Étude des valeurs extrêmes univariées

El Mazzouji Wahel, Mariac Damien, Condamy Fabian

2 avril 2025

Table des matières

1 Introduction

Le théorème central limite, formulé par Pierre-Simon de Laplace en 1809, garantit que, sous des conditions raisonnables, la somme normalisée de ces variables suit asymptotiquement une loi normale. Cette convergence est utile pour étudier le comportement **global** des observations, mais elle ne renseigne pas sur le comportement des valeurs extrêmes.

Il est donc naturel de se demander quelle peut être la convergence en loi de ses dernières.
Autrement dit, pour $X = (X_1, ..., X_n)$ un échantillon de variables aléatoires i.i.d , on pose :

$$M_n = \max\{X_i \mid i \in \{1, ..., n\}\}$$

et on s'intéresse à la convergence de M_n, ainsi qu'aux hypothèses sous lesquelles cette convergence a lieu.

Remarque : Etudier le minimum est totalement analogue dans ce qui suit.

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 :

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

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

Mais on rencontre un probleme ici, puisque si $n \to +\infty$, $F(t)^n$ converge vers 0 (ou 1 si t est la borne sup du support des X_i).

L'idée est donc d'introduire 2 suites (b_n) et (a_n) (avec $a_n > 0$ pour tout n) afin de pouvoir contrôler M_n.

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

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

Si il existe de tel suite a_n et b_n alors on dit que F est dans le domaine d'attraction de G.

à ce stade la, 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 (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 astucieux. On va poser :

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

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

Or, on a $\lim_{n\to\infty}(1-\frac{v}{n})^{n-1} = e^{-v}$, et on a $\lim_{n\to\infty}\int_0^n = \int_0^{+\infty}$.

2.2 Paramètre b_n

On en déduit une bonne valeur pour b_n. En effet,

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} < t\right) \xrightarrow[n\to+\infty]{} G(t) \in]0:1[$$

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

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

$$\iff n(-F(a_n t + b_n) + 1) \xrightarrow[n\to+\infty]{} ln(G(t)) \qquad (\text{car} \quad \lim_{x\to 0}\frac{ln(1-x)}{x} = -1)$$

$$\iff n\,\mathbb{P}(X_1 > a_n t + b_n) \xrightarrow[n\to+\infty]{} -ln(G(t))$$

On obtient alors pour paramètre d'échelle :

$$n\mathbb{P}(X_1 > b_n) = 1 \iff \mathbb{P}(X_1 < b_n) = 1 - \frac{1}{n}$$

$$\iff F(b_n) = 1 - \frac{1}{n}$$

$$b_n = Q(1 - \frac{1}{n}) = K(n)$$

Dans la dernière équivalence, on a composé par la fonction quantile.

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 $\lim_{x\to\infty}\frac{K(xu)-K(x)}{a(x)}$ converge **vers une fonction** $h(u)$.

Proposition :
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.

Preuve de la Proposition
Soient $u, v > 0$. Alors :

$$\frac{K(xuv)-K(x)}{a(x)} \;=\; \frac{K(xuv)-K(xu)}{a(xu)}\,\frac{a(xu)}{a(x)} \;+\; \frac{K(xu)-K(x)}{a(x)}. \tag{2.3}$$

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 (2.3) et en utilisant ce qui précède, on obtient :

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

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

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

2.4.1 Nature du support

En reprenant l'équation (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 :

2.4.2 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 :

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

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

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

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

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

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

2.4.3 Si $\gamma < 0$

Pour $\gamma < 0$, la loi est définie quand :

$$1 + \gamma u > 0 \quad \Longrightarrow \quad u < -\frac{1}{\gamma}$$

Cela signifie que la distribution a son support dans $]-\infty, -\frac{1}{\gamma}[$
On pose alors $x_{\max} = -\frac{1}{\gamma}$.
Par conséquent, la fonction de survie $S(x) = 1 - G(x) = 0$ 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**.

On dit alors que que queue de distribution est bornée.

2.4.4 2. Cas $\gamma = 0$

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.

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) :

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

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 1000 fois chaque loi puis ensuite effectué 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 :

