Chapitre 2

Lois de valeurs extrêmes

Le 1erFévrier 1953, lors d'une forte tempête la mer passe par-dessus plusieurs digues aux Pays-Bas, les détruit et inonde la région. Il s'agit d'un accident majeur. Un comité est mis en place pour étudier le phénomène et proposer des recommandations sur les hauteurs de digues. Il tient compte des facteurs économiques (coût de construction, coût des inondations,...), des facteurs physiques (rôle du vent sur la marée,...), et aussi des données enregistrées sur les hauteurs de marées. En fait il est plus judicieux de considérer les surcotes, c'est-à-dire la différence entre la hauteur réelle et la hauteur prévue de la marée, que les hauteurs des marées. En effet, on peut supposer, dans une première approximation, que les surcotes des marées lors des tempêtes sont des réalisations de variables aléatoires de même loi. Si on regarde les surcotes pour des marées de tempêtes séparées par quelques jours d'accalmie, on peut même supposer que les variables aléatoires sont indépendantes. L'étude statistique sur des surcotes a pour but de répondre aux questions suivantes :

- Soit $q \in]0,1[$ fixé, trouver h tel que la probabilité pour que la surcote soit supérieure à h est q.
- Soit $q \in]0,1[$ fixé, typiquement de l'ordre de 10^{-3} ou 10^{-4}, trouver h tel que la probabilité pour que la plus haute surcote annuelle soit supérieure à h est q.

Nous recommandons la lecture de l'article écrit par de Haan [4] sur ce cas particulier.

Si F désigne la fonction de répartion de la loi des surcotes ($F(x)$ est la probabilité que la surcote soit plus petite que x), alors dans la première question, h est le quantile d'ordre $p = 1 - q \in]0,1[$ de la loi des surcotes : $h = \inf\{x \in \mathbb{R}; F(x) \geq 1 - q\}$. Bien sûr la loi des surcotes et donc les quantiles associés sont inconnus.

Pour simplifier l'écriture on notera $z_q = x_p$ le quantile d'ordre $p = 1 - q$. Il existe plusieurs méthodes pour donner un estimateur $\hat{x}_p$ (resp. $\hat{z}_q$) de x_p (resp. z_q). Nous en présentons trois familles : l'estimation paramétrique, les quantiles empiriques et l'utilisation des lois de valeurs extrêmes.

Estimation paramétrique. Supposons que l'on sache **a priori** que la loi des surcotes appartient à une famille paramétrique de lois, par exemple la famille des lois exponentielles. Bien sûr la vraie valeur du paramètre est inconnue. Le quantile d'ordre $1 - q$ de la loi exponentielle de paramètre $\lambda > 0$ est donné par $z_q = -\log(q)/\lambda$. Dans ce modèle élémentaire, on observe que l'estimation du quantile peut se réduire à l'estimation du paramètre λ. On peut vérifier que si $(X_k, k \geq 1)$ est une suite de variables aléatoires indépendantes de loi exponentielle de paramètre $\lambda > 0$, alors l'estimateur du maximum de vraisemblance de $1/\lambda$ est donné par la moyenne empirique : $\frac{1}{n}\sum_{k=1}^{n} X_k$. De plus cet estimateur est convergent d'après la loi forte des grands nombres. On en déduit que $\hat{z}_q = -\dfrac{\log(q)}{n}\displaystyle\sum_{k=1}^{n} X_k$ est un estimateur convergent de z_q : p.s. $\lim_{n\to\infty} \hat{z}_q = z_q$. Ces résultats semblent satisfaisants, mais ils reposent très fortement sur le choix initial de la famille paramétrique. Supposons que ce choix soit erroné, et que la loi de X_k soit par exemple la loi de $|G|$, où G suit la loi gaussienne centrée réduite, $\mathcal{N}(0,1)$. Dans ce cas, on a par la loi forte des grands nombres $\lim_{n\to\infty} \hat{z}_q = -\log(q)\mathbb{E}[|G|] =$

$-\log(q)\sqrt{2/\pi}$, alors que la vraie valeur de z_q est définie par $2\int_0^{z_q} \mathrm{e}^{-u^2/2}\,\frac{du}{\sqrt{2\pi}} = 1-q$. Si on trace la valeur du quantile en fonction de $q = 1/y$, pour le vrai modèle qui correspond à la loi de $|G|$, et le modèle erroné, qui correspond à la loi exponentielle de paramètre $\lambda = \frac{1}{\mathbb{E}[|G|]} = \sqrt{\pi/2}$, on obtient la figure 2.1. Pour les faibles valeurs de $q = 1/y$, c'est-à-dire pour le comportement de la "queue" de la distribution, l'erreur sur le modèle entraîne des erreurs très importantes sur l'estimation des quantiles.

En conclusion l'estimation des quantiles à partir d'un modèle paramétrique est très sensible au choix a priori de la famille paramétrique de lois. Cette méthode n'est pas fiable.

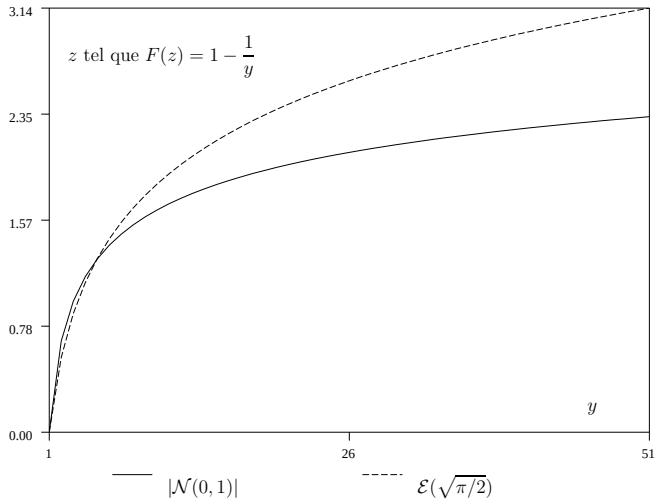

FIG. 2.1 – Quantile d'ordre $1-\frac{1}{y}$ pour la loi de $|G|$, où G suit la loi $\mathcal{N}(0,1)$, et pour la loi exponentielle de paramètre $\sqrt{\pi/2}$.

Quantile empirique. Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes de même loi. On note $X_{(1,n)} \leq \cdots \leq X_{(n,n)}$ leur réordonnement aléatoire croissant appelé statistique d'ordre. Nous suivons ici l'usage en vigueur pour les notations sur les statistiques d'ordre. Dans certains livres sur les lois de valeurs extrêmes, on considère le réordonnement décroissant.

Soit $p \in]0,1[$. Nous montrerons, voir la proposition 2.1.8, que le quantile empirique $X_{([pn]+1,n)}$, où $[pn]$ désigne la partie entière de pn, est un estimateur qui converge presque sûrement vers le quantile d'ordre p : x_p. En outre, on connaît les comportements possibles de cet estimateur en fonction des caractéristiques de la fonction de répartition : convergence, comportement asymptotique, intervalle de confiance. La figure 2.2 représente une simulation de l'évolution de la médiane empirique, i.e. $X_{([n/2],n)}$, et de l'intervalle de confiance associé, en fonction de la taille $n \geq 2$ de l'échantillon pour des variables aléatoires indépendantes de loi de Cauchy (de paramètre $a = 1$). L'intervalle de confiance est donné par la proposition 2.1.13. Rappelons que pour une loi de Cauchy, la densité est $1/(\pi(1+x^2))$, et la médiane est $x_{1/2} = 0$.

Pour tout $p > 0$ tel que $pn < 1$, l'estimateur de x_p est $X_{(1,n)}$. L'estimation est clairement mauvaise. Intuitivement, si $n < 1/p$, il n'y a pas assez d'observations pour apprécier des événements de probabilité p. Un problème similaire se pose si $n < 1/(1-p)$. L'estimation du quantile d'ordre p, par le quantile empirique, est pertinente lorsque l'on dispose de nombreuses données. Ceci correspond aux cas où $1 - \dfrac{1}{n} \gg p \gg \dfrac{1}{n}$.

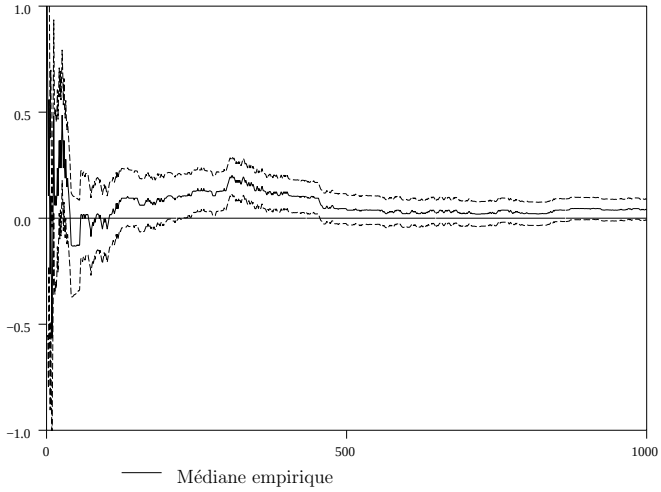

FIG. 2.2 – Médiane empirique, $n \to X_{([n/2],n)}$, et bornes (en pointillés) de l'intervalle de confiance de niveau asymptotique 95% pour la médiane de la loi de Cauchy.

Pour les digues des Pays-Bas, il faut remonter jusqu'en 1570 pour retrouver une marée comparable à celle du 1$^{\text{er}}$Février 1953. Une estimation de la probabilité d'observer une marée au moins aussi forte pendant une année est grossièrement de l'ordre de $\dfrac{1}{1953-1570} \sim 3.10^{-3}$. Les valeurs typiques pour p ou q sont de l'ordre de 10^{-3} ou 10^{-4}. On cherche donc à estimer la probabilité d'un événement que l'on n'a jamais vu ! De plus on ne dispose de données fiables que depuis la fin du $XIX^{\text{ème}}$ siècle. La méthode du quantile empirique est inutilisable en pratique.

Lois de valeurs extrêmes. Les deux méthodes précédentes ne sont pas fiables. Mais il ne faut pas se faire d'illusions : **il n'existe pas de solution miracle pour répondre aux questions posées quand on dispose de peu ou pas de données** dans la région d'intérêt. L'utilisation des lois de valeurs extrêmes repose sur les propriétés des statistiques d'ordre et sur des méthodes d'extrapolation. Plus précisément, elle repose sur des convergences en loi des maximums de variables aléatoires convenablement renormalisés. Les lois limites possibles sont connues. Elles sont appelées les lois de valeurs extrêmes. L'estimation de x_p se déroulera en deux étapes :

- Identification de la loi de valeurs extrêmes associée aux données.
- Estimation des paramètres de renormalisation.

Dans le paragraphe 2.1 nous donnons des résultats sur les statistiques d'ordre et les quantiles empiriques. Puis, dans le paragraphe 2.2, nous étudions sur des exemples la convergence du maximum renormalisé de variables aléatoires indépendantes et de même loi. Au paragraphe 2.3, nous caractérisons les lois limites, à un facteur de translation et d'échelle près, appelées lois de valeurs extrêmes. Nous donnons ensuite, au paragraphe 2.4, des critères pour que la limite en loi du maximum renormalisé suive telle ou telle loi de valeurs extrêmes. Au paragraphe 2.5, nous donnons deux méthodes statistiques qui permettent d'identifier la loi limite. Puis, nous présentons au paragraphe 2.6 des estimateurs des paramètres de renormalisation qui permettent de fournir des estimations des quantiles : les estimateurs de Hill et les estimateurs de Pickand. Enfin nous répondons aux deux

questions initiales de l'introduction au paragraphe 2.7.

Il existe de nombreux autres estimateurs des quantiles extrêmes. Tous reposent sur des méthodes d'extrapolation. Il faut être conscient des limites de la théorie. En particulier, il est conseillé de ne pas reposer son analyse sur un seul estimateur, mais plutôt de les utiliser ensemble et de vérifier s'ils donnent des résultats concordants.

La recherche autour des lois de valeurs extrêmes est particulièrement active depuis les années 1970. Nous renvoyons aux ouvrages de Beirlant and al. [1], Coles [3], Embrecht and al. [6] et Falk and al. [7] pour une approche détaillée de la théorie des lois de valeurs extrêmes, ainsi que pour des références concernant les applications de cette théorie : en hydrologie, comme le montre l'exemple du début de ce chapitre, en assurance et en finance pour les calculs de risques, en météorologie pour les événements extrêmes, etc.

2.1 Statistique d'ordre, estimation des quantiles

La fonction caractéristique est un des outils fondamentaux pour démontrer la convergence de sommes renormalisées de variables aléatoires indépendantes, comme dans les démonstrations classiques du théorème central limite. Pour l'analyse des statistiques d'ordre et des convergences en loi des maximums renormalisés, un des outils fondamentaux est la fonction de répartition.

Soit $(X_n, n \geq n)$ une suite de variables aléatoires réelles indépendantes identiquement distribuées de fonction de répartition F ($F(x) = \mathbb{P}(X_n \leq x)$, pour $x \in \mathbb{R}$). Soit $\mathcal{S}_n$ l'ensemble des permutation de $\{1, \ldots, n\}$.

Définition 2.1.1. *La **statistique d'ordre** de l'échantillon $(X_1, \ldots, X_n)$ est le réarrangement croissant de $(X_1, \ldots, X_n)$. On la note $(X_{(1,n)}, \ldots, X_{(n,n)})$. On a $X_{(1,n)} \leq \ldots \leq X_{(n,n)}$, et il existe une permutation aléatoire $\sigma_n \in \mathcal{S}_n$ telle que $(X_{(1,n)}, \ldots, X_{(n,n)}) = (X_{\sigma_n(1)}, \ldots, X_{\sigma_n(n)})$.*

En particulier on a $X_{(1,n)} = \min_{1 \leq i \leq n} X_i$ et $X_{(n,n)} = \max_{1 \leq i \leq n} X_i$.

On note F^{-1} l'inverse généralisé de F. Le lemme suivant découle de la croissance de la fonction F.

Lemme 2.1.2. *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 sur $[0, 1]$. Alors $(F^{-1}(U_{(1,n)}), \ldots, F^{-1}(U_{(n,n)}))$ a même loi que $(X_{(1,n)}, \ldots, X_{(n,n)})$.*

On suppose dorénavant que F est continue.

