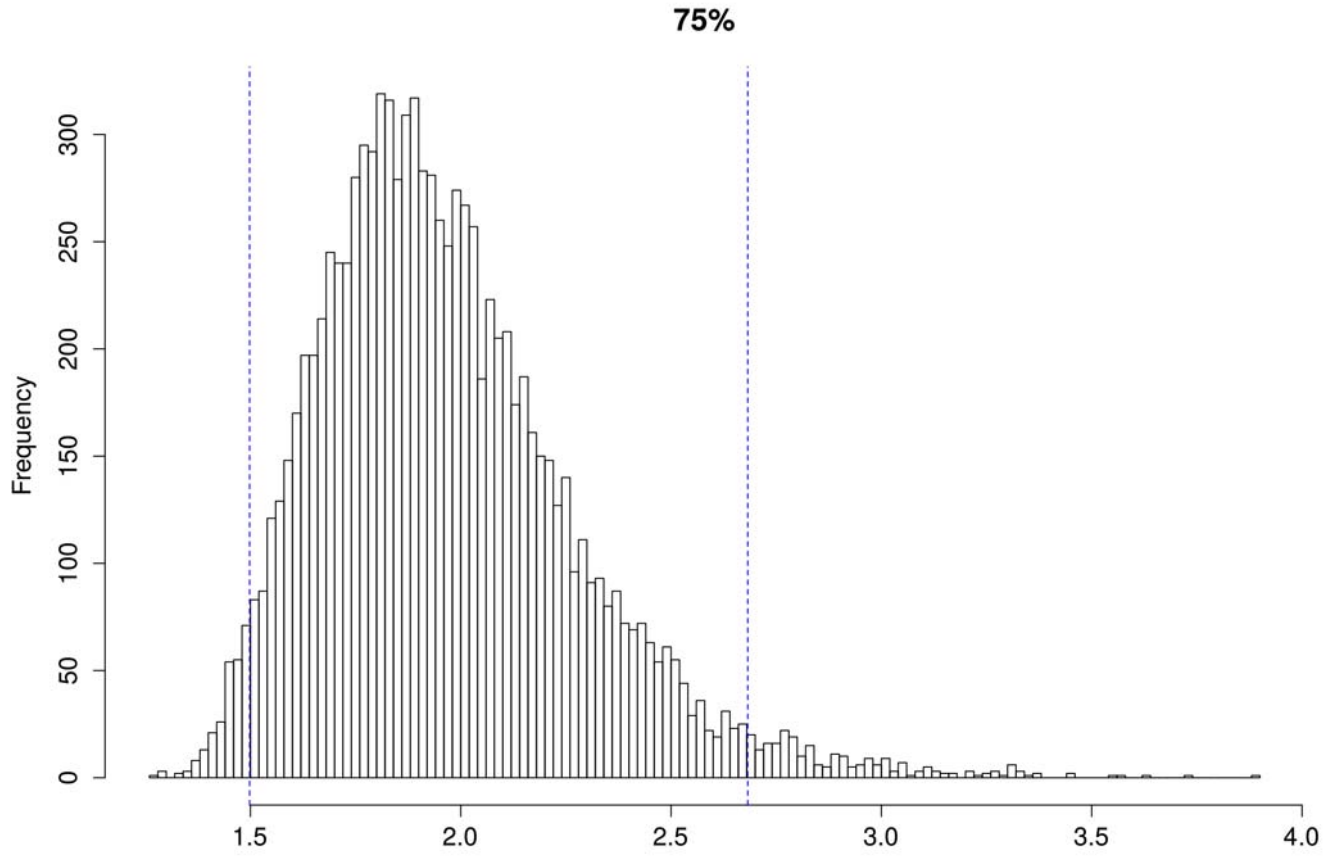

Figure: Quantile à 75%

Bootstrap paramétrique

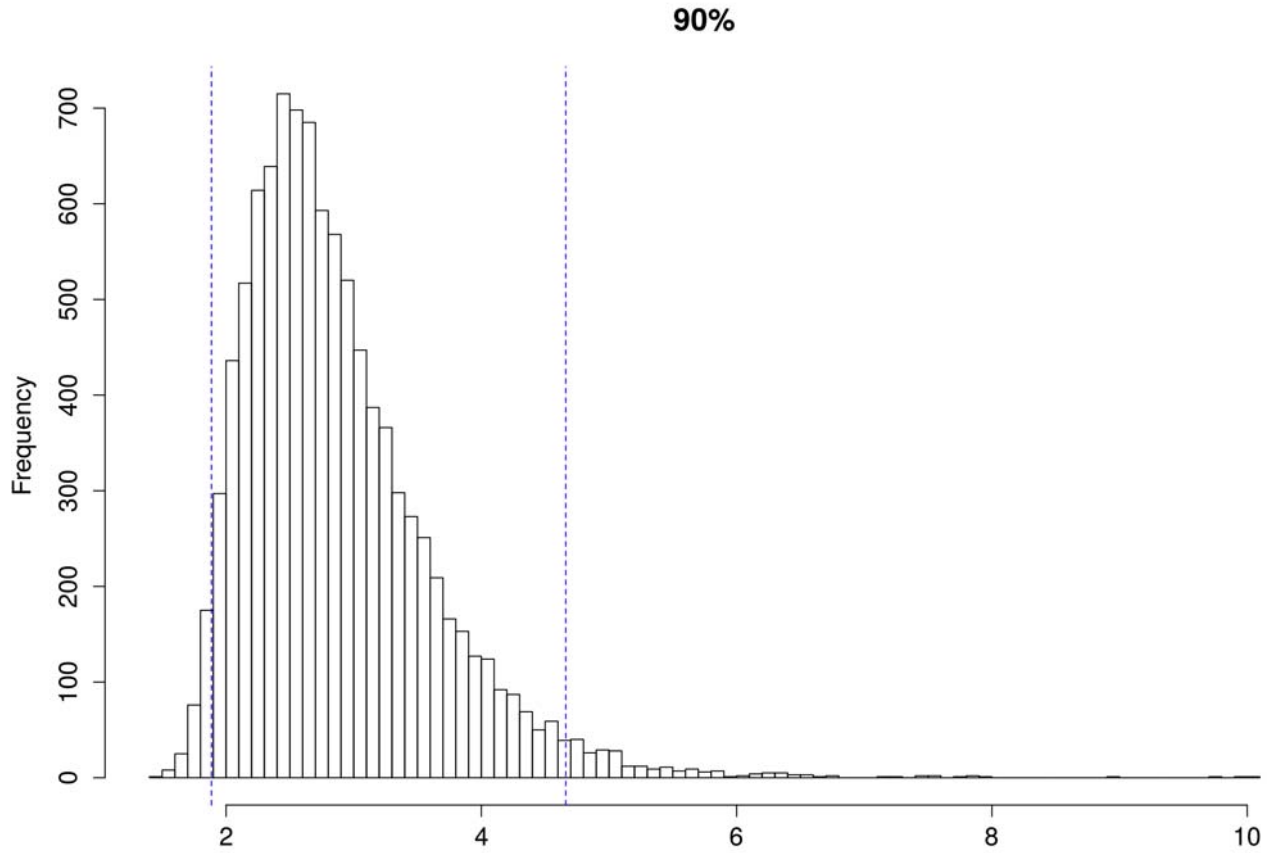

Figure: Quantile à 90%

Bootstrap paramétrique

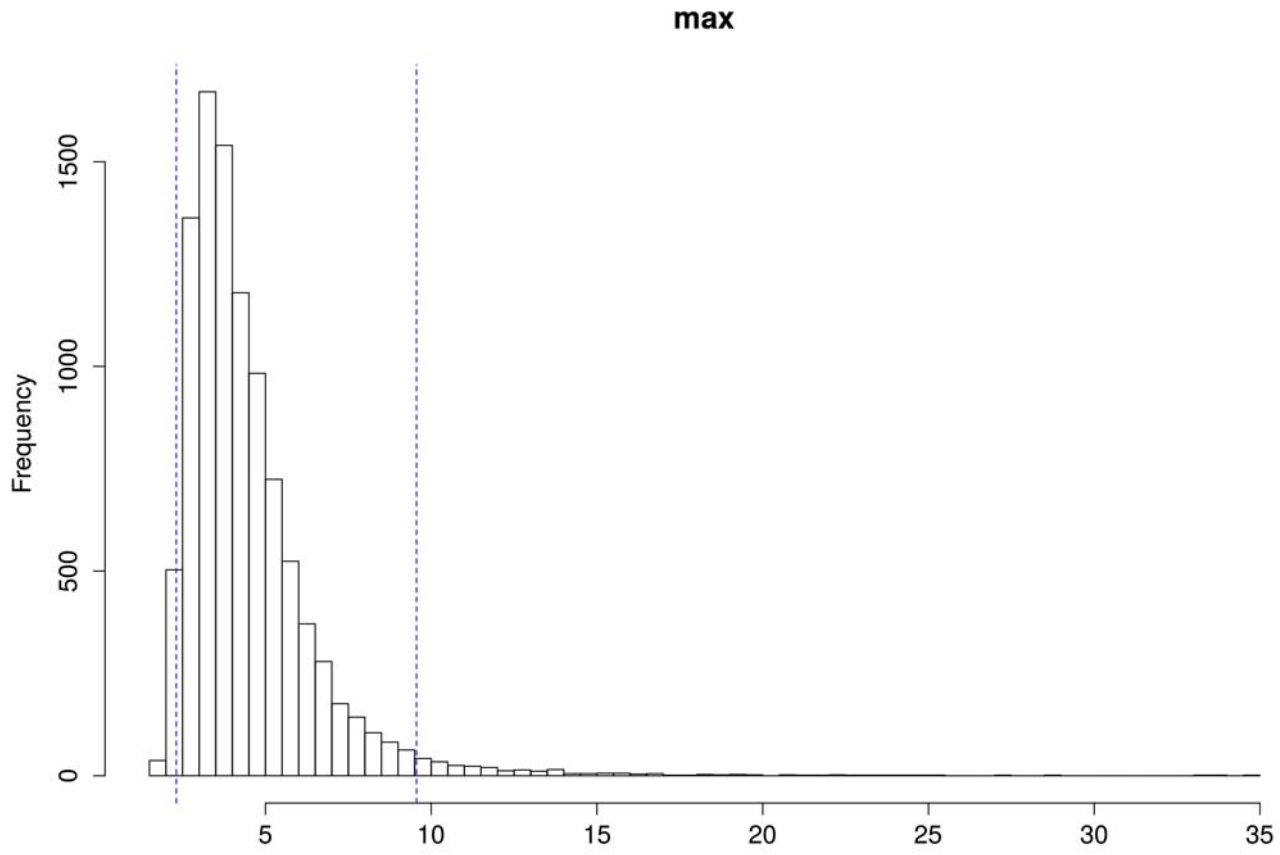

Figure: Quantile à 99/100

	Quantile théorique	IC Oracle	IC Asympt.	IC B. param. (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: Quantiles estimés et intervalles de confiance à 95% par les différentes méthodes (n = 100)

Estimateur Hill

Estimateur de Hill

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

avec $X_{(1)}, .., X_{(k)}$ les k plus grandes valeurs de l'échantillon $X^{(n)}$

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

Bootstrap naïf avec estimateur de Hill

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 :