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

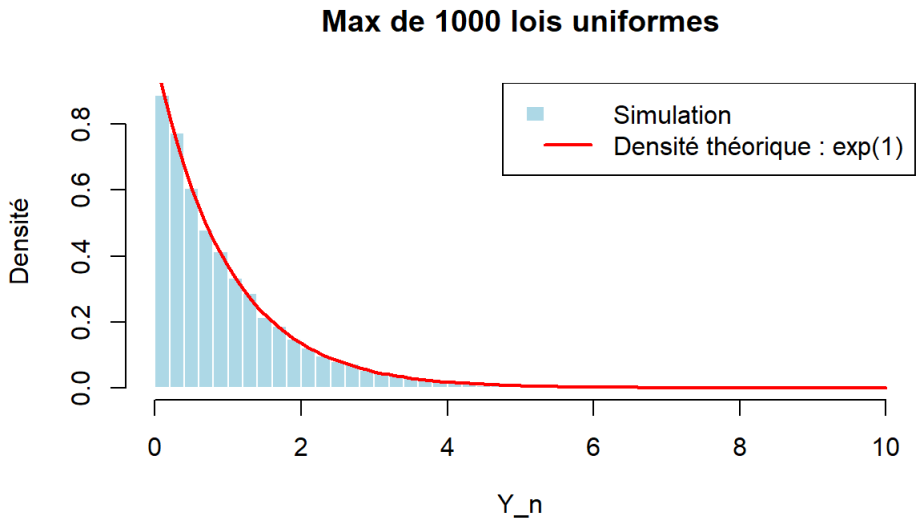
Pour n grand, on a : $(1 - y/n)^n \approx e^{-y}$. Donc, $P(M_n \leq 1 - y/n) \approx e^{-y}$.
Or, par définition, la loi exponentielle de paramètre 1 a pour fonction de répartition : $P(Y \leq y) = 1 - e^{-y}, \quad y > 0$.
Ainsi, on a donc montré que :

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

ce qui établit la convergence en loi :

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

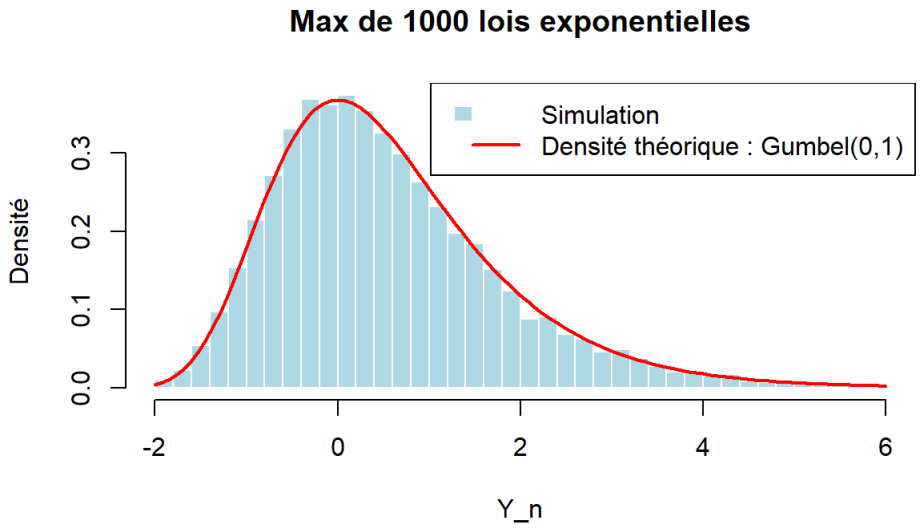
Ainsi, on trouve que $a_n = \frac{1}{n}$ et $b_n = 1$.
Avec notre machine, nous obtenons le graphe suivant :



Remarquons que l'on obtient une loi de Gumbell, 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

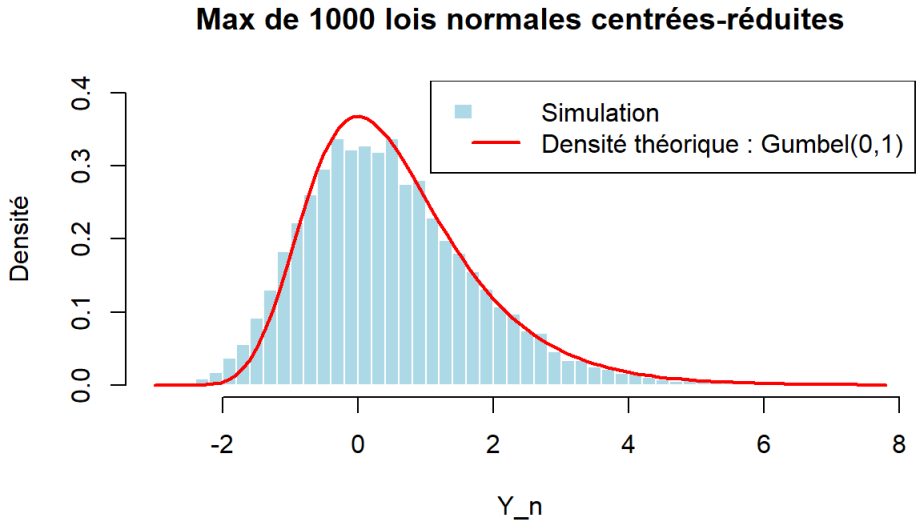
Pour une loi exponentielle de paramètre 1, la loi limite est une loi de Gumbel. Théoriquement, on trouve $a_n = 1$ et $b_n = \log(n)$.