Lemme 2.1.3. *Si F est continue, alors p.s. on a $X_{(1,n)} < \ldots < X_{(n,n)}$.*

Démonstration. Il suffit de vérifier que $\mathbb{P}(\exists i \neq j \quad \text{tel que} \quad X_i = X_j) = 0$. On a

$$\begin{aligned}\mathbb{P}(\exists i \neq j \quad \text{tel que} \quad X_i = X_j) &\leq \mathbb{P}(\exists i \neq j \quad \text{tel que} \quad F(X_i) = F(X_j)) \\ &\leq \sum_{i \neq j} \mathbb{P}(F(X_i) = F(X_j)).\end{aligned}$$

Les variables $F(X_i)$ et $F(X_j)$ sont pour $i \neq j$ des variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. On en déduit que

$$\mathbb{P}(F(X_i) = F(X_j)) = \int_{[0,1]^2} \mathbf{1}_{\{u=v\}} du dv = 0.$$

Donc p.s. pour tous $i \neq j$, on a $X_i \neq X_j$. □

Corollaire 2.1.4. *Si F est continue, la permutation aléatoire σ_n de la définition 2.1.1 est p.s. unique.*

Lemme 2.1.5. *On suppose F continue. La loi de σ_n est la loi uniforme sur $\mathcal{S}_n$. De plus la permutation σ_n est indépendante de la statistique d'ordre.*

Démonstration. Soit $\sigma \in \mathcal{S}_n$. On a $\mathbb{P}(\sigma_n = \sigma) = \mathbb{P}(X_{\sigma(1)} < \ldots < X_{\sigma(n)})$. Les variables $X_{\sigma(1)}, \ldots, X_{\sigma(n)}$ sont indépendantes et de même loi. En particulier le vecteur $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ a même loi que $(X_1, \ldots, X_n)$. Il vient

$$\mathbb{P}(\sigma_n = \sigma) = \mathbb{P}(X_1 < \ldots < X_n).$$

Le membre de droite est indépendant de σ. La loi de σ_n est donc la loi uniforme sur $\mathcal{S}_n$, et on a $\mathbb{P}(\sigma_n = \sigma) = \dfrac{1}{n!}$.

Soit g une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$, mesurable bornée. Comme le vecteur $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ a même loi que $(X_1, \ldots, X_n)$, on a

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\{\sigma_n=\sigma\}}\, g(X_{(1,n)}, \ldots, X_{(n,n)})\right] &= \mathbb{E}\left[\mathbf{1}_{\{X_{\sigma(1)}<\ldots<X_{\sigma(n)}\}}\, g(X_{\sigma(1)}, \ldots, X_{\sigma(n)})\right] \\ &= \mathbb{E}\left[\mathbf{1}_{\{X_1<\ldots<X_n\}}\, g(X_1, \ldots, X_n)\right].\end{aligned}$$

On en déduit, en sommant sur $\sigma \in \mathcal{S}_n$, que

$$\mathbb{E}\left[g(X_{(1,n)}, \ldots, X_{(n,n)})\right] = n!\, \mathbb{E}\left[\mathbf{1}_{\{X_1<\ldots<X_n\}}\, g(X_1, \ldots, X_n)\right]. \tag{2.1}$$

Enfin, on remarque que

$$\mathbb{E}\left[\mathbf{1}_{\{\sigma_n=\sigma\}} g(X_{(1,n)}, \ldots, X_{(n,n)})\right] = \mathbb{P}(\sigma_n = \sigma)\mathbb{E}\left[g(X_{(1,n)}, \ldots, X_{(n,n)})\right].$$

Cela implique que la permutation σ_n est indépendante de la statistique d'ordre. □

Le corollaire suivant découle de (2.1).

Corollaire 2.1.6. *Si la loi de X_1 possède une densité f, alors, la statistique d'ordre $(X_{(1,n)}, \ldots, X_{(n,n)})$ possède la densité $n!\mathbf{1}_{\{x_1<\cdots<x_n\}}\, f(x_1) \ldots f(x_n)$.*

On peut déduire de (2.1) la loi de $X_{(k,n)}$. Mais nous préférons présenter une méthode que nous utiliserons plusieurs fois dans ce paragraphe. Soit $x \in \mathbb{R}$ fixé. Les variables aléatoires $(\mathbf{1}_{\{X_i \le x\}}, i \ge 1)$ sont des variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $\mathbb{P}(X_i \le x) = F(x)$. La variable aléatoire $S_n(x) = \sum_{i=1}^n \mathbf{1}_{\{X_i \le x\}}$ suit donc la loi binomiale de paramètre (n, p), où $p = F(x)$. Remarquons enfin que l'on a $S_n(x) \ge k$ si et seulement si parmi les variables $X_1, \ldots, X_n$, au moins k sont plus petites que x, c'est-à-dire si et seulement si $X_{(k,n)} \le x$. Ainsi il vient

$$\{X_{(k,n)} \le x\} = \{k \le S_n(x)\}. \tag{2.2}$$

On en déduit que

$$\mathbb{P}(X_{(k,n)} \le x) = \mathbb{P}(S_n(x) \ge k) = \sum_{r=k}^{n} \binom{n}{r} F(x)^r \left(1 - F(x)\right)^{n-r}. \tag{2.3}$$

Intuitivement, pour que $X_{(k,n)}$ appartienne à $[y, y + dy]$, il faut :
- qu'il existe i, parmi n choix possibles, tel que $X_i \in [y, y + dy]$; ceci est de probabilité $nf(y)dy$, si la loi de X_k possède une densité f ;
- choisir $k - 1$ variables aléatoires parmi les $n - 1$ restantes, qui sont plus petites que y, ceci est de probabilité $\binom{n-1}{k-1} F(y)^{k-1}$;
- que les $n-k$ autres variables aléatoires soient plus grandes que y, ceci est de probabilité $F(y)^{n-k}$.

Il vient $\mathbb{P}(X_{(k,n)} \le x) = \dfrac{n!}{(k-1)!(n-k)!} \displaystyle\int_0^x F(y)^{k-1}(1 - F(y))^{n-k} f(y)\, dy$. Avec le changement de variable $t = F(y)$, on obtient

$$\mathbb{P}(X_{(k,n)} \le x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k}\, dt. \tag{2.4}$$

Le but de l'exercice suivant est de donner une démonstration de (2.4).

Exercice 2.1.7. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires réelles indépendantes et de même loi possédant une fonction de répartition continue.

1. Montrer, par récurrence descendante, la formule (2.4) en utilisant (2.3).
2. Vérifier que $F(X_{(k,n)})$ suit la loi béta de paramètre $(k, n-k+1)$.
3. Dans le cas où la loi de X_1 possède la densité f, retrouver ces résultats en calculant la densité de loi marginale de $X_{(k,n)}$ à partir de la densité de la loi de la statistique d'ordre donnée dans le corollaire 2.1.6. Puis vérifier que la densité de la loi de $X_{(n,n)}$ est donnée par $nF(x)^{n-1}f(x)$.

⧫

Nous donnons le résultat principal de ce paragraphe sur la convergence du quantile empirique.

Proposition 2.1.8. *Soit $p \in]0,1[$. Supposons que F est continue et qu'il existe une seule solution x_p à l'équation $F(x) = p$. Soit $(k(n), n \geq 1)$ une suite d'entiers telle que $1 \leq k(n) \leq n$ et $\lim_{n\to\infty} \frac{k(n)}{n} = p$. Alors, la suite des quantiles empiriques $(X_{(k(n),n)}, n \geq 1)$ converge presque sûrement vers x_p.*

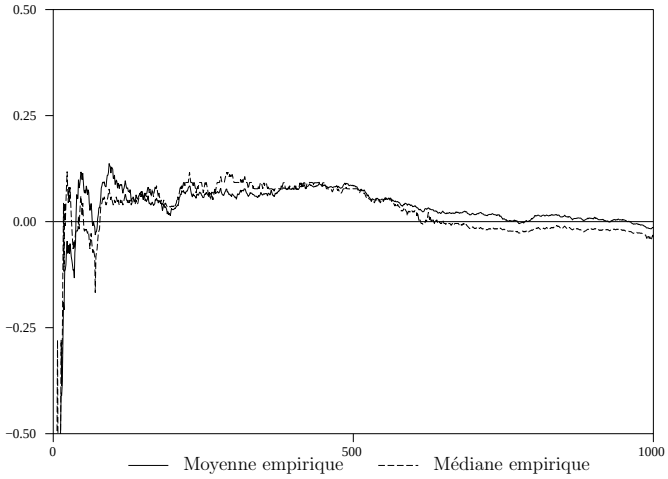

FIG. 2.3 – Moyenne empirique et médiane empirique pour la loi $\mathcal{N}(0,1)$ en fonction de la taille de l'échantillon.

Exemple 2.1.9. Soit $(X_i, i \geq 1)$ une suite de variables aléatoires gaussiennes de loi $\mathcal{N}(m, \sigma^2)$, où la moyenne m et la variance σ^2 sont inconnues. On désire estimer m. Comme $\mathbb{E}[X_1] = m$, on en déduit, par la loi forte des grands nombres, que la moyenne empirique $\frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur convergent de m. On peut aussi remarquer que m est la médiane, i.e. le quantile d'ordre $1/2$, de la loi de X_1. On en déduit que la médiane empirique $X_{([n/2],n)}$ est un estimateur convergent de m. La figure 2.3 représente l'évolution de la moyenne empirique et de la médiane empirique en fonction de la taille n de l'échantillon, pour la loi gaussienne centrée réduite. Remarquons que pour calculer la médiane empirique en fonction de n, il faut conserver toutes les valeurs de l'échantillon en mémoire, ce qui n'est pas le cas pour la moyenne empirique. En revanche si, suite à une erreur, une donnée erronée se glisse dans les données, on peut alors vérifier que la médiane empirique est moins sensible à cette erreur que la moyenne empirique. ◊

Démonstration de la proposition 2.1.8. Soit $x \in \mathbb{R}$ fixé. On rappelle la notation $S_n(x) = \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}$. On déduit de l'égalité (2.2) que

$$\left\{X_{(k(n),n)} \leq x \quad \text{à partir d'un certain rang}\right\} = \left\{1 \leq \frac{S_n(x)}{k(n)} \quad \text{à partir d'un certain rang}\right\}.$$

La loi forte des grands nombres assure que $\lim_{n\to\infty} \dfrac{S_n(x)}{n} = \mathbb{E}\left[\mathbf{1}_{\{X_i \leq x\}}\right] = F(x)$ presque sûrement. De plus on a $\lim_{n\to\infty} \dfrac{n}{k(n)} = \dfrac{1}{p}$ et donc $\lim_{n\to\infty} \dfrac{S_n(x)}{k(n)} = \lim_{n\to\infty} \dfrac{S_n(x)}{n}\dfrac{n}{k(n)} = \dfrac{F(x)}{p}$ presque sûrement. En particulier, on a

$$\mathbb{P}\left(1 \leq \frac{S_n(x)}{k(n)} \text{ à partir d'un certain rang}\right) = \begin{cases} 0 & \text{si } F(x) < p, \text{ i.e. si } x < x_p, \\ 1 & \text{si } F(x) > p, \text{ i.e. si } x > x_p. \end{cases}$$

Cela implique donc que

$$\mathbb{P}\left(X_{(k(n),n)} \leq x \quad \text{à partir d'un certain rang}\right) = \begin{cases} 0 & \text{si } x < x_p, \\ 1 & \text{si } x > x_p. \end{cases}$$

Cela signifie que p.s. $\lim_{n\to\infty} X_{(k(n),n)} = x_p$. □

Proposition 2.1.10. *Soit $p = 1$ (resp. $p = 0$). Soit $(k(n), n \geq 1)$ une suite d'entiers telle que $1 \leq k(n) \leq n$ et $\lim_{n\to\infty} \dfrac{k(n)}{n} = p$. Alors la suite $(X_{(k(n),n)}, n \geq 1)$ converge presque sûrement vers $x_F = \inf\{x; F(x) = 1\}$ (resp. $\tilde{x}_F = \sup\{x; F(x) = 0\}$), avec la convention $\inf \emptyset = +\infty$ (resp. $\sup \emptyset = -\infty$).*

Démonstration. Pour $p = 1$, un raisonnement similaire à celui de la démonstration de la proposition précédente assure que p.s. $\liminf_{n\to\infty} X_{(k(n),n)} \geq x_F$. Par définition de x_F, on a p.s. $X_n \leq x_F$ pour tout $n \geq 1$. Ceci assure donc que $\lim_{n\to\infty} X_{(k(n),n)} = x_F$.

Pour $p = 0$, en considérant les variables $Y_k = -X_k$, on est ramené au cas $p = 1$. □

Dans certains cas, on peut donner un intervalle de confiance pour le quantile empirique.

Proposition 2.1.11. *Soit $p \in]0,1[$. Supposons que la loi de X_1 possède une densité, f continue en x_p et telle que $f(x_p) > 0$. On suppose de plus que $k(n) = np + o\left(\sqrt{n}\right)$. On a la convergence en loi suivante :*

$$\boxed{\sqrt{n}\left(X_{(k(n),n)} - x_p\right) \xrightarrow[n\to\infty]{\textbf{Loi}} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right).}$$

Soit $\alpha > 0$. L'intervalle aléatoire $\left[X_{(k(n),n)} \pm \dfrac{a_\alpha \sqrt{p(1-p)}}{f(X_{(k(n),n)})\sqrt{n}}\right]$, où a_α est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$, est un intervalle de confiance pour x_p, de niveau asymptotique $1 - \alpha$.

Le graphique 2.2 représente l'évolution de la médiane empirique et de l'intervalle de confiance de niveau asymptotique 95% associé pour des variables aléatoires indépendantes de loi de Cauchy en fonction de la taille de l'échantillon. Rappelons que pour une loi de Cauchy, la densité est $1/(\pi(1+x^2))$, et la médiane est 0.

Démonstration. Notons $k = k(n)$. Soit $x \in \mathbb{R}$ fixé. On pose $y_n = x_p + \dfrac{x}{\sqrt{n}}$ et $p_n = F(y_n)$. Comme la densité est continue, on a $p_n - p = \displaystyle\int_{x_p}^{y_n} f(u)du = \frac{x f(x_p)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$.

