MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université Paris Cité

Exploring the Extremes : A Deep Dive into Extreme Value Theory

Shamir JAILANY, Seifeldine ABDALLA

Année universitaire 2023/2024

Master Mathématiques, Modélisations, Apprentissages

Responsable universitaire : Anne Sabourin

Remerciements

Nous tenons à exprimer notre sincère gratitude envers Madame Anne Sabourin, qui nous a non seulement proposé un sujet de projet enrichissant et stimulant, mais qui nous a également guidés avec dévouement et expertise tout au long de notre travail. Son soutien constant, ses encouragements et ses conseils précieux ont été essentiels à notre réussite. Nous sommes profondément reconnaissants pour son engagement sans faille qui a fait de cette expérience une période d'apprentissage très enrichissante et agréable pour nous deux.

Résumé

La théorie des valeurs extrêmes est un domaine captivant à l'intersection des mathématiques pures et des applications pratiques, offrant des perspectives précieuses sur les événements rares et impactants. Ses origines remontent aux travaux pionniers du début du vingtième siècle, où des statisticiens ont élaboré des modèles pour comprendre et prévoir le comportement des valeurs situées aux extrémités des distributions de données.

Au cœur de cette théorie se trouve l'analyse des données extrêmes, celles qui dépassent les seuils habituels d'observation. Les méthodes employées, telles que les estimateurs du maximum de vraisemblance et les modèles asymptotiques, permettent d'estimer la probabilité d'occurrences extrêmes. En particulier, les approches asymptotiques modélisent les distributions aux queues lourdes, où les événements extrêmes, bien que rares, peuvent avoir une intensité significative.

Les applications de ces principes sont vastes. En ingénierie civile, par exemple, la théorie aide à concevoir des structures résistantes à des événements naturels catastrophiques, tels que des tempêtes extrêmes. En science des matériaux, elle permet de comprendre la probabilité de défaillance sous des charges extrêmes. Dans la finance et les télécommunications, elle évalue les risques de pertes sévères et anticipe les pics de trafic.

Ce rapport est structuré en trois chapitres principaux, afin de montrer comment la théorie des valeurs extrêmes peut être appliquée de manière systématique et rigoureuse pour résoudre des problèmes pratiques.

Le premier chapitre introduit les bases théoriques nécessaires pour comprendre l'EVT. Nous y aborderons les statistiques d'ordre, les lois des valeurs extrêmes, et les distributions limites des plus grandes observations.

Le deuxième chapitre s'appuie sur les concepts introduits précédemment pour examiner les techniques d'estimation des paramètres critiques de l'EVT. Nous y détaillerons l'estimation par maximum de vraisemblance (MLE), l'estimateur de Hill, et la méthode de Weissman pour estimer les quantiles extrêmes. Ces méthodes sont cruciales pour analyser et interpréter les événements rares.

Enfin, le troisième chapitre illustre l'application pratique des concepts et méthodes étudiés. Nous utiliserons les techniques d'estimation des valeurs extrêmes pour analyser un jeu de données hormonales et détecter des cas potentiels de dopage. Cette étude concrète démontre la pertinence de l'EVT dans des contextes réels et fournit un exemple tangible de son utilité.

Table des matières

1 Théorie des Valeurs Extrêmes

1.1 Introduction

Les statistiques classiques des valeurs extrêmes se concentrent sur le comportement asymptotique de la distribution des maximums d'une séquence de variables aléatoires. Considérons une séquence de variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées avec une fonction de distribution F. Pour un échantillon de taille n, le maximum est défini par $M_n = \max(X_1, X_2, \ldots, X_n)$. L'objectif est de dériver la distribution limite de M_n à mesure que n tend vers l'infini, sous un rééquilibrage approprié.La théorie des valeurs extrêmes (EVT) a été développée pour estimer les probabilités d'occurrences d'événements rares et extrêmes. Elle permet d'extrapoler le comportement des queues de distribution en utilisant les valeurs les plus élevées observées, c'est-à-dire les observations extrêmes dans un échantillon.

Ce chapitre propose une introduction aux concepts des statistiques d'ordre, des valeurs extrêmes et des queues de distribution. Nous y présentons les résultats principaux concernant les distributions limites des plus grandes observations d'un échantillon ainsi que les domaines d'attraction correspondants. En règle générale, les résultats relatifs aux minima peuvent être déduits de ceux concernant les maxima par la relation suivante :

$$\min(X_1, ..., X_n) = -\max(-X_1, ..., -X_n)$$

1.2 Concepts et définitions

1.2.1 Définition : Fonction de répartition

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) définies sur le même espace de probabilité. Leur fonction de répartition F est :

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

En plus, notons par $\bar{F}$ la fonction de survie (ou la fonction des queues) :

$$\bar{F}(x) = P(X > x) = 1 - F(x), \quad x \in \mathbb{R}.$$

1.2.2 Définition : Statistiques d'ordre

Soient $X_1, \ldots, X_n$ des variables aléatoires (v.a) indépendantes et identiquement distribuées (i.i.d.) avec une fonction de distribution commune F et une fonction de densité f. Considérons les variables aléatoires $X_{1,n}, \ldots, X_{n,n}$ triées par ordre croissant, telles que :

$$X_{1,n} \leq \ldots \leq X_{n,n}$$

Les variables aléatoires $(X_{1,n}, \ldots, X_{n,n})$ sont appelées les statistiques d'ordre de l'échantillon $(X_1, \ldots, X_n)$.

Pour $1 \leq k \leq n$, la variable $X_{k,n}$ est connue sous le nom de k-ième statistique d'ordre, ou statistique d'ordre k. Deux statistiques d'ordre sont particulièrement intéressantes pour l'étude des événements extrêmes : ce sont les statistiques d'ordre extrêmes, définies comme les termes correspondant aux maximum et minimum des n variables aléatoires $X_1, \ldots, X_n$. La variable $X_{1,n}$ est la plus petite statistique d'ordre (ou statistique du minimum), et $X_{n,n}$ est la plus grande statistique d'ordre (ou statistique du maximum).

1.2.3 Définition : Variation régulière

Une fonction $U : \mathbb{R}^+ \to \mathbb{R}^+$ est régulièrement variable (RV) s'il existe $\rho \in \mathbb{R}$ tel que

$$\forall x > 0, \lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\rho.$$

Le paramètre ρ est appelé l'indice de variation régulière. Nous écrivons « $U \in RV(\rho)$ », ce qui signifie que U est régulièrement variable avec l'indice de variation régulière ρ. Si $\rho = 0$, U est dit lentement variable.

1.3 Lois des valeurs extrêmes

La question centrale de la théorie classique des valeurs extrêmes consiste à déterminer la nature de la distribution limite de $M_n = \max(X_1, \ldots, X_n)$, où $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes suivant tous une fonction de répartition F commune. L'objectif est de trouver des constantes de mise à l'échelle $a_n > 0$ et b_n telles que la distribution de $\frac{M_n - b_n}{a_n}$ converge vers une distribution limite non dégénérée lorsque $n \to \infty$. Cette recherche mène à l'expression suivante (admise) :

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x)$$

où $G(x)$ est une fonction de répartition non dégénérée. Cette convergence permet de mieux comprendre le comportement asymptotique des maximums d'une séquence de variables aléatoires indépendantes et identiquement distribuées (i.i.d.).

1.3.1 Théorème des Types Extrêmes et Distribution Extrême Généralisée (GEV)

Le Théorème stipule que la distribution limite $G(x)$ appartient à l'une des trois familles de distributions : la distribution de Gumbel, de Fréchet, ou la distribution inversée de Weibull. Cela implique que, indépendamment de la forme exacte de F, la distribution des maximums normalisés ne peut converger que vers ces types spécifiques de distributions.

$$\text{Famille Gumbel : } G(z) = \exp\left\{-\exp\left\{-\frac{(z-b)}{a}\right\}\right\}, \quad -\infty < z < \infty \tag{1}$$

$$\text{Famille Fréchet : } G(z) = \begin{cases} 0, & \text{si } z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & \text{si } z > b \end{cases} \tag{2}$$

$$\text{Famille Weibull : } G(z) = \begin{cases} 1 & \text{si } z \geq b, \\ \exp\left\{-\left[-\left(\frac{z-b}{\alpha}\right)\right]^a\right\}, & \text{si } z < b \end{cases} \tag{3}$$

a > 0, b et $\alpha > 0$ sont des paramètres qui indiquent respectivement : où la distribution est centrée ; la dispersion ou la largeur de la distribution ; la forme de la distribution de probabilité.