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

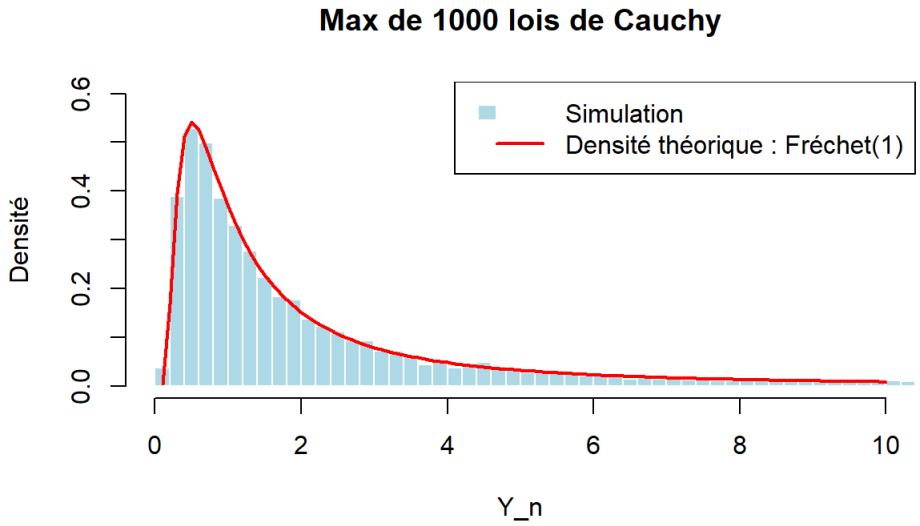
Pour maintenant une loi normale centrée-réduite, on peut montrer que la loi limite est encore une fois une loi de Gumbel. On trouve les paramètres généralisés $a_n = \frac{1}{\sqrt{(2*\log(n)}}$ et $b_n = \frac{1}{a_n} - \frac{\log(\log(n))+\log(4*pi)}{2*\sqrt{2*\log(n)}}$.



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 $b_n = n$.



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 nous intéressons aux différentes méthodes d'estimation du paramètre γ, intervenant dans la distribution des valeurs extrêmes généralisée.
D'une part, des approches non paramétriques sont dédiées à l'estimation de l'indice de queue, notamment les estimateurs de Hill et de Pickands. D'autre part, des méthodes paramétriques ont été développées, parmi lesquelles la méthode du maximum de vraisemblance, la méthode des moments et les approches bayésiennes.

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 :

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

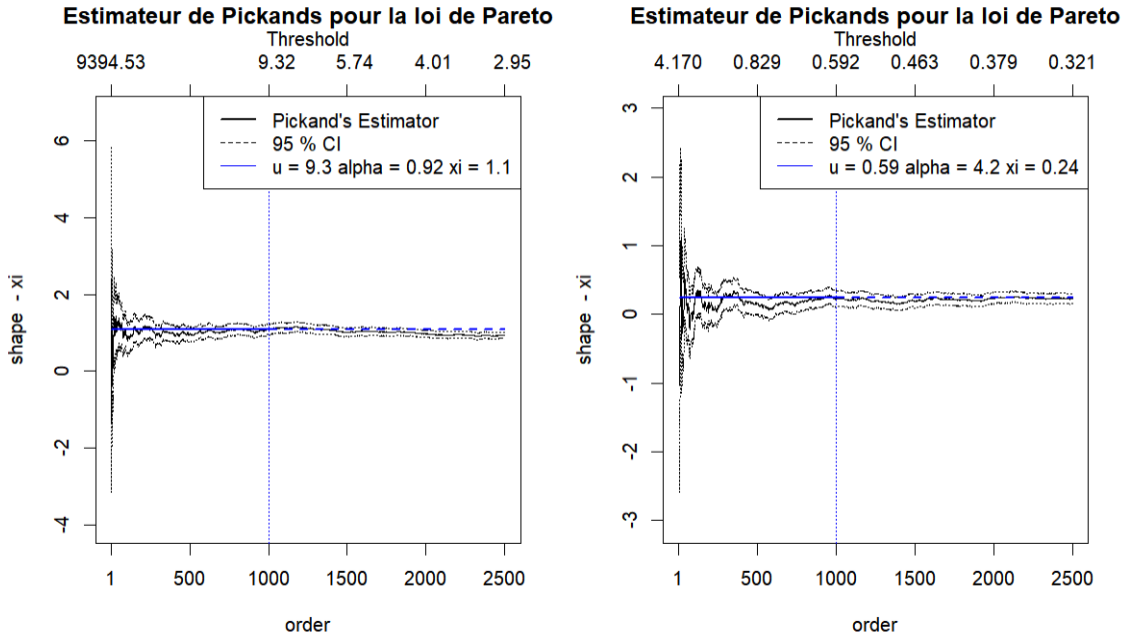
4.1 Estimateur de Pickands

L'estimateur de Pickands est défini par la statistique

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

Cet estimateur présente l'avantage d'être applicable quelle que soit la distribution des extrêmes.

Cependant, la représentation graphique de cet estimateur en fonction du nombre k d'observations considérées révèle généralement un comportement volatil au départ, ce qui peut nuire à la lisibilité du graphique. De plus, cet estimateur est particulièrement sensible à la taille de l'échantillon sélectionné, ce qui le rend peu robuste. Afin d'illustrer ces observations, nous appliquons l'estimateur de Pickands à des données simulées suivant une loi de Pareto avec différents paramètres de forme α.