On rappelle (2.2). La démonstration repose sur le fait que

$$\sqrt{n}\left(X_{(k,n)}-x_p\right)\le x \Leftrightarrow X_{(k,n)}\le y_n \Leftrightarrow S_n(y_n)\ge k \Leftrightarrow V_n \ge \sqrt{n}\left(\frac{k}{n}-p_n\right),$$

où $V_n=\sqrt{n}\left(\dfrac{S_n(y_n)}{n}-p_n\right)$. En utilisant les fonctions caractéristiques, on a

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{iuV_n}\right] &= \mathbb{E}\left[\mathrm{e}^{iu\sqrt{n}\left(\frac{S_n(y_n)}{n}-p_n\right)}\right]\\
&= \mathbb{E}\left[\mathrm{e}^{iu\sum_{j=1}^n\left(\mathbf{1}_{\{X_j\le y_n\}}-p_n\right)/\sqrt{n}}\right]\\
&= \mathbb{E}\left[\mathrm{e}^{iu(\mathbf{1}_{\{X_1\le y_n\}}-p_n)/\sqrt{n}}\right]^n\\
&= \left[p_n\,\mathrm{e}^{iu(1-p_n)/\sqrt{n}}+(1-p_n)\,\mathrm{e}^{-iup_n/\sqrt{n}}\right]^n.
\end{aligned}$$

On fait un développement limité du dernier terme entre crochets. On a

$$\begin{aligned}
&p_n\,\mathrm{e}^{iu(1-p_n)/\sqrt{n}}+(1-p_n)\,\mathrm{e}^{-iup_n/\sqrt{n}}\\
&\qquad= p_n\left(1+\frac{iu}{\sqrt{n}}(1-p_n)-\frac{u^2}{2n}(1-p_n)^2+o(n^{-2})\right)\\
&\qquad\qquad+(1-p_n)\left(1-\frac{iu}{\sqrt{n}}p_n-\frac{u^2}{2n}p_n^2+o(n^{-2})\right)\\
&\qquad= 1-\frac{u^2}{2n}p_n(1-p_n)+o(n^{-2})\\
&\qquad= 1-\frac{u^2}{2n}p(1-p)+O(n^{-3/2}).
\end{aligned}$$

Il en découle donc que

$$\mathbb{E}\left[\mathrm{e}^{iuV_n}\right]=\left[1-\frac{u^2}{2n}p(1-p)+O(n^{-3/2})\right]^n.$$

On en déduit que

$$\lim_{n\to\infty}\left[1-\frac{u^2}{2n}p(1-p)+O(n^{-3/2})\right]^n=\lim_{n\to\infty}\left[1-\frac{u^2}{2n}p(1-p)\right]^n=\mathrm{e}^{-u^2p(1-p)/2}.$$

Donc la suite $(V_n, n\ge 1)$ converge en loi vers V, de loi gaussienne $\mathcal{N}(0,p(1-p))$. Comme de plus on a $\lim\limits_{n\to\infty}\sqrt{n}\left(\dfrac{k}{n}-p_n\right)=-xf(x_p)$, on déduit du théorème de Slutsky que pour $x\neq 0$, $V_n/\sqrt{n}\left(\frac{k}{n}-p_n\right)$ converge en loi vers $V/(-xf(x_p))$, qui est une variable aléatoire continue. On a donc

$$\mathbb{P}\left(\sqrt{n}\left(X_{(k,n)}-x_p\right)\le x\right)=\mathbb{P}\left(V_n\ge\sqrt{n}\left(\frac{k}{n}-p_n\right)\right)\underset{n\to\infty}{\longrightarrow}\mathbb{P}\left(V\ge -xf(x_p)\right).$$

En utilisant le fait que V et $-V$ ont même loi il vient $\mathbb{P}\left(V\ge -xf(x_p)\right)=\mathbb{P}\left(\frac{V}{f(x_p)}\le x\right)$. Cela implique que la fonction de répartition de la loi de $\sqrt{n}\left(X_{(k,n)}-x_p\right)$ converge vers celle de $V/f(x_p)$ et donc vers la fonction de répartition de la loi $\mathcal{N}(0,p(1-p)/f(x_p)^2)$. Ceci implique la première partie de la proposition.

Comme la densité f est continue en x_p, on déduit de la proposition 2.1.8 que p.s. $\lim_{n\to\infty} f(X_{(k,n)})=f(x_p)$. Le théorème de Slutsky assure que

$$\sqrt{n}\frac{f(X_{(k,n)})}{\sqrt{p(1-p)}}(X_{(k,n)}-x_p)\ \xrightarrow[n\to\infty]{\textbf{Loi}}\ \mathcal{N}(0,1).$$

On en déduit que l'intervalle aléatoire $\left[X_{(k(n),n)} \pm \frac{a_\alpha\sqrt{p(1-p)}}{f(X_{(k(n),n)})\sqrt{n}}\right]$, où a_α est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$, est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour x_p. □

Remarque 2.1.12. Si la loi ne possède pas de densité, ou si la densité est irrégulière, la vitesse de convergence du quantile empirique vers le quantile peut être beaucoup plus rapide que $1/\sqrt{n}$. En revanche, si la densité existe, est continue et si $f(x_p) = 0$, alors la vitesse de convergence peut être plus lente que $1/\sqrt{n}$. ◊

Dans la proposition 2.1.11 intervient la densité de la loi. Or, en général, si on cherche à estimer un quantile, il est rare que l'on connaisse la densité. On peut construire un autre intervalle de confiance pour x_p sous des hypothèses plus générales, qui ne fait pas intervenir la densité. Si les hypothèses de la proposition 2.1.11 sont vérifiées, alors la largeur aléatoire de cet intervalle de confiance est de l'ordre de $1/\sqrt{n}$.

Proposition 2.1.13. *Soit $p \in]0,1[$. Soit a_α le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. On considère les entiers $i_n = [np - \sqrt{n}a_\alpha\sqrt{p(1-p)}]$ et $j_n = [np + \sqrt{n}a_\alpha\sqrt{p(1-p)}]$. Pour n assez grand les entiers i_n et j_n sont compris entre 1 et n. De plus l'intervalle aléatoire $[X_{(i_n,n)}, X_{(j_n,n)}]$ est un intervalle de confiance pour x_p de niveau asymptotique $1-\alpha$.*

Démonstration. On pose $Z_n = \sqrt{n}\frac{\frac{1}{n}S_n - p}{\sqrt{p(1-p)}}$, où $S_n = \sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x_p\}}$. On déduit du théorème central limite que la suite $(Z_n, n \ge 1)$ converge en loi vers Z de loi gaussienne centrée réduite. Pour n suffisamment grand, on a $1 \le i_n \le j_n \le n$, et

$$\begin{aligned}\mathbb{P}(X_{(i_n,n)} \le x_p \le X_{(j_n,n)}) &= \mathbb{P}(i_n \le S_n \le j_n)\\ &= \mathbb{P}\left(\sqrt{n}\frac{\frac{1}{n}i_n - p}{\sqrt{p(1-p)}} \le Z_n \le \sqrt{n}\frac{\frac{1}{n}j_n - p}{\sqrt{p(1-p)}}\right).\end{aligned}$$

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

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

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

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\frac{\frac{1}{n}i_n - p}{\sqrt{p(1-p)}} \le Z_n \le \sqrt{n}\frac{\frac{1}{n}j_n - p}{\sqrt{p(1-p)}}\right) = \mathbb{P}(-a_\alpha \le Z \le a_\alpha) = 1-\alpha.$$

On a donc obtenu

$$\lim_{n\to\infty} \mathbb{P}(X_{(i_n,n)} \le x_p \le X_{(j_n,n)}) = 1-\alpha.$$

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

La suite de ce paragraphe est consacrée à des résultats qui seront utiles dans les paragraphes suivants. Le lemme 2.1.2 assure que l'étude de la statistique d'ordre associée à une loi quelconque

peut se déduire de l'étude de la statistique d'ordre associée à la loi uniforme sur $[0,1]$. Nous donnons une représentation de cette dernière à l'aide de variables aléatoires de loi exponentielle.

Soit $(E_i, i \geq 1)$ une suite de variables aléatoires de loi exponentielle de paramètre 1. On note $\Gamma_n = \sum_{i=1}^n E_i$. La variable aléatoire Γ_n suit la loi gamma de paramètre $(1,n)$. Soit $(U_1, \ldots, , U_n)$ une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$.

Lemme 2.1.14. *La variable aléatoire $(U_{(1,n)}, \ldots, U_{(n,n)})$ à même loi que $\left(\dfrac{\Gamma_1}{\Gamma_{n+1}}, \ldots, \dfrac{\Gamma_n}{\Gamma_{n+1}}\right)$.*

Démonstration. Soit g une fonction réelle mesurable bornée définie sur $\mathbb{R}^n$. On a

$$\mathbb{E}\left[g\left(\frac{\Gamma_1}{\Gamma_{n+1}}, \ldots, \frac{\Gamma_n}{\Gamma_{n+1}}\right)\right] = \int_{(\mathbb{R}_+^*)^{n+1}} g\left(\frac{x_1}{\sum_{i=1}^{n+1} x_i}, \ldots, \frac{\sum_{j=1}^n x_j}{\sum_{i=1}^{n+1} x_i}\right) \mathrm{e}^{-\sum_{i=1}^{n+1} x_i}\, dx_1 \ldots dx_{n+1}.$$

En considérant les changements successifs de variables $y_1 = x_1, y_2 = x_1 + x_2, \ldots, y_{n+1} = \sum_{i=1}^{n+1} x_i$ puis $z_1 = y_1/y_{n+1}, \ldots, z_n = y_n/y_{n+1}, z_{n+1} = y_{n+1}$, on obtient après intégration sur z_{n+1},

$$\mathbb{E}\left[g\left(\frac{\Gamma_1}{\Gamma_{n+1}}, \ldots, \frac{\Gamma_n}{\Gamma_{n+1}}\right)\right] = n! \int \mathbf{1}_{\{0<z_1<\cdots<z_n<1\}}\, g(z_1, \ldots, z_n)\, dz_1 \ldots dz_n.$$

On déduit alors du corollaire 2.1.6 que le membre de droite est en fait égal à $\mathbb{E}\left[g(U_{(1,n)}, \ldots, U_{(n,n)})\right]$. □

Soit $(V_i, i \geq 1)$ une suite de variables aléatoires indépendantes de loi de **Pareto** dont la fonction de répartition est $F(x) = 1 - \dfrac{1}{x}$ pour $x \geq 1$. La loi de Pareto interviendra au paragraphe 2.5 lors de la construction de l'estimateur de Pickand. D'après la proposition 2.1.10, si $(k(n), n \geq 1)$ une suite d'entiers telle que $1 \leq k(n) \leq n$, $\lim\limits_{n\to\infty} k(n) = \infty$ et $\lim\limits_{n\to\infty} \dfrac{k(n)}{n} = 0$ alors on a p.s. $\lim_{n\to\infty} V_{(n-k(n)+1,n)} = +\infty$. Le lemme suivant précise la vitesse de cette convergence.

Proposition 2.1.15. *Soit $(k(n), n \geq 1)$ une suite d'entiers telle que $1 \leq k(n) \leq n$, $\lim\limits_{n\to\infty} k(n) = \infty$ et $\lim\limits_{n\to\infty} \dfrac{k(n)}{n} = 0$. Alors la suite de variables aléatoires $\left(\dfrac{k(n)}{n} V_{(n-k(n)+1,n)}, n \geq 1\right)$ converge en probabilité vers 1.*

En fait on peut démontrer que si $\lim\limits_{n\to\infty} \dfrac{k(n)}{\log n} = \infty$, alors la convergence de la proposition 2.1.15 est une convergence presque sûre.

Démonstration. On écrit k pour $k(n)$. Remarquons que $F^{-1}(u) = \frac{1}{1-u}$ pour $u \in [0,1[$. On déduit du lemme 2.1.2, que $V_{(n-k+1,n)}$ a même loi que $F^{-1}(U_{(n-k+1,n)})$ et donc, grâce au lemme 2.1.14, que $F^{-1}\left(\dfrac{\Gamma_{n-k+1}}{\Gamma_{n+1}}\right) = \dfrac{\Gamma_{n+1}}{\Gamma_{n+1} - \Gamma_{n-k+1}}$.

Remarquons que Γ_{n+1} est la somme de $\Gamma_{n+1} - \Gamma_{n-k+1}$ et de Γ_{n-k+1} qui sont deux variables aléatoires indépendantes de loi gamma de paramètre respectif $(1,k)$ et $(1, n-k+1)$. Ainsi la variable $\dfrac{k}{n} V_{(n-k+1,n)}$ a même loi que

$$J_{k,n} = \frac{k}{n} + \frac{k}{\Gamma'_k} \frac{\Gamma_{n-k+1}}{n-k+1} \frac{n-k+1}{n},$$

où $\Gamma'_k = \sum_{i'=1}^{k} E'_{i'}$, $\Gamma_{n-k+1} = \sum_{i=1}^{n-k+1} E_i$ et les variables $(E_i, E_{i'}; i \geq 1, i' \geq 1)$ sont indépendantes et de loi exponentielle de paramètre 1. La loi forte des grands nombres assure que p.s.

$$\lim_{n\to\infty} \frac{\Gamma_{n-k+1}}{n-k+1} = \lim_{n\to\infty} \frac{\Gamma'_k}{k} = \mathbb{E}[E_1] = 1.$$

On en déduit que presque sûrement $\lim_{n\to\infty} J_{k,n} = 1$. Ainsi pour tout $\varepsilon > 0$, on a $\lim_{n\to\infty} \mathbb{P}(|J_{k,n} - 1| \geq \varepsilon) = 0$. Comme $\frac{k}{n} V_{(n-k+1,n)}$ a même loi que $J_{k,n}$, on en déduit que $\frac{k}{n} V_{(n-k+1,n)}$ converge en probabilité vers 1. □

2.2 Exemples de convergence du maximum renormalisé

Dans ce paragraphe, on considère des variables aléatoires $(X_n, n \geq 1)$ indépendantes de même loi, ainsi que leur maximum $M_n = \max_{i\in\{1,\dots,n\}} X_i$. On recherche des suites $(a_n, n \geq 0)$ et $(b_n, n \geq 1)$, avec $a_n > 0$, telles que la suite $(a_n^{-1}(M_n - b_n), n \geq 1)$ converge en loi vers une limite non dégénérée. Nous considérons des variables de loi uniforme, exponentielle, de Cauchy et de Bernoulli.