Distribution Extrême Généralisée (GEV)

La distribution Extrême Généralisée (GEV) encapsule les trois familles de distributions du théorème des types extrêmes en une seule forme de distribution, avec la fonction de distribution :

$$G(x;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{4}$$

où μ est le paramètre de localisation, $\sigma > 0$ est le paramètre d'échelle, et ξ est le paramètre de forme (aussi appelé indice de queue, qui détermine le type de distribution limite), qui jouera un rôle central dans la suite de notre étude.

La GEV fournit un modèle pour la distribution des maxima de blocs. Le principe est de diviser les données en blocs de taille égale et d'ajuster la GEV à l'ensemble des maxima de blocs. Le choix de la taille du bloc est crucial : des blocs trop petits conduisent à une mauvaise approximation et des blocs trop grands génèrent une grande variance d'estimation.

Les maxima de blocs sont désignés par Z_1,..., Z_m et sont supposés être des variables indépendantes d'une distribution GEV. Même si les données initiales X_i sont dépendantes, cela peut être utile dans des situations où les conditions du Théorème ne sont pas satisfaites.

Pour estimer les paramètres dans les modèles de valeurs extrêmes, il existe des méthodes basées sur les moments et sur les statistiques d'ordre, ainsi que des méthodes basées sur la vraisemblance. Les techniques basées sur la vraisemblance sont généralement préférées pour leur utilité globale et leur adaptabilité à la construction de modèles complexes (Chapitre 2).

Bien que la modélisation des maxima par blocs soit une approche efficace pour certaines analyses de valeurs extrêmes, elle peut ne pas être optimale lorsque d'autres données sur les extrêmes sont disponibles. Une alternative consiste à modéliser les excès par rapport à un seuil élevé, une méthode qui peut mieux utiliser l'information disponible. Cette approche se concentre sur les observations qui dépassent un seuil donné.

1.3.2 La Distribution Généralisée de Pareto (GPD)

Soient $X_1, X_2, \ldots$, une séquence de variables aléatoires indépendantes et identiquement distribuées, ayant une fonction de répartition marginale F. Les événements extrêmes sont ceux des X_i qui dépassent un seuil élevé u. En désignant un terme arbitraire dans la séquence des X_i par X, la probabilité conditionnelle des événements extrêmes est donnée par :

$$\Pr\{X > u + y \mid X > u\} = \frac{1 - F(u+y)}{1 - F(u)}, \quad y > 0. \tag{5}$$

Si nous connaissions la distribution d'origine F, nous pourrions déterminer exactement la distribution des valeurs qui dépassent un certain seuil u. Mais, en pratique, nous ne connaissons pas toujours cette distribution d'origine. Nous utilisons donc des approximations pour les valeurs qui dépassent des seuils élevés u. Cette méthode est similaire à celle où nous utilisons la distribution GEV pour estimer la distribution des maximums d'une série de données lorsque nous ne connaissons pas la distribution d'origine.

Théorème : Soient $X_1, X_2, \ldots$ une séquence de variables aléatoires indépendantes ayant une fonction de distribution commune F, et soit $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Désignons un terme

arbitraire dans la séquence des X_i par X, et supposons que F satisfait les conditions du Théorème des Types Extrêmes, de sorte que pour n suffisamment grand,

$$\Pr\{M_n \leq z\} \approx G(z),$$

où

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{1/\xi}\right\}$$

pour certains $\mu, \sigma > 0$ et ξ.

Alors, pour u suffisamment grand, la fonction de distribution de $X - u$, conditionnellement à $X > u$, est approximativement donnée par :

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}, \tag{6}$$

définie sur $\{y : y > 0 \text{ et } (1 + \xi y/\tilde{\sigma}) > 0\}$, où

$$\tilde{\sigma} = \sigma + \xi(u - \mu) \tag{7}$$

Le Théorème peut également être rendu plus précis, justifiant (5) comme une distribution limite lorsque u augmente.

La famille de distributions définie par l'Équation (6) est appelée la famille de Pareto généralisée. Le Théorème implique que si les maxima par blocs ont une distribution d'approximation G, alors les excès de seuil ont une distribution approximative correspondante dans la famille de Pareto généralisée. De plus, les paramètres de la distribution de Pareto généralisée des excès de seuil sont uniquement déterminés par ceux de la distribution GEV associée des maxima par blocs. En particulier, le paramètre ξ dans (6) est égal à celui de la distribution GEV correspondante. Le choix d'une taille de bloc différente, mais toujours grande, affecterait les valeurs des paramètres GEV, mais pas ceux de la distribution de Pareto généralisée correspondante des excès de seuil : ξ est invariant par rapport à la taille du bloc, tandis que le calcul de ξ dans (7) n'est pas perturbé par les changements de μ et σ qui se compensent mutuellement.

La dualité entre les familles GEV et de Pareto généralisée signifie que le paramètre de forme ξ est dominant pour déterminer le comportement qualitatif de la distribution de Pareto généralisée, tout comme il l'est pour la distribution GEV. Retenons que :

- Si $\xi > 0$, la distribution a une queue lourde, typique des distributions de Pareto, on a une probabilité relativement élevée d'observer des valeurs extrêmes.
- Si $\xi = 0$, la distribution a une queue exponentielle : la probabilité de valeurs extrêmes décroît exponentiellement avec la taille de l'événement.
- Si $\xi < 0$, la distribution a une queue bornée, ce qui signifie qu'il existe une limite supérieure au-delà de laquelle aucune observation ne peut se produire.

Maintenant que nous avons établi une base solide en présentant les concepts fondamentaux de la théorie des valeurs extrêmes, nous pouvons nous tourner vers les méthodes d'estimation des paramètres clés de cette théorie. Ces méthodes sont essentielles pour appliquer l'EVT à des données réelles, en particulier pour estimer les quantiles extrêmes et les indices de queue qui caractérisent la probabilité des événements rares. Le prochain chapitre se concentrera donc sur ces techniques d'estimation, en explorant des approches telles que l'estimation par maximum de vraisemblance et l'estimateur de Hill. Ces outils nous permettront de modéliser et de quantifier

les risques associés aux événements extrêmes de manière précise et rigoureuse, préparant ainsi le terrain pour l'application à des données réelles dans le cadre de la détection de dopage dans le sport.

2 Méthodes d'Estimation des Paramètres et des Quantiles Extrêmes dans la Théorie des Valeurs Extrêmes

Ce chapitre constitue la dernière étape avant de passer à l'application concrète sur le jeu de données. Nous y détaillerons les méthodes d'estimation des paramètres essentiels de la théorie des valeurs extrêmes, nécessaires pour analyser les événements rares et extrêmes. Les techniques telles que l'estimation par maximum de vraisemblance (MLE) et l'estimateur de Hill sont au cœur de cette analyse. Ces méthodes permettent non seulement de quantifier les valeurs extrêmes, mais aussi de modéliser de manière précise les comportements aux queues des distributions. Nous explorerons en particulier comment ces estimations peuvent être appliquées pour déterminer les quantiles extrêmes et les indices de queue. Ce chapitre prépare ainsi le terrain pour l'application concrète de ces techniques dans l'étude de cas sur la détection de dopage sportif, présentée au chapitre suivant.

2.1 Estimation des Paramètres

2.1.1 Estimation par Maximum de Vraisemblance (MLE)

L'estimation par maximum de vraisemblance (MLE) est une méthode statistique utilisée pour estimer les paramètres des distributions de probabilité. Cette méthode consiste à maximiser la fonction de vraisemblance, qui représente la probabilité des données observées en fonction des paramètres du modèle.

MLE pour la Distribution GEV :

Pour la distribution GEV, les paramètres à estimer sont μ (paramètre de localisation), σ (paramètre d'échelle) et ξ (paramètre de forme). La fonction de vraisemblance est définie comme suit :

$$\ell(\mu, \sigma, \xi) = -n \log(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{n} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}$$

sous réserve que :

$$1 + \xi \left(\frac{z_i - \mu}{\sigma}\right) > 0, \quad \text{pour} \quad i = 1, \ldots, n.$$

Si $\xi = 0$, on utilise la limite de Gumbel de la distribution GEV, ce qui donne la log-vraisemblance suivante :

$$\ell(\mu, \sigma, 0) = -n \log(\sigma) - \sum_{i=1}^{n} \left[z_i - \mu + \sigma \log \left(1 + \frac{z_i - \mu}{\sigma}\right)\right].$$

La maximisation de la fonction de log-vraisemblance permet d'obtenir les estimations des paramètres $\hat{\mu}$, $\hat{\sigma}$ et $\hat{\xi}$.