Les résultats obtenus montrent que pour $\alpha = 1$, nous obtenons $\gamma \approx 1.1$, tandis que pour $\alpha = 5$, l'estimateur converge vers $\gamma \approx 0.24$, proche de la valeur théorique de 0.2. Ces différences illustrent la sensibilité de l'estimateur aux paramètres choisis et aux fluctuations des données extrêmes.

Cet estimateur est asymptotiquement normal, avec :

$$\sqrt{k}\frac{\hat{\gamma_{k,n}} - \gamma}{\sigma(\gamma)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Lorsque $k \to +\infty$, la variance asymptotique est donnée par :

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

Une généralisation de l'estimateur de Pickands a été introduite comme suit :

$$\hat{\gamma}_{(k,u,v)} = \frac{1}{\ln(v)} \ln\left(\frac{X_{n-k+1,n} - X_{n-[uk]+1,n}}{X_{n-[vk]+1,n} - X_{n-[uvk]+1,n}}\right)$$

où u et v sont des réels positifs différents de 1, de sorte que les indices $[vk]$, $[uk]$ et $[uvk]$ ne dépassent pas n. Lorsque $u = v = 2$, on retrouve l'estimateur de Hill $\hat{\gamma}_{k,n}$.

4.2 Construction de l'estimateur de Pickands

Proposition : (Caractérisations de $D(H_\gamma)$)

Pour $\gamma \in \mathbb{R}$, les affirmations suivantes sont équivalentes.

(a) $F \in D(H_\gamma)$

(b) Pour une certaine fonction positive $c(t) = a\left(\frac{1}{t}\right)$:

$$\lim_{t\to 0} \frac{U(tx) - U(t)}{c(t)} = \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{si } \gamma \neq 0, \\ \log(x) & \text{si } \gamma = 0, \end{cases} \quad \text{pour } x > 0.$$

La dernière affirmation est équivalente à :

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

pour $x, y > 0$ et $y \neq 1$.

Lemme A : 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 $(X_{1,n}, \ldots, X_{n,n})$

Preuve de la construction de l'estimateur de Pickands :

On déduit de la proposition précédente que 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 fait, 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 $k(n), n \geq 1$ 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

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

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

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

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

Donc la convergence suivante a lieu en probabilité :

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

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

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

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

4.3 Estimateur de Hill

Tout d'abord, l'estimateur de Hill est applicable uniquement aux distributions de Fréchet ($\gamma > 0$), où il permet d'obtenir un estimateur de l'indice de queue plus efficace que celui de Pickands. Cet estimateur est défini par la statistique suivante :

$$\hat{\gamma}_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \ln\left(\frac{X_{n-i+1,n}}{X_{n-k,n}}\right)$$

pour $k \in \{1, \ldots, n-1\}$. Si l'on choisit $k, n \to +\infty$, de sorte que $\frac{k}{n} \to 0$, alors on peut montrer que $\lim_{k\to\infty} \hat{\gamma}_{k,n} = \gamma$. Cet estimateur possède la propriété d'être asymptotiquement normal, ce qui signifie que :

$$\sqrt{k}\frac{\hat{\gamma}_{k,n} - \gamma}{\gamma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Il existe plusieurs approches pour construire l'estimateur de Hill. Une approche possible consiste à utiliser la méthode du maximum de vraisemblance.

Tout d'abord, on considère une suite de variables aléatoires $X_1, \ldots, X_n$ i.i.d. suivant une loi de Pareto de paramètre $\lambda > 0$, dont la fonction de répartition est donnée par :

$$F(x) = 1 - x^{-\lambda}, \quad \text{pour } x \geq 1.$$

La densité de probabilité associée est alors :

$$f(x) = \lambda x^{-\lambda-1}, \quad \text{pour } x \geq 1.$$

La fonction de vraisemblance est donnée par :

$$L(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^{n} f(x_i) = \lambda^n \prod_{i=1}^{n} x_i^{-\lambda-1}.$$

En prenant le logarithme, on obtient la log-vraisemblance :

$$\log L(x_1, \ldots, x_n, \lambda) = n \log \lambda - (\lambda + 1) \sum_{i=1}^{n} \log x_i.$$

En dérivant cette expression par rapport à λ, on obtient :

$$\frac{d \log L}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} \log x_i.$$

En dérivant une seconde fois, nous obtenons :

$$\frac{d^2 \log L}{d\lambda^2} = -\frac{n}{\lambda^2} < 0,$$

ce qui confirme qu'il s'agit bien d'un maximum.
Ainsi, l'estimateur du maximum de vraisemblance de $\frac{1}{\lambda}$ est donné par :