Loi uniforme.

On suppose que la loi de X_1 est la loi uniforme sur $[0, \theta]$, $\theta > 0$. La fonction de répartition de la loi est $F(x) = x/\theta$ pour $x \in [0, \theta]$. Par la proposition 2.1.10, la suite $(M_n, n \geq 1)$ converge p.s. vers θ.

Lemme 2.2.1. *La suite* $\Big(n\Big(\frac{M_n}{\theta} - 1\Big), n \geq 1\Big)$ *converge en loi vers* W *de fonction de répartition définie par*

$$\mathbb{P}(W \leq x) = \mathrm{e}^x, \quad x \leq 0.$$

La loi de W est une loi de **Weibull**. La famille des lois de Weibull sera définie au paragraphe 2.3. Dans ce cas particulier, la loi de $-W$ est la loi exponentielle de paramètre 1.

Démonstration. On note F_n la fonction de répartition de $n(\frac{M_n}{\theta} - 1)$. Comme $M_n < \theta$, on a $F_n(x) = 1$ si $x \geq 0$. Considérons le cas $x < 0$:

$$F_n(x) = \mathbb{P}\Big(M_n \leq \theta + \theta\frac{x}{n}\Big) = \mathbb{P}\Big(X_1 \leq \theta + \theta\frac{x}{n}\Big)^n = \Big(1 + \frac{x}{n}\Big)^n.$$

Il vient $\lim_{n\to\infty} F_n(x) = \mathrm{e}^x$ pour $x < 0$. On en déduit que $n\left(\frac{M_n}{\theta} - 1\right)$ converge en loi vers W de fonction de répartition $x \to \min(\mathrm{e}^x, 1)$. □

Loi exponentielle.

On suppose X_1 de loi exponentielle de paramètre $\lambda > 0$. La fonction de répartition de cette loi est $F(x) = 1 - \mathrm{e}^{-\lambda x}$ pour $x \geq 0$. Comme $x_F = +\infty$, la suite $(M_n, n \geq 1)$ diverge vers l'infini d'après la proposition 2.1.10.

Lemme 2.2.2. *La suite* $(\lambda M_n - \log(n), n \geq 1)$ *converge en loi vers* G *de fonction de répartition définie par*

$$\mathbb{P}(G \leq x) = \mathrm{e}^{-\,\mathrm{e}^{-x}}, \quad x \in \mathbb{R}.$$

La loi de G est la loi de **Gumbel**.

Démonstration. On note F_n la fonction de répartition de $\lambda M_n - \log(n)$. On a

$$F_n(x) = \mathbb{P}(\lambda M_n - \log(n) \leq x) = \mathbb{P}(M_n \leq (x + \log(n))/\lambda)$$
$$= \mathbb{P}(X_1 \leq (x + \log(n))/\lambda)^n = \left(1 - \frac{\mathrm{e}^{-x}}{n}\right)^n.$$

On a alors $\lim_{n\to\infty} F_n(x) = \mathrm{e}^{-\mathrm{e}^{-x}}$, $x \in \mathbb{R}$. On en déduit que la suite $(\lambda M_n - \log(n), n \geq 1)$ converge en loi vers G, de fonction de répartition $x \to \mathrm{e}^{-\mathrm{e}^{-x}}$. □

L'exemple suivant donne une application bien connue de ce résultat.

Loi de Cauchy.

On suppose que X_1 suit la loi de Cauchy (de paramètre $a = 1$). La densité de la loi est $f(x) = \frac{1}{\pi(1+x^2)}$. Comme le support de la densité est non borné, il est clair que la suite $(M_n, n \geq 1)$ diverge.

Lemme 2.2.3. *La suite* $\left(\frac{\pi M_n}{n}, n \geq 1\right)$ *converge en loi vers* W *de fonction de répartition définie par*

$$\mathbb{P}(W \leq x) = \mathrm{e}^{-1/x}, \quad x > 0.$$

La loi de W appartient à la famille des lois de **Fréchet**.

Démonstration. On note F_n la fonction de répartition de $\frac{\pi M_n}{n}$. On a

$$F_n(x) = \mathbb{P}\Big(M_n \leq \frac{nx}{\pi}\Big) = \mathbb{P}\Big(X_1 \leq \frac{nx}{\pi}\Big)^n = \left(1 - \int_{nx/\pi}^{\infty} \frac{1}{\pi(1+y^2)}\ dy\right)^n.$$

Pour $x > 0$, on a

$$\int_{nx/\pi}^{\infty} \frac{1}{\pi(1+y^2)}\ dy = \int_{nx/\pi}^{\infty} \frac{1}{\pi y^2}\ dy + \int_{nx/\pi}^{\infty} \left[\frac{1}{\pi(1+y^2)} - \frac{1}{\pi y^2}\right] dy$$
$$= \frac{1}{nx} + O((nx)^{-3}).$$

On a alors pour $x > 0$, $F_n(x) = \left(1 - \frac{1}{nx} + O((nx)^{-3})\right)^n$. On en déduit que $\lim_{n\to\infty} F_n(x) = \mathrm{e}^{-1/x}$ pour $x > 0$. Ainsi la suite $(\pi M_n/n, n \geq 1)$ converge en loi vers W de fonction de répartition définie par

$$\mathbb{P}(W \leq x) = \mathrm{e}^{-1/x}, \quad x > 0.$$

□

Exercice 2.2.4. Montrer que le maximum judicieusement renormalisé d'un échantillon de variables aléatoires indépendantes de loi de Pareto converge en loi. Identifier la limite. ♦

Loi de Bernoulli.

On suppose X_i de loi de Bernoulli de paramètre $p \in]0,1[$: $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$. On a $M_n = 1$ si $n \geq T = \inf\{k \geq 1; X_k = 1\}$. La loi de T est une loi géométrique de paramètre p. Donc T est fini p.s. et M_n est donc constant égal à 1 p.s. à partir d'un certain rang. Il n'existe donc pas de suite $((a_n, b_n), n \geq 1)$, avec $a_n > 0$, telle que la suite de terme $a_n^{-1}(M_n - b_n)$ converge en loi vers une limite non triviale, c'est-à-dire une limite différente d'une variable aléatoire constante.

On peut démontrer qu'il n'existe pas non plus de limite non triviale pour les limites des maximums renormalisés des lois géométriques et de Poisson.

Exercice 2.2.5. On suppose que X_1 suit la loi de Yule de paramètre $\rho > 0$, c'est-à-dire X_1 est à valeurs dans $\mathbb{N}^*$, et on a pour $k \in \mathbb{N}^*$, $\mathbb{P}(X_1 = k) = \rho B(k, \rho + 1)$, où $B(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. De la définition de la loi béta, on obtient que $B(a, b) = \int_{]0,1[} x^{a-1}(1-x)^{b-1}\, dx$.

On pose $p(k) = \mathbb{P}(X_1 = k)$

1. Vérifier que $\sum_{k\geq 1} p(k) = 1$.
2. Calculer la moyenne et la variance de la loi de Yule.
3. En utilisant la formule de Stirling, $\lim_{a\to\infty} \dfrac{\Gamma(a)}{a^{a-\frac{1}{2}}\,\mathrm{e}^{-a}\sqrt{2\pi}} = 1$, donner un équivalent de $\sum_{i=k}^{\infty} p(i)$ quand k tend vers l'infini. On dit que la loi de Yule, comme la loi de Pareto, est une loi en puissance.
4. En déduire que la suite $\left(\frac{M_n}{(\Gamma(\rho+1)\ n)^{1/\rho}}, n \geq 1\right)$ converge en loi vers une variable aléatoire de fonction de répartition $x \to \mathrm{e}^{-x^{-\rho}}$, $x > 0$. Cette loi limite appartient à la famille des lois de Fréchet.

Les lois en puissance semblent correspondre à de nombreux phénomènes, voir [8] : taille de la population des villes, nombre de téléchargements des pages internet, nombre d'occurrences des mots du langage, etc. Le théorème 2.4.9 assure que les limites des maximums convenablement renormalisés correspondant à ces lois suivent des lois de Fréchet. ♦

2.3 Limites des maximums renormalisés

Définition 2.3.1. *La loi $\mathcal{L}_0$ est dite max-stable si pour tout $n \geq 2$, $(W_1, \ldots, W_n)$ étant des variables aléatoires indépendantes de loi $\mathcal{L}_0$, il existe $a_n > 0$ et $b_n \in \mathbb{R}$, tels que $a_n^{-1}\left(\max_{i\in\{1,\ldots,n\}} W_i - b_n\right)$ suit la loi $\mathcal{L}_0$.*

On peut montrer que si $(X_i, i \geq 1)$ est une suite de variables aléatoires indépendantes et de même loi, telle que la suite $\left(a_n^{-1}\big(\max_{i\in\{1,\ldots,n\}} X_i - b_n\big), n \geq 1\right)$ converge en loi pour une suite appropriée $a_n > 0$ et $b_n \in \mathbb{R}$ vers une limite non triviale, c'est-à-dire vers une variable aléatoire non constante, alors la limite est une loi max-stable. Bien sûr la suite $((a_n, b_n), n \geq 1)$ n'est pas unique. Le théorème suivant permet d'identifier l'ensemble des lois max-stables.

Théorème 2.3.2. *Soit $(X_i, i \geq 1)$ une suite de variables aléatoires indépendantes et de même loi. Supposons qu'il existe une suite $((a_n, b_n), n \geq 1)$ telle que $a_n > 0$ et la suite $\left(a_n^{-1}\left(\max_{i\in\{1,\ldots,n\}} X_i - b_n\right), n \geq 1\right)$ converge en loi vers une limite non triviale. Alors à une translation et un changement d'échelle près la **fonction de répartition** de la limite est de la forme suivante :*

$$\textit{Loi de Weibull} \quad \Psi_\alpha(x) = \begin{cases} \mathrm{e}^{-(-x)^\alpha}, & x \leq 0 \\ 1, & x > 0 \end{cases} \quad \textit{et} \quad \alpha > 0.$$

$$\textit{Loi de Gumbel} \quad \Lambda(x) = \mathrm{e}^{-\,\mathrm{e}^{-x}}, x \in \mathbb{R}.$$

$$\textit{Loi de Fréchet} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \mathrm{e}^{-x^{-\alpha}}, & x > 0 \end{cases} \quad \textit{et} \quad \alpha > 0.$$

L'ensemble des lois limites s'obtient donc en considérant les lois de $cW + d$, où W suit une loi de Weibull, de Gumbel ou de Fréchet. L'exercice suivant permet de vérifier que les lois de Weibull, Gumbel et Fréchet sont max-stables.

Exercice 2.3.3. Soit $(X_i, i \geq 1)$ une suite de variables aléatoires indépendantes, de même loi que X. On pose $M_n = \max_{i\in\{1,\dots,n\}} X_i$. Montrer que si X suit la loi de
- Weibull de paramètre α, alors M_n a même loi que $n^{-1/\alpha}X$;
- Gumbel, alors M_n a même loi que $X + \log(n)$;
- Fréchet de paramètre α, alors M_n a même loi que $n^{1/\alpha}X$.

♦

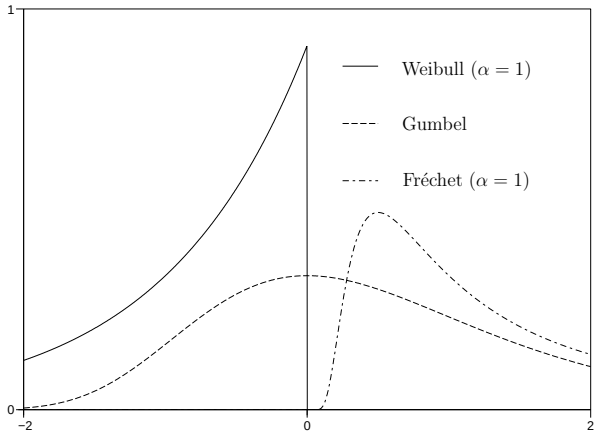

FIG. 2.4 – Densité des lois de valeurs extrêmes

Il est possible de rassembler les trois familles de lois en une seule famille paramétrique $(H(\xi), \xi \in \mathbb{R})$ dite famille des lois de valeurs extrêmes généralisées. Elle est paramétrée par une seule variable $\xi \in \mathbb{R}$, mais toujours à un facteur de changement d'échelle et de translation près. La fonction de répartition est pour $\xi \in \mathbb{R}$

$$\boxed{H(\xi)(x) = \mathrm{e}^{-(1+\xi x)^{-1/\xi}}, \quad \text{si } 1 + \xi x > 0.}$$

Pour $\xi = 0$, il faut lire $H(0)(x) = \mathrm{e}^{-\mathrm{e}^{-x}}$, $x \in \mathbb{R}$, qui s'obtient dans la formule précédente en faisant tendre ξ vers 0. Les lois de valeurs extrêmes généralisées correspondent à une translation et changement d'échelle près aux lois de valeurs extrêmes. Plus précisément, on a :
- $\Psi_\alpha(x) = H\Big(-\dfrac{1}{\alpha}\Big)(\alpha(x+1))$ pour $x \in \mathbb{R}$. Ainsi, si W suit la loi de Weibull de paramètre $\alpha > 0$, alors $\alpha(W + 1)$ suit la loi de valeurs extrêmes généralisées de paramètre $\xi = -1/\alpha$.
- $\Lambda = H(0)$. La loi de Gumbel correspond à la loi de valeurs extrêmes généralisées de paramètre $\xi = 0$.
- $\Phi_\alpha(x) = H\Big(\dfrac{1}{\alpha}\Big)(\alpha(x-1))$ pour $x \in \mathbb{R}$. Ainsi, si W suit la loi de Fréchet de paramètre $\alpha > 0$, alors $\alpha(W - 1)$ suit la loi de valeurs extrêmes généralisées de paramètre $\xi = 1/\alpha$.