MLE pour la Distribution GPD :

Pour la distribution GPD, les paramètres à estimer sont σ (paramètre d'échelle) et ξ (paramètre de forme). Une fois le seuil u sélectionné, les excès $Y_i = X_i - u$ peuvent être modélisés par une GPD. La log-vraisemblance pour les paramètres de la GPD lorsque $\xi \neq 0$ est donnée par :

$$\ell(\sigma, \xi) = -k \log(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log \left[1 + \xi \left(\frac{y_i}{\sigma}\right)\right]$$

sous réserve que :

$$1 + \xi \left(\frac{y_i}{\sigma}\right) > 0, \quad \text{pour} \quad i = 1, \ldots, k.$$

Si $\xi = 0$, la log-vraisemblance est obtenue en utilisant la distribution exponentielle :

$$\ell(\sigma) = -k \log(\sigma) - \sum_{i=1}^{k} \left(\frac{y_i}{\sigma}\right).$$

La maximisation de ces expressions par rapport aux paramètres (σ, ξ) donne les estimations du maximum de vraisemblance pour la distribution GPD. En pratique, on utilise en général des techniques numériques pour cette optimisation, en veillant à ce qu'il n'y ait pas d'instabilités numériques et en vérifiant que les conditions sur ξ et σ sont respectées.

2.1.2 Niveaux de Retour

Le terme "niveau de retour" se réfère à un événement extrême dont la probabilité d'occurrence est égale ou inférieure à une certaine valeur p. Il est souvent utilisé dans l'analyse des risques naturels ou des phénomènes climatiques pour estimer la fréquence à laquelle des événements extrêmes peuvent se produire. Les preuves des résultats théoriques à ce sujet ne sont pas ce qui nous intéresse ici. Les résultats sont énoncés à but informatif. En pratique, nous utiliserons des outils numériques, tels que les fonctions `fit.gev` ou `gpd.diag`, pour vérifier les différents modèles et assurer la validité de nos estimations.

Niveaux de Retour pour la Distribution GEV :

Une fois que les paramètres μ, σ et ξ ont été estimés par MLE, ils peuvent être utilisés pour calculer les niveaux de retour. Le niveau de retour z_p pour une probabilité p (avec $0 < p < 1$) est donné par :

$$\hat{z}_{y_p} = \begin{cases} \mu - \frac{\hat{\sigma}}{\hat{\xi}} \left(1 - y_p^{-\hat{\xi}}\right) & \text{pour } \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}(-\log(y_p)) & \text{pour } \hat{\xi} = 0, \end{cases}$$

où $y_p = -\log(1 - p)$.

La variance de z_p peut être estimée par la méthode du delta (admise) :

$$\text{Var}(z_p) \approx (\nabla z_p)^T V (\nabla z_p),$$

où V est la matrice de variance-covariance des paramètres estimés $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ et

$$\nabla z_p = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1, -\xi^{-1}(1 - y_p^{-\xi}), -\sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1}y_p^{-\xi}\log y_p\right]$$

est évaluée en utilisant les valeurs estimées $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Pour les longues périodes de retour, correspondant à de petites valeurs de p, ce sont les plus intéressantes. Si $\hat{\xi} < 0$, il est également possible de faire des estimations sur le point de fin supérieur de la distribution (la valeur extrême la plus élevée que la distribution peut atteindre), correspondant à z_p avec $p = 0$.

L'estimation du maximum de vraisemblance est :

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}.$$

et la formule de variance reste valable avec

$$\nabla z_0^T = [1, -\xi^{-1}, \sigma\xi^{-2}],$$

encore une fois évaluée à $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Lorsque $\hat{\xi} \geq 0$, l'estimation du maximum de vraisemblance du point de fin supérieur est infinie.

Niveaux de Retour pour la Distribution GPD :

Pour la GPD, le niveau de retour X_m correspondant à m observations est défini comme :

$$X_m = u + \frac{\sigma}{\xi}\left[(m\bar{F}(u))^{\xi} - 1\right]$$

où $\bar{F}(u)$ est l'estimation de la probabilité d'excès au-dessus du seuil u.

2.2 Estimateur de Hill

L'estimateur de Hill est un estimateur classique utilisé pour déterminer l'indice de queue α d'une distribution de probabilité, en utilisant les k plus grandes valeurs de l'échantillon. Il est largement utilisé en théorie des valeurs extrêmes, particulièrement dans le contexte des queues lourdes pour estimer la décroissance de la probabilité des événements extrêmes. Cet estimateur est central pour la suite de notre étude puisque c'est celui que nous utiliserons pour nos analyses. En particulier, nous utiliserons l'estimateur de Hill et sa version améliorée pour estimer des quantiles extrêmes, ce qui nous permettra de mieux vérifier le caractère anormal de certaines données dans notre application pratique sur le dopage, et ainsi affiner nos analyses et prévisions des comportements extrêmes.

2.2.1 Proposition : Estimateur de Hill

Considérons une série de variables aléatoires i.i.d $X_1, \ldots, X_n$ tirées d'une distribution F, dont la queue suit une loi régulière à variation lente $(1 - F(x) \in \mathrm{RV}(-\frac{1}{\xi}))$. Et soit un entier k (nombre de plus grandes observations à considérer) qui augmente avec n (le nombre total d'observations), mais de telle sorte que le rapport k/n tende vers 0 quand n tend vers l'infini. On ordonne les valeurs de l'échantillon de manière décroissante $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$. Alors l'estimateur de Hill est donné par :

$$\hat{\xi}_n = \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{(i)}}{X_{(k)}} \right),$$

où $\hat{\xi}_n$ est l'estimation consistante de ξ.

Le choix du paramètre k est crucial : un k trop petit peut entraîner une estimation biaisée en raison de la faible taille de l'échantillon, tandis qu'un k trop grand peut inclure des valeurs qui ne se trouvent pas dans la queue de la distribution, augmentant ainsi la variance de l'estimation. En pratique on le chosit tel que $1 \ll k \ll n$.

Lorsqu'on augmente n la taille de l'échantillon, l'estimation de $1/\xi$ devient de plus en plus précise.

2.2.2 Estimateur de Hill Ajusté (Trimmed Hill Estimator)

L'estimateur de Hill ajusté, aussi appelé trimmed Hill estimator ou BLUE estimator (Best Linear Unbiased Estimator), améliore l'estimateur de Hill classique en réduisant l'impact des valeurs très élevées. Cet estimateur est conçu pour être sans biais et minimiser la variance de l'indice de queue ξ. Il offre donc une estimation plus fiable de ξ.

Cette estimation est cruciale pour notre application sur le dopage. Les données des sportifs peuvent contenir des valeurs extrêmes, indiquant potentiellement des cas de dopage. En utilisant l'estimateur de Hill ajusté, nous réduisons l'impact de ces très grandes valeurs (en excluant certaines valeurs très élevées qui pourraient fausser les résultats.) et obtenons une estimation plus précise des quantiles extrêmes.

Ces quantiles extrêmes nous permettent d'établir des seuils critiques pour identifier les performances anormales.

Proposition : Estimateur de Hill Ajusté : Considérons $X_1, \ldots, X_n$ des variables aléatoires i.i.d suivant une distribution de Pareto(σ, ξ). L'estimateur de Hill ajusté pour $0 \leq k_0 < k < n$ est donné par :

$$\hat{\xi}_{k_0,k}(n) = \frac{k_0}{k - k_0} \log \left(\frac{X_{(n-k_0,n)}}{X_{(n-k,n)}} \right) + \frac{1}{k - k_0} \sum_{i=k_0+1}^{k} \log \left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}} \right),$$

où le deuxième terme est l'estimateur de Hill classique appliqué aux observations $X_{(n-k_0,n)} \geq \cdots \geq X_{(n-k,n)}$, excluant les k_0 plus grandes valeurs.

Cet estimateur est le meilleur parmi les estimateurs linéaires car il a la plus petite variance. Il est robuste parce qu'il exclut les valeurs les plus extrêmes, réduisant ainsi l'impact des valeurs aberrantes.

