Étude des valeurs extrêmes univariées

El Mazzouji Wahel, Mariac Damien, Condamy Fabian

9 mai 2025

Table des matières

1 Introduction

Les événements extrêmes, comme les crues majeures, les canicules intenses ou les krachs financiers, sont rares mais peuvent avoir des conséquences très importantes. Les étudier est devenu essentiel dans des domaines variés comme la climatologie, l'assurance, la finance ou encore l'ingénierie.

La théorie des valeurs extrêmes est un outil statistique qui permet de modéliser ces phénomènes rares, en se concentrant sur les valeurs situées aux extrémités d'un jeu de données : les plus grandes ou les plus petites. Contrairement aux méthodes classiques qui s'intéressent surtout à la moyenne ou à la variance, la théorie des valeurs extrêmes cherche à comprendre le comportement des valeurs extrêmes.

Dans les statistiques classiques, beaucoup de résultats reposent sur le théorème central limite, qui explique que, sous certaines conditions, la moyenne d'un grand nombre de variables aléatoires suit une loi normale. C'est un fondement de nombreuses méthodes, mais il ne dit rien sur les valeurs les plus extrêmes d'un échantillon, qui sont pourtant cruciales dans l'analyse du risque.

La théorie des valeurs extrêmes s'intéresse donc à des grandeurs comme le maximum d'un échantillon $(X_1, X_2, \ldots, X_n)$:

$$M_n = \max\{X_1, X_2, \ldots, X_n\}.$$

On cherche à savoir si, en normalisant ce maximum, il converge vers une loi limite. Les travaux de Fréchet, Fisher, Tippett et Gnedenko ont montré que cette convergence est possible, et qu'il n'existe que trois lois limites possibles : la loi de Fréchet, la loi de Gumbel et la loi de Weibull, qui correspondent à différents types de comportements des queues de distribution. En s'intéressant aux comportements les plus rares, la théorie des valeurs extrêmes fournit un cadre rigoureux pour mieux anticiper l'intensité et la fréquence des événements extrêmes, et ainsi mieux s'y préparer.

2 Les lois de M_n

2.1 Quelques notations

On commence par faire une remarque sur la fonction de repartion de M_n en utilisant le fait que les X_i sont i.i.d.
En effet, si on note F_{M_n} la fonction de repartition de M_n, et F_{X_i} la fonction de repartition de X_i on a :

$$\forall t \in \mathbb{R} \quad F_{M_n}(t) = \mathbb{P}(M_n \leq t) = \mathbb{P}(X_1 \leq t, ..., X_n \leq t) = \mathbb{P}(X_1 \leq t)^n = F_{X_1}^n(t)$$

Dans la suite, on notera $F(t)$, la fonction de repartition des X_i.

On remarque que, pour tout t strictement inférieur à la borne supérieure du support des X_i, on a $F(t) < 1$ et donc

$$F(t)^n \xrightarrow[n\to+\infty]{} 0$$

et dans le cas ou t est égal à la borne supérieure du support des X_i, on a

$$F(t)^n \xrightarrow[n\to+\infty]{} 1$$

L'idée est donc d'introduire deux suites (b_n) et (a_n) (avec $a_n > 0$ pour tout n) afin de pouvoir contrôler M_n et avoir une limite non dégénérée.

Puis étudier la loi de la limite de $\frac{M_n - b_n}{a_n}$. Comme la fonction de repartition caractérise la loi, il nous suffit d'étudier la fonction G définie pour tout t dans le support des X_i comme :

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq t\right) \xrightarrow[n\to+\infty]{} G(t)$$

Si il existe de telle suites a_n et b_n, on dit que F est dans le domaine d'attraction de G.

Il nous faut donc trouver les distributions G qui peuvent apparaître comme limite dans l'équation ci-dessus.

Pour ce faire, nous allons utiliser le théoreme suivant :

Théorème 2.1. ***(méthode de la fonction muette)*** *: Soit Y_n une variable aléatoire de fonction de répartition F_n, et soit Y une variable aléatoire de fonction de répartition F. Alors $Y_n \xrightarrow{\mathcal{L}} Y$ si et seulement si pour toute fonction z réelle, bornée et continue :*

$$\mathbb{E}[z(Y_n)] \to \mathbb{E}[z(Y)].$$

En prenant ici $Y_n = \frac{M_n - b_n}{a_n}$, on obtient :

$$\mathbb{E}[z(\frac{M_n - b_n}{a_n})] = \int_{-\infty}^{\infty} z(\frac{x - b_n}{a_n})\, n\, F^{n-1}(x) dF(x)$$

L'astuce ici va être de faire un changement de variable. On introduit alors la fonction quantile que l'on définit ci-dessous :

Définition 2.2. *La fonction quantile Q associée à une fonction de répartition F est définie par :*

$$Q(p) \;=\; F^{-1}(p) \;=\; \inf\big\{x \in \mathbb{R} \;\big|\; F(x) \geq p\big\}, \quad p \in (0,1).$$

On pose alors comme changement de variable :

$$x = Q(1 - \frac{1}{y}) = U(y) \qquad \text{avec } Q \text{ la fonction quantile}$$

$$\text{Donc,} \quad \int_{-\infty}^{\infty} z\Big(\frac{x - b_n}{a_n}\Big)\, n\, F^{n-1}(x)\, dF(x) = \int_0^n z\Big(\frac{U\left(\frac{n}{v}\right) - b_n}{a_n}\Big)\Big(1 - \frac{v}{n}\Big)^{n-1} dv. \tag{1}$$

Or, on a $\lim_{n\to\infty}(1 - \frac{v}{n})^{n-1} = e^{-v}$, et on a $\lim_{n\to\infty} \mathbf{1}_{[0;n]}(x) = \mathbb{R}_+$.

Remarquons que, pour tout $t \in \mathbb{R}$, la probabilité

$$\mathbb{P}\Big(\frac{M_n - b_n}{a_n} \leq t\Big) \;=\; \mathbb{P}\big(M_n \leq a_n\, t + b_n\big) \;=\; F\big(a_n\, t + b_n\big)^n.$$

Ainsi, la convergence en loi

$$\frac{M_n - b_n}{a_n} \;\xrightarrow{L}\; Y \quad \Longleftrightarrow \quad \mathbb{P}\big(M_n \leq a_n\, t + b_n\big) \;\longrightarrow\; G(t),$$

se traduit exactement par

$$F\big(a_n t + b_n\big)^n \;\xrightarrow[n\to\infty]{}\; G(t).$$

On en déduit explicitement une forme pour b_n.

2.2 Paramètre b_n

En reprenant l'expression ci-dessus et en passant au logarithme, on a :

$$n\, log(F(a_n t + b_n)) \xrightarrow[n\to+\infty]{} log(G(t))$$
$$\Longleftrightarrow n(-F(a_n t + b_n) + 1) \xrightarrow[n\to+\infty]{} ln(G(t))$$

On a utilisé un développement limité de la fonction logarithme.
En particulier, pour déterminer b_n, on choisit $t = 0$.
Cela donne :

$$n\big(F(b_n) - 1\big) \;\xrightarrow[n\to\infty]{}\; \ln G(0) \;=\; -\ln G(0) \quad (\text{puisque } G(0) \in (0,1)),$$

Comme on souhaite une limite non-dégénérée, on impose

$$n\big[1 - F(b_n)\big] \;=\; 1.$$

Il vient alors

$$F(b_n) = 1 - \frac{1}{n} \quad \Longleftrightarrow \quad b_n \;=\; Q\Big(1 - \tfrac{1}{n}\Big) \;=\; U(n),$$

2.3 Paramètre a_n

Avec le parametre b_n définie au dessus et en posant $u = \frac{1}{v}$ on obtient alors une condition, il faut qu'il existe une fonction a tel que $\lim_{x\to\infty} \frac{U(xu)-U(x)}{a(x)}$ converge **vers une fonction** $h(u)$.

Proposition 2.3. *Les limites possibles sont données par :*

$$c\,h_\gamma(u) \;=\; c\int_1^u v^{-\gamma-1}\,dv \;=\; c\,\frac{u^\gamma-1}{\gamma}. \tag{2}$$

Nous interprétons $h_0(u) = \log(u)$ *lorsque* $\gamma = 0$.

Remarque : On ne veut pas que $c = 0$, car il conduit à une limite dégénérée pour $\frac{M_n-b_n}{a_n}$. Ensuite, le cas $c > 0$ peut être ramené au cas $c = 1$ en incorporant c dans la fonction a.

Démonstration. Soient $u, v > 0$. Alors :

$$\frac{U(xuv)-U(x)}{a(x)} \;=\; \frac{U(xuv)-U(xu)}{a(xu)}\,\frac{a(xu)}{a(x)} \;+\; \frac{U(xu)-U(x)}{a(x)}. \tag{3}$$

Si la limite dans F est dans le domaine d'attraction de G (*ce qu'on suppose depuis le début*), alors le rapport $\frac{a(ux)}{a(x)}$ converge vers $g(u)$.

De plus,

$$\frac{a(xuv)}{a(x)} \;=\; \frac{a(xuv)}{a(xv)}\,\frac{a(xv)}{a(x)}.$$

Par passage à la limite pour x, la fonction g satisfait l'*équation fonctionnelle de Cauchy* :

$$g(uv) = g(u)\,g(v).$$

Les solutions de cette équation sont de la forme $g(u) = u^\gamma$ avec γ un réel.
Donc, on a $\lim_{x\to\infty} \frac{a(ux)}{a(x)} \;=\; x^\gamma l(x)$, on dit dans ce cas que a est une fonction à variation régulière.

En réécrivant l'expression (2.3) avec cette convergence, on en déduit que la fonction limite est de la forme

$$h_\gamma(u) = c\,\frac{u^\gamma-1}{\gamma},$$

avec la convention $h_0(u) = \ln u$.
Ainsi, nous concluons que

$$h_\gamma(u) = \frac{u^\gamma-1}{\gamma} \quad (\text{avec } h_0(u) = \ln u),$$

□

2.4 Les lois limites

En reprenant (3) et en utilisant ce qui précède, on obtient :

$$\lim_{x\to\infty} \frac{U(xuv)-U(x)}{a(x)} = u^\gamma h(v) + h(u)$$

$$\text{autrement dit :} \quad h_\gamma(uv) = u^\gamma h_\gamma(v) + h_\gamma(u)$$

On fait alors une disjonction de cas sur la valeur de γ.