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

Bootstrap naïf sur le quantile estimé par Hill

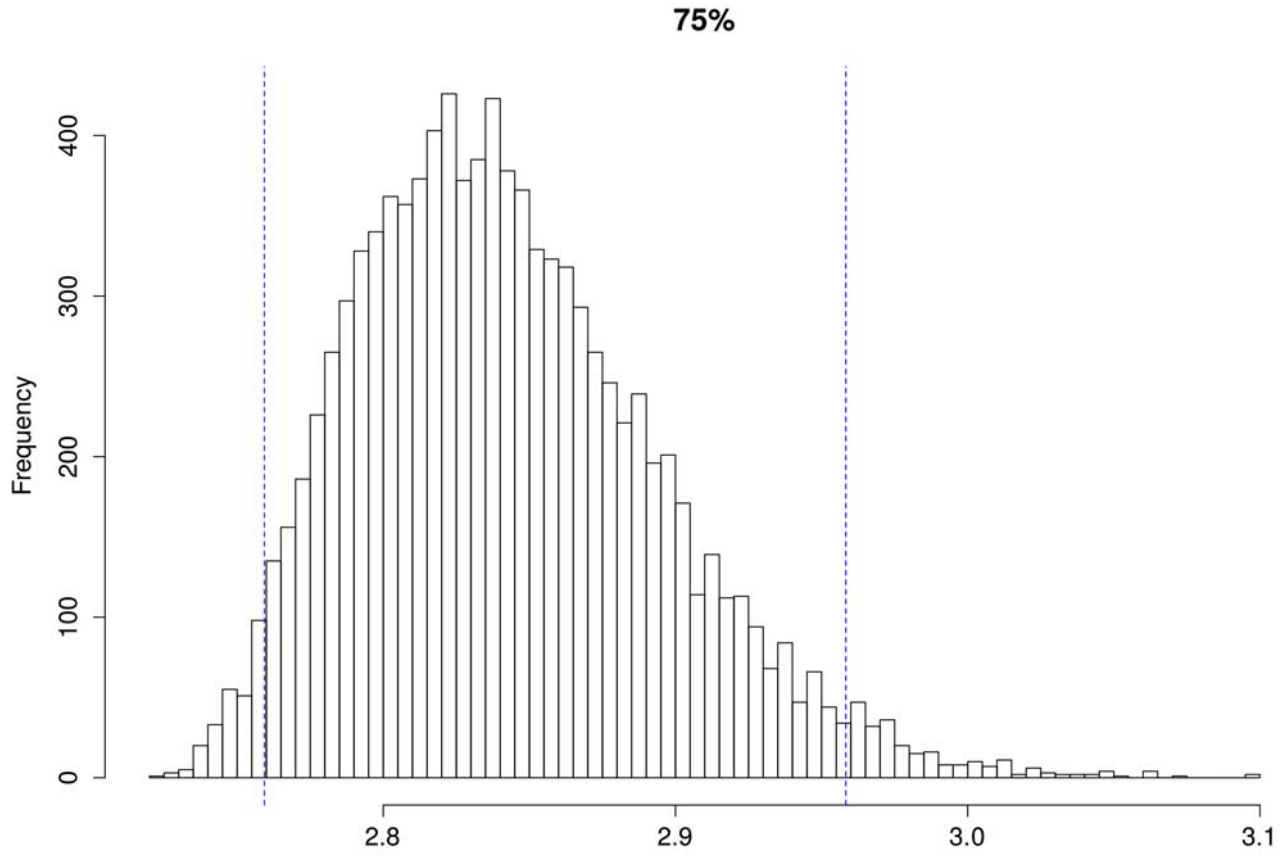

Figure: Quantile à 75%

Bootstrap naïf sur le quantile estimé par Hill

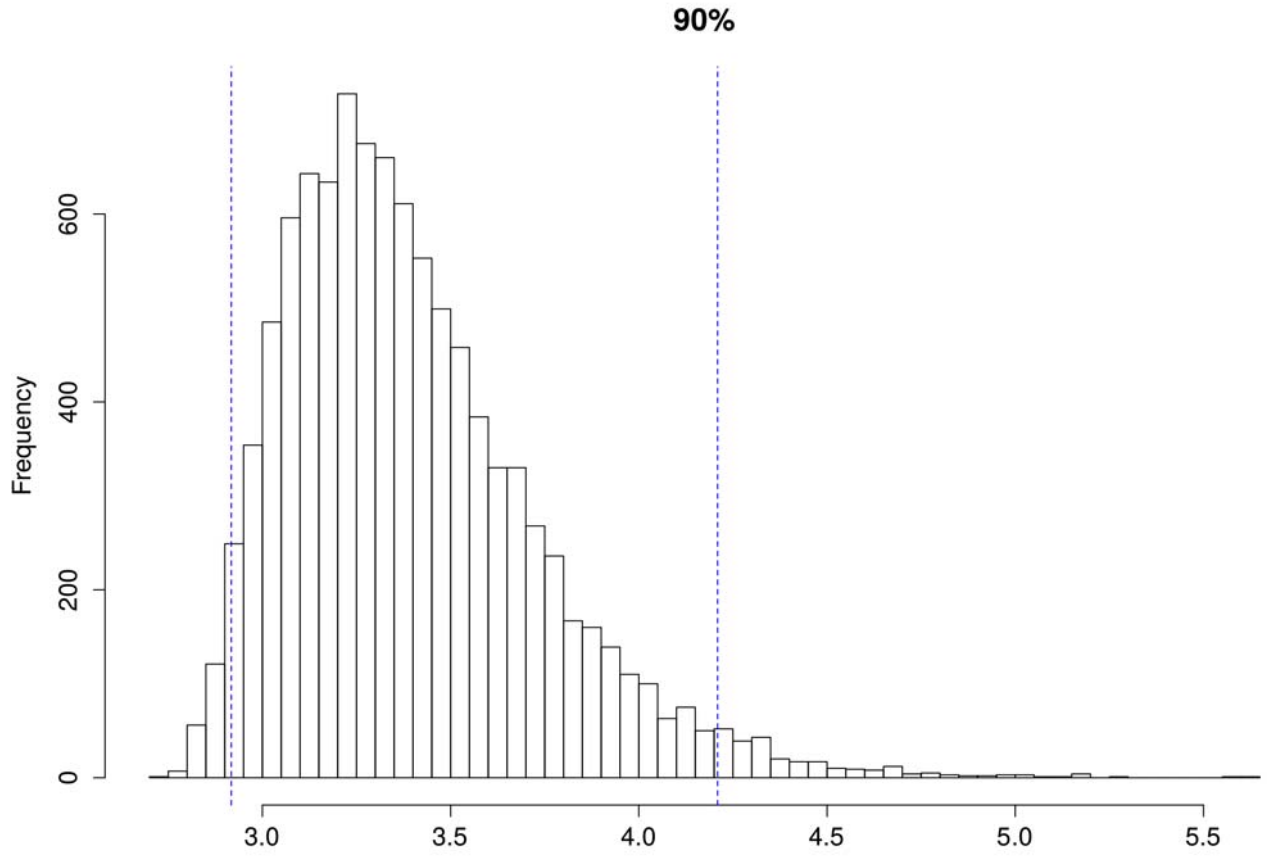

Figure: Quantile à 90%

Bootstrap naïf sur le quantile estimé par Hill

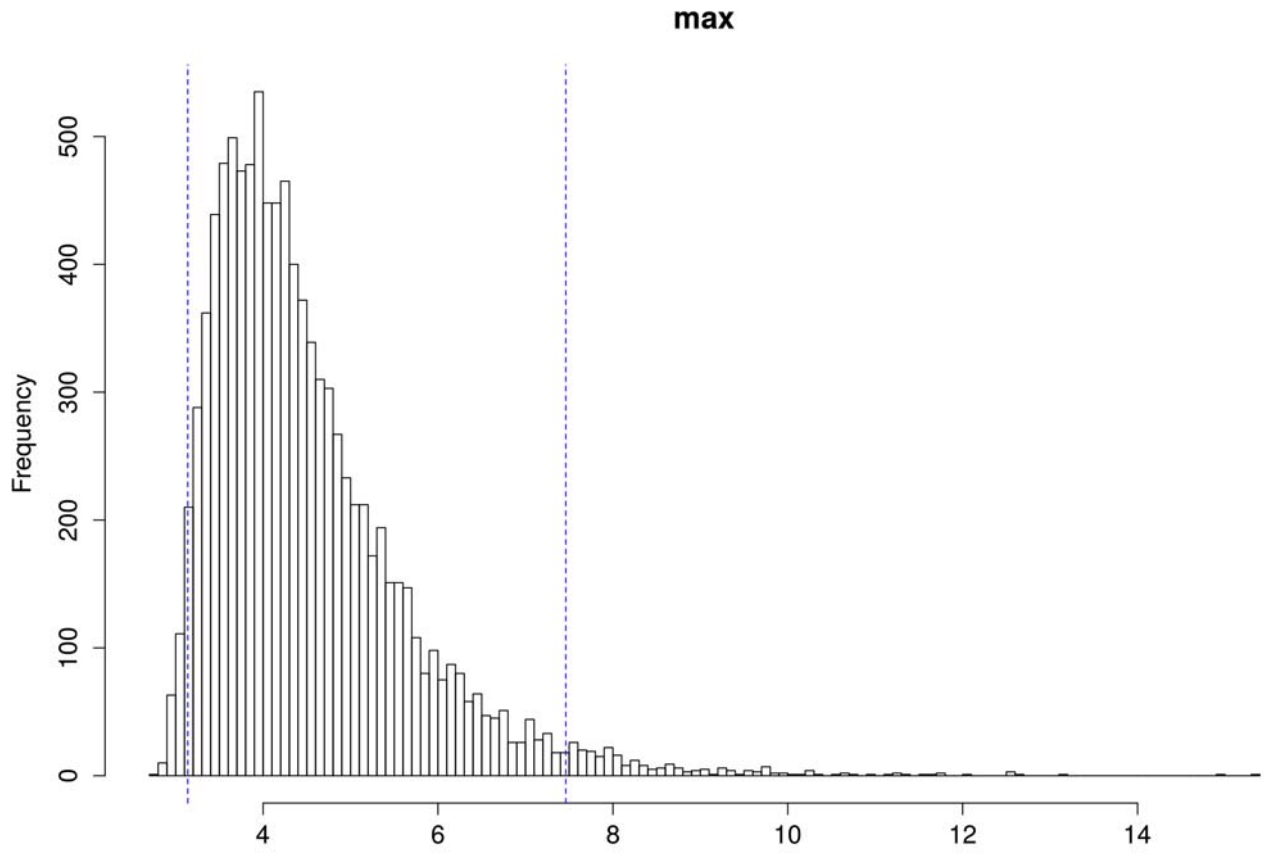

Figure: Quantile à 99/100

	Quantile théorique	IC Oracle	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 $[2.23, 3.11]$
pseudo max	10	6.57 $[-44.2, 57.33]$	6.57 $[5.17, 7.97]$	6.65 $[4.1, 10.79]$

Table: Quantiles estimés et intervalles de confiance à 95% par les différentes méthodes (n = 100)

Conclusion

Résultats principaux

- Médiocrité du bootstrap naïf
- Améliorations par lissage pour les quantiles "peu élevés"
- Utilisation la forme paramétrique de la loi très pertinente
- Estimateur de Hill permet de s'affranchir des problèmes de spécification

Quantiles élevés d'une loi à queue lourde $\Rightarrow$ l'estimateur de Hill est plus fiable que les approches non paramétriques