Le choix du paramètre k_0 est essentiel. En pratique, nous utiliserons une estimation de ξ avec l'estimateur de Hill classique pour estimer un quantile extrême et déterminer un seuil. Cela nous donnera une idée des plus grandes valeurs extrêmes qui pourraient biaiser l'estimation. Ensuite, nous pourrons tester différents k_0 pour l'utilisation de l'estimateur de Hill ajusté.

2.3 Choix du Seuil

Un seuil approprié équilibre le biais et la variance des estimations des paramètres. Un seuil trop bas peut violer les hypothèses asymptotiques du modèle, tandis qu'un seuil trop élevé réduit le nombre d'excès disponibles, augmentant ainsi la variance des estimations.

Le seuil doit être suffisamment élevé pour satisfaire les hypothèses du modèle, mais suffisamment bas pour garantir un nombre adéquat d'excès pour des estimations fiables.

Nous ne présenteons que le calcul du seuil obtenu par l'estimation d'un quantile extrême, dit estimateur de Weissman. Les points au-dessus de ce quantile extrême pourront être déclarés comme des possibles anomalies. C'est ce fameux seuil qui sera utilisé dans l'étude pratique de nos données sur les sportifs.

2.3.1 Estimation des quantiles extrêmes par la méthode Weissman

Considérons le cas des queues lourdes. La méthode présentée ci-dessous utilise l'estimation de $1/\xi > 0$, afin d'estimer un quantile extrême.

Considérons la fonction $U = (1/(1-F))^{-1}$, on dispose du résultat (admis) :

$$\frac{U(ty)}{U(t)} \xrightarrow[t\to\infty]{} y^{\xi}.$$

Notre objectif est d'estimer le quantile extrême $U(1/p)$ en utilisant l'approximation asymptotique $\frac{U(ty)}{U(t)} \approx y^{\xi}$. Avec k choisi tel que $p < k/n \ll 1$ pour que les quantiles estimés sont dans la queue de la distribution, et $t = n/k$, une estimation de $U(t) = U(n/k)$ est $X_{(k)}$, le k-ième (plus grand) ordre statistique de l'échantillon. En choisissant y tel que $ty = 1/p$, qui donne $y = k/(np)$, on obtient :

$$\frac{U(1/p)}{U(n/k)} \approx \left(\frac{k}{np}\right)^{\xi}.$$

On remplace les quantités inconnues par leurs versions empiriques, ce qui donne l'estimateur appelé "Estimateur de Weissman" :

$$\hat{z}_{p,w} = \hat{U}(1/p) = X_{(k)} \left(\frac{k}{np}\right)^{\hat{\xi}}$$

où $\hat{\xi}$ est un estimateur consistant de l'indice de queue (on utilisera l'estimateur de Hill) construit avec un échantillon i.i.d. $(X_i)_{i\leq n} \sim F$.

Armés des méthodes d'estimation détaillées précédemment, nous sommes maintenant prêts à passer à l'application sur nos données réelles. La troisième partie se consacre à une étude de cas concrète où nous utiliserons les techniques d'estimation des valeurs extrêmes pour détecter d'éventuelles anomalies indicatives de dopage dans le sport. En analysant un jeu de données sur les quantités d'hormones chez les athlètes, nous appliquerons les outils et méthodes développés précédemment pour identifier des valeurs anormales. Cette application pratique illustrera la pertinence et l'efficacité de la théorie des valeurs extrêmes dans des contextes réels.

3 Application concrête : Détection de dopage dans le milieu du sport

Dans cette partie, nous nous intéressons à l'analyse de données hormonales pour la détection de cas potentiels de dopage chez les athlètes. Nous disposons d'un jeu de données contenant les quantités de deux hormones spécifiques mesurées chez des sportifs, hommes et femmes. Ces données peuvent contenir des valeurs anormalement élevées, qui pourraient indiquer l'utilisation de substances dopantes.

L'objectif de cette étude est d'identifier les anomalies dans les valeurs extremes, afin de détecter les cas potentiels de dopage. Nous nous restreindrons ici à l'études de l'hormone REC_ng_mL pour les athlètes femmes. L'étude pour les hommes est similaires.

Le processus méthodologique s'articule autour des étapes suivantes :

- **Préparation des données :** Importation et nettoyage des données des hormones pour les sportifs hommes et femmes.
- **Estimation de l'indice de queue ξ :** Utilisation de l'estimateur de Hill pour obtenir une estimation initiale de ξ et détermination du meilleur k en utilisant le Hill Plot.
- **Validation de k avec FitRange :** Vérification de la stabilité des paramètres estimés pour différents seuils u afin de confirmer le choix de k.
- **Ajustement de la GPD :** Application du modèle GPD aux données au-delà du seuil déterminé et analyse des diagnostics du modèle.
- **Estimation des quantiles extrêmes :** Calcul des quantiles extrêmes à l'aide de la méthode de Weissman en utilisant l'estimateur de Hill.
- **Vérification avec Hill Trimmed :** Réévaluation de ξ en retirant les plus grandes valeurs extrêmes pour s'assurer qu'elles n'influencent pas de manière disproportionnée l'estimation initiale.
- **Détection des anomalies :** Identification des valeurs potentiellement anormales en comparant les quantiles extrêmes estimés aux données observées.

Cette approche nous permettra de fixer des seuils objectifs et scientifiquement fondés pour la détection des anomalies hormonales. Les sections suivantes décrivent en détail le processus de préparation des données, l'application des modèles statistiques, et l'interprétation des résultats obtenus. En appliquant ces techniques, nous visons à offrir un cadre analytique solide pour la détection des cas potentiels de dopage dans le milieu sportif.

Afin d'illustrer l'application pratique de cette méthodologie, nous présentons ci-dessous le code R utilisé pour réaliser l'analyse.

3.1 Préparation des Données

Tout d'abord, nous installons les packages nécessaires et chargeons les données :

```r
install.packages("extRemes")
install.packages("ismev")
install.packages("ggplot2")
install.packages("evmix")
install.packages("pid")
install.packages("ReIns")

library(pid)
library(extRemes)
```

```r
library(evmix)
library(ismev)
library(ggplot2)
library(evir)
library(ReIns)

female <-
    read.csv("/Users/seifeldineabdalla/Documents/ProjetS2/sport_femmes.csv")
dff <- as.numeric(female$REC_ng_mL)
data <- na.omit(dff)
length(data)
```

3.2 Le jeu de données

Nous avons un jeu de données de n = 4542 valeurs. Nous commençons par examiner les 10 valeurs les plus élevées de notre jeu de données :

```r
n <- length(data)
tail(sort(data), n = 10) # On regarde les plus grandes valeurs
```

3.3 Hill Plot

Le Hill Plot est utilisé pour trouver la région de stabilité de k, ce qui nous permet de choisir k pour une estimation fiable de ξ.

```r
dev.new(width=8, height=6) # Ajuster la taille de la fenêtre graphique
data <- data[data > 0]
hh <- hillplot(data, orderlim = c(2, 500), xlab = 'k', y.alpha = FALSE,
    try.thresh = NULL, main = 'Hill plot')

abline(v = 100, lty = 2)
abline(v = 150, lty = 2)
abline(v = 200, lty = 2)
abline(v = 300, lty = 2)
abline(h = 0, lty = 4)
```