2.4.1 Nature du support

En reprenant le résultat de la proposition 2.3 et ce qui précède (2), on obtient :

$$h_\gamma\Big(\frac{1}{v}\Big) = \frac{(1/v)^\gamma - 1}{\gamma} = \frac{v^{-\gamma} - 1}{\gamma}$$

Posons $u = \frac{v^{-\gamma}-1}{\gamma}$. On résout alors pour v :

$$v^{-\gamma} = 1 + \gamma u \quad \Longrightarrow \quad v = (1+\gamma u)^{-1/\gamma}$$

Le changement de variable de v à u permet de réécrire l'intégrale limite sous la forme

$$\int_{u \in S_\gamma} z(u)\, d\Big\{\exp\Big[-(1+\gamma u)^{-1/\gamma}\Big]\Big\}$$

ce qui conduit à identifier la loi limite par

$$G_\gamma(u) = \exp\Big\{-(1+\gamma u)^{-1/\gamma}\Big\}$$

Il reste alors à étudier la nature du support S_γ, mais celui-ci dépend du signe de γ :

Cas si $\gamma > 0$:

L'inversion montre que $v \in [0,1]$ correspond à $u > -\frac{1}{\gamma}$.

De plus, pour de grandes valeurs x on a :

$$S_\gamma \approx \exp\Big[-\big(1+\gamma x\big)^{-1/\gamma}\Big]$$

Or, par un développement asymptotique, $\big(1+\gamma x\big)^{-1/\gamma}$ est proportionnel à $x^{-1/\gamma}$ pour x grand. On obtient alors

$$S_\gamma \approx \exp\big[-C\, x^{-1/\gamma}\big] \quad \text{(pour une constante } C > 0).$$

Par croissance comparé, comme $x^{-1/\gamma}$ tend vers 0 moins vite que $\exp(-\alpha x)$. On a alors :

$$S_\gamma \;\sim\; K\, x^{-1/\gamma} \quad (\text{pour } x \to \infty),$$

ce qui caractérise une **queue lourde** : la probabilité d'observer des valeurs très grandes est plus élevée que dans un modèle à décroissance exponentielle.

Cas si $\gamma < 0$:

Pour $\gamma < 0$, la loi est définie si $1 + \gamma u > 0$ c'est à dire $u < -\frac{1}{\gamma}$. Cela signifie que la distribution a son support dans $]-\infty, -\frac{1}{\gamma}[$, et on pose alors $x_{\max} = -\frac{1}{\gamma}$.
Par conséquent, la fonction de survie $S_\gamma = 1 - G(x) = 0$ pour $x \geq -\frac{1}{\gamma}$.

Autrement dit, il n'y a aucune probabilité d'observer une valeur au-delà de $x_{\max}$. Dans ce cas, on dit que la distribution est à **queue bornée**.

Cas si $\gamma = 0$:

Lorsque $\gamma = 0$, on a posé $h_0(u) = \ln u$.
Donc, le changement de variable s'adapte :

$$u = h_0\Big(\frac{1}{v}\Big) = \ln\Big(\frac{1}{v}\Big) = -\ln v,$$

ce qui implique

$$v = e^{-u}.$$

Le changement de variable transforme alors l'intégrale limite en

$$\int_{-\infty}^{\infty} z(u)\, d\Big\{\exp\Big[-e^{-u}\Big]\Big\},$$

et la loi limite est alors donnée par

$$G_0(u) = \exp\Big\{-e^{-u}\Big\}, \qquad u \in \mathbb{R},$$

On retrouve ici une queue à décroissance exponentielle, ce qui est caractéristique d'une **queue légère** : la probabilité d'observer des valeurs extrêmes est faible mais pas impropable. Il s'agit d'un cas intermédiaire entre les deux cas précédents.

2.5 Résumé

Les lois limites qui s'imposent dependent d'un parametre γ et sont les suivantes :

— Si $\gamma > 0$ (loi de Fréchet) :

$$G_\gamma(u) = \exp\left\{-\left(1+\gamma u\right)^{-1/\gamma}\right\}, \quad u > -\frac{1}{\gamma}.$$

— Si $\gamma = 0$ (loi de Gumbel) :

$$G_0(u) = \exp\left\{-e^{-u}\right\}, \quad u \in \mathbb{R}.$$

— Si $\gamma < 0$ (loi de Weibull) :

$$G_\gamma(u) = \exp\left\{-\left(1+\gamma u\right)^{-1/\gamma}\right\}, \quad u < -\frac{1}{\gamma}.$$

La loi se généralise pour toute valeur de gamma et on l'appelle GEV (Generalized Extreme Value), et donne :

$$G_\gamma(x) \;=\; \exp\Big\{-\big[1+\gamma\, u\big]^{-1/\gamma}\Big\}.$$

3 Quelques exemples numériques

Voici maintenant quelques applications numériques sur des lois usuelles de ce que nous avons vu dans cette section. Pour chacune des représentations suivantes, nous avons simulé $n = 1000$ fois chaque loi puis ensuite effectué $N = 10000$ simulations pour le maximum afin d'avoir une précision correcte.

3.0.1 Loi uniforme

Pour la loi uniforme sur [0,1], on peut montrer théoriquement que la limite du max est une loi exponentielle de paramètre 1 (loi de Weibull bien particulière).
Soient $U_1, U_2, \ldots, U_n$ des variables aléatoires indépendantes et identiquement distribuées selon la loi uniforme sur $[0, 1]$.
On a, pour $x \in [0, 1]$:

$$\begin{aligned} P(M_n \leq x) &= P(U_1 \leq x, \ldots, U_n \leq x) \\ &= P(U_1 \leq x)^n \text{ par indépendance des } U_i \\ &= x^n \end{aligned}$$

Nous allons maintenant effectuer le changement de variable $x = 1 - y/n$ avec $y > 0$ pour examiner la queue de la distribution :

$$P(M_n \leq 1 - y/n) = (1 - y/n)^n.$$

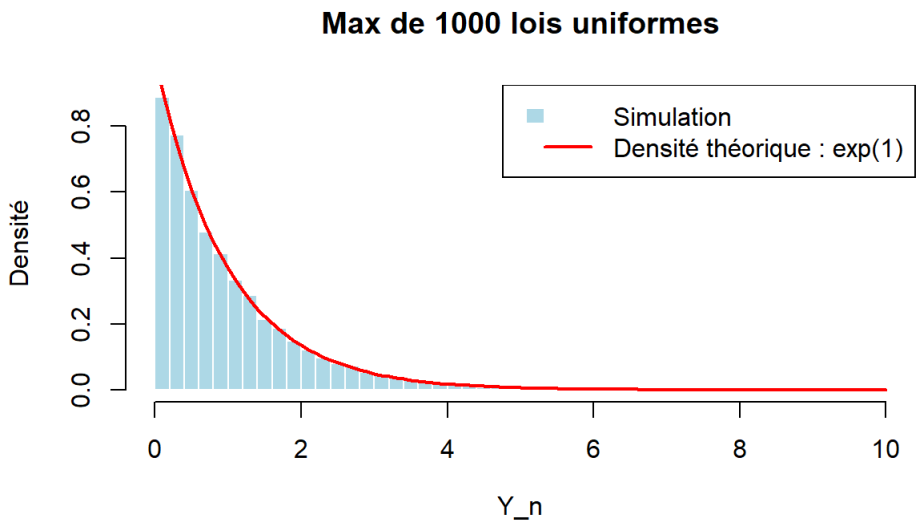
De plus, on a : $(1 - y/n)^n \xrightarrow[n\to+\infty]{} e^{-y}$. Donc, $P(M_n \leq 1 - y/n) \xrightarrow[n\to+\infty]{} e^{-y}$.
Or, par définition, la loi exponentielle de paramètre 1 a pour fonction de répartition : $P(Y \leq y) = 1 - e^{-y}, \quad y > 0$.
Ainsi, on a donc montré que :

$$P(n(1 - M_n) \leq y) \to P(Y \leq y) = 1 - e^{-y},$$

ce qui établit la convergence en loi :

$$Y_n = n(1 - M_n) \xrightarrow{\mathcal{L}} \mathcal{E}(1).$$

Ainsi, on prendra ici $a_n = \frac{1}{n}$ et $b_n = 1$.
Une simulation sur Rstudio donne le graphe suivant :



Remarquons que l'on obtient une loi de Gumbel, ce qui est assez logique au vu du fait que ce soit une loi à queue très légère (elle n'en a tout simplement pas car son support est borné).

3.0.2 Loi exponentielle

Pour une loi exponentielle de paramètre 1, la loi limite est une loi de Gumbel. Théoriquement, on trouve $a_n = 1$ et $b_n = \log(n)$.
En effet, soient $X_1, \dots, X_n$ des variables aléatoires iid de loi $\mathcal{E}(1)$. On a pour $x \geq 0$:

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, \dots, X_n \leq x) \\ &= P(X_1 \leq x)^n \\ &= (1 - \exp(-x))^n \end{aligned}$$

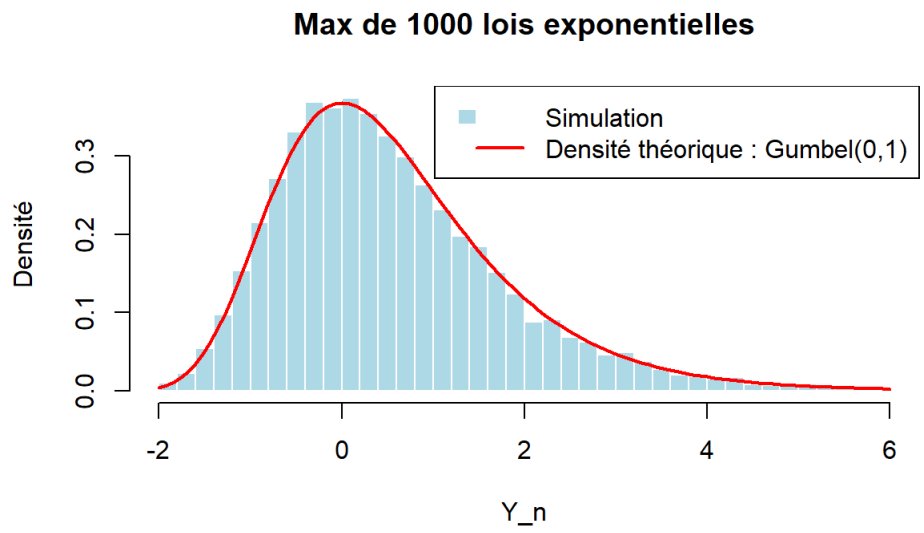
Passons donc à la normalisation, posons $z = \frac{x - b_n}{a_n}$, on va utiliser le fait que $(1-\exp(-x))^n \xrightarrow[x \to +\infty]{} \exp(-n\exp(-x))$.
Ainsi on veut choisir b_n de telle sorte que $n\exp(-b_n) \xrightarrow[n \to +\infty]{} 1$. D'où $b_n = \log(n)$
On a donc $x = a_n z + \log(n)$. On peut donc choisir $a_n = 1$ afin d'avoir :

$$\begin{aligned} P(\frac{M_n - \log(n)}{1} \leq z) &= (1 - \exp(-(z + \log(n))))^n \\ &= (1 - \frac{\exp(-z)}{n})^n \end{aligned}$$