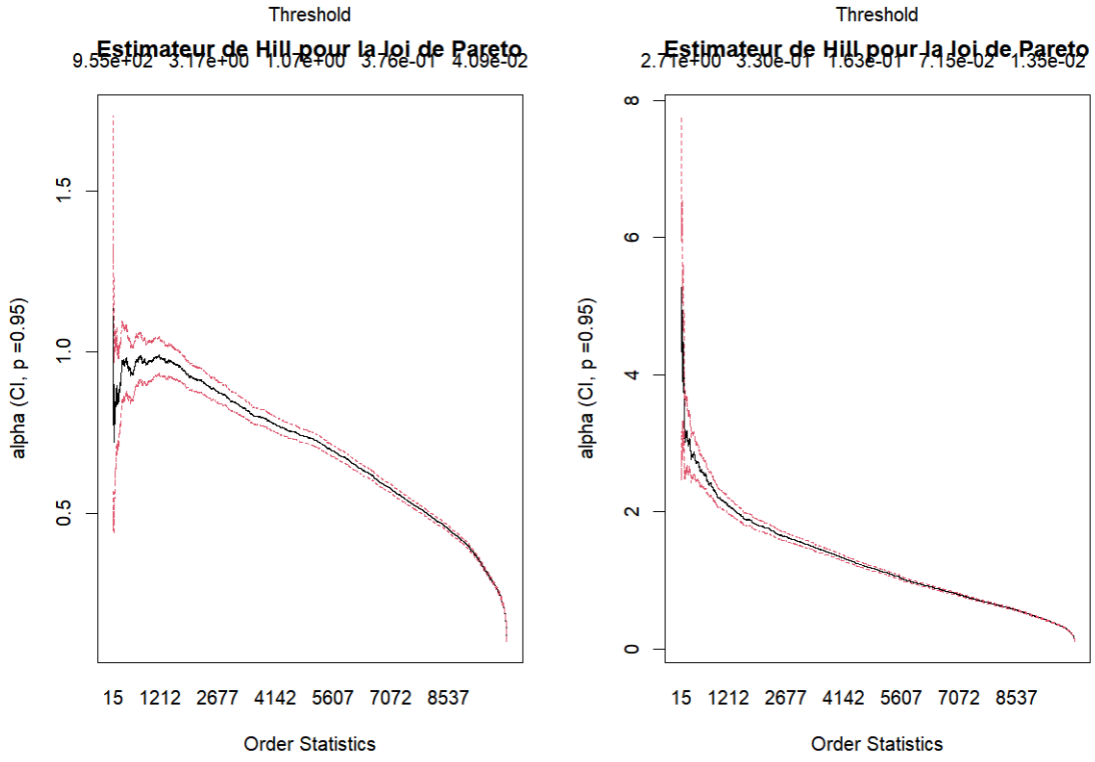
$$\hat{\lambda}^{-1} = \frac{1}{n} \sum_{i=1}^{n} \log X_i.$$

Cela implique que l'estimateur du paramètre λ est :

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} \log X_i \right)^{-1}.$$

4.3.1 Graphique pour l'estimateur de Hill

Dans cette section nous allons analyser le comportement de l'estimateur de Hill sur des données simulées à partir de la distribution de Pareto.



L'estimateur de Hill, appliqué à des données simulées à partir de la distribution de Pareto avec des paramètres de forme de 1 et 5, montre une décroissance globale et présente une forte volatilité, particulièrement pour les faibles valeurs du paramètre de forme $\alpha = 1$. Cette volatilité est due au fait que l'estimateur repose sur un nombre limité d'observations extrêmes, ce qui peut entraîner des fluctuations importantes. À mesure que le paramètre de forme augmente $\alpha = 5$, l'estimateur tend à se stabiliser plus rapidement, illustrant ainsi la convergence vers la valeur théorique de l'indice de forme.

4.4 Estimateur de DEDH

Le troisième estimateur de l'indice de queue est celui proposé par Dekkers, Einmahl et De Haan. Il s'agit d'une généralisation de l'estimateur de Hill, applicable à tous les domaines d'attraction. Il est défini par :

$$\hat{\gamma}_n^{(DEdH)}(k_n) = \mathcal{M}_{k_n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(\mathcal{M}_{k_n}^{(1)})^2}{\mathcal{M}_{k_n}^{(2)}}\right)^{-1}$$

où

$$\mathcal{M}_{k_n}^{(r)} = \frac{1}{k_n}\sum_{i=1}^{k_n}(\ln(X_{(n-i+1)}) - \ln(X_{(n-k_n)}))^r.$$

La valeur de $\mathcal{M}_{k_n}^{(1)}$ correspond à l'estimateur de Hill.