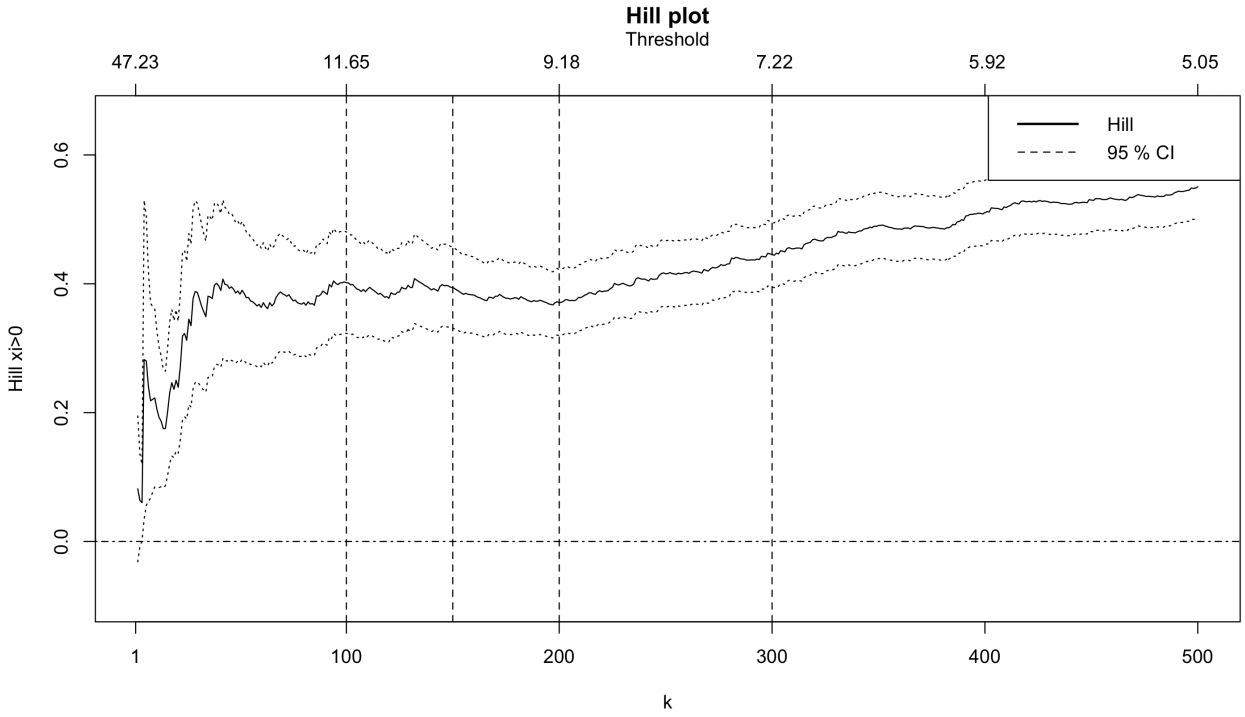

FIGURE 1 – FitRange Plot pour l'estimation de ξ et de l'échelle

Résultats et Analyse :

Le Hill plot ci-dessus montre l'estimation de l'indice de queue ξ en fonction de k, le nombre de valeurs les plus grandes utilisées dans l'estimation. Les points importants à noter dans ce graphique sont :

- **Stabilité de l'Estimation :** Nous observons que l'estimation de l'indice de queue ξ devient relativement stable pour des valeurs de k comprises entre 100 et 200. Cela signifie que dans cette plage de k, l'estimateur de Hill fournit une estimation fiable de ξ. Pour ce k, l'estimateur de Hill nous donne $\hat{\xi}$ environ égale à 0,38 (on le calculera exactement plus tard lorsqu'on validera ce k)
- **Choix de k :** Pour nos analyses ultérieures, nous choisissons $k = 200$, car il offre une bonne estimation de ξ sans trop dériver de ce qui se passe pour des k plus petits.
- **Valeur de ξ :** Pour $k = 200$, l'indice de queue ξ est estimé à environ 0.38. Cela indique que la distribution des performances sportives présente des queues lourdes, signifiant une probabilité plus élevée d'observer des valeurs extrêmes.

En choisissant ce k, nous pouvons établir un seuil d'entraînement pour sélectionner les données extrêmes, ce qui nous permettra d'estimer les paramètres pour la Generalized Pareto Distribution (GPD) comme le quantile $(1 - k/n)$.

3.4 Vérification du Seuil avec FitRange

Nous utilisons ensuite 'fitrange' pour vérifier si notre choix de k est approprié pour estimer les paramètres de la GPD, ξ et σ :

```r
fr <- gpd.fitrange(data, umin= quantile(data,1-800/n), umax =
    quantile(data,1-15/n))
```

Interprétation du FitRange :

Le FitRange nous permet de vérifier si notre choix de k est bon pour l'estimation des paramètres de la Generalized Pareto Distribution (GPD). L'objectif est de trouver le seuil le plus petit possible tel que les estimations des paramètres restent stables.

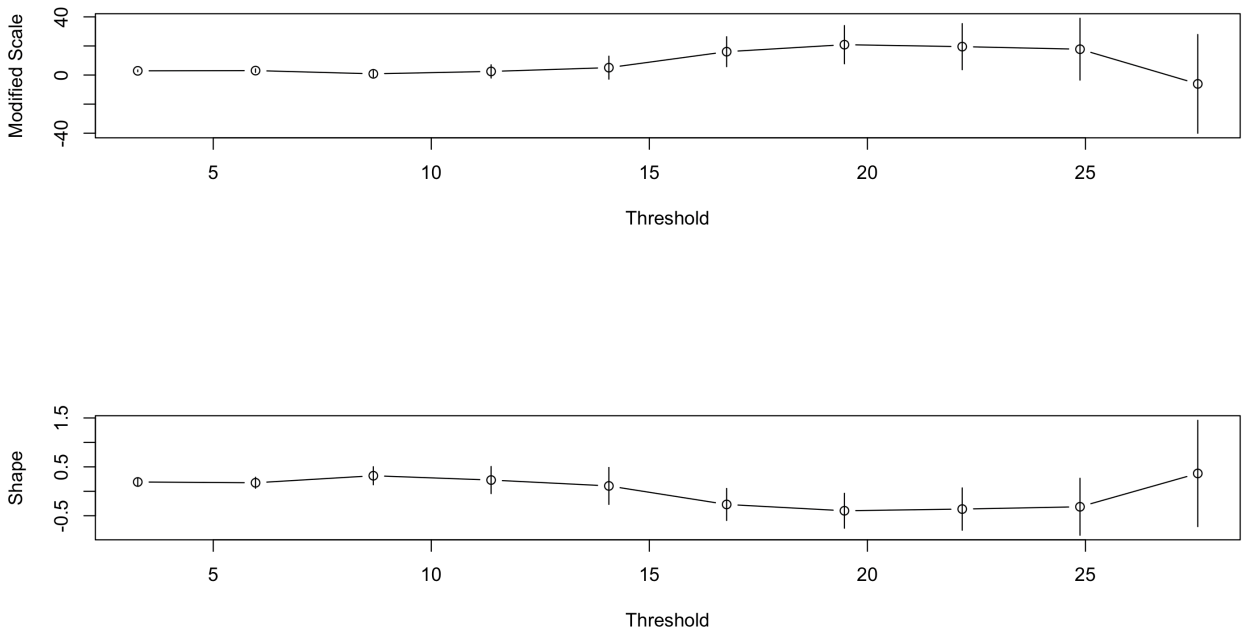

FIGURE 2 – FitRange Plot pour l'estimation de ξ et de l'échelle

Stabilité des Estimations : Les courbes du FitRange montrent que les estimations des paramètres restent relativement stables dans la plage des seuils choisis.
Choix du Seuil : Notre quantile est $1 - 200/n$, ce qui équivaut à un seuil de 13,9 environ pour nos données. Le FitRange montre une stabilité raisonnable autour de ce choix de seuil, confirmant qu'il est approprié pour l'estimation des paramètres de la GPD.
Observation de la Plage de Seuils : À partir de 15, un biais commence à apparaître, représenté par la présence d'une forme de U autour de 20. Cela signifie que les intervalles de confiance commencent à croître dangereusement et que les estimations dévient. Il y a donc trop de variance à droite de cette plage.

En résumé, le FitRange valide notre choix de k en montrant que les estimations des paramètres restent stables autour du seuil choisi, et il indique que des biais apparaissent à partir de 15, confirmant que notre choix initial de seuil est adéquat.