Et donc : $(1 - \frac{\exp(-z)}{n})^n \xrightarrow[n \to +\infty]{} \exp(\exp(-z))$

Ainsi, on a bien montré que $\frac{M_n - \log(n)}{1} \xrightarrow{\mathcal{L}} \text{Gumbel}(0, 1)$

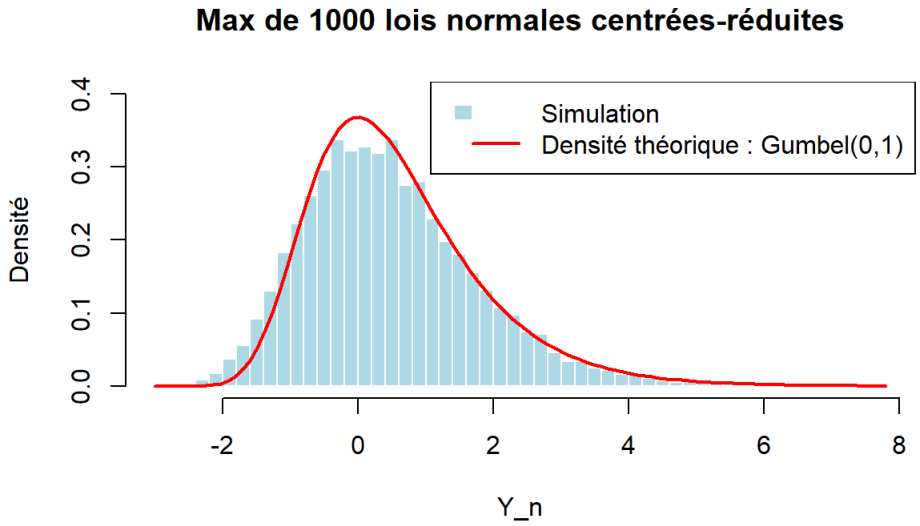
On obtient alors numériquement le graphe suivant :



Cette fois-ci, on avait une loi à queue fine, et on obtient loi de Gumbel, ce qui était attendu.

3.0.3 Loi normale

Pour maintenant une loi normale centrée-réduite, on peut montrer que la loi limite est encore une fois une loi de Gumbel. Théoriquement, on trouve les paramètres généralisés suivants : $a_n = \frac{\log\left(\frac{4\log^2(2)}{\log^2\left(\frac{4}{3}\right)}\right)}{2\sqrt{2\log(n)}}$ et $b_n = \sqrt{2\log(n)} - \frac{\log(\log(n)) + \log\left(4\pi\log^2(2)\right)}{2\sqrt{2\log(n)}}$. (Gnedenko, *On The Limiting Distribution of the Maximum Term in a Random Series*).
On obtient cette fois-ci le graphe suivant :



Notons ainsi que l'on a la même loi limite que pour la loi exponentielle de paramètre 1, les graphes sont quasiment identiques.

3.0.4 Loi de Cauchy

Enfin, pour une loi de Cauchy (de paramètres 0 et 1 ici), la loi limite est une loi de Fréchet. On a les coefficients suivants : $a_n = \pi$ et $b_n = n$.

On s'intéresse au comportement asymptotique du maximum $M_n = \max(X_1, \dots, X_n)$. Pour la loi de Cauchy, on sait que la fonction de répartition est donnée par $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$. Donc, la fonction de survie est $\bar{F}(x) = 1 - F(x) = \frac{1}{2} - \frac{1}{\pi}\arctan(x)$.

On va utiliser le développement asymptotique suivant, pour $x \to +\infty$: $\arctan(x) = \frac{\pi}{2} - \frac{1}{x} + o\left(\frac{1}{x}\right)$.

Ainsi, on a $\bar{F}(x) \sim \frac{1}{\pi x}$ lorsque $x \to +\infty$.
Et donc la loi du maximum M_n s'écrit alors : $\mathbb{P}(M_n \leq x) = F(x)^n \approx \left(1 - \frac{1}{\pi x}\right)^n$
On cherche donc $a_n > 0$, b_n tels que : $\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) = \exp(-1/z)$ où G est la fonction de répartition d'une loi de Fréchet de paramètre 1.

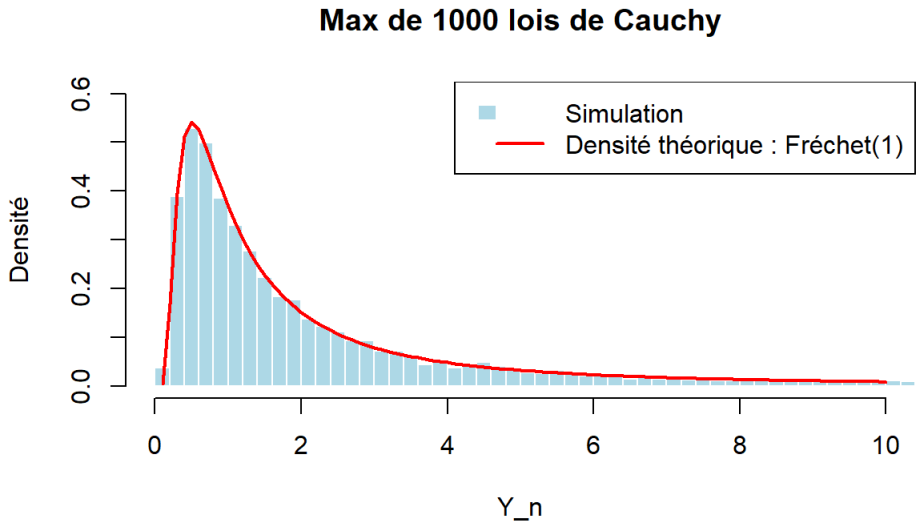
Posons $b_n = n$, $a_n = \pi$. Alors :

$$\mathbb{P}\left(\frac{M_n - n}{\pi} \leq z\right) = \mathbb{P}\left(M_n \leq \pi z + n\right) \approx \left(1 - \frac{1}{\pi(\pi z + n)}\right)^n$$

En développant, on obtient : $\left(1 - \frac{1}{\pi(\pi z+n)}\right)^n \approx \exp\left(-\frac{n}{\pi(\pi z+n)}\right)$
Et lorsque $n \to \infty$, on a $\frac{n}{\pi(\pi z+n)} \to \frac{1}{z}$
Donc : $\mathbb{P}\left(\frac{M_n - n}{\pi} \leq z\right) \to \exp\left(-\frac{1}{z}\right)$

Ce qui montre que :

$$\frac{M_n - n}{\pi} \xrightarrow{d} Z \sim \text{Fréchet}(1)$$



Enfin ici, on avait une loi à queue lourde, et on obtient bien la loi de Fréchet attendue.

4 Méthodes d'estimation de l'indice de valeurs extrêmes

Dans cette section, nous cherchons à estimer le paramètre γ directement à partir des observations X_i, sans nous limiter aux maxima M_n, afin de tirer parti de l'ensemble des données. Nous nous concentrons sur deux approches non paramétriques classiques : les estimateurs de Hill et de Pickands. Ces méthodes reposent uniquement sur les statistiques d'ordre et sont particulièrement utilisées en pratique. D'autres estimateurs existent, notamment des versions généralisées ou des approches paramétriques basées sur la vraisemblance ou les moments, mais ils ne seront pas abordés ici.

Définition : On appelle *statistique d'ordre* la permutation aléatoire de l'échantillon $X_1, \ldots, X_n$, qui ordonne les valeurs de l'échantillon par ordre croissant :

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

Définition : On dit qu'une suite $(k_n)_{n\geq 0}$ d'entiers est intermédiaire si :

$$\lim_{n\to\infty} k_n = \infty \quad et \lim_{n\to\infty} \frac{k_n}{n} = 0$$

Définition : On dit qu'un estimateur $\hat{\gamma_n}$ est convergent s'il converge en probabilité vers γ, soit :

$$\lim_{n\to\infty} P(|\hat{\gamma_n} - \gamma| > \epsilon) = 0 \quad \forall \epsilon > 0$$

4.1 Estimateur de Pickands

L'estimateur de Pickands est construit à partir de trois statistiques d'ordre dans un échantillon. Il constitue l'un des premiers estimateurs non paramétriques proposés pour estimer l'indice des valeurs extrêmes γ. Son principal avantage réside dans le fait qu'il est valide quel que soit le domaine d'attraction de la loi sous-jacente : Fréchet ($\xi > 0$), Gumbel ($\xi = 0$) ou Weibull ($\xi < 0$). Il n'est donc pas restreint à une famille particulière de distributions et reste applicable dans un cadre très général.

Néanmoins, cet estimateur est connu pour être assez sensible à la taille de l'échantillon, et en particulier au choix du paramètre intermédiaire k, ce qui peut entraîner une certaine instabilité dans les estimations. Cela limite parfois sa robustesse, en particulier pour des tailles d'échantillon modestes.

En 1975, Pickands a démontré la consistance faible de son estimateur, c'est-à-dire la convergence en probabilité vers le vrai paramètre lorsque la taille de l'échantillon tend vers l'infini. Plus tard en 1989, Dekkers et de Haan ont établi la convergence forte ainsi que la normalité asymptotique de cet estimateur sous des conditions plus générales.