L'estimateur de DEDH possède la propriété de convergence en loi :

$$\sqrt{k_n}(\frac{\hat{\gamma}_n^{(DEdH)}(k_n) - \gamma}{\sigma_M}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \text{quand } n \to \infty.$$

où :

$$\sigma_M^2 = \begin{cases} 1+\gamma^2, & \text{si } \gamma \geq 0, \\ (1-\gamma^2)(1-2\gamma)\left(4 - \frac{8(1-2\gamma)}{1-3\gamma} - \frac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right), & \text{si } \gamma < 0. \end{cases}$$

En pratique, il est difficile de comparer ces estimateurs de manière tranchée. Toutefois, l'estimateur de Hill se distingue par une variance asymptotique plus faible, ce qui justifie son choix dans la suite. Étant donné que cet estimateur n'est valide uniquement pour les distributions appartenant au domaine d'attraction de Fréchet, c'est-à-dire dans le cas où $\gamma > 0$, il est essentiel de vérifier cette hypothèse.

5 Sélection des estimateurs de l'indice de valeurs extrêmes

Le choix de l'estimateur dépend du type de distribution sous-jacente. L'estimateur de Hill est spécifiquement adapté aux distributions de Fréchet ($\gamma > 0$), caractérisées par des queues lourdes. Il est donc plus efficace dans ce cas et sera préféré à l'estimateur de Pickands.

Cependant, pour les distributions de Weibull ($\gamma < 0$) et Gumbel ($\gamma = 0$), l'estimateur de Hill n'est pas applicable. Dans ces cas, on utilise l'estimateur de Pickands, qui est valide quel que soit le signe de γ.

L'estimateur de Pickands est basé sur les distances entre deux statistiques d'ordre, sans tenir compte du maximum de l'échantillon, ce qui entraîne une perte d'information sur la queue de distribution. Par conséquent, il présente une plus grande volatilité que l'estimateur de Hill, qui repose sur la moyenne des logarithmes des observations.

6 Détermination du domaine d'attraction

Une approche graphique qui permet de déterminer à quel domaine d'attraction appartiennent les données consiste à tracer le quantile plot généralisé repris de Anis Borchani (2010). Le quantile plot généralisé est défini par :

$$\left(\ln\left(\frac{n+1}{j}\right), \ln\left(\hat{\gamma}_{j,n}^{(UH)}\right)\right) \quad \text{pour tout } j \in [1; k_n]$$

avec

$$\hat{\gamma}_{j,n}^{(UH)} = X_{(n-j)}\hat{\gamma}_n^{(H)}(k_n).$$

La difficulté pratique dans le calcul de ces estimateurs réside dans le choix du nombre d'excès k_n à prendre en compte. Si les estimateurs sont calculés en utilisant un trop grand nombre d'observations, leur biais sera élevé. À l'inverse, si le nombre d'observations est trop faible, cela entraînera une variance importante.

7 Application sur des données réelles

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 précisement wooster et rain. Wooster contient les données de temperature minimal (en Fahrenheit) annuelle à Wooster de 1983 à 1988. Tandis que Rain contient les données de pluie journalière dans en Angletterre de 1914 à 1962.
Nous allons utiliser deux méthodes d'estimation sur les paramêtres a_n, b_n et γ afin de d'estimer la valeur extrêmes.

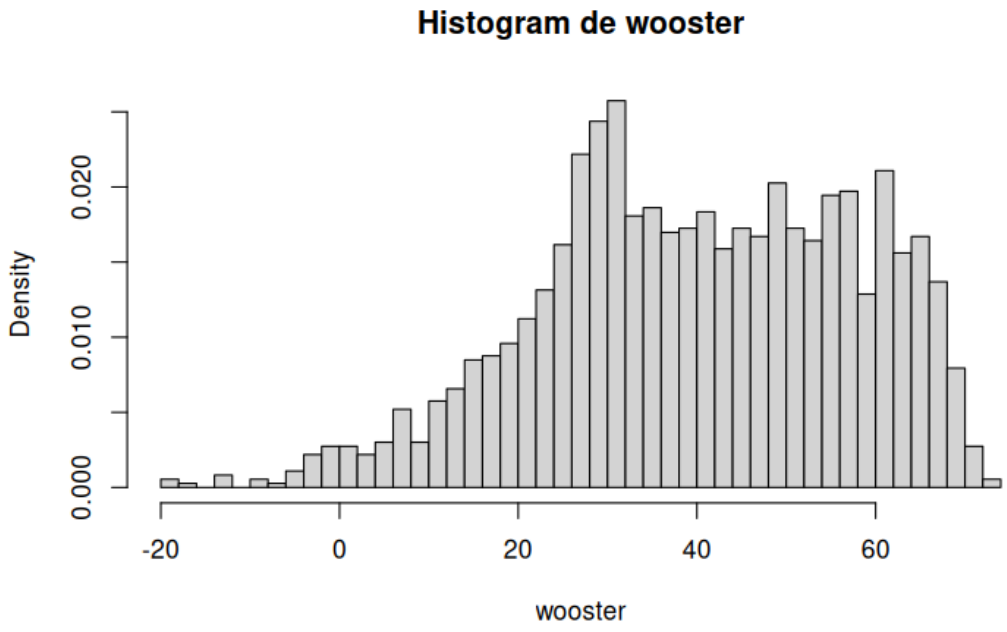
7.1 Première méthode (Méthode des maxima en bloc) avec Wooster

7.1.1 Principe

La première étape consiste à decouper nos données en blocs de taille k et de calculer le maximum sur chaque bloc. Le parametre k est choisit en fonction de l'interpretation des données. (par exemple, si on a des données journalières, on peut choisir $k = 365$ pour avoir des maximums annuelle). Ensuite, pour chaque bloc on calcule le maximum. Cela nous donne une suite de maximum. Une fois les maximums obtenus, on estime a_n,b_n et γ en utilisant la méthode du maximum de vraisemblance.