Dans les exemples du paragraphe précédent, on a obtenu la convergence en loi du maximum renormalisé vers une variable qui suit :
- La loi de Weibull de paramètre $\alpha = 1$, pour une suite de variables aléatoires de loi uniforme.
- La loi de Gumbel, pour une suite de variables aléatoires de loi exponentielle.
- La loi de Fréchet de paramètre $\alpha = 1$, pour une suite de variables aléatoires de loi de Cauchy.

Rappelons que la convergence en loi du maximum renormalisé n'a pas lieu pour toutes les lois, cf. l'exemple de la loi de Bernoulli au paragraphe 2.2.

Démonstration du théorème 2.3.2. La démonstration de ce théorème est longue, voir [9] proposition 0.3. Néanmoins le raisonnement présenté dans [1] permet de se forger une intuition des résultats et d'introduire des quantités qui nous seront utiles par la suite.

Supposons qu'il existe une suite $((a_n, b_n), n \geq 1)$ telle que $a_n > 0$ et la suite de terme $a_n^{-1}(M_n - b_n)$ converge en loi vers une limite W non constante. Pour toute fonction g continue bornée, on a

$$\lim_{n\to\infty} \mathbb{E}[g(a_n^{-1}(M_n - b_n))] = \mathbb{E}[g(W)].$$

Supposons par simplicité que la loi de X_1 possède la densité $f > 0$. Alors la loi de M_n possède la densité $nF(x)^{n-1}f(x)$ d'après la question 3 de l'exercice 2.1.7. On a donc

$$I_n = \mathbb{E}[g(a_n^{-1}(M_n - b_n))] = \int_{\mathbb{R}} g\left(\frac{x - b_n}{a_n}\right) nF(x)^{n-1}f(x)\, dx.$$

Comme $f > 0$, la fonction F est inversible et d'inverse continue. On pose pour $t > 1$

$$U(t) = F^{-1}\left(1 - \frac{1}{t}\right).$$

En particulier, on a $U(t) = x \iff 1 - \frac{1}{t} = F(x) \iff \mathbb{P}(X > x) = \frac{1}{t}$. On effectue le changement de variable $F(x) = 1 - \frac{v}{n}$, i.e. $x = U\left(\frac{n}{v}\right)$. On obtient

$$I_n = \int_{\mathbb{R}} g\left(\frac{U(n/v) - b_n}{a_n}\right)\left(1 - \frac{v}{n}\right)^{n-1} \mathbf{1}_{]0,n]}(v)\, dv.$$

Remarquons que $\left(1 - \frac{v}{n}\right)^{n-1} \mathbf{1}_{]0,n]}(v)$ converge en croissant vers $\mathrm{e}^{-v}\,\mathbf{1}_{\{v>0\}}$. Comme par hypothèse I_n converge pour tout g, il est naturel, mais erroné a priori, de penser que pour tout $v > 0$, la suite de terme $J_n(v) = \dfrac{U(n/v) - b_n}{a_n}$ converge. Supposons malgré tout que cette convergence ait lieu. On en déduit en considérant $J_n(1/w) - J_n(1)$ que pour tout $w > 0$, $\dfrac{U(wn) - U(n)}{a_n}$ converge quand n tend vers l'infini, vers une limite que l'on note $h(w)$. Comme la variable aléatoire W est non triviale, cela implique que la fonction h n'est pas égale à une constante. Comme la fonction U est croissante, la fonction h est également croissante. Supposons que plus généralement, on ait pour tout $w > 0$,

$$\frac{U(wx) - U(x)}{a(x)} \underset{x\to\infty}{\longrightarrow} h(w),$$

où $a(x) = a_{[x]}$ pour $x \geq 1$, $[x]$ désignant la partie entière de x. Soit $w_1, w_2 > 0$. On a

$$\frac{U(xw_1w_2) - U(x)}{a(x)} = \frac{U(xw_1w_2) - U(xw_1)}{a(xw_1)}\frac{a(xw_1)}{a(x)} + \frac{U(xw_1) - U(x)}{a(x)}.$$

En faisant tendre x vers l'infini dans l'égalité ci-dessus, on obtient que $\dfrac{a(xw_1)}{a(x)}$ converge pour tout $w_1 > 0$. On note $l(w_1)$ la limite, et il vient

$$h(w_1w_2) = h(w_2)l(w_1) + h(w_1). \tag{2.5}$$

La fonction l est mesurable et localement bornée. Comme la fonction h est croissante et non constante, on en déduit que l est strictement positive. De plus en posant $yw' = x$, on a pour $w' > 0$

$$l(w) = \lim_{x\to\infty} \frac{a(xw)}{a(x)} = \lim_{y\to\infty} \frac{a(yw'w)}{a(y)}\frac{a(y)}{a(yw')} = \frac{l(w'w)}{l(w')}.$$

Ainsi on a pour tous $w, w' > 0$,

$$l(w'w) = l(w)l(w'), \tag{2.6}$$

où l est une fonction strictement positive mesurable localement bornée. Vérifions que les solutions non nulles de cette équation fonctionnelle sont : $l(w) = w^\xi$, où $\xi \in \mathbb{R}$. En intégrant (2.6) pour w' entre 1 et 2, il vient en effectuant le changement de variable $ww' = u$, $\frac{1}{w}\int_w^{2w} l(u)du = l(w)\int_1^2 l(w')dw'$. On en déduit que l est continu puis dérivable. On obtient alors en dérivant (2.6) par rapport à w' et en évaluant en $w' = 1$ que $wl'(w) = \xi l(w)$ où $\xi \in \mathbb{R}$. Ceci implique que $l(w) = cw^\xi$. Comme d'après (2.6), $l(1) = l(1)^2$ et que l est strictement positive, on en déduit que $l(1) = 1$ et que $l(w) = w^\xi$ pour $w > 0$. On retrouve ξ l'indice de la loi de valeurs extrêmes généralisées. L'équation (2.5) se récrit

$$h(w_1w_2) = h(w_2)w_1^\xi + h(w_1) \quad \text{pour tous } w_1, w_2 > 0.$$

Pour $\xi = 0$ on obtient l'équation fonctionnelle $h(w_1w_2) = h(w_2) + h(w_1)$. Un raisonnement semblable à celui effectué à partir de l'équation fonctionnelle (2.6) assure que les solutions mesurables localement bornées sur $]0, \infty[$ de cette équation fonctionnelle sont $h(w) = c\log w$, avec $c > 0$.

Pour $\xi \neq 0$, par symétrie, on a

$$h(w_1w_2) = h(w_2)w_1^\xi + h(w_1) = h(w_1)w_2^\xi + h(w_2).$$

En particulier, on a

$$h(w_1)(1 - w_2^\xi) = h(w_2)(1 - w_1^\xi).$$

Cela implique $h(w) = 0$ si $w = 1$, et sinon $\dfrac{h(w)}{w^\xi - 1}$ est constant.

Donc on obtient $h(w) = c(w^\xi - 1)$. À un changement d'échelle près, on peut choisir $c = \frac{1}{\xi}$. À une translation près, on peut choisir $U(n) = b_n$. En définitive, il vient

$$\lim_{n\to\infty} \frac{U(n/v) - b_n}{a_n} = h\left(\frac{1}{v}\right) = \begin{cases} \dfrac{v^{-\xi} - 1}{\xi} & \text{si } \xi \neq 0, \\ -\log(v) & \text{si } \xi = 0. \end{cases}$$

On peut maintenant calculer la limite de $I_n = \mathbb{E}[g(a_n^{-1}(M_n - b_n))]$. Il vient par convergence dominée

$$\begin{aligned} \lim_{n\to\infty} I_n &= \int g\left(\frac{v^{-\xi} - 1}{\xi}\right) \mathrm{e}^{-v}\, \mathbf{1}_{\{v>0\}}\, dv \\ &= \int g(y)\mathbf{1}_{\{1+\xi y>0\}}\, d\left(\mathrm{e}^{-(1+\xi y)^{-1/\xi}}\right), \end{aligned}$$

où on a posé $y = \dfrac{v^{-\xi} - 1}{\xi}$ si $\xi \neq 0$ et $y = \log(v)$ si $\xi = 0$. La fonction de répartition de la loi limite est donc $H(\xi)$. □

2.4 Domaines d'attraction

Après avoir caractérisé les lois limites, il nous reste à déterminer les lois $\mathcal{L}$ pour lesquelles la loi du maximum renormalisé converge vers une loi max-stable donnée $\mathcal{L}_0$. On dit alors que la loi $\mathcal{L}$ (ou sa fonction de répartition F) appartient au bassin d'attraction de $\mathcal{L}_0$ (ou de sa fonction de répartition F_0) pour la convergence du maximum renormalisé. On le notera $\mathcal{L} \in D(\mathcal{L}_0)$ (ou $F \in D(F_0)$).

2.4.1 Caractérisations générales

On définit la fonction U par

$$U(t) = F^{-1}\left(1 - \frac{1}{t}\right), \quad t > 1,$$

où F^{-1} est l'inverse généralisé de F. Les calculs de la démonstration du théorème 2.3.2 suggèrent que si $F \in D(H(\xi))$, alors on a, à un changement d'échelle près,

$$\lim_{s\to\infty} \frac{U(sw) - U(s)}{a(s)} = \frac{w^\xi - 1}{\xi}.$$

En particulier, si $x, y > 0$ et $y \neq 1$, on a

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

En fait la proposition suivante, voir [6] théorème 3.4.5 pour une démonstration, assure que cette condition est suffisante pour que $F \in D(H(\xi))$.

Proposition 2.4.1. *Soit $\xi \in \mathbb{R}$. Il y a équivalence entre $F \in D(H(\xi))$ et pour tous $x > 0$, $y > 0$, $y \neq 1$,*

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

Nous utiliserons ce résultat pour construire l'estimateur de Pickand de ξ au paragraphe 2.5.

Exercice 2.4.2. Calculer la fonction U pour la loi exponentielle de paramètre $\lambda > 0$, la loi uniforme sur $[0, \theta]$ et la loi de Cauchy. Calculer $\lim_{s\to\infty} \dfrac{U(sx) - U(s)}{U(sy) - U(s)}$, et retrouver ainsi les résultats du paragraphe 2.2. ♦

Il est d'usage d'utiliser la notation $\bar{F}$ pour la distribution de la queue de la loi de X : $\bar{F}(x) = \mathbb{P}(X > x) = 1 - F(x$ En fait, au paragraphe 2.2, pour démontrer la convergence en loi du maximum renormalisé, on a recherché une suite $((a_n, b_n), n \geq 1)$, avec $a_n > 0$, telle que

$$\mathbb{P}\left(a_n^{-1}(M_n - b_n) \leq x\right) = F(xa_n + b_n)^n = (1 - \bar{F}(xa_n + b_n))^n$$

converge vers une limite non triviale.

Proposition 2.4.3. *On a $F \in D(H(\xi))$ si et seulement si*

$$\boxed{n\bar{F}(xa_n + b_n) \underset{n\to\infty}{\longrightarrow} -\log H(\xi)(x).}$$

pour une certaine suite $((a_n, b_n), n \geq 1)$ où $a_n > 0$ et $b_n \in \mathbb{R}$. On a alors la convergence en loi de $(a_n^{-1}(M_n - b_n), n \geq 1)$ vers une variable aléatoire de fonction de répartition $H(\xi)$.

Démonstration. Si $F \in D(H(\xi))$, alors on a $\big(1-\bar{F}(xa_n+b_n)\big)^n \underset{n\to\infty}{\longrightarrow} H(\xi)(x)$ pour une certaine suite $((a_n,b_n), n\geq 1)$ où $a_n>0$ et $b_n\in\mathbb{R}$. En prenant le logarithme de cette expression il vient

$$n\log\big(1-\bar{F}(xa_n+b_n)\big) \underset{n\to\infty}{\longrightarrow} \log H(\xi)(x).$$

Ceci implique que pour $1+\xi x>0$, $\bar{F}(xa_n+b_n)$ tend vers 0 et

$$n\bar{F}(xa_n+b_n) \underset{n\to\infty}{\longrightarrow} -\log H(\xi)(x).$$

La réciproque est claire : si on a la convergence ci-dessus, alors $F\in D(H(\xi))$. □

Rappelons que $x_F=\inf\{x\in\mathbb{R}; F(x)=1\}$ désigne le quantile d'ordre 1 de la loi de fonction de répartition F. Nous avons le résultat plus général suivant.

Proposition 2.4.4. *Soit $\xi\in\mathbb{R}$. Il y équivalence entre $F\in D(H(\xi))$ et il existe une fonction mesurable a telle que pour $1+\xi x>0$, on a*

$$\lim_{u\to x_F^-}\frac{\bar{F}(xa(u)+u)}{\bar{F}(u)} = \begin{cases}(1+\xi x)^{-1/\xi} & \text{si } \xi\neq 0,\\ \mathrm{e}^{-x} & \text{si } \xi=0.\end{cases}$$

Démonstration. Supposons que la fonction a existe et que la limite, quand u tend en croissant vers x_F, de $\bar{F}(xa(u)+u)/\bar{F}(u)$ soit celle décrite dans la proposition. On suppose par simplicité que $\bar{F}$ est continue. Alors en choisissant $b_n=U(n)$, on a $\bar{F}(b_n)=1/n$. En prenant $u=b_n$, on a donc $\lim_{n\to\infty} n\bar{F}(xa(b_n)+b_n) = -\log H(\xi)(x)$. Cela assure que $F\in D(H(\xi))$ d'après la proposition 2.4.3. La réciproque, plus difficile, est admise, voir [6] théorème 3.4.5. □

2.4.2 Domaines d'attraction des lois de Fréchet et Weibull

Pour décrire plus en détail les domaines d'attraction, il est nécessaire de décrire précisément le comportement de $\bar{F}(x)$ quand x converge vers x_F.

Définition 2.4.5. *On dit qu'une fonction L est à variation lente si $L(t)>0$ pour t assez grand et si pour tout $x>0$, on a*

$$\lim_{t\to\infty}\frac{L(tx)}{L(t)}=1.$$

Par exemple $\log(x)$ est une fonction à variation lente.

Les fonctions à variation lente jouent un rôle prépondérant dans l'étude des lois de valeurs extrêmes. Les résultats concernant ces fonctions sont difficiles. Nous renvoyons au livre très complet de Bingham et al. [2]. En particulier, on sait caractériser les fonctions à variation lente, voir [2] théorème 1.3.1.