Définition. Soit $X_1, \ldots, X_n$ une suite de variables aléatoires i.i.d. de loi F, appartenant à l'un des domaines d'attraction des lois de valeurs extrêmes. On note $X_{1,n} \leq \cdots \leq X_{n,n}$ les statistiques d'ordre croissantes. Soit $(k_n)_{n\geq 1}$ une suite intermédiaire telle que $k_n \to \infty$ et $k_n/n \to 0$, l'estimateur de Pickands est défini par :

$$\hat{\gamma}_{k_n,n} = \frac{1}{\ln(2)} \ln\left(\frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-2k_n+1,n} - X_{n-4k_n+1,n}}\right)$$

L'estimateur de Pickands repose sur l'idée que, dans les queues d'une distribution extrême, les plus grandes observations suivent un comportement régulier. En considérant des statistiques d'ordre décroissantes, on peut approximer la structure de la queue à l'aide de différences successives entre grandes valeurs. L'utilisation d'une transformation logarithmique permet alors d'isoler l'indice de queue γ, sous des conditions d'attraction à une loi limite.

Propriété de consistance. Si (k_n) est une suite intermédiaire, alors :

$$\hat{\gamma}_{k_n,n} \xrightarrow{\mathbb{P}} \gamma \quad \text{lorsque } n \to \infty.$$

De plus, sous hypothèses régulières, l'estimateur est asymptotiquement normal :

$$\sqrt{k_n}\,(\hat{\gamma}_{k_n,n} - \gamma) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\gamma))$$

où la variance asymptotique est donnée par :

$$\sigma(\gamma) = \frac{\gamma\sqrt{2^{2\gamma+1}+1}}{2(2^\gamma - 1)\ln(2)}.$$

Cette formule théorique permet de construire des intervalles de confiance pour l'estimation de γ, bien qu'en pratique la variance soit souvent estimée par simulation.

4.1.1 Construction de l'estimateur de Pickands

Afin de construire l'estimateur de Pickands, nous allons reprendre la proposition (2.3). Pour $x, y > 0$ et $y \neq 1$, la proposition est équivalente à :

$$\lim_{s\to 0} \frac{U(sx) - U(s)}{U(sy) - U(s)} = \begin{cases} \frac{x^\gamma - 1}{y^\gamma - 1} & \text{si } \gamma \neq 0, \\ \frac{\log(x)}{\log(y)} & \text{si } \gamma = 0. \end{cases}$$

Lemme 4.1. *Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de fonction de répartition F. Soit $U_1, \ldots, U_n$ des variables aléatoires indépendantes de loi uniforme $[0, 1]$. Alors $F^{-1}(U_{1,n}), \ldots, F^{-1}(U_{n,n})$ a même loi que $(X_{1,n}, \ldots, X_{n,n})$.*

Démonstration. admis □

On en déduit la construction de l'estimateur de Pickands.

Construction de l'estimateur de Pickands :

Par la proposition précédente, pour $\gamma \in \mathbb{R}$ et α on a avec le choix $t = 2s$, $x = 2$ et $y = \frac{1}{2}$,

$$\lim_{t\to\infty} \frac{U(t) - U(t/2)}{U(t/2) - U(t/4)} = 2^\gamma.$$

En utilisant la croissance de U qui se déduit de la croissance de F, on obtient

$$\lim_{t\to\infty} \frac{U(t) - U(t_{c_1}(t))}{U(t_{c_1}(t)) - U(t_{c_2}(t))} = 2^\gamma$$

dès que $\lim_{t\to\infty} c_1(t) = \frac{1}{2}$ et $\lim_{t\to\infty} c_2(t) = \frac{1}{4}$. Il reste donc à trouver des estimateurs pour $U(t)$.

Soit k_n, pour $n \geq 1$ une suite d'entiers telle que $1 \leq k_n \leq \frac{n}{4}$ et $\lim_{n\to\infty} \frac{k_n}{n} = 0$ et $\lim_{n\to\infty} k_n = \infty$.
Soit $(V_{1,n}, \ldots, V_{n,n})$ la statistique d'ordre d'un échantillon de variables aléatoires indépendantes de loi de Pareto.
On note $F_V(x) = 1 - x^{-1}, x \geq 1$.

On déduit avec certains résultats de bases liés à $(V_{1,n}, \ldots, V_{n,n})$ que les suites

$$\frac{k_n}{n} V_{n-k_n+1,n}, \quad \frac{2k_n}{n} V_{n-2k_n+1,n}, \quad \frac{4k_n}{n} V_{n-4k_n+1,n}$$

pour $n \geq 1$ convergent en probabilité vers 1.

On en déduit en particulier, les convergences en probabilité suivantes :

$$V_{n-k_n+1,n} \to \infty, \quad \frac{V_{n-2k_n+1,n}}{V_{n-k_n+1,n}} \to \frac{1}{2}, \quad \frac{V_{n-4k_n+1,n}}{V_{n-k_n+1,n}} \to \frac{1}{4}.$$

Donc la convergence suivante a lieu en probabilité :

$$\frac{U(V_{n-k_n+1,n}) - U(V_{n-2k_n+1,n})}{U(V_{n-2k_n+1,n}) - U(V_{n-4k_n+1,n})} \to 2^\gamma.$$

Remarquons que si $x \geq 1$, alors $U(x) = F^{-1}(F_V(x))$. On a donc

$$(U(V_{1,n}), \ldots, U(V_{n,n})) = (F^{-1}(F_V(V_{1,n})), \ldots, F^{-1}(F_V(V_{n,n}))).$$

Or F_V est la fonction de répartition de la loi de Pareto.
On déduit de la croissance de F_V que $(F^{-1}(F_V(V_{1,n})), \ldots, F^{-1}(F_V(V_{n,n})))$ a la même loi qu'une suite de n variables aléatoires uniformes sur $[0, 1]$ indépendantes.

On déduit du lemme A que le vecteur aléatoire $(F^{-1}(F_V(V_{1,n})), \ldots, F^{-1}(F_V(V_{n,n})))$ a la même loi que $(X_1, \ldots, X_n)$.

Donc la variable aléatoire $\frac{U(V_{n-k_n+1,n}) - U(V_{n-2k_n+1,n})}{U(V_{n-k_n+1,n}) - U(V_{n-4k_n+1,n})}$ a la même loi que :

$$\frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-k_n+1,n} - X_{n-4k_n+1,n}}$$

Ainsi cette quantité converge en loi vers 2^γ quand n tend vers l'infini.

4.1.2 Représentation graphique de l'estimateur de Pickands

Pour mieux comprendre le comportement de l'estimateur de Pickands dans différents contextes, nous l'appliquons à des données simulées qui sont les mêmes que dans la partie 3, c'est à dire : Pareto, Exponentielle, Uniforme et Cauchy. Ces lois sont choisies pour représenter les trois domaines d'attraction des lois de valeurs extrêmes :

Chaque échantillon comporte $n = 100000$ observations, et l'estimation est faite en fonction de k, où k représente le nombre de plus grandes valeurs utilisées pour estimer l'indice de queue γ.

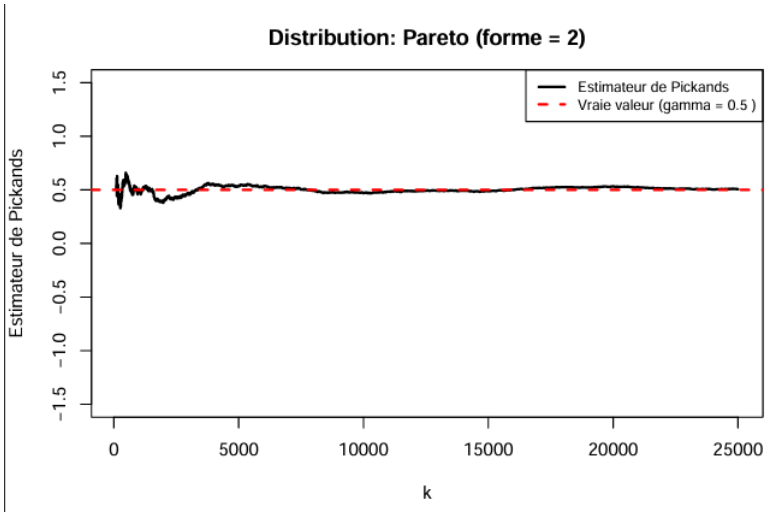


FIGURE 1 – Estimateur de Pickands appliqué à la loi de Pareto ($\gamma = 0.5$)

Tout d'abord, pour la loi de Pareto, nous remarquons que l'estimateur de Pickands converge rapidement et de manière stable vers la valeur théorique $\gamma = 0.5$, dès que k devient raisonnablement grand. Cela confirme la bonne performance de l'estimateur pour des lois à queue lourde. On observe cependant une légère variabilité pour les très petites valeurs de k, comme attendu.

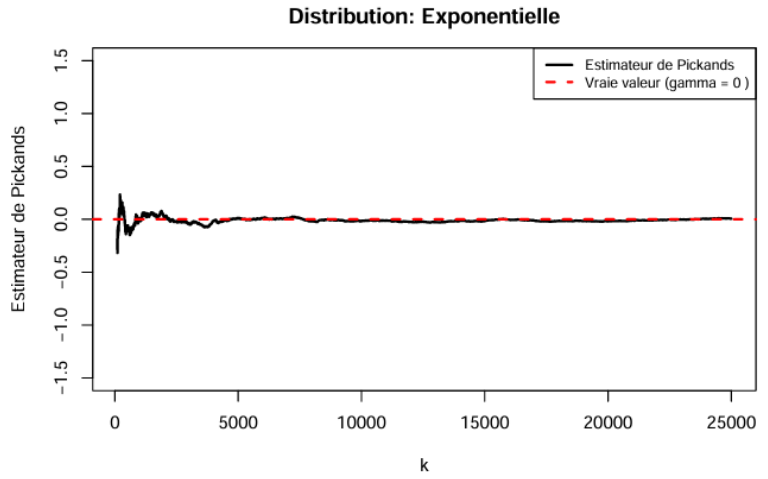


FIGURE 2 – Estimateur de Pickands appliqué à la loi Exponentielle ($\gamma = 0$)

Dans le cas de la loi exponentielle, qui appartient au domaine de Gumbel, l'estimateur se stabilise très rapidement autour de $\gamma = 0$. Le résultat est ainsi conforme aux attentes.