7.1.2 Application sur les données de Wooster

L'objectif sur ces données est de savoir si il existe (et dans le cas écheant de le calculer) un seuil tel que les températures ne puisse pas depasser. Chercher cette valeur seuil serait utilse en agriculure par exemple pour savoir si les températures ne sont pas trop élevées pour les cultures.



Nous avons découper ici nos données en blocs de taille 60 et de calculer le maximum sur chaque bloc. L'estimation numériques par l'algorithme de Nelder-Mead des paramètres a_n,b_n (qu'on note pour la suite σ et μ) et γ nous donne :

$$\sigma = 36.02, \quad \mu = 18.82, \quad \gamma = -0.5$$

Nous obtenons une valeur de γ négative, ce qui signifie que la distribution des températures max à Wooster est de type Weibull. Nous pouvons donc conclure que les températures à Wooster sont pas limitées par un seuil. Ce seuil étant donnée dans la partie 1, il vaut : $x_{max} = \mu - \frac{\sigma}{\gamma} = 74.93$

Il est alors raisonnables de penser que les températures à Wooster ne depassent pas 74.93.
v

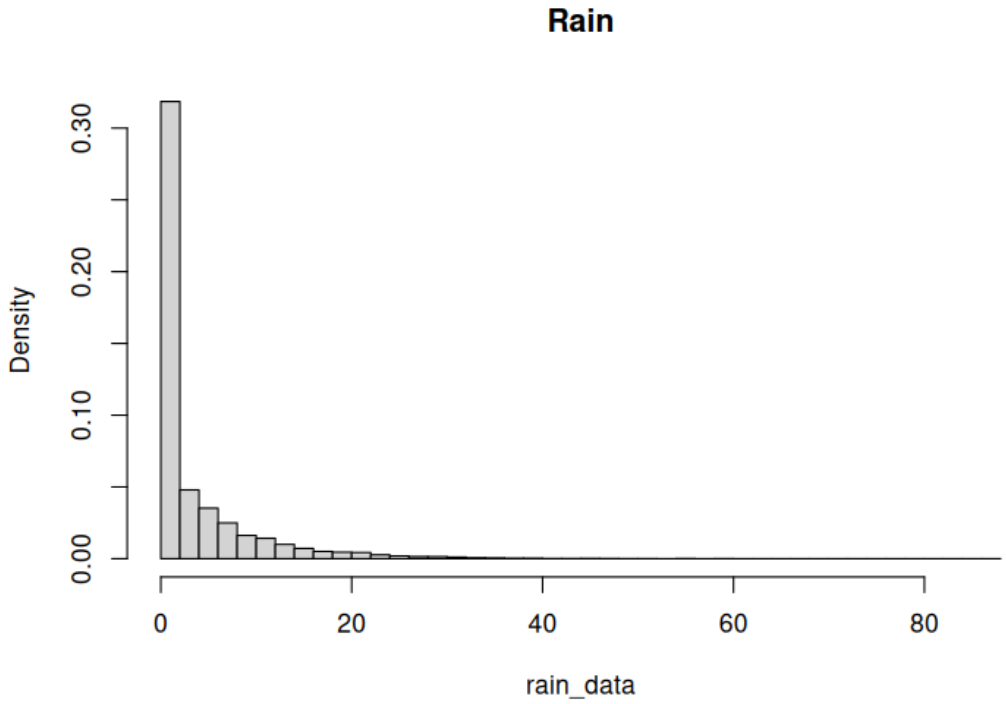
7.2 Méthode de depassement de seuil avec Rain

7.2.1 Principe

La deuxieme méthode consiste à fixer un seuil u et de considèrer les données qui depassent ce seuil. C'est à dire X_i tel que $X_i > u$. Ensuite, on stocke les excès $X_i - u$. Cela nous donne un jeu de données positifs. La clé de cette méthode est que pour un seuil u bien choisit, les excès suivent une loi de Pareto de parametres σ (échelle) et γ (le gamma qu'on estime dans toute la théorie). C'est alors qu'on ajuste les paramêtres σ et γ par maximum de vraisemblance.

7.2.2 Application sur les données de Rain

L'objectif sur ces données est de savoir si il existe (et dans le cas écheant de le calculer) un seuil tel que les pluies ne puisse pas depasser. Chercher cette valeur seuil serait utilse en agriculure par exemple pour savoir si les pluies ne sont pas trop élevées pour les cultures.