Proposition 2.4.6. *Soit L une fonction à variation lente. Il existe deux fonctions mesurables $c>0$ et κ telles que :*

$$\lim_{x\to\infty} c(x)=c_0\in]0,\infty[\quad \textit{et} \quad \lim_{x\to\infty}\kappa(x)=0,$$

et $a\in\mathbb{R}$, tels que pour tout $x\geq a$,

$$L(x)=c(x)\exp\int_a^x\frac{\kappa(u)}{u}\,du. \tag{2.7}$$

Exercice 2.4.7.

1. Vérifier que la fonction $\log(x)$ se met sous la forme (2.7).
2. Vérifier que toute fonction de la forme (2.7) est à variation lente.

♦

Nous donnerons des résultats complémentaires sur les fonctions à variation lente dans le lemme 2.5.2.

Remarque 2.4.8. Si $g(t)$ est positive pour t assez grand et si pour tout $x > 0$, $\lim_{t\to\infty} g(tx)/g(t) = x^\beta$, alors on a $g(x) = x^\beta L(x)$, où L est une fonction à variation lente. On dit que la fonction g est à variation d'ordre β. $\Diamond$

Théorème 2.4.9. *La fonction de répartition F appartient au domaine d'attraction de la loi de Fréchet de paramètre $\alpha > 0$ si et seulement si* $\boxed{\bar{F}(x) = x^{-\alpha}L(x)}$ *où la fonction L est à variation lente. En particulier $x_F = +\infty$. De plus si $F \in D(\Phi_\alpha)$, alors avec $a_n = U(n) = F^{-1}\left(1 - \frac{1}{n}\right)$, la suite $(a_n^{-1}M_n, n \geq 1)$ converge en loi vers une variable aléatoire de fonction de répartition Φ_α.*

Démonstration. Supposons que $\bar{F}(x) = x^{-\alpha}L(x)$, où L est à variation lente. On conserve les notations de la proposition 2.4.6. On a $\bar{F}(x) \sim g(x)$ en $+\infty$, où $g(x) = x^{-\alpha}c_0 \exp\int_a^x \frac{\kappa(u)}{u}\,du$ est une fonction continue. Posons $a_n = U(n)$. On a $\bar{F}(a_n) \leq \dfrac{1}{n} \leq \bar{F}(a_n^-)$ et a_n tend vers l'infini avec n. Si F est continue en a_n, alors on a $\bar{F}(a_n) = 1/n$, sinon comme $\bar{F}$ est équivalente en $+\infty$ à une fonction continue, on en déduit que $\bar{F}(a_n) \sim 1/n$ quand n tend vers l'infini. Pour $x > 0$, on a donc

$$\lim_{n\to\infty} n\bar{F}(a_n x) = \lim_{n\to\infty} \frac{\bar{F}(a_n x)}{\bar{F}(a_n)} = x^{-\alpha}.$$

Un raisonnement similaire à celui de la démonstration de la proposition 2.4.3 assure que $F \in D(\Phi_\alpha)$. On admettra la réciproque, voir [6] théorème 3.3.7. □

Théorème 2.4.10. *La fonction de répartition F appartient au domaine d'attraction de la loi de Weibull de paramètre $\alpha > 0$ si et seulement si $x_F < \infty$ et* $\boxed{\bar{F}\Big(x_F - \dfrac{1}{x}\Big) = x^{-\alpha}L(x)}$ *où la fonction L est à variation lente. De plus si $F \in D(\Psi_\alpha)$, alors avec $a_n = x_F - U(n) = x_F - F^{-1}\left(1 - \frac{1}{n}\right)$, la suite $(a_n^{-1}(M_n - x_F), n \geq 1)$ converge en loi vers une variable aléatoire de fonction de répartition Ψ_α.*

Démonstration. La démonstration est similaire à celle du théorème précédent, voir [6] théorème 3.3.12 pour la réciproque. □

Les résultats concernant le domaine d'attraction de la loi de Gumbel sont plus délicats. Nous renvoyons à [1], chapitre 2, ou [6], chapitre 3.3, pour plus de détails. En particulier les lois gamma, gaussiennes, exponentielles et lognormales appartiennent au domaine d'attraction de la loi de Gumbel.

Enfin dans le cas particulier où la loi de X_1 possède une densité, on obtient des caractérisations pour appartenir au domaine d'attraction d'une loi de valeurs extrêmes, voir [6] théorèmes 3.3.8 et 3.3.13.

Proposition 2.4.11 (Critère de von Mises)**.** *Soit F la fonction de répartition d'une loi de densité f.*

1. *Si on a*
$$\lim_{x\to\infty} \frac{xf(x)}{\bar{F}(x)} = \alpha > 0,$$
alors F appartient au domaine d'attraction de la loi de Fréchet de paramètre α.
2. *On suppose la loi de densité f strictement positive sur un intervalle (z, x_F), avec $x_F < \infty$. Si on a*
$$\lim_{x\to x_F^-} \frac{(x_F - x)f(x)}{\bar{F}(x)} = \alpha > 0,$$
alors F appartient au domaine d'attraction de la loi de Weibull de paramètre α.

Exercice 2.4.12. Montrer les résultats suivants à l'aide des théorèmes 2.4.9 et 2.4.10.

1. La loi de Pareto d'indice $\alpha > 0$, de fonction de répartition $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$, appartient au domaine d'attraction de la loi de Fréchet de paramètre α.
2. La loi de Cauchy appartient au domaine d'attraction de la loi de Fréchet de paramètre 1.
3. La loi béta de paramètre (a, b) appartient au domaine d'attraction de la loi de Weibull de paramètre b.

♦

2.5 Estimation du paramètre de la loi de valeurs extrêmes

Nous donnons deux familles d'estimateurs du paramètre de la loi de valeurs extrêmes généralisées. Il en existe de nombreux autres, voir les monographies [1] et [6].

2.5.1 Estimateur de Pickand

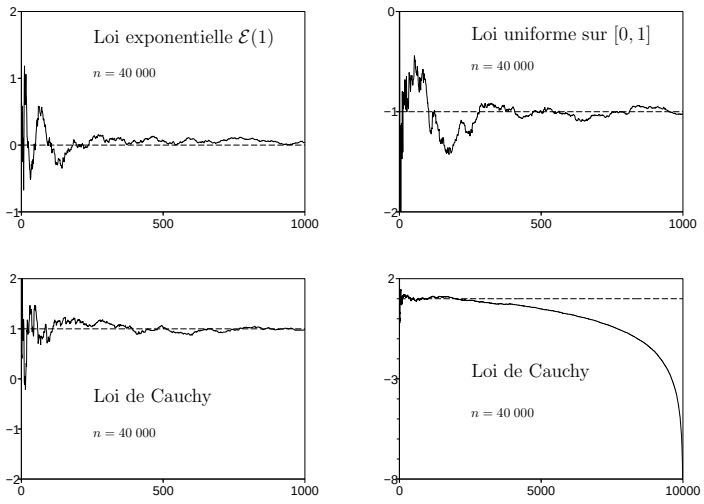

FIG. 2.5 – Estimateur de Pickand : $k \to \xi^P_{(k,n)}$ à n fixé pour différentes lois

Théorème 2.5.1. *Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de même loi de fonction de répartition $F \in D(H(\xi))$, où $\xi \in \mathbb{R}$. Si $\lim_{n\to\infty} k(n) = \infty$ et $\lim_{n\to\infty} \frac{k(n)}{n} = 0$, alors l'**estimateur de Pickand***

$$\boxed{\xi^P_{(k(n),n)} = \frac{1}{\log 2} \log\left(\frac{X_{(n-k(n)+1,n)} - X_{(n-2k(n)+1,n)}}{X_{(n-2k(n)+1,n)} - X_{(n-4k(n)+1,n)}}\right)}$$

converge en probabilité vers ξ.

Supposons les hypothèses du théorème précédent satisfaites ainsi que $\lim_{n\to\infty} \frac{k(n)}{\log\log n} = \infty$, alors on admet que l'estimateur de Pickand est fortement convergent : la convergence de l'estimateur est presque sûre et non plus seulement en probabilité. Sous certaines hypothèses supplémentaires sur la suite $(k(n), n \geq 1)$ et sur F, on peut montrer qu'il est également asymptotiquement normal, voir

[5] : la suite $\left(\sqrt{k(n)}\left(\xi^P_{(k(n),n)} - \xi\right), n \geq 1\right)$ converge en loi vers une variable gaussienne centrée de variance

$$\frac{\xi^2(2^{2\xi+1}+1)}{(2(2^\xi-1)\log(2))^2}.$$

Cela permet donc de donner un intervalle de confiance pour l'estimation. Mais attention, l'estimateur de Pickand est biaisé. Pour un échantillon de taille n fixé, on trace le diagramme de Pickand : $\xi^P_{(k,n)}$ en fonction de k. On est alors confronté au dilemme suivant :

- Pour k petit, on a une estimation avec un intervalle de confiance large. On observe de grandes oscillations de la trajectoire $k \to \xi^P_{(k,n)}$ pour k petit dans la figure 2.5.
- Pour k grand, l'intervalle de confiance est plus étroit, mais il faut tenir compte d'un biais inconnu. Ce biais peut être important, comme on peut le voir sur la figure 2.5 pour la loi de Cauchy. Le comportement de $\xi^P_{(k,n)}$ pour les grandes valeurs de k, observé pour la loi de Cauchy sur les figures 2.5 et 2.6, se retrouve pour de nombreuses lois.

Enfin, l'estimateur de Pickand possède une grande variance. De nombreux auteurs ont proposé des variantes de cet estimateur, avec des variances plus faibles, construites à partir de combinaisons linéaires des logarithmes des accroissements de la statistique d'ordre, voir [1] chapitre 5.

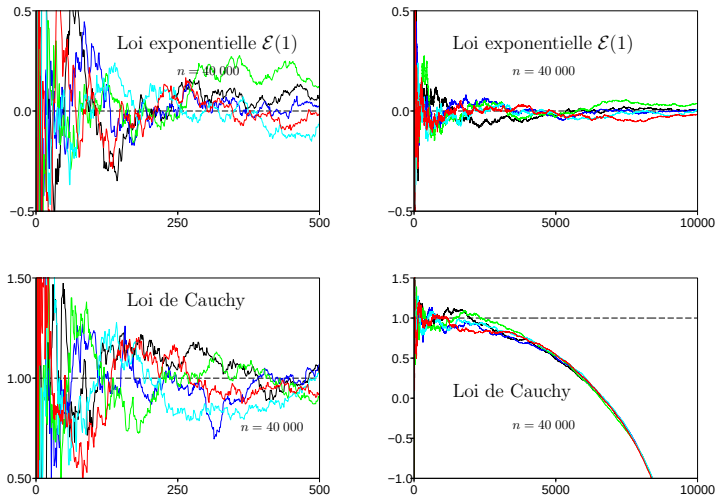

FIG. 2.6 – Estimateur de Pickand : $k \to \xi^P_{(k,n)}$, à n fixé, pour différentes lois et plusieurs réalisations

Démonstration du théorème 2.5.1. On déduit de la proposition 2.4.1 que pour $\xi \in \mathbb{R}$, on a avec le choix $t = 2s$, $x = 2$ et $y = 1/2$,

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

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

$$\lim_{t\to\infty} \frac{U(t) - U(tc_1(t))}{U(tc_1(t)) - U(tc_2(t))} = 2^\xi.$$

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

Soit $(k(n), n \geq 1)$ une suite d'entiers telle que $1 \leq k(n) \leq n/4$, $\lim_{n\to\infty} k(n) = \infty$ et $\lim_{n\to\infty} \frac{k(n)}{n} = 0$. Nous écrivons k pour $k(n)$. 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 $F_V(x) = 1 - \frac{1}{x}$, pour $x \geq 1$, la fonction de répartition de la loi de Pareto. On déduit de la proposition 2.1.15 que les suites $(\frac{k}{n}V_{(n-k+1,n)}, n \geq 1)$, $(\frac{2k}{n}V_{(n-2k+1,n)}, n \geq 1)$ et $(\frac{4k}{n}V_{(n-4k+1,n)}, n \geq 1)$ convergent en probabilité vers 1. En particulier, on a les convergences en probabilité suivantes :

$$V_{(n-k+1,n)} \underset{n\to\infty}{\longrightarrow} \infty, \quad \frac{V_{(n-2k+1,n)}}{V_{(n-k+1,n)}} \underset{n\to\infty}{\longrightarrow} 1/2, \quad \text{et} \quad \frac{V_{(n-4k+1,n)}}{V_{(n-k+1,n)}} \underset{n\to\infty}{\longrightarrow} 1/4.$$

On en déduit donc que la convergence suivante a lieu en probabilité :

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

Il reste maintenant à déterminer la loi de $\big(U(V_{(1,n)}), \ldots, U(V_{(n,n)})\big)$. Remarquons que si $x \geq 1$, alors $U(x) = F^{-1}(F_V(x))$. On a donc

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

où F_V est la fonction de répartition de la loi de Pareto. On déduit de la croissance de F_V que $\big(F_V(V_{(1,n)}), \ldots, F_V(V_{(n,n)})\big)$ a même loi que la statistique d'ordre de n variables aléatoires uniformes sur $[0,1]$ indépendantes. On déduit du lemme 2.1.2 que le vecteur aléatoire $\big(F^{-1}(F_V(V_{(1,n)})), \ldots, F^{-1}(F_V(V_{(n,n)}))\big)$ a même loi que $(X_{(1,n)}, \cdots, X_{(n,n)})$, la statistique d'ordre d'un échantillon de n variables aléatoires indépendantes dont la loi a pour fonction de répartition F. Donc la variable aléatoire $\dfrac{U(V_{(n-k+1,n)}) - U(V_{(n-2k+1,n)})}{U(V_{(n-2k+1,n)}) - U(V_{(n-4k+1,n)})}$ a même loi que

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

Ainsi cette quantité converge en loi vers 2^{ξ} quand n tend vers l'infini. Comme la fonction logarithme est continue sur $\mathbb{R}_+^*$, on déduit que l'estimateur de Pickand converge en loi vers ξ. Mais comme ξ est une constante, on a également la convergence en probabilité. □

2.5.2 Estimateur de Hill

Dans tout ce paragraphe, on suppose que $\xi > 0$. Nous avons besoin d'un résultat préliminaire sur les fonctions à variation lente.