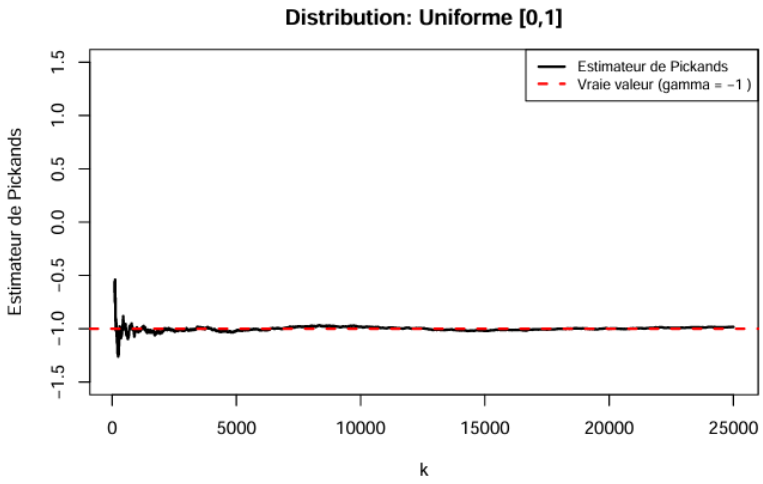


FIGURE 3 – Estimateur de Pickands appliqué à la loi Uniforme ($\gamma = -1$)

Avec la loi uniforme, dont le support est borné, l'estimateur converge vers $\gamma = -1$, ce qui est cohérent avec la théorie. On observe une plus grande variabilité dans les faibles valeurs de k, mais une convergence raisonnablement bonne lorsque k augmente.

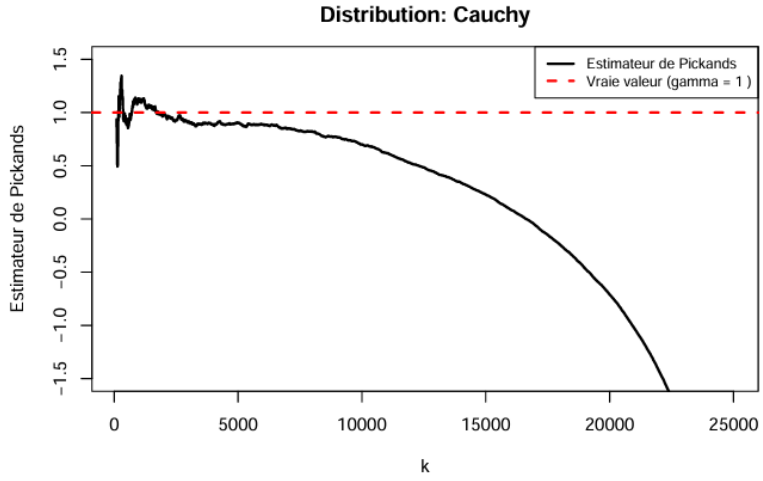


FIGURE 4 – Estimateur de Pickands appliqué à la loi de Cauchy ($\gamma = 1$)

La loi de Cauchy, caractérisée par une queue extrêmement lourde, présente un comportement plus instable. Pour les faibles valeurs de k, l'estimateur est instable. Il converge vers $\gamma = 1$ mais, passé un certain seuil (vers $k = 10000$), il commence à décroître, indiquant une dégradation de l'estimation. Cela met en évidence la sensibilité de l'estimateur de Pickands au choix de k, en particulier pour des lois aux queues extrêmes. Lorsque trop d'observations non-extrêmes sont incluses, l'estimation devient biaisée.

Ainsi, ces résultats illustrent bien les limites de l'estimateur de Pickands : si k est trop petit, on observe une forte variance, tandis que si k est trop grand, l'estimation est biaisée car elle intègre des observations non extrêmes. D'un point de vue théorique, la validité asymptotique de l'estimateur repose sur les conditions suivantes : $k \to \infty$ et $k/n \to 0$ lorsque $n \to \infty$. Cela signifie qu'on doit considérer suffisamment de valeurs extrêmes, tout en veillant à ce que k reste petit devant n. Dans cette optique, nous limitons volontairement k à $n/4$ dans nos simulations, afin de respecter ce cadre théorique et d'éviter que l'estimation ne soit dégradée.

4.2 Estimateur de Hill

Cet estimateur a été introduit par Hill en 1975 dans le but d'estimer, de manière non paramétrique, le paramètre de queue des lois appartenant au domaine d'attraction de Fréchet. Il offre une estimation de l'indice de queue généralement plus efficace que celle fournie par l'estimateur de Pickands. La construction de cet estimateur repose sur l'utilisation des k_n plus grandes statistiques d'ordre de l'échantillon.

4.2.1 La construction de l'estimateur de Hill

Dans le cadre de l'analyse des valeurs extrêmes, l'estimation de l'indice de queue γ est cruciale pour comprendre le comportement des queues de distribution. L'estimateur de Hill est une méthode largement utilisée pour estimer cet indice de queue.

Soient α_n et β_n deux suites de nombres positifs. La construction de l'estimateur de Hill repose sur une relation fondamentale entre les quantiles d'une distribution à queue lourde appartenant au domaine d'attraction de Fréchet. Cette relation est donnée par :

$$q_{\beta_n} \simeq q_{\alpha_n} \left(\frac{\alpha_n}{\beta_n} \right)^{\gamma}.$$

Ici, q_α représente le quantile d'ordre α de la distribution, et γ est l'indice de queue que nous cherchons à estimer. Cette relation exprime que le quantile d'ordre β_n peut être approximé par le quantile d'ordre α_n multiplié par un facteur dépendant de γ.

Pour déterminer l'estimateur de Hill, nous commençons par prendre le logarithme des deux côtés de l'équation du dessus. Nous obtenons :

$$\log(q_{\beta_n}) - \log(q_{\alpha_n}) \simeq \gamma \log \left(\frac{\alpha_n}{\beta_n} \right).$$

En posant $\alpha_n = k_n/n$, où k_n est un nombre d'observations extrêmes que nous considérons, et en prenant plusieurs valeurs pour β_n, avec $\beta_n = i/n$ pour $i = 1, \ldots, k_n - 1$ et $\beta_n < \alpha_n$, nous obtenons :

$$\log(q_{i/n}) - \log(q_{k_n/n}) \simeq \gamma \log(k_n/i).$$

Ensuite, nous estimons les quantiles par leurs équivalents empiriques, c'est-à-dire les statistiques d'ordre de l'échantillon. Cela conduit à :

$$\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) \simeq \gamma \log(k_n/i).$$

En sommant sur $i = 1, \ldots, k_n - 1$, nous obtenons :

$$\gamma = \frac{\sum_{i=1}^{k_n-1} \log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})}{\sum_{i=1}^{k_n-1} \log(k_n/i)}.$$

Le dénominateur peut être réécrit comme $\log(k_n^{k_n-1}/(k_n-1)!)$. En utilisant la formule de Stirling, nous trouvons que ce dénominateur est équivalent à k_n au voisinage de l'infini. Cela nous permet d'obtenir l'estimateur de Hill.

Soit $(k_n)_{n\geq 1}$ une suite d'entiers avec $1 \leq k_n \leq n$, l'estimateur de Hill est défini par :

$$\hat{\gamma}^H_{k_n} = \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} \log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}).$$

L'estimateur de Hill satisfait la propriété de consistance faible. Plus précisément, si $(k_n)_{n\geq 1}$ est une suite intermédiaire, alors l'estimateur $\hat{\gamma}^H_{k_n}$ converge en probabilité vers le paramètre de queue γ, c'est-à-dire :

$$\hat{\gamma}^H_{k_n} \xrightarrow{\mathbb{P}} \gamma.$$

Remarque 4.2. Dans la pratique, déterminer une valeur appropriée pour le paramètre k_n, c'est-à-dire le nombre de plus grandes observations à retenir, constitue une étape délicate. Il faut en effet trouver un compromis entre la variance et le biais : utiliser suffisamment de données pour obtenir une estimation fiable, tout en s'assurant que ces données proviennent bien de la queue de la distribution. Diverses approches ont été développées dans la littérature pour guider ce choix.

4.2.2 Comportement empirique de l'estimateur de Hill

Pour analyser le comportement de l'estimateur de Hill, nous l'appliquons à des échantillons simulés de taille $n = 40000$, issus de plusieurs distributions : la loi de Lévy, la loi exponentielle, la loi uniforme sur $[0, 1]$, et la loi de Cauchy. Ces distributions couvrent les trois domaines d'attraction des lois de valeurs extrêmes. Les figures suivantes présentent l'évolution de l'estimateur de Hill en fonction de k, le nombre de plus grandes valeurs

utilisées. La ligne rouge indique la valeur théorique de γ, tandis que la courbe noire représente l'estimation, accompagnée d'une bande de confiance à 95%.

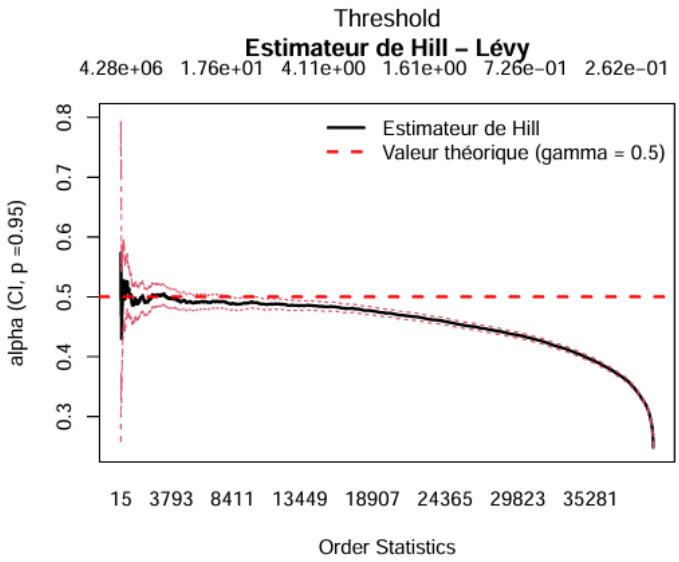


FIGURE 5 – Estimateur de Hill appliqué à la loi de Lévy ($\gamma = 0.5$)

Pour la loi de Lévy, à queue lourde, l'estimateur de Hill converge de manière assez stable vers $\gamma = 0.5$. Au-delà d'un certain seuil de k, on note cependant une dégradation due à l'inclusion d'observations moins extrêmes.