3.5 Ajustement de la GPD

Pour ajuster une Generalized Pareto Distribution (GPD) aux données, nous utilisons le seuil trouvé précédemment :

```
gpdfit <- gpd.fit(sort(data, decreasing = TRUE), threshold = quantile(data,
    1 - 200/n))
gpd.diag(gpdfit)
xi_mle <- gpdfit$mle[2]
xi_mle
```

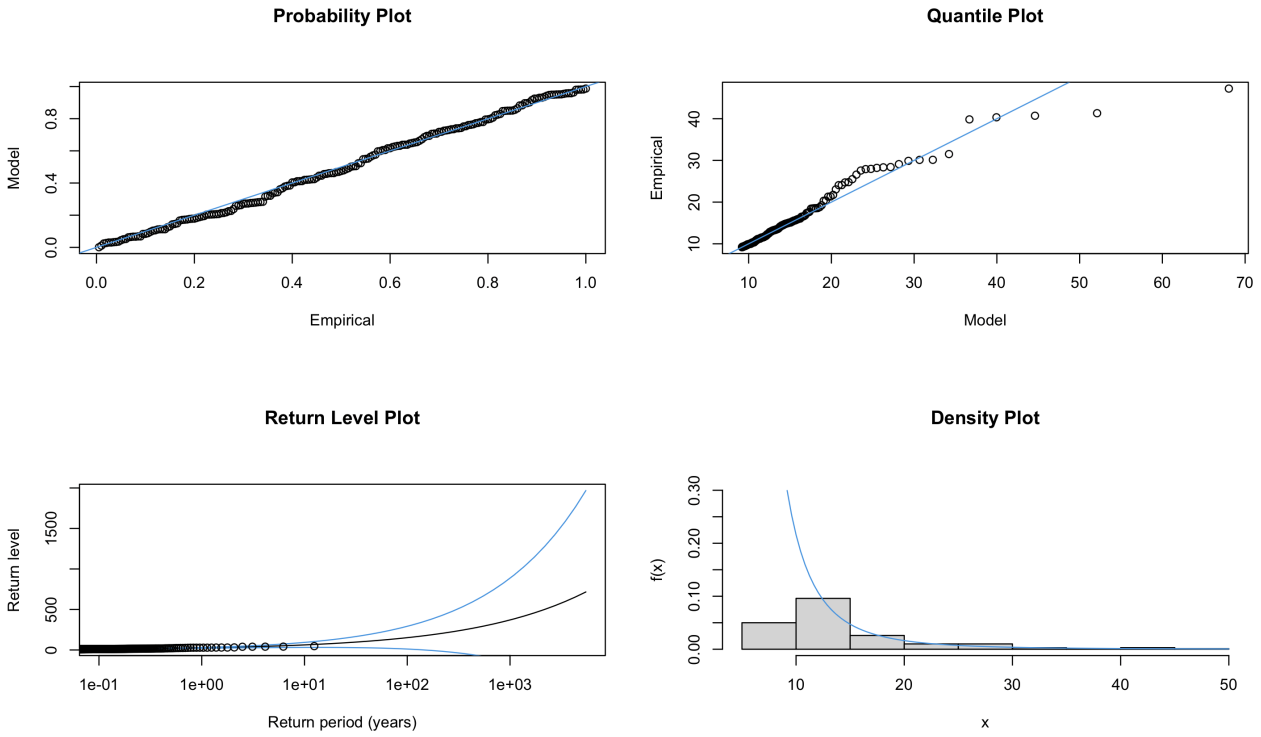

FIGURE 3 – Gpd diag

Résultats et Analyse :

Stabilité des Ajustements : En observant le Return Level Plot, nous ne voyons aucun problème majeur. Les intervalles de confiance indiquent que les données ne sortent pas des bornes prévues.
Estimation de ξ : L'estimation de ξ par Maximum Likelihood Estimation (MLE) donne une valeur de 0.3883702 Cette valeur est importante car elle indique une queue lourde, signifiant une probabilité plus élevée d'observer des valeurs extrêmes.

3.6 Analyse du Profil Likelihood

Pour nous assurer que notre estimation de ξ est robuste, nous effectuons une analyse du profil likelihood. Étant donné que la valeur de ξ est proche de 0, nous voulons tester l'hypothèse nulle $H_0 : \xi \leq 0$.

```r
gpd.profxi(gpdfit, xlow = -0.05 , xup = 1, conf = 0.95, nint = 100)
```

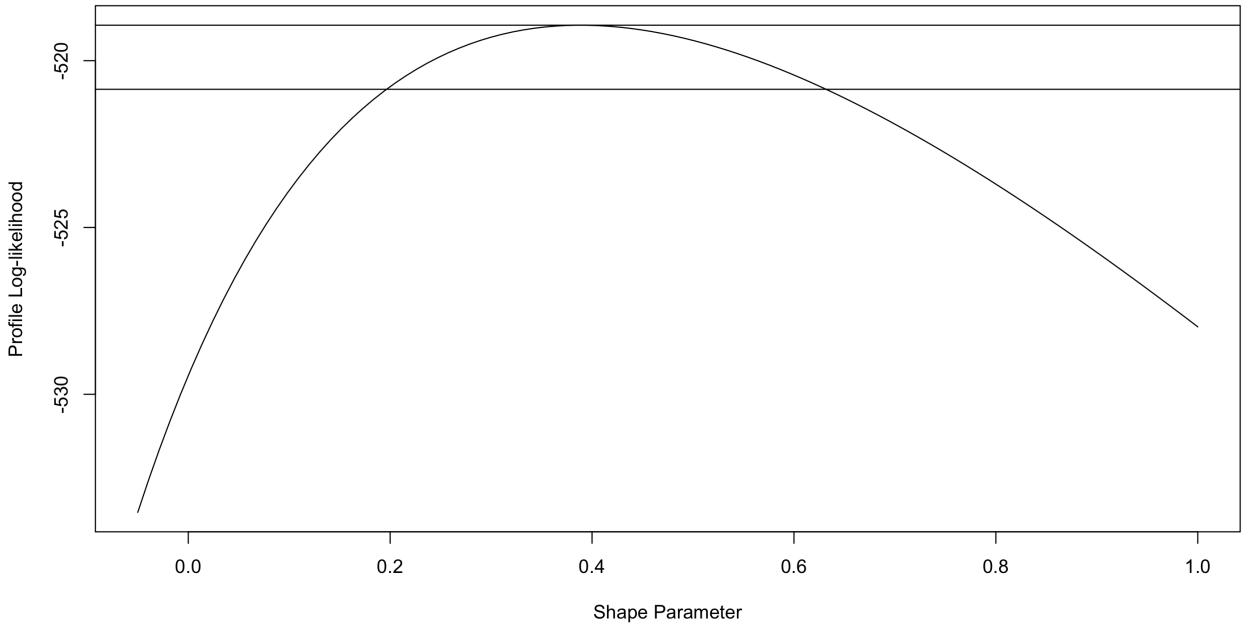

FIGURE 4 – Profil Likelihood pour l'estimation de ξ

Résultats et Analyse :

Test d'Hypothèse : Le profil likelihood nous permet de construire un intervalle de confiance pour $\hat{\xi}$. L'analyse montre que $\hat{\xi}$ appartient à l'intervalle de confiance [0.2, 0.6].

Conclusion : Étant donné que 0 n'appartient pas à cet intervalle de confiance, nous pouvons rejeter l'hypothèse nulle $H_0 : \xi \leq 0$. Cela nous rassure sur le fait que ξ est positif et que notre estimation est correcte.

En résumé, l'ajustement de la GPD et l'analyse du profil likelihood confirment que notre choix de k = 200 pour l'estimation est robuste et fiable. Les tests montrent que notre ξ est significativement différent de 0, validant ainsi notre approche et nos estimations.

Nous pouvons alors déterminer l'estimateur de Hill pour k=200.

```r
hill_estimator <- Hill(data_clean)
xi_hill <- hill_estimator$gamma[200]
xi_hill
```

3.7 Estimation du Quantile Extrême avec Weissman

Nous estimons ensuite le quantile extrême à l'aide de l'estimateur de Hill et de la méthode de Weissman. Ce quantile nous serivra de seuil pour notre détermination de valeurs extrêmes anormales :

```r
hill_estimator <- Hill(data_clean)
xi_hill <- hill_estimator$gamma[200]
xi_hill

weissman_quantile <- function(data, gamma_hat, p, k) {
  n <- length(data)
  X_k <- sort(data, decreasing = TRUE)[k]
  q_extreme <- X_k * ((k / (n * p)) ^ gamma_hat)
  return(q_extreme)
}

p = 0.001 # Quantile 1-1/1000 (0.1%)
q_extreme <- weissman_quantile(data_clean, xi_hill, p, k)
print(q_extreme)
```

Nous déterminons les valeurs au-dessus de ce seuil (les valeurs extremes) et les affichons :

```r
extreme_values <- data_clean[data_clean > q_extreme]
print(extreme_values)
length(extreme_values)
```

```
> extreme_values <- data_clean[data_clean > q_extreme]
> print(extreme_values)
[1] 47.225 41.324 40.713 40.340 39.835
> length(extreme_values)
[1] 5
>
```