Merci de votre attention

Bootstrap naïf sur le quantile estimé par MV

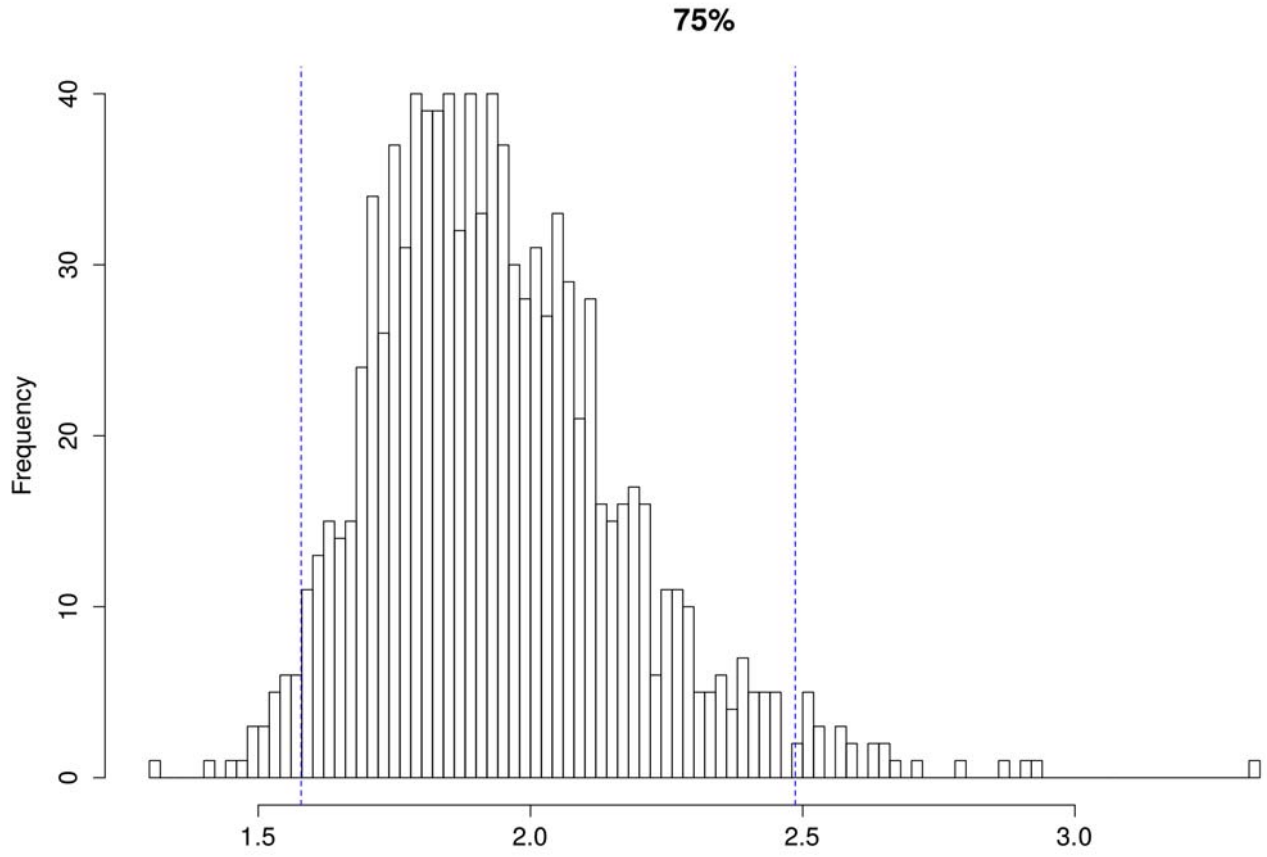

Figure: Quantile à 75%

Bootstrap naïf sur le quantile estimé par MV

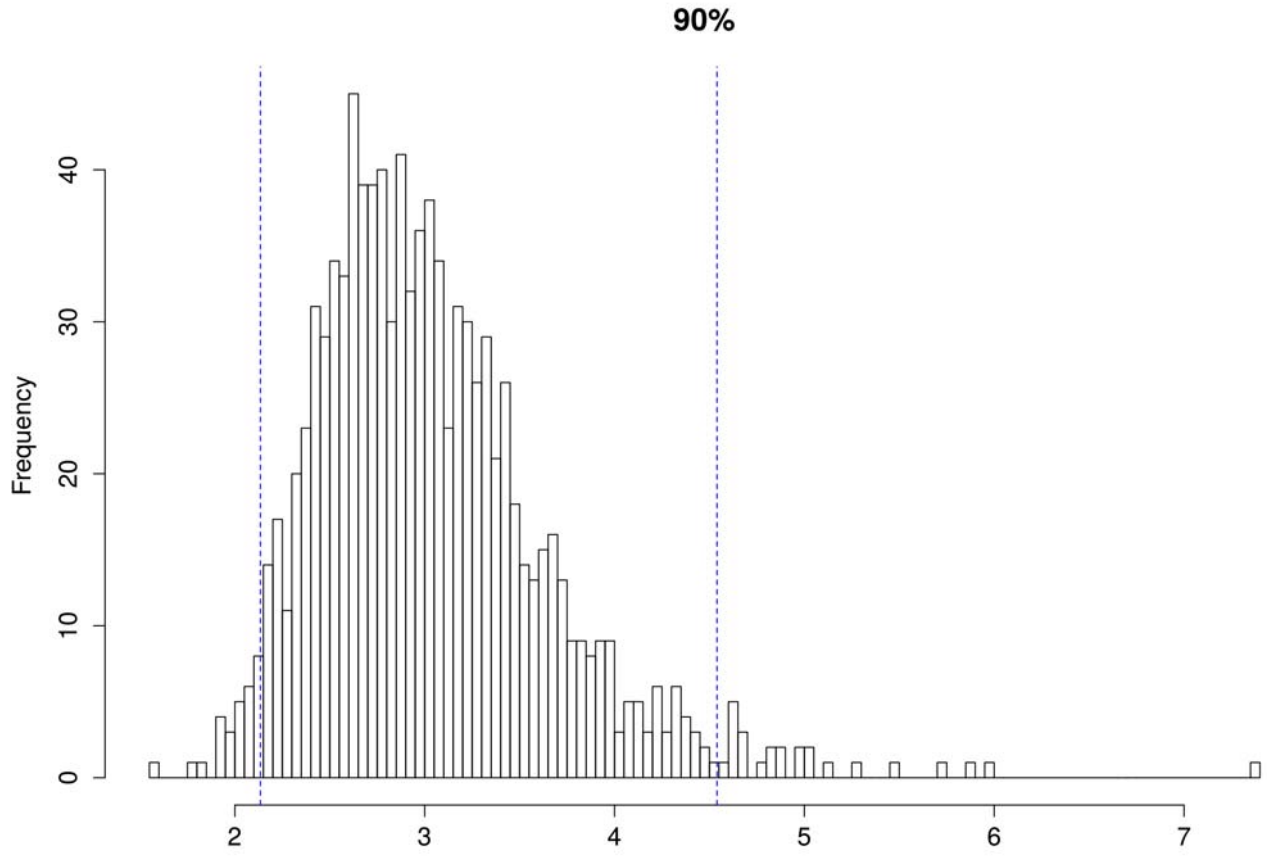

Figure: Quantile à 90%

Bootstrap naïf sur le quantile estimé par MV