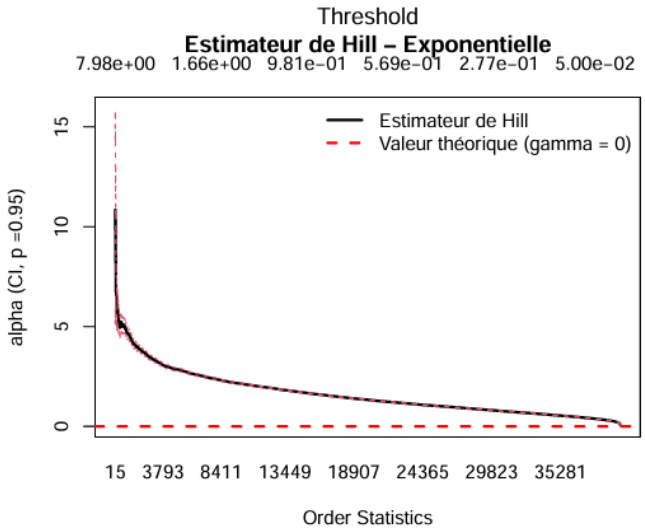


FIGURE 6 – Estimateur de Hill appliqué à la loi exponentielle ($\gamma = 0$)

La loi exponentielle appartient au domaine de Gumbel, avec un indice de queue nul. L'estimateur de Hill n'est pas conçu pour ce cas, ce que montre bien le graphique : les valeurs estimées sont nettement surestimées et décroissent lentement sans atteindre la valeur théorique. Ce comportement traduit l'inadéquation de Hill pour ce type de loi.

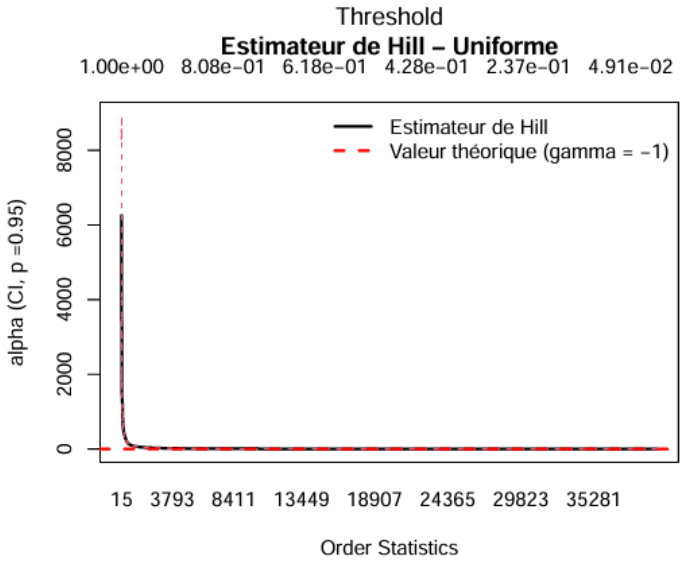


FIGURE 7 – Estimateur de Hill appliqué à la loi uniforme ($\gamma = -1$)

La loi uniforme a un support borné et un indice théorique de $\gamma = -1$, ce qui sort du domaine d'application de Hill, qui suppose $\gamma > 0$. L'estimateur produit ici des résultats très instables et incohérents, confirmant qu'il ne doit pas être utilisé sur ce type de distribution.

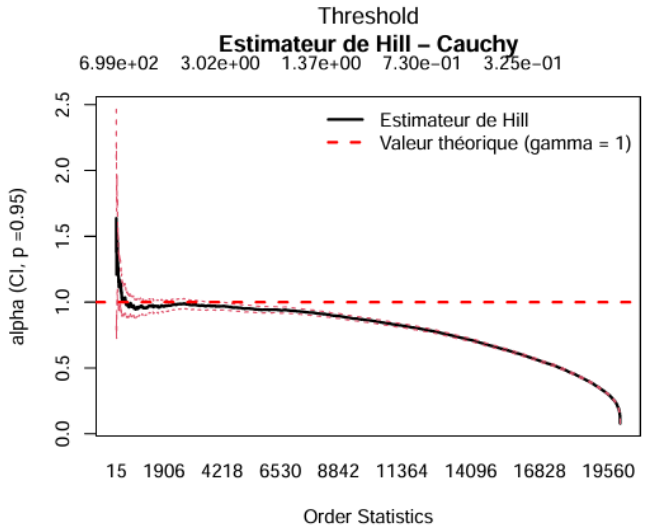


FIGURE 8 – Estimateur de Hill appliqué à la loi de Cauchy ($\gamma = 1$)

La loi de Cauchy, avec un indice $\gamma = 1$, appartient au domaine de Fréchet. L'estimateur fonctionne bien pour de faibles valeurs de k, avec une estimation proche de la valeur théorique. Mais dès que k devient trop grand, la courbe chute, signe d'un biais induit par les valeurs non extrêmes. Cela illustre la sensibilité de l'estimateur au choix du seuil.

Ainsi, ces résultats montrent que l'estimateur de Hill est très dépendant du choix du paramètre k, et de la nature de la distribution. Un compromis est nécessaire entre biais (si k est trop grand) et variance (si k est trop petit), pour garantir une estimation fiable.

5 Méthode des maxima par blocs

L'approche des maxima par blocs (en anglais Blocks Maxima) consiste à diviser les n observations en N blocs de taille k. Concrètement, la suite $X_1, ..., X_n$ est divisée en N blocs, le premier bloc est $X_1, ..., X_k$, le second $X_{k+1}, ..., X_{2k}$, etc. On obtient ainsi une suite de maxima $M_1, ..., M_n$ définis sur chacun des blocs.
En général, on considère une période temporelle, comme une journée ou bien une année pour refléter le sens des observations.
On peut alors déterminer la loi limite des maxima, en vertu du théorème de Fisher-Tippett-Gnedenko c'est une distribution GEV classique de la forme :

$$G_{\mu,\sigma,\gamma}(x) \;=\; \exp\Big\{-\big[1+\gamma\,u\big]^{-1/\gamma}\Big\}.$$

De la même manière que ce que l'on avait sans les blocs, il faut alors déterminer les valeurs des paramètres en les approximant par des méthodes comme le maximum de vraisemblance. Des auteurs comme Ferreira et de Haan (2006 et 2015) ont alors démontré l'existence d'estimateurs pertinents pour cette méthode, nommés PWM (pour "probability weighted moment"). Pour les définir, on part de la statistique suivante, soient $X_{1,k}, ..., X_{k,k}$ les observations ordonnées du bloc $X_1, ..., X_k$, on définit :

$$\beta_r = \frac{1}{k}\sum_{i=1}^{k} \frac{(i-1)...(i-r)}{(k-1)...(k-r)} X_{i,k} \quad \text{pour } r = 1, 2, 3, ..., k > r$$

A partir de β_r, on peut ensuite définir les trois estimateurs PVM suivants pour γ, a_n et b_n qui possèdent de bonnes propriétés asymptotiques sous certaines conditions (Γ est la fonction gamma d'Euler).

Pour γ : $\hat{\gamma}_{k,m}$ est solution de $\frac{3\hat{\gamma}_{k,m}-1}{2\hat{\gamma}_{k,m}-1} = \frac{3\beta_2-\beta_0}{2\beta_1-\beta_0}$
Pour a_n : $\hat{a}_{k,m} = \frac{\hat{\gamma}_{k,m}}{2\hat{\gamma}_{k,m}-1} \cdot \frac{2\beta_1-\beta_0}{\Gamma(1-\hat{\gamma}_{k,m})}$
Pour b_n : $\hat{b}_{k,m} = \beta_0 + \hat{a}_{k,m} \cdot \frac{1-\Gamma(1-\hat{\gamma}_{k,m})}{\hat{\gamma}_{k,m}}$

Sous certaines conditions, on peut enfin démontrer que les quantiles élevés sont estimables par cette méthode. On a ainsi :

$$\frac{\sqrt{k}\left(\hat{X}_{k,m} - X_n\right)}{a_n\, q_\gamma(c_n)} \xrightarrow{d} \Delta + (\gamma^-)^2 B - \gamma^- \Lambda - \lambda \frac{\gamma^-}{\gamma^- + \rho}$$

où :
- $\hat{X}_{k,m}$ est l'estimateur du quantile extrême
- X_n est le vrai quantile à estimer
- a_n est le paramètre d'échelle
- Δ, Λ, λ sont des paramètres issus de la théorie asymptotique de Ferreira et de Haan (2015)
- B est un pont brownien
- $q_\gamma(c_n)$ est une fonction définie par $q_\gamma(t) = \int_1^t s^{\gamma-1} \log s \, ds$
- $\gamma^- = \min(0, \gamma)$

Cette approche possède tout de même un défaut car lorsque l'on prend le maximum sur un bloc, on fait potentiellement disparaître des valeurs élevées, on perd des données intéressantes.

6 Méthode des excès

La méthode des excès, également appelée approche par dépassement de seuil (en anglais *Peaks Over Threshold*, ou POT), a été introduite par Pickands en 1975. Elle constitue une alternative à l'approche classique par blocs pour modéliser les phénomènes extrêmes.

Le principe est de ne conserver que les observations excédant un seuil élevé u. Si ce seuil est bien choisi la distribution des excès définis par :

$$Y_i = X_i - u \quad \text{pour} \quad X_i > u$$

peut être approximée par une distribution de Pareto généralisée (GPD).

Cette approche repose sur un résultat fondamental de Balkema et de Haan (1974), et de Pickands (1975), selon lequel, pour une grande classe de lois de probabilité F, la loi des excès conditionnels au-delà d'un seuil élevé converge vers une loi de Pareto généralisée lorsque le seuil u tend vers la borne supérieure de F.

Formellement, on considère une suite de variables aléatoires i.i.d. $X_1, \ldots, X_n$ de fonction de répartition F, et x_F le point terminal de F. Pour tout seuil $u < x_F$, on définit la fonction de répartition des excès par :

$$F_u(x) := \mathbb{P}(X - u \leq x \mid X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad \text{pour } 0 \leq x \leq x_F - u.$$

Et sa version en fonction de survie :

$$\overline{F}_u(x) := \mathbb{P}(X - u > x \mid X > u) = \frac{\overline{F}(x+u)}{\overline{F}(u)}.$$

Lorsque le seuil u est suffisamment élevé, F_u peut être bien approchée par une distribution de Pareto généralisée $G_{\gamma,\beta(u)}$, définie comme suit :

6.1 Loi de Pareto généralisée (GPD)

La fonction de répartition de la GPD est donnée par :

$$G_{\gamma,\beta}(y) = \begin{cases} 1 - \left(1 + \dfrac{\gamma y}{\beta}\right)^{-1/\gamma}, & \text{si } \gamma \neq 0, \\ 1 - \exp\left(-\dfrac{y}{\beta}\right), & \text{si } \gamma = 0, \end{cases}$$

avec $y \geq 0$, sous la condition $1 + \gamma y/\beta > 0$. Le paramètre $\beta > 0$ représente l'échelle et γ le paramètre de forme (indice de queue).