Lemme 2.5.2. *Soit L une fonction à variation lente. Alors on a : pour tout $\rho > 0$, $L(x) = o(x^{\rho})$ en $+\infty$ et*

$$\int_x^{\infty} t^{-\rho-1} L(t)\, dt \sim \frac{1}{\rho}\, x^{-\rho} L(x) \quad \text{en } +\infty.$$

Démonstration. La démonstration repose sur la formule de représentation (2.7). Soit $\rho > 0$. Il existe $x_0, M > 0$ tel que pour $x \geq x_0$, on a $\kappa(x) \leq \rho/2$ et $c(x)\, \mathrm{e}^{\int_a^{x_0} \frac{\kappa(u)}{u} du} \leq M$. On en déduit que pour $x \geq x_0$, on a

$$L(x) \leq M\, \mathrm{e}^{\int_{x_0}^{x} \frac{\rho}{2u} du} \leq M' \left(\frac{x}{x_0}\right)^{\rho/2}.$$

On obtient $L(x) = o(x^{\rho})$ en $+\infty$.

Soit $u \geq 1$. La fonction $h_x(u) = \left(\dfrac{L(ux)}{L(x)} - 1\right) u^{-\rho-1}$ est majorée en valeur absolue par $\left(1 + \dfrac{c(ux)}{c(x)} \exp\Big[\int_x^{ux} \dfrac{\kappa(v)}{v} dv\Big]\right) u^{-}$

En utilisant les convergences de c et de κ, on en déduit que pour $x \geq x_0$, la fonction $|h_x(u)|$ est majorée par la fonction

$$g(u) = (1 + A\, \mathrm{e}^{\int_x^{ux} \frac{\rho}{2v} dv}) u^{-\rho-1} \leq A' u^{-\frac{\rho}{2}-1},$$

où A et A' sont des constantes qui ne dépendent pas de u. La fonction g est intégrable sur $[1,\infty[$. De plus on a $\lim_{x\to\infty}\left(\frac{L(ux)}{L(x)}-1\right)u^{-\rho-1}=0$, car L est à variation lente. Par le théorème de convergence dominée, on en déduit que

$$\lim_{x\to\infty}\int_1^\infty\left(\frac{L(ux)}{L(x)}-1\right)u^{-\rho-1}\,du=0.$$

Ce qui implique que $\lim_{x\to\infty}\int_1^\infty\frac{L(ux)}{L(x)}u^{-\rho-1}\,du=\frac{1}{\rho}$ et en posant le changement de variable $v=ux$,

$$\lim_{x\to\infty}\frac{1}{x^{-\rho}L(x)}\int_x^\infty v^{-\rho-1}L(v)\,dv=\frac{1}{\rho}\,.$$

On obtient bien la dernière propriété du lemme. □

Lemme 2.5.3. *Soit $F\in D(\Phi_\alpha)$. On a*

$$\frac{1}{\bar F(t)}\,\mathbb{E}\left[(\log X-\log t)\mathbf{1}_{\{X>t\}}\right]\underset{t\to\infty}{\longrightarrow}\frac{1}{\alpha}=\xi.$$

Démonstration. On déduit de la définition des fonctions à variation lente et du théorème 2.4.9, que $F\in D(\Phi_\alpha)$, où $\xi=1/\alpha$, si et seulement si $\lim_{t\to\infty}\frac{\bar F(tx)}{\bar F(t)}=x^{-\alpha}$ pour tout $x>0$. Supposons par simplicité que la loi de X possède la densité f. Par intégration par partie, on a pour $t>1$

$$\mathbb{E}\left[(\log X-\log t)\mathbf{1}_{\{X>t\}}\right]$$
$$=\int_t^{+\infty}(\log x-\log t)f(x)dx=\left[-\bar F(x)(\log x-\log t)\right]_t^{+\infty}+\int_t^{+\infty}\frac{\bar F(x)}{x}dx.$$

En fait le membre de gauche est égal au membre de droite en toute généralité.

Grâce au lemme 2.5.2, on a $\bar F(x)=x^{-\alpha}L(x)=o(x^{-\alpha+\rho})$, avec $-\alpha+\rho<0$. Le membre de droite de l'équation ci-dessus se réduit donc à $\int_t^{+\infty}\frac{\bar F(x)}{x}dx$. On a d'après la deuxième partie du lemme 2.5.2,

$$\int_t^{+\infty}\frac{\bar F(x)}{x}dx=\int_t^\infty x^{-\alpha-1}L(x)\,dx\sim\frac{1}{\alpha}\,t^{-\alpha}L(t)=\frac{1}{\alpha}\,\bar F(t).$$

On en déduit donc le lemme. □

Il nous faut maintenant trouver un estimateur de $\bar F(t)=\mathbb{E}\left[\mathbf{1}_{\{X>t\}}\right]$ et un estimateur de $\mathbb{E}\left[(\log X-\log t)\mathbf{1}_{\{X>t\}}\right]$. La loi forte des grands nombres assure que $\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\{X_i>t\}}$ converge p.s. vers $\bar F(t)$. Il reste à remplacer t par une quantité qui tende vers $+\infty$ avec n. Comme pour l'estimateur de Pickand, il est naturel de remplacer t par $X_{(n-k(n)+1,n)}$, où la suite $(k(n),n\geq 1)$ satisfait les hypothèses suivantes : $\lim_{n\to\infty}k(n)=+\infty$, et $\lim_{n\to\infty}k(n)/n=0$. Cette dernière condition assure d'après la proposition 2.1.10 et le théorème 2.4.9 que p.s. $X_{(n-k(n)+1,n)}$ diverge vers l'infini.

Pour alléger les notations, notons $k(n)=k$. Si l'on suppose que F est continue, la statistique d'ordre est strictement croissante p.s., et on a pour estimation de $\bar F(X_{(n-k+1,n)})$:

$$\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\{X_i>X_{(n-k+1,n)}\}}=\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\{X_{(i,n)}>X_{(n-k+1,n)}\}}=\frac{k-1}{n}.$$

La loi forte des grands nombres assure que $\frac{1}{n}\sum_{i=1}^{n}(\log X_i - \log t)\,\mathbf{1}_{\{X_i>t\}}$ converge p.s. vers $g(t) = \mathbb{E}\left[(\log X - \log t)\mathbf{1}_{\{X>t\}}\right]$. On remplace à nouveau t par $X_{(n-k+1,n)}$, et on obtient comme estimation de $g(X_{(n-k+1,n)})$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\log X_i - \log X_{(n-k+1,n)}\right)\mathbf{1}_{\{X_i>X_{(n-k+1,n)}\}} = \frac{1}{n}\left(\sum_{i=n-k+2}^{n}\log X_{(i,n)} - (k-1)\log X_{(n-k+1,n)}\right).$$

On en déduit que

$$\frac{1}{k-1}\sum_{i=n-k+2}^{n}\log X_{(i,n)} - \log X_{(n-k+1,n)}$$

est un bon candidat pour l'estimation de ξ. Il est d'usage de remplacer $k-1$ par k sauf dans le dernier terme, ce qui ne change rien au résultat asymptotique. Nous admettrons le théorème suivant, voir [6] théorème 6.4.6.

Théorème 2.5.4. *Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi $F \in D(H(\xi))$, où $\xi > 0$. Si $\lim_{n\to\infty} k(n) = \infty$ et $\lim_{n\to\infty}\frac{k(n)}{n} = 0$, alors l'**estimateur de Hill** défini par*

$$\boxed{\xi^H_{(k(n),n)} = \frac{1}{k(n)}\sum_{i=n-k(n)+1}^{n}\log X_{(i,n)} - \log X_{(n-k(n)+1,n)}.}$$

converge en probabilité vers ξ.

Si l'on suppose de plus que $\lim_{n\to\infty}\frac{k(n)}{\log\log n} = \infty$, alors l'estimateur de Hill est fortement convergent : i.e on a la convergence presque sûre. Sous certaines hypothèses supplémentaires sur la suite $(k(n), n \geq 1)$ et sur F, on peut montrer qu'il est également asymptotiquement normal : la suite $\left(\sqrt{k(n)}\left(\xi^H_{(k(n),n)} - \xi\right), n \geq 1\right)$ converge en loi vers une variable de loi gaussienne centrée de variance ξ^2. Cela permet donc de donner un intervalle de confiance pour l'estimation. Mais attention, tout comme l'estimateur de Pickand, l'estimateur de Hill est biaisé.

Pour un échantillon de taille n fixé, on trace le diagramme de Hill : $\xi^H_{(k,n)}$ en fonction de k. On est alors confronté au dilemme suivant : si k est petit, on a une estimation avec un intervalle de confiance large ; si k est grand, l'intervalle de confiance est plus étroit, mais il faut tenir compte d'un biais inconnu.

Il existe des variantes de l'estimateur de Hill qui estiment ξ pour tout $\xi \in \mathbb{R}$, voir par exemple [6] p. 339 ou [1] p. 107.

2.6 Estimation des quantiles extrêmes

On désire estimer z_q le quantile d'ordre $1-q$ quand q est petit. Si la fonction de répartition F est continue strictement croissante, cela revient à résoudre l'équation $F(z_q) = 1-q$. On suppose que $F \in D(H(\xi))$ pour un certain $\xi \in \mathbb{R}$. Si q est fixé, alors un estimateur de z_q est le quantile empirique $X_{(n-[qn],n)}$. Nous avons déjà vu, au paragraphe 2.1, son comportement asymptotique : convergence p.s., normalité asymptotique, intervalle de confiance. Or notre problématique correspond plutôt à l'estimation de z_q quand on a peu d'observations, c'est-à-dire pour q de l'ordre de $\frac{1}{n}$. Donc

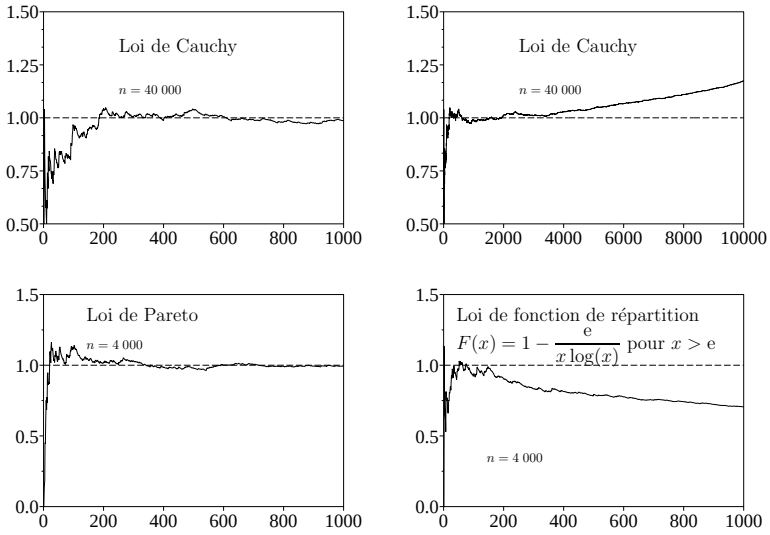

FIG. 2.7 – Estimateur de Hill : $k \to \xi^H_{(k,n)}$ à n fixé pour différentes lois.

on recherche un estimateur de z_{q_n} lorsque $q_n n$ admet une limite $c \in]0, \infty[$ quand n tend vers l'infini, et on est intéressé par son comportement asymptotique. On dit que l'estimation est à l'intérieur des données si $c > 1$, et que l'estimation est hors des données si $c < 1$.

Soit M_n le maximum de n variables aléatoires indépendantes de fonction de répartition $F \in D(H(\xi))$. Il existe donc une suite $((a_n, b_n), n \geq 1)$, telle que pour n grand,

$$\mathbb{P}(a_n^{-1}(M_n - b_n) \leq x) = F(xa_n + b_n)^n \approx H(\xi)(x).$$

Nous utiliserons en fait l'approximation plus générale suivante : pour k fixé et n grand, on a

$$F(xa_{n/k} + b_{n/k})^{n/k} \approx H(\xi)(x).$$

Ainsi, on a intuitivement

$$\begin{aligned} q_n &= 1 - F(z_{q_n}) \\ &\approx 1 - H(\xi)\left(\frac{z_{q_n} - b_{n/k}}{a_{n/k}}\right)^{k/n} \\ &= 1 - \exp\left\{-\frac{k}{n}\left(1 + \xi\frac{z_{q_n} - b_{n/k}}{a_{n/k}}\right)^{-1/\xi}\right\} \\ &\approx \frac{k}{n}\left(1 + \xi\frac{z_{q_n} - b_{n/k}}{a_{n/k}}\right)^{-1/\xi}. \end{aligned}$$

On en déduit donc que

$$z_{q_n} \approx \frac{\left(\frac{k}{nq_n}\right)^{\xi} - 1}{\xi} a_{n/k} + b_{n/k}, \tag{2.8}$$

où $q_n n \approx c$. Les estimateurs de Pickand et de Hill que nous présentons s'écrivent sous cette forme. Il reste donc à donner des estimations pour les paramètres de normalisation $a_{n/k}$ et $b_{n/k}$.

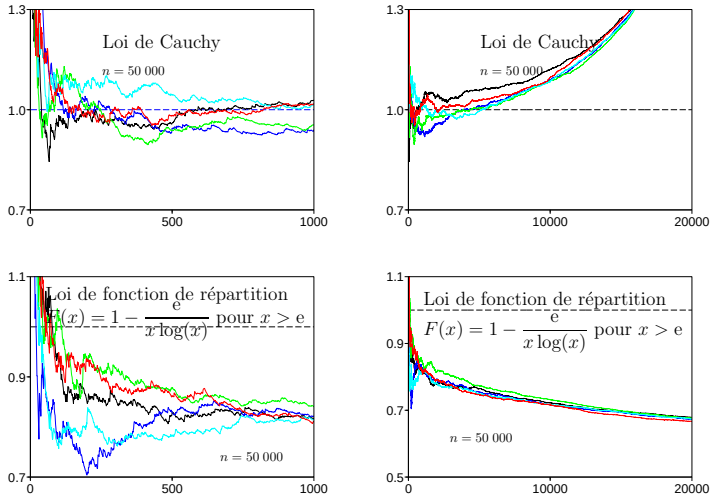

FIG. 2.8 – Estimateur de Hill : $k \to \xi^H_{(k,n)}$, à n fixé, pour différentes lois et plusieurs réalisations

2.6.1 À l'aide de l'estimateur de Pickand.