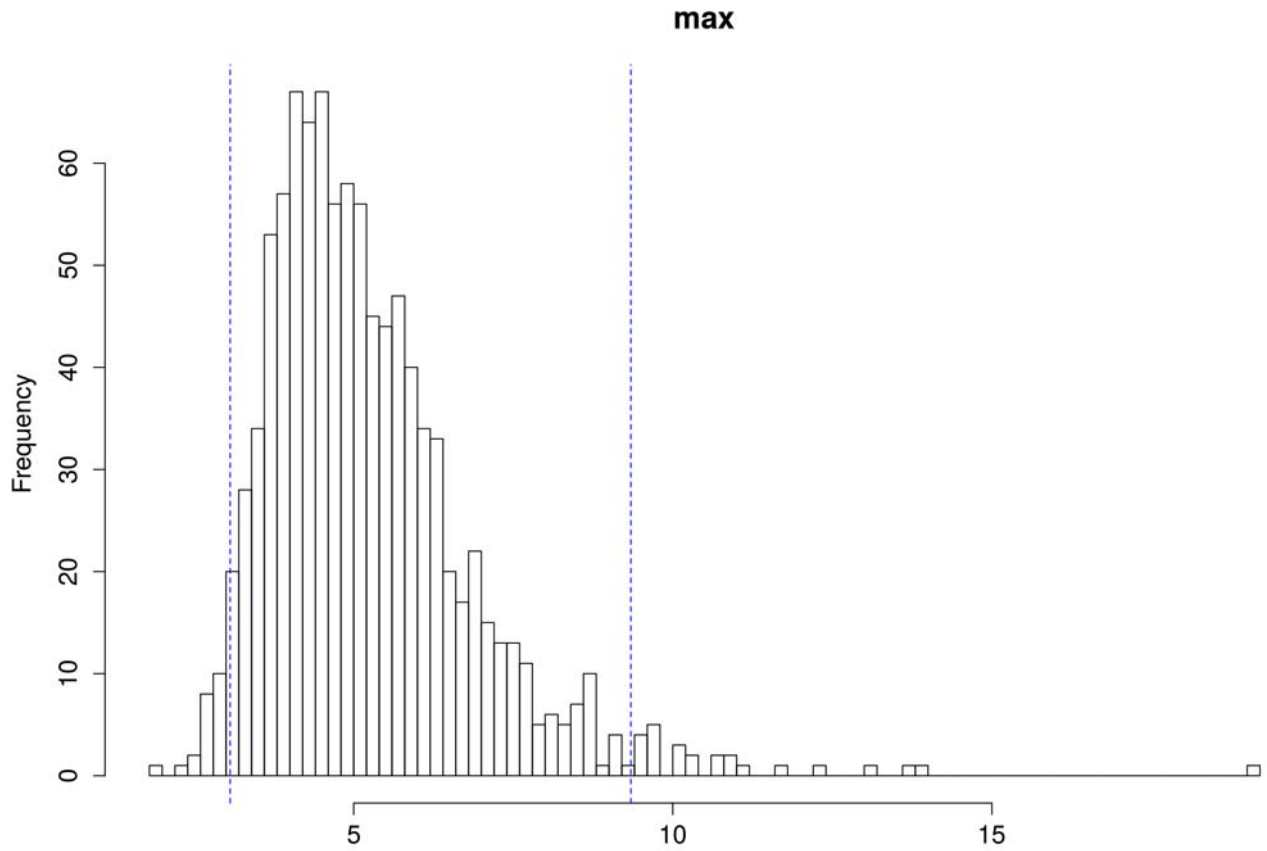

Figure: Quantile à 99/100

Bootstrap naïf sur le quantile estimé par MV

	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: Quantiles estimés et intervalles de confiance à 95% par les différentes méthodes (n = 100)