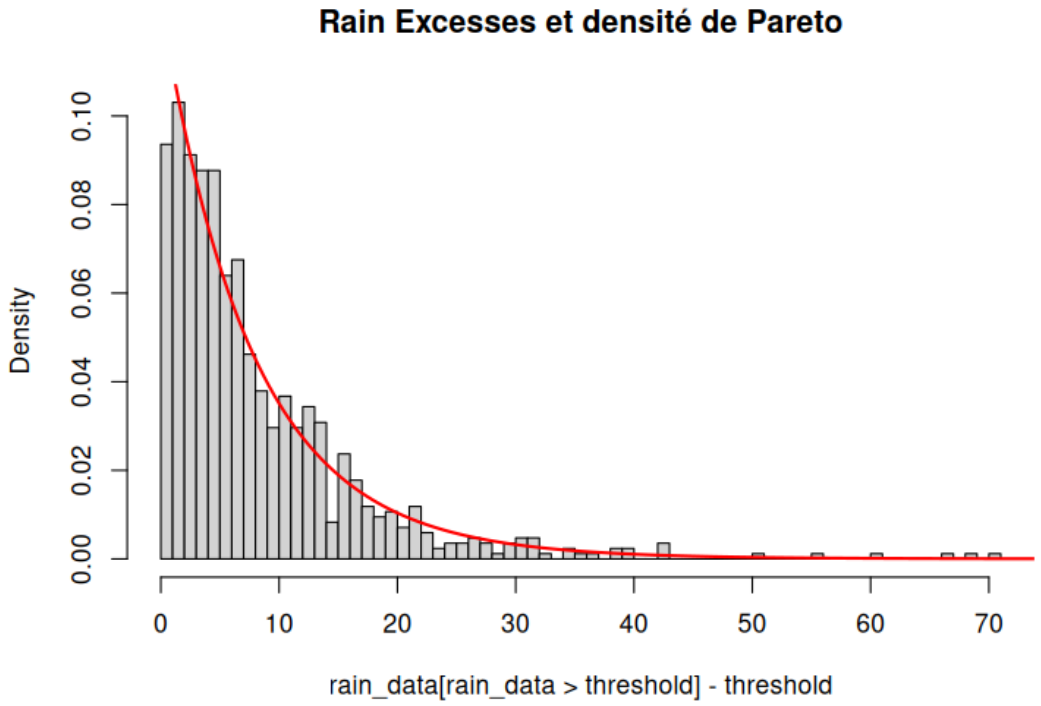


On remarque dans un premier temps que les données sont concentrés autour de 0. Mais qu'elles sont capable de prendre des données très élevées. Il est alors raisonnables de penser qu'apres estimation, on va obtenir une valeur de gamma positive ou nulle. En effet, il n'apparait pas de cassure dansla distribution des données. De plus, la queue de distribution est longue mais ne parait pas lourde. Ce qui suggererait une valeur de gamma proche de 0.

Après estimation numériques, on obtient : $\sigma = 7.94$ et $\gamma = 0.034$ avec pour γ un intervalle de confiance : $[-0.022; 0.102]$.

Une valeur de gamma aussi proche de 0 doit nous conduire à une étude plus approfondie. Plusieurs méthode s'offre à nous pour améliorer l'estimation de gamma.
ON Peut considere la premiere méthode afin de comparer les résultats.
On peut faire varier le seuil u est juger de l'impact sur l'estimation de gamma.
Ou alors de façon plus arbitraire, on peut considère de la valeur de gamma en fonction du type de donnée qu'on étudie et de la cohérence que cela apporte.

Pour notre exemple, on considère que $\gamma > 0$



La distribution de Pareto (courbe rouge) avec les parametres estimmés semblent bien coller avec les données. Cela signifie qu'on s'attend à des excès au-delà du seuil de plus en plus rares, sans pour autant exclure la survenue de précipitations sensiblement élevées,

L'avantage de cette méthode est qu'elle est plus efficace car elle utilise plus de données. Cependant, elle est plus difficile à mettre en place car il faut choisir un seuil u qui est crucial pour l'estimation de gamma.

7.2.3 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 γ ¿ 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 $x_c = x_0 + \rho(x_3 - x_0)$. 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 σ et de centre x_1 : on remplace ainsi x_i par $x_1 + \sigma(x_i - x_1)$ et on retourne à l'étape 2

On répète cela jusqu'à atteinte du critère d'arrêt, à définir.

8 Annexe

8.1 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)

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)
```

```r

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)
```

```r
#################### 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")

# on trace l'histogramme des données en excés par rapport au seuil et la loi de pareto
hist(rain_data[rain_data > threshold] - threshold, breaks = 50, freq = FALSE, main = "
    Rain Excesses et densité de Pareto")

# on trace la loi de gpd avec les paramètres estimés
xseq <- seq(min(rain), max(rain), length.out = 200)
lines(xseq, pareto(xseq, gamma, sigma),col='red', lwd=2)



# pour le qq-plot et residus
gpd.diag(gpd_result)
```