Remarquons que $z_{q_n} = U\left(\frac{1}{q_n}\right)$. De la proposition 2.4.1, on déduit que pour $\xi \neq 0$ et s assez grand, on a

$$U(sx) = \frac{x^\xi - 1}{1 - y^\xi}(U(s) - U(sy))(1 + o(1)) + U(s). \tag{2.9}$$

En faisant un choix pour s, x et y tels que $sx = \frac{1}{q_n}$ et s grand, de sorte que l'on puisse négliger $o(1)$, on désire retrouver une estimation de z_{q_n} de la forme (2.8). Pour cela, il nous faut fournir un estimateur de la fonction U. Il est naturel de choisir la fonction empirique U_n, l'inverse généralisé de la fonction de répartition empirique : $U_n(t) = F_n^{-1}\left(1 - \frac{1}{t}\right)$ où $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$. Comme p.s. $F_n\left(X_{(i,n)}\right) = \frac{i}{n}$ pour tout $i \in \{1, \ldots, n\}$ et que F_n est constante sur $[X_{(i,n)}, X_{(i+1,n)}[$, on a

$$U_n\left(\frac{n}{k}\right) = F_n^{-1}\left(\frac{n-k}{n}\right) = X_{(n-k,n)}.$$

On a (théorème de Glivenko-Cantelli) que $(F_n, n \geq 1)$ converge p.s. vers F pour la norme de la convergence uniforme. Cela implique que $U_n(x)$ converge p.s. vers $U(x)$ pour presque tout x.

Choisissons alors $s = \frac{n}{k-1}, x = \frac{1}{sq_n} = \frac{k-1}{nq_n} \approx \frac{k}{nq_n}$ et $y = 1/2$, où k **est fixé**, c'est-à-dire k ne dépend pas de n. En remplaçant $U(s)$ par $U_n\left(\frac{n}{k-1}\right) = X_{(n-k+1,n)}$ et $U(sy)$ par $U_n\left(\frac{n}{2k-2}\right) \approx U_n\left(\frac{n}{2k-1}\right) = X_{(n-2k+1,n)}$, on obtient à partir de (2.9) qu'un candidat pour estimer $z_{q_n} = U(sx)$ est :

$$\boxed{z^P_{k,q_n} = \frac{\left(\frac{k}{nq_n}\right)^{\xi^P} - 1}{1 - 2^{-\xi^P}}\left(X_{(n-k+1,n)} - X_{(n-2k+1,n)}\right) + X_{(n-k+1,n)},} \tag{2.10}$$

où ξ^P est l'estimateur de Pickand de ξ. Il faut bien sûr remplacer $\dfrac{\left(\frac{k}{nq_n}\right)^{\xi^P}-1}{1-2^{-\xi^P}}$ par $\dfrac{\log\left(\frac{k}{nq_n}\right)}{\log(2)}$ si $\xi^P=0$. On retrouve (2.8) avec $b_{n/k}=X_{(n-k+1,n)}$ et $a_{n/k}=\frac{\xi^P}{1-2^{-\xi^P}}\left(X_{(n-k+1,n)}-X_{(n-2k+1,n)}\right)$.

Nous admettrons le résultat suivant, voir [5], où deux coquilles se sont glissées dans la description de la loi de Q_k et H_k.

Théorème 2.6.1. *Soit $(X_n, n\geq 1)$ une suite de variables aléatoires indépendantes de fonction de répartition $F\in D(H(\xi))$, $\xi\in\mathbb{R}$. Supposons que $\lim_{n\to\infty} nq_n = c\in]0,\infty[$. Soit z^P_{k,q_n} l'estimateur défini par (2.10). Pour $k>c$, **fixé**, on a la convergence en loi de la suite*

$$\left(\frac{X_{(n-k+1,n)}-z_{q_n}}{X_{(n-k+1,n)}-X_{(n-2k+1,n)}}, n\geq 1\right)$$

vers $1+\dfrac{1-\left(\frac{Q_k}{c}\right)^{\xi}}{\mathrm{e}^{\xi H_k}-1}$, où Q_k et H_k sont indépendants, la loi de Q_k est la loi gamma de paramètre $(1,2k)$ et H_k a même loi que $\sum_{i=k}^{2k-1} E_i/i$, les variables E_i étant indépendantes de loi exponentielle de paramètre 1.

Le théorème ci-dessus permet de donner un intervalle aléatoire de la forme

$$\Big[a_-\left(X_{(n-k+1,n)}-X_{(n-2k+1,n)}\right)+X_{(n-k+1,n)},\\ a_+\left(X_{(n-k+1,n)}-X_{(n-2k+1,n)}\right)+X_{(n-k+1,n)}\Big],$$

où a_- et a_+ sont des quantiles de la loi de $1+\dfrac{1-\left(\frac{Q_k}{c}\right)^{\xi}}{\mathrm{e}^{\xi H_k}-1}$, qui contient z_{q_n} avec une probabilité asymptotique fixée.

Les exercices suivants permettent d'étudier directement la loi asymptotique de $\dfrac{X_{(n-k+1,n)}-z_{q_n}}{X_{(n-k+1,n)}-X_{(n-2k+1,n)}}$ dans le cas élémentaire où $k=1$, $q_n=c/n$.

Exercice 2.6.2. Soit $(X_n, n\geq 1)$ une suite de variables aléatoires indépendantes de loi exponentielle de paramètre $\lambda>0$. On pose $q_n=c/n$. On suppose que λ est inconnu.

1. Vérifier que $z_{q_n}=\log(n/c)/\lambda$.
2. Montrer en utilisant la densité de la statistique d'ordre que $X_{(n-1,n)}$ et $X_{(n,n)}-X_{(n-1,n)}$ sont indépendants.
3. Vérifier que $\lambda(X_{(n,n)}-X_{(n-1,n)})$ suit une loi exponentielle de paramètre 1.
4. Montrer, en utilisant (2.3), que $(\lambda X_{(n-1,n)}-\log(n), n\geq 2)$ converge en loi vers une variable aléatoire, $\tilde{G}$, de fonction de répartition $\mathrm{e}^{-\mathrm{e}^{-x}}(1+\mathrm{e}^{-x})$.
5. Soit $(T_n, n\geq 1)$ et $(U_n, n\geq 1)$ deux suites de variables aléatoires qui convergent en loi vers T et U, telles que T_n et U_n sont indépendantes pour tout $n\geq 1$. Montrer, en utilisant les fonctions caractéristiques par exemple, que la suite $((T_n,U_n), n\geq 1)$ converge en loi vers (T',U') où T' et U' sont indépendants et ont même loi que T et U.
6. Montrer que $(X_{(n,n)}-z_{q_n}, n\geq 2)$ converge en loi vers $\dfrac{1}{\lambda}\left[Y+\tilde{G}+\log(c)\right]$, où Y est une variable aléatoire de loi exponentielle de paramètre 1 indépendante de $\tilde{G}$.
7. Vérifier que $\left(\dfrac{X_{(n,n)}-z_{q_n}}{X_{(n,n)}-X_{(n-1,n)}}, n\geq 2\right)$ converge en loi et que la limite ne dépend pas de λ. Vérifier que la loi limite correspond à celle donnée dans le théorème 2.6.1.

8. En déduire un intervalle aléatoire qui contient z_{q_n} avec une probabilité asymptotique fixée. Vérifier que la largeur de cet intervalle aléatoire ne tend pas vers 0 quand n tend vers l'infini.

♦

Exercice 2.6.3. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi uniforme sur $[0, \theta]$, avec $\theta > 0$. On pose $q_n = c/n$. On suppose que θ est inconnu.

1. Vérifier que $z_{q_n} = \theta(1 - \frac{c}{n})$.
2. Montrer en utilisant la densité de la statistique d'ordre que $X_{(n,n)}$ et $X_{(n-1,n)}/X_{(n,n)}$ sont indépendants.
3. Montrer que $\left(n(1 - \frac{X_{(n,n)}}{\theta}), n \geq 2\right)$ converge en loi vers une variable aléatoire V_1 de loi exponentielle de paramètre 1.
4. En déduire que $\left(n(X_{(n,n)} - z_{q_n}), n \geq 2\right)$ converge en loi vers $\theta(c - V_1)$.
5. Vérifier que $\left(n(1 - \frac{X_{(n-1,n)}}{X_{(n,n)}}), n \geq 2\right)$ converge en loi vers V_2 de loi exponentielle de paramètre 1.
6. Soit $(T_n, n \geq 1)$ et $(U_n, n \geq 1)$ deux suites de variables aléatoires qui convergent en loi vers T et U, telles que T_n et U_n sont indépendantes pour tout $n \geq 1$. Montrer, en utilisant les fonctions caractéristiques par exemple, que la suite $((T_n, U_n), n \geq 1)$ converge en loi vers (T', U') où T' et U' sont indépendants et ont même loi que T et U.
7. En déduire que $\left(\frac{X_{(n,n)} - z_{q_n}}{X_{(n,n)} - X_{(n-1,n)}}, n \geq 2\right)$ converge en loi vers $\frac{c - V_1}{V_2}$, où V_1 et V_2 sont indépendantes de loi exponentielle de paramètre 1.
8. Montrer que si Y_1 et Y_2 sont deux variables aléatoires indépendantes de loi exponentielle de paramètre 1, alors $Y_1 + Y_2$ suit la loi gamma de paramètre $(1, 2)$, $\frac{Y_1}{Y_1 + Y_2}$ suit la loi uniforme sur $[0, 1]$ et ces deux variables aléatoires sont indépendantes.
9. Soit Q_1 de loi gamma de paramètre $(1, 2)$ et H_1 de loi exponentielle de paramètre 1. Déterminer la loi de e^{-H_1}. Déduire de la question précédente que (Q_1, e^{-H_1}) a même loi que $(Y_1 + Y_2, Y_1/(Y_1 + Y_2))$, puis que $(Q_1\,\mathrm{e}^{-H_1}, Q_1(1 - \mathrm{e}^{-H_1}))$ a même loi que (Y_1, Y_2).
10. Vérifier que la loi de $\frac{c - V_1}{V_2}$ correspond à celle $1 + \frac{1 - \frac{c}{Q_1}}{\mathrm{e}^{-H_1} - 1}$ avec les notations du théorème 2.6.1.
11. En déduire un intervalle aléatoire qui contient z_{q_n} avec probabilité asymptotique fixée. Vérifier que la largeur de cet intervalle aléatoire est d'ordre $1/n$ quand n tend vers l'infini.

♦

2.6.2 À l'aide de l'estimateur de Hill

On suppose que $\xi > 0$. L'estimateur du quantile associé à l'estimateur de Hill est donné par

$$\boxed{z^H_{k,q_n} = \left(\frac{k}{nq_n}\right)^{\xi^H} X_{(n-k+1,n)},} \tag{2.11}$$

où ξ^H est l'estimateur de Hill de ξ. On retrouve la forme donnée par (2.8) avec $b_{n/k} = a_{n/k}/\xi^H = X_{(n-k+1,n)}$. Nous renvoyons à [1] paragraphe 4.6 et [6] page 348 pour les propriétés de cet estimateur et les références correspondantes.

2.7 Conclusion

Le paragraphe précédent a permis de répondre à la première question posée : Trouver une estimation de z_q telle que la probabilité que la surcote de la marée durant une tempête soit plus haute que z_q est q. On peut utiliser l'estimateur de Pickand ou l'estimateur de Hill et fournir un intervalle de confiance.

Rappelons la deuxième question : Soit q fixé, typiquement de l'ordre de 10^{-3} ou 10^{-4}, trouver y_q tel que la probabilité pour que la plus grande surcote **annuelle** soit supérieure à y_q est q. Pour cela il faut estimer le nombre moyen de tempêtes par an, disons k_0. La probabilité pour que parmi k_0 tempêtes, il y ait une surcote supérieure à h est $1 - F(h)^{k_0}$. On en déduit donc que y_q est solution de $F(y_q)^{k_0} = 1 - q$. On trouve $y_q = z_{q'}$ où $q' = 1 - (1 - q)^{1/k_0}$. On peut utiliser l'estimateur de Pickand ou l'estimateur de Hill et fournir un intervalle de confiance pour la réponse. Pour l'exemple précis concernant les Pays-Bas, Haan [4] observe que le paramètre de forme est très légèrement négatif, et il choisit de l'estimer à 0, car cela permet d'avoir des résultats plus conservateurs, dans le sens où les probabilités que la marée dépasse un niveau donné sont majorées. Il semble que de manière générale, les phénomènes observés dans les domaines de la finance et de l'assurance correspondent à des paramètres ξ positifs. En revanche les phénomènes météorologiques correspondent plutôt à des paramètres ξ négatifs.

Soulignons en guise de conclusion que pour l'estimation des quantiles, il est vivement recommandé d'utiliser plusieurs méthodes. Ainsi dans l'article [4], pas moins de huit méthodes différentes sont utilisées pour fournir une réponse et la commenter.

Bibliographie

[1] J. Beirlant, Y. Goegebeur, J. Teugels et J. Segers. *Statistics of extremes*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2004.

[2] N. Bingham, C. Goldie et J. Teugels. *Regular variation*. Cambridge University Press, Cambridge, 1987.

[3] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001.

[4] L. de Haan. Fighting the arch-enemy with mathematics. *Stat. Neerl.*, 44(No.2) :45–68, 1990.

[5] A. L. M. Dekkers et L. de Haan. On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.*, 17(4) :1795–1832, 1989.

[6] P. Embrechts, C. Klueppelberg et T. Mikosch. *Modelling extremal events for insurance and finance*, volume 33 d'*Applications of Mathematics*. Springer, Berlin, 1997.

[7] M. Falk, J. Hüsler et R.-D. Reiss. *Laws of small numbers : extremes and rare events*, volume 23 de *DMV Seminar*. Birkhäuser Verlag, Basel, 1994.

[8] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46 :323–351, 2005.

[9] S. Resnick. *Extreme values, regular variation, and point processes*. Applied Probability. Springer-Verlag, New York, 1987.