Exemple (cas exponentiel).
Soit $F(x) = 1 - e^{-x}$ la loi exponentielle standard. On a pour tout $y > 0$:

$$\mathbb{P}(X - u > y \mid X > u) = \frac{e^{-(u+y)}}{e^{-u}} = e^{-y}.$$

On retrouve donc une loi exponentielle, qui correspond à une GPD avec $\gamma = 0$ et $\beta = 1$. Cela montre que l'exponentielle est un cas particulier de GPD.

6.2 Théorème de Balkema–de Haan–Pickands

Le résultat central qui justifie l'utilisation de la GPD pour modéliser les excès est le suivant :

> Soit F une fonction de répartition appartenant au domaine d'attraction d'une loi de valeur extrême $\mathcal{H}_\gamma$. Alors, lorsque $u \to x_F$, il existe une fonction $\beta(u)$ telle que :
>
> $$\sup_{0 \leq x \leq x_F - u} \left| F_u(x) - G_{\gamma,\beta(u)}(x) \right| \to 0.$$

Autrement dit, plus le seuil u est élevé, plus la loi des excès au-dessus de ce seuil est bien approchée par une GPD.

Cette propriété est essentielle en statistique des valeurs extrêmes, car elle permet d'exploiter pleinement les données situées dans les queues de distribution, sans se limiter au maximum d'un bloc.

6.3 quantile de retour

En pratique, on cherche une valeur qui donnerait une certaine confiance quand à l'apparition de données la depassant. Dans le cas d'une distribution à queue bornée il suffit de prendre la bornes supérieure de la distribution.
Or, dans le cas d'une distribution à queue lourde et exponentiel ce n'est pas possible. On introduit alors la notion de **quantile de retour** qui est une valeur que l'on dépasse en moyenne une fois tous les T ans.
On pose z_T cette quantité, et elle est solution de l'équation suivante :

$$P\big(M \leq z_T\big) \;=\; G_{\mu,\sigma,\gamma}(z_T) \;=\; 1 - \frac{1}{T},$$

où $G_{\mu,\sigma,\gamma}$ est la fonction de répartition de la GEV. En résolvant cette équation que l'on admet, on obtient

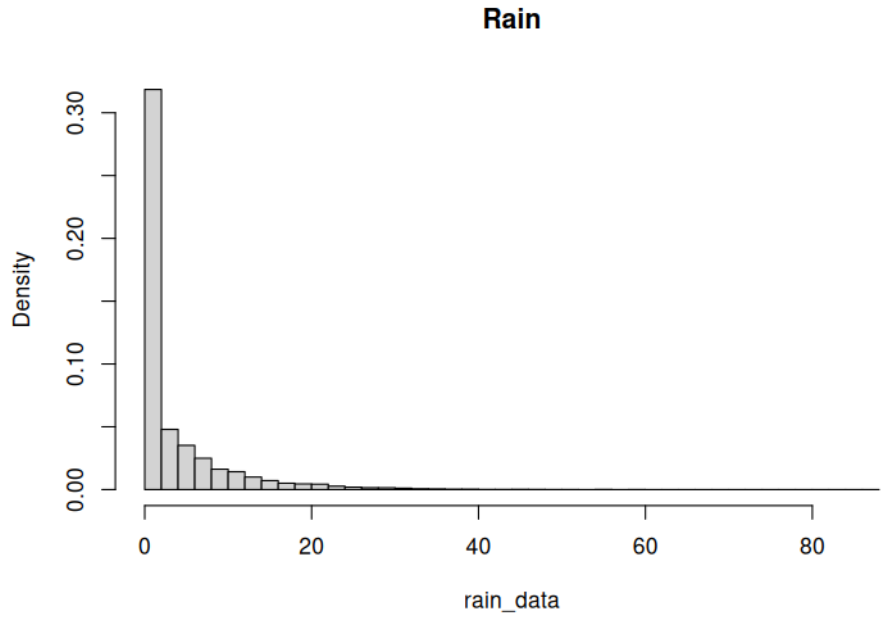
$$z_T = \begin{cases} \mu \;+\; \dfrac{\sigma}{\gamma}\Big[\big(-\ln(1-1/T)\big)^{-\gamma} - 1\Big], & \gamma \neq 0, \\ \mu \;-\; \sigma\, \ln\big(-\ln(1-1/T)\big), & \gamma = 0. \end{cases}$$

7 Application sur des données réelles

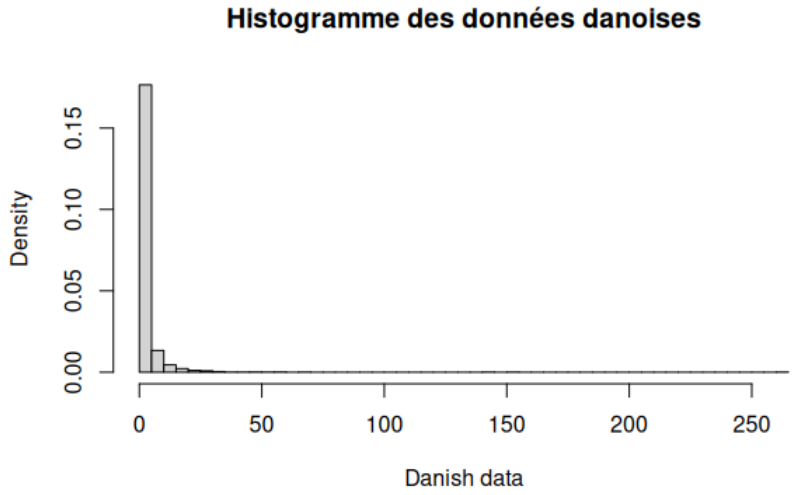
7.1 Description des données

Afin d'illustrer les méthodes d'estimation de l'indice de valeurs extrêmes, nous allons appliquer ces techniques sur des données réelles. Nous allons utiliser les données du package *ismev* de R. Plus précisément celui de rain et de danish. Ils contienent respectivement les données de pluie journalière en Angleterre de 1914 à 1962 et les grands sinistres incendie survenus au Danemark entre 1980 et 1990.

L'objectif sur ces données est de savoir s'il existe (et le cas échéant de le calculer) un seuil tel que à l'avenir on dépasse cette valeur rarement tout en restant raisonable.



Dans le cas de rain, on remarque que les données sont concentrées autour de 0 mais qu'elles sont capables de prendre des valeurs très élevées jusqu'à 90. Il est alors raisonnable de penser qu'après estimation, on va obtenir une valeur de gamma positive ou nulle. En effet, il n'apparait pas de cassure dans la distribution des données. De plus, les données prennent des valeurs grandes mais perdent rapidement en densité pour celle-ci. Ce qui suggèrerait une valeur de gamma proche de 0 et positive.



De même, pour les sinistres, on remarque que la majorité des sinistres sont de faible intensité mais qu'il existe des sinistres de grande intensité. On peut donc s'attendre à une valeur de gamma positive ou nulle.

7.2 Méthode des maxima par bloc

7.2.1 Application sur les données de Rain

Les données étant journalières, on va choisir des blocs de taille 365, comme les données vont de 1914 à 1962, on se retrouve avec 48 blocs de 365 jours.

Via le package **evd**, on obtient via une estimations par maximum de vraisemblance les paramètres suivants : $\mu = 40,8$, $\sigma = 9,73$ et $\gamma = 0,107$.
L'estimation des paramètres sont calculées par l'algorithme de Nelder-Mead (voir annexe).

On suppose alors que $\gamma \geq 0$, donc il n'existe pas de valeur maximale finie : la probabilité de très gros maxima décroît mais n'est pas finie, selon un comportement polynomial ou exponentiel ce qui est embettant dans le cas pratique surtout quand on cherche des seuils rarement atteints.

7.2.2 Application sur les données danish

On applique la même méthode sur les données de sinistres mais avec une approche légèrement différente puisque les données ne sont pas dans le même format. En effet, on ne dispose pas de données journalières mais l'ensemble des sinistres sur une période de 10 ans. C'est à dire qu'il existe des jours sans sinistres qui ne sont pas comptabilisés. On regroupe alors les sinistres par mois et on prend le sinistre le plus élevé. On se retrouve alors avec des blocs de taille irrégulière.

Après analyse numérique par maximum de vraisemblance, on obtient les paramètres suivants : $\mu = 8,376$, $\sigma = 5,971$ et $\gamma = 0,623$.

On possède une valeur de gamma positive ce qui renforce l'hypothèse d'une queue lourde pour la distribution du max.

7.2.3 synthèse sur la méthode des maxima en bloc

Quand on dispose des données sur une longue période, la méthode des maxima en bloc est efficace. D'autant plus quand on a des données temporelles (journalières, mensuelles, annuelles). En revanche, elle utilise moins de données ce qui la rend plus difficile à utiliser en pratique. En effet, on ne garde que les maximums et on perd donc une partie des données qui peuvent etre consequents en fonction du choix de k. Par ailleurs, le choix de k est important car il faut choisir un nombre de blocs suffisant pour avoir une estimation fiable mais pas trop grand pour ne pas perdre trop d'informations.

7.3 Méthode de dépassement de seuil

7.3.1 Application sur les données de Rain

On a calculé la valeur de γ par la méthode des maxima en bloc et obtenu une valeur de γ positive mais proche de 0. On souhaite s'assurer du signe de γ en procédant par la méthode de dépassement de seuil. On va deplus estimer la valeur de γ avec l'estimateur de Pickands.
On decide de prendre un seuil correspondant au quantile d'ordre 0,95.

On obtient numériquement : $\gamma = -0.027$

On a obtenu dans la méthode des maxima en bloc une valeur de γ postive mais proche de 0, tandis qu'avec la méthode de dépassement de seuil, on obtient une valeur négative mais tout aussi proche de 0.
Ce qui nous amène à conclure que la valeur de γ est de 0.
Autrement dit, la distribution du maximum de pluie suit une loi de Gumbel.