FIGURE 5 – Valeurs extrêmes

```r
plot(data, main = "Données avec seuil de quantile extrême avec p = 0.001",
    ylab = "Niveaux d'hormones", xlab = "Index")
abline(h = q_extreme, col = "red", lwd = 2, lty = 2)
```

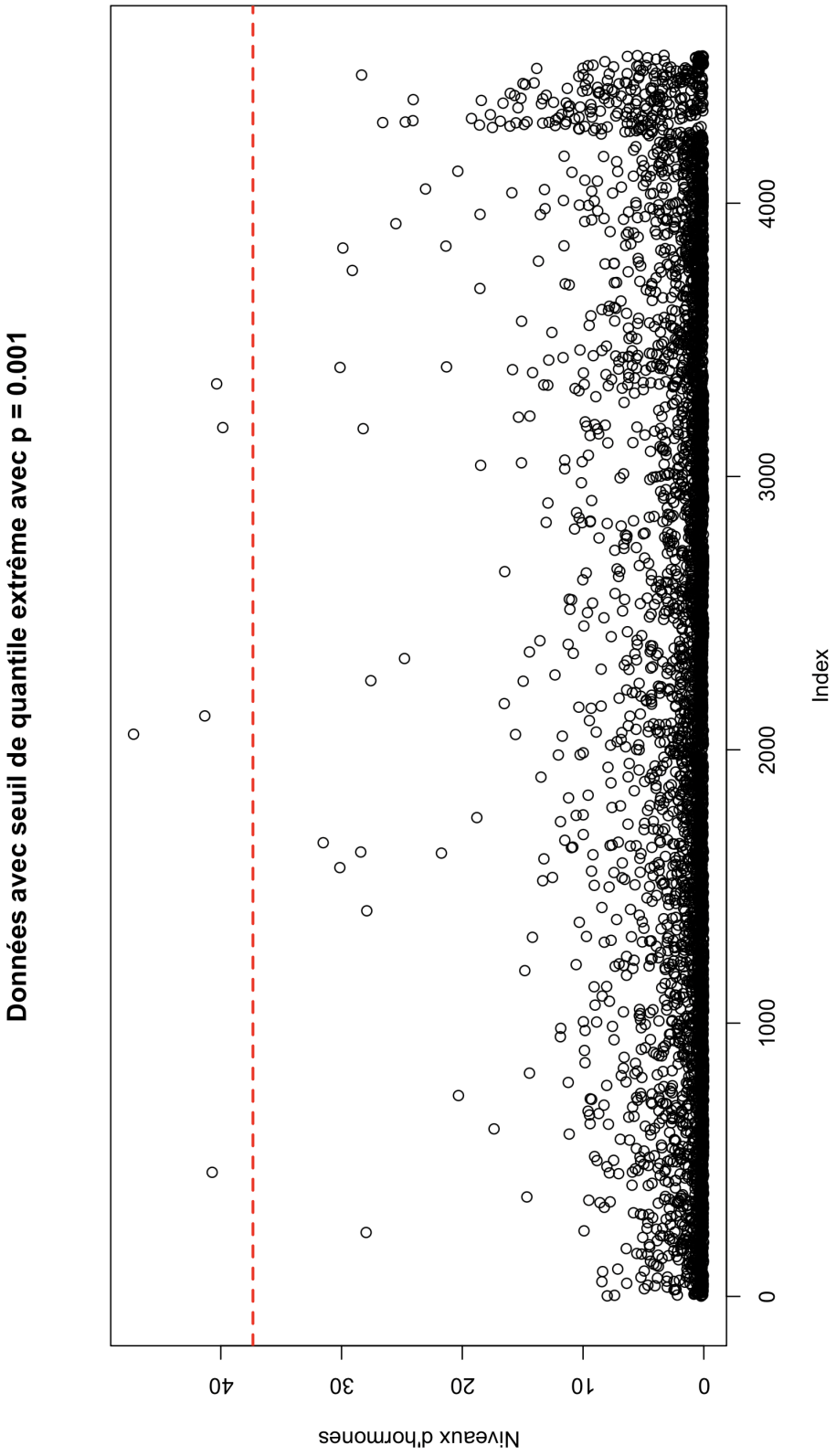

FIGURE 6 – Visualisation des données avec le seui de quantiles extrèmes déterminé

3.8 Conclusion sur l'estimation du quantile extrême avec l'estimateur de Hill

L'estimation du quantile extrême avec la méthode de Weissman et l'estimateur de Hill nous fournit un seuil au-delà duquel les valeurs sont considérées comme extrêmes. Pour $p = 0.001$, nous trouvons que le seuil est de 37.33673. Nous observons que 5 valeurs dépassent ce seuil, ce qui est conforme aux attentes théoriques de $n/1000$. Par conséquent, ces valeurs ne sont pas considérées comme des anomalies.

Cette analyse montre que, bien que certaines valeurs soient au-dessus du seuil, elles sont en accord avec ce que l'on peut théoriquement attendre d'un échantillon de cette taille. Ainsi, avec cette méthode, nous n'avons pas détecté de dopage parmi les valeurs extrêmes de notre jeu de données.

Nous nous interrogeons maintenant sur les données qui dépassent le seuil : dans quelle mesure sont-elles extrêmes et à quel quantile extrême correspondent-elles ? Nous constatons que le quantile extrême de chaque valeur détectée est toujours au-dessus du seuil. Cela signifie que nous ne pouvons rien conclure de concret sur l'anormalité, ni établir d'intervalle de confiance fiable pour ces données.

3.9 Estimation avec Hill Trimmed

Avec l'estimateur de Hill classique, nous n'avons pas pu détecter d'anomalies, car les valeurs dépassant le quantile extrême étaient conformes aux attentes théoriques. Cela pourrait indiquer que les valeurs les plus grandes ont biaisé notre estimation de ξ. En effet, si nous nous basons sur des valeurs potentiellement anormales, il est logique que nous trouvions des valeurs normales et donc pas de valeurs extrêmes douteuses.

Pour vérifier si ces grandes valeurs influencent notre estimation, nous allons utiliser un estimateur de Hill ajusté. Cet estimateur, appelé Hill Trimmed, retire les plus grandes valeurs extrêmes de l'estimation. Par exemple, avec l'estimateur de Hill classique et le quantile choisi, nous avons identifié 5 valeurs extrêmes. Bien que retirer $k0 = 5$ ne change pas grand-chose, nous allons supposer que les 50 plus grandes valeurs sont douteuses. Nous retirons donc ces 50 valeurs et recalculons l'estimateur.

Cette approche nous permet d'obtenir une estimation de ξ sans l'influence des données extrêmes, et nous espérons ainsi obtenir une estimation plus fiable. Voici le code utilisé pour cette méthode :

```r
trimmed_hill <- function(data, k, k0) {
  n <- length(data)
  data_sorted <- sort(data, decreasing = FALSE)
  log_ratio1 <- log(data_sorted[n - k0] / data_sorted[n - k + 1])
  log_ratio2 <- sum(log(data_sorted[(n - k0):(n - k - 1)] / data_sorted[n -
      k + 1]))

  hill_estimator <- (k0 / (k - k0)) * log_ratio1 + (1 / (k - k0)) *
      log_ratio2
  return(hill_estimator)
}
k0 <- 50
xii_trimmed <- trimmed_hill(data, k, k0)
xii_trimmed
```

```
13
14 qextremet <- weissman_quantile(data_clean, xii_trimmed, p, k)
15 qextremet
16
17 extreme_values <- data_clean[data_clean > qextremet]
18 print(extreme_values)
19 length(extreme_values)
20
21 plot(data, main = "Données avec seuil de quantile extrême pour Hill
       transformé avec k0 = 50", ylab = "Niveaux d'hormones", xlab = "Index")
22 abline(h = qextremet, col = "red", lwd = 2, lty = 2)
```

```
> #xi_trimmed <- trimmed_hill(data, k, k0)
> #xi_trimmed
> xii_trimmed <- trimmed_hill(data,k,k0)
> xii_trimmed
[1] 0.3657922
> qextremet <- weissman_quantile(data_clean, xii_trimmed, p, k)
> qextremet
[1] 36.65732
>
```

FIGURE 7 – Estimateur de Hill ajusté ainsi que son quantile extrême pour p = 0,001

```
- -
> qextremet <- weissman_quantile(data_clean, xii_trimmed, p, k)
> qextremet
[1] 36.65732
> extreme_values <- data_clean[data_clean > qextremet]
> print(extreme_values)
[1] 47.225 41.324 40.713 40.340 39.835
> length(extreme_values)
[1] 5
>
```

FIGURE 8 – Valeurs extrêmes pour le quantile extrême avec Hill ajusté

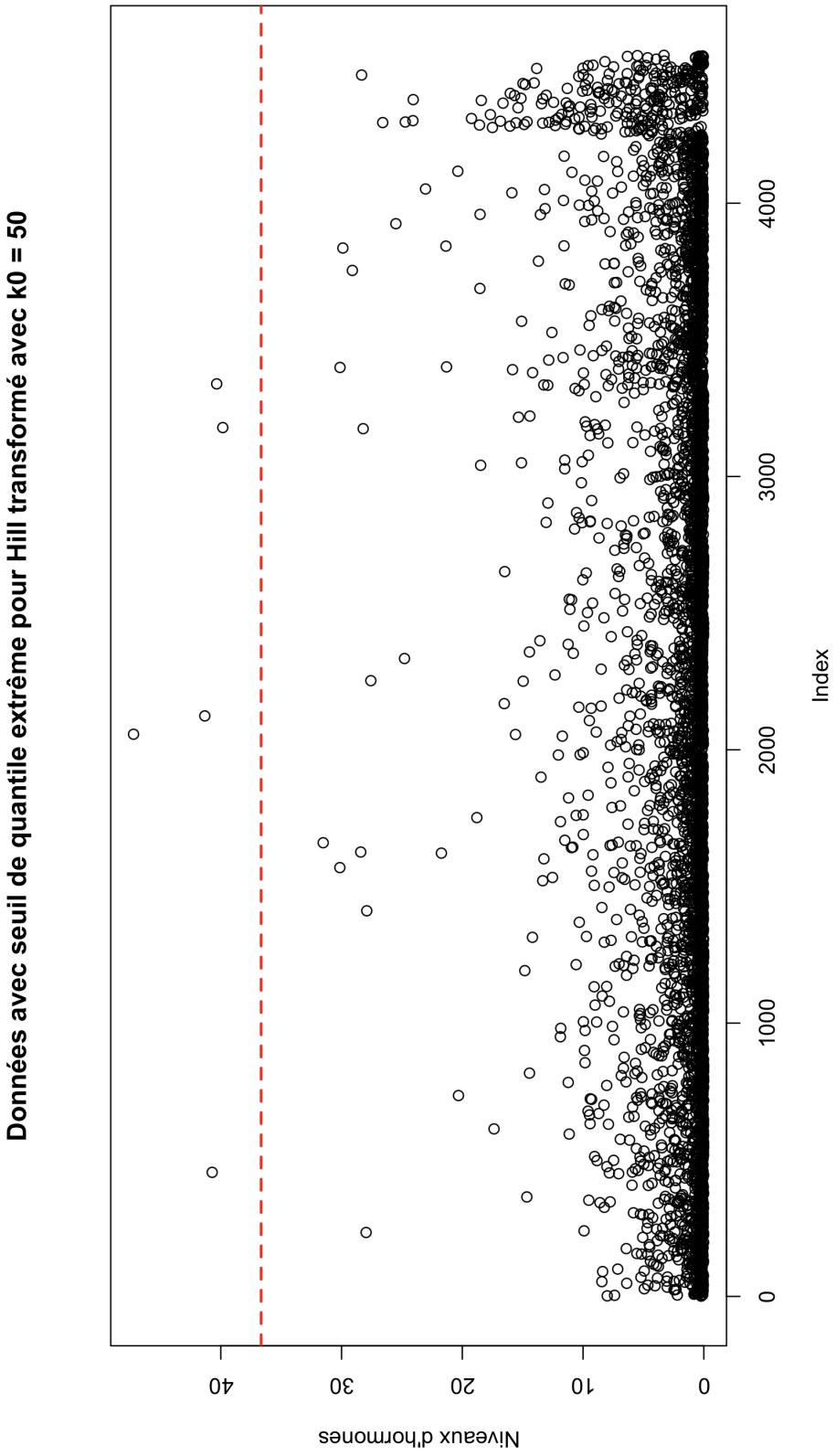

FIGURE 9 – Visualisation des données avec le seuil de quantiles extrêmes déterminé avec l'estimateur de Hill ajusté

Résultats et Analyse :

Nous utilisons un estimateur de Hill ajusté pour retirer les plus grandes valeurs extrêmes de l'estimation. En retirant les 50 plus grandes valeurs ($k0 = 50$), nous recalculons l'estimateur. Cependant, cette modification ne change pas grand-chose, car l'estimateur reste très similaire au premier. Cela suggère que les valeurs extrêmes initiales n'ont pas biaisé notre estimation.

En utilisant un estimateur de Hill ajusté pour retirer les plus grandes valeurs extrêmes, nous espérons obtenir une estimation plus fiable de ξ. Après avoir retiré les 50 plus grandes valeurs ($k0 = 50$), nous recalculons l'estimateur. Cependant, cette modification ne change pas grand-chose, car l'estimateur reste très similaire au premier. Cela suggère que les valeurs extrêmes initiales n'ont pas biaisé notre estimation.

3.10 Conclusion

Dans cette étude, nous avons analysé les niveaux d'hormones chez des sportifs pour détecter d'éventuelles anomalies pouvant suggérer un dopage. Nous avons utilisé diverses méthodes statistiques, telles que le Hill Plot, le FitRange et l'ajustement de la GPD, pour estimer l'indice de queue ξ et identifier les valeurs extrêmes.

Le Hill Plot nous a permis de déterminer que $k = 200$ était un choix approprié pour l'estimation de ξ. Cette estimation a été validée par le FitRange, qui a confirmé la stabilité des paramètres jusqu'à un certain seuil. L'ajustement de la GPD a donné une estimation de ξ à 0,38, validée par un test de profil de vraisemblance excluant zéro de l'intervalle de confiance.

L'estimation des quantiles extrêmes via la méthode de Weissman a révélé que cinq valeurs dépassaient le seuil de 37,33 pour $p = 0.001$, ce qui correspond aux attentes théoriques. Nous avons également utilisé un estimateur de Hill ajusté pour vérifier l'impact des valeurs extrêmes, mais cette modification n'a pas significativement altéré nos résultats.

Bien que nos méthodes aient montré leur efficacité théorique, elles se sont révélées inefficaces pour détecter des anomalies dans notre étude spécifique. Les valeurs identifiées comme extrêmes correspondaient aux attentes théoriques, ne permettant pas de détection de comportements de dopage.

Ce travail offre néanmoins une base méthodologique solide pour des applications futures dans d'autres contextes de détection d'anomalies.

4 Conclusion générale

En conclusion, notre document montre l'importance de la théorie des valeurs extrêmes (EVT) dans la compréhension et la gestion des événements rares. Le Théorème des Types Extrêmes et les distributions de Gumbel, Fréchet et Weibull ont été présentés comme bases théoriques essentielles. Les méthodes présentées, telles que l'estimation par maximum de vraisemblance et l'estimateur de Hill, permettent d'analyser de manière rigoureuse les données extrêmes et de tirer des conclusions fiables. L'application concrète à la détection de dopage montre que, bien que les outils de l'EVT soient puissants, ils doivent être utilisés avec soin pour garantir des résultats pertinents et fiables.

Nous avons appliqué les concepts et les méthodes de l'EVT à une étude de cas réelle concernant la détection de dopage chez les athlètes. En analysant les données hormonales, nous avons utilisé les outils développés pour identifier des valeurs anormalement élevées pouvant indiquer un usage de substances dopantes. Les analyses ont montré que, bien que certaines valeurs soient extrêmes, elles n'étaient pas suffisantes pour être considérées comme des anomalies indiquant un dopage. Cette application pratique a illustré la pertinence et l'efficacité de l'EVT dans des contextes réels, fournissant un exemple tangible de son utilité.

En somme, l'EVT offre un cadre analytique solide et adaptable pour diverses applications, de l'ingénierie à la finance, en passant par le sport, permettant de mieux comprendre et gérer les risques associés aux événements rares.

Références

[1] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag.

[2] Clémençon, S., Sabourin, A. (2023). *Statistical Learning with Extreme Values Master MVA*. École Normale Supérieure Paris-Saclay, October 30, 2023.

[3] Bhattacharya, S., Kallitsis, M. (2024). Data-adaptive trimming of the Hill estimator and detection of outliers in the extremes of heavy-tailed data.