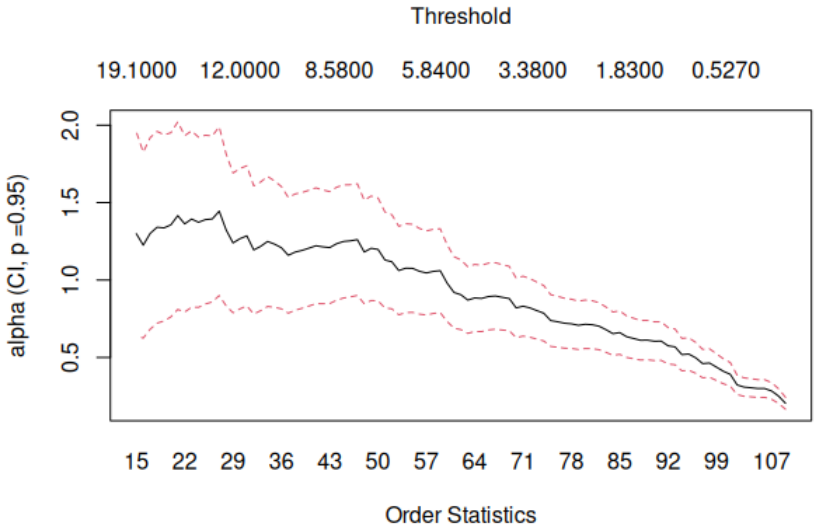
On peut néamoins donner une valeur "seuil" qui nous assurerait que la probabilité de dépasser cette valeur est très faible. On calcule alors le quantile de retour pour $T = 100$ afin d'avoir une valeur de pluie qui ne devrait pas être dépassée plus d'une fois tous les 100 ans.

Dans notre cas, on obtient alors : $z_t = 98,636$. Autrement dit, une fois tous les 100 ans, on peut s'attendre à avoir une pluie de plus de 98.636 mm.

7.3.2 Application sur les données danish

On prend un seuil correspondant au quantile d'ordre 0,95. Apres calcul numérique, on obtient $\sigma = 7.038$ et $\gamma = 0.492$.
On a obtenue une valeur de γ positive ce qui renforce l'hypothèse d'une queue lourde pour la distribution du max, ce qui est cohérent avec la méthode des maxima en bloc. On peut pousser l'analyse de gamma en calculant son estimation via l'estimateur de Hill afin d'avoir une valeur plus précise.

On trace alors la courbe de l'estimateur de Hill en fonction de k.



On observe alors un plateau pour k entre 30 et 45. On obtient alors apres estimation numérique une valeur de $\gamma = 0,66$.
On peut alors calculer le quantile de retour pour $T = 50$ afin d'avoir une valeur de sinistre qui ne devrait pas être dépassée plus d'une fois tous les 50 ans.
On obtient alors : $z_t = 515,22$. Autrement dit, une fois tous les 50 ans, on peut s'attendre à avoir un sinistre depassant plus de $515,22$.

7.3.3 Synthèse sur la méthode de dépassement de seuil

Cette méthode à le bont goût d'exploiter toutes les valeurs supérieures à un seuil, pas seulement un maximum par bloc. Elle utilise donc plus de données.
En revanche le choix du seuil est important car il faut choisir un seuil suffisamment élevé pour ne pas avoir trop peu de données mais pas trop élevé pour ne pas perdre d'informations sur les max. Il faut donc faire le compromis entre le biais et la variance.

8 Annexe

8.1 Méthode de Nelder-Mead

Le package "evd", que nous avons utilisé pour réaliser les méthodes de dépassement de seuil et des maxima en bloc, utilise l'algorithme de Nelder-Mead pour calculer les paramètres de la fonction limite et ainsi savoir dans quel cas où se trouve : Fréchet, Gumbel ou Weibull.

Nelder-Mead est un algorithme d'optimisation non linéaire, il consiste en la chose suivante dans le cadre des valeurs extrêmes :

- **Etape 1** : on commence par choisir 3 premiers points x_1, x_2, x_3 par une rapide estimation des paramètres σ, μ et γ de nos données. Ce seront nos points de départs de l'algorithme et ils définissent notre premier simplexe (triangle ici) dans R^2.
- **Etape 2** : on calcule ensuite la valeur de la fonction en ces 3 points : f est la fonction GEV généralisée (à définir plus précisément) et on les trie par valeurs décroissantes.
- **Etape 3** : on cherche le centre de gravité x_0 de nos premiers points : $x_0 = \frac{x_1+x_2+x_3}{3}$.
- **Etape 4** : on fait ensuite une réflexion en calculant $x_r = x_0 + \alpha(x_0 - x_3)$ où $\alpha > 0$ est appelé le coefficient de réflexion
- **Etape 5** : si $f(x_1) \leq f(x_r) \leq f(x_3)$: on remplace x_3 par x_r et on retourne à l'étape 2.
- **Etape 6** : si $f(x_r) \leq f(x_1)$: on procède à une expansion du simplexe, on calcule $x_3 = x_0 + \gamma(x_r - x_0)$ où $\gamma > 1$. Si $f(x_e) \leq f(x_r)$, on remplace x_3 par x_e sinon on remplace x_3 par x_r et on retourne à l'étape 2
- **Etape 7** : si $f(x_r) \geq f(x_3)$: on procède à une contraction du simplexe, on cherche $x_c = x_0 + \rho(x_3 - x_0)$ où $0 < \rho < 0.5$. Si $f(x_c) \leq f(x_3)$, on remplace x_3 par x_c et on retourne à l'étape 2, sinon on continue jusqu'à l'étape 8.
- **Etape 8** : on effectue une homothétie de rapport ω et de centre x_1 : on remplace ainsi x_i par $x_1 + \omega(x_i - x_1)$ où $0 < \omega < 1$et on retourne à l'étape 2

On répète cela jusqu'à atteinte du critère d'arrêt, en général : $\sqrt{\sum_{i=1}^{n+1} \frac{(f_i - \bar{f})^2}{n}} < \epsilon$ où $\bar{f} = \frac{1}{n+1} \sum_{i=1}^{n+1} f_i$ et ϵ est un réel proche de 0.

8.2 Codes R

Voici un exemple de code R utilisé dans la première section :

```r
# Paramètres
n <- 1000        # Taille de l'échantillon pour la simulation des lois uniformes
N <- 10000       # Nombre de simulations pour le maximum

# Simulation des maxima de lois uniformes(0,1)
set.seed(123)    # fixation de l'aléa
M_n <- replicate(N, max(runif(n))) # M_n = max / X_n = runif

# Normalisation pour observer la convergence
Y_n <- n * (1 - M_n)

# Histogramme des valeurs transformées
hist(Y_n, breaks = 50, probability = TRUE,
col = "lightblue", border = "white", ylab = "Densité",
xlab = expression(Y_n), main = "Max de 1000 lois uniformes")

# Densité théorique de la loi exponentielle (paramètre = 1)
curve(dexp(x, rate = 1), col = "red", lwd = 2, add = TRUE)

# Légende
legend("topright", legend = c("Simulation", "Densité théorique : exp(1)"),
fill = c("lightblue", NA), border = c("white", NA),
lty = c(NA, 1), col = c(NA, "red"), lwd = c(NA, 2))
```

```r
#################### CODE POUR WOOSTER ####################
library(ismev)
library(evd)
data("wooster")

gev_fit <- fgev(wooster)
```

```r

mu <- as.numeric(gev_fit$param[1])
sigma <- as.numeric(gev_fit$param[2])
gamma <- as.numeric(gev_fit$param[3])

# estimation de gamma avec pickands (juste pour comparer)

x <- sort(wooster)
n <- length(x)
k <- floor(0.1 * length(wooster))
X1 <- x[n - k + 1]
X2 <- x[n - 2*k + 1]
X3 <- x[n - 4*k + 1]
pickands_est <- (1 / log(2)) * log((X1 - X2) / (X2 - X3))
print(pickands_est)


# gamma est < 0 donc on calcule la borne max
x_max <- mu - sigma / gamma

# Définir la densité de la loi (pour gamma < 0)
dgev <- function(x, mu, sigma, gamma) {
  t <- 1 + gamma * ((x - mu) / sigma)
  dens <- ifelse(t > 0,
                                  (1/sigma) * t^(-1/gamma - 1) * exp(-t^(-1/gamma)),
                                  0)
  return(dens)
}

xseq <- seq(min(wooster), max(wooster), length.out = 200)

# PLOT

hist(wooster,main = "Histogram de wooster" ,breaks = 60, probability = TRUE, col = "
    lightgray")

lines(xseq, dgev(xseq, mu, sigma, gamma), col = "blue", lwd = 2)


abline(v = x_max, col = "red", lwd = 2, lty = 2)
legend("topright", legend = paste("x_max =", round(x_max, 2)), col = "red", lwd = 2,
    lty = 2)

# plot plus détaillé
plot(gev_fit)

#################### CODE POUR RAIN ####################
library(ismev)
library(evd)
data(rain)
rain_data <- rain

# seuil
threshold <- quantile(rain_data, probs = 0.95)
gpd_result <- gpd.fit(rain_data, threshold)

# on stocke la parametre d'échelle et de forme
sigma <- gpd_result$mle[1]
gamma <- gpd_result$mle[2]
SE <- gpd_result$se[2]
IC <- c(gamma - 1.96 * SE, gamma + 1.96 * SE)    # contient 0 (oups)


# On code la fonction de pareto généralisée parametre echel sigma et de forme gamma
pareto <- function(x, gamma, sigma) {
  if (gamma == 0) {
        return(1/sigma * exp(-x/sigma))
  } else {
        return(1/sigma * (1 + gamma * x/sigma)^(-1/gamma - 1))
  }
}

# on trace l'histogramme des données
hist(rain_data, breaks = 50, freq = FALSE, main = "Rain")
```

```
29
30          # on trace l'histogramme des données en excés par rapport au seuil et la loi de pareto
31          hist(rain_data[rain_data > threshold] - threshold, breaks = 50, freq = FALSE, main = "
                Rain␣Excesses␣et␣densité␣de␣Pareto")
32
33          # on trace la loi de gpd avec les paramètres estimés
34          xseq <- seq(min(rain), max(rain), length.out = 200)
35          lines(xseq, pareto(xseq, gamma, sigma),col='red', lwd=2)
36
37
38
39          # pour le qq-plot et residus
40          gpd.diag(gpd_result)
```