République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Statistique**

Par

ZEROUG Meriem

Titre :

Statistique d'ordre : Estimation et test statistique

Membres du Comité d'Examen :

Pr.	**SAYAH Abdallah**	UMKB	Président
Pr.	**NECIR Abdelhakim**	UMKB	Encadreur
Dr.	**BERKANE Hassiba**	UMKB	Examinatrice

Juin 2022

Dédicace

Je dédie ce modeste travail

À mes parents.

À mon frère et ma soeur.

À ma famille.

À mes amies.

À mes collègues de mathématiques 2021/2022.

REMERCIEMENTS

Tout d'abord, je remercie "**Allah**" Le Tout-Puissant de m'avoir aidé et donné la santé et volonté pour arriver à ce stade.

Mes vifs remerciements, sont adressés à mon encadreur **Mr. NECIR ABDELHAKIM** pour ses précieux conseils, ses orientations pertinentes et sa patience tout au long de la réalisation de ce mémoire.

Je tiens à remercier : **Mr. Sayah Abdallah**. et **Mme**. **Berkane Hassiba** qui m'ont fait l'honneur de faire partie du jury de soutenance.

Je remercie tous les enseignants qui ont contribué à ma formation, ainsi que tous les employés du département de mathématiques.

Je remercie tout particulièrement mes parents, pour leur encouragement et soutien sur tous les aspects, ainsi que toute ma famille.

Je n'oublie pas l'ensemble de mes amis mes proches et aussi mes collègues d'études.

À ceux qui ont contribué, de près ou de loin, à la réalisation de ce modeste travail.

Un grand merci à vous tous.

Table des matières

Table des figures

Introduction

Les statistiques d'ordre sont un concept très utile en sciences statistiques. En effet, elles ont de nombreuses applications, à savoir l'estimation paramétrique et non paramétrique et la modélisation des phénomènes aléatoires de façon générale telle que les catastrophes naturelles, les fluctuations financières, les coûts de sinistres (polices d'assurance). Dans ce mémoire nous présentons une synthèse sur le sujet en question et nous présentons aussi quelques outils de base concernant les propriétés asymptotiques des statistiques d'ordre et leurs applications. Le mémoire est organisé comme suit :

Le premier chapitre : est consacré à l'étude des distributions d'une statistique d'ordre, Nous commençons par présenter les définitions et les propriétés générales des statistiques d'ordre associées à un échantillon de n variables aléatoires $\{(X_1, X_2, ..., X_n)\,, \forall n \in \mathbb{N}^*\}$. En outre nous donnons les propriétés des lois des statistiques d'ordre ainsi que leurs distributions jointes. Nous étudions aussi les représentations de celles-ci les relations générales sur les moments et en particulier les moments des statistiques d'ordre uniforme et exponentiel. Nous terminons notre étude par le comportement asymptotique des statistiques d'ordre modérées et intermédiaires.

Le deuxième chapitre : dans ce chapitre, nous discutons l'estimation optimale de quelques paramètres statistiques des modèles probabilistes en se basant sur les statistiques d'ordre. Plus précisément nous intéressons par les estimateurs linéaires sans biais de variance minimale $(MVLUE)\,s$ (en anglais Minimum Variance Linear

Unbiased Estimators). En d'autre part, nous énonçons les caractérisations générales des valeurs extrêmes, et les domaines d'attraction du maximum d'un échantillon. Ces lois sont indexées par un paramètre appelé indice des valeurs extrême noté γ. Nous distinguons trois domaines d'attraction : Weibull ($\gamma < 0$), Fréchet ($\gamma > 0$) et Gumbel ($\gamma = 0$). Nous énonçons ensuite les conditions d'appartenance d'une fonction de répartition à l'un de ces trois domaines d'attraction. Nous présentons aussi la distribution des valeurs extrêmes généralisées (GEV), la distribution de Pareto généralisée (GPD) et le théorème de Balkema-de Haan et Pickands.

Le troisième chapitre : Ce chapitre est réservé à l'application des statistiques d'ordre aux tests statistiques à savoir le test de normalité des lois des probabilités. En s'intéresse ici aux tests de Kolmogorov-Smirnov, test de Shapiro-Wilk, test de Lilliefors, test de d'Anderson-Darling, test de Cramer-von Mises. Dans ce chapitre, nous mettons l'accent sur le test de Shapiro-Wilk dû à puissance aux cas des petits effectifs ($n \leq 50$). Nous terminons cette partie par deux exemples de ce test.

Le quatrième chapitre : Nous nous intéressons ici à la simulation des statistiques d'ordre et nous présentons des résultats numériques du test de Kolmogrov-Smirnov à l'aide de langage **R**.

Nous terminerons ce mémoire avec une conclusion générale résume notre travail.

Chapitre 1

Théorie de base des statistiques d'ordre

1.1 Statistiques d'ordres

Les statistiques d'ordre sont des échantillons de variables placés par ordre croissant. L'étude des statistiques d'ordre traite des applications de ces valeurs ordonnées et de leurs fonctions.

Si X_1, X_2, ..., X_n sont des variables aléatoires $(v.a)$ indépendentes identiquement distribuées $(i.i.d)$, d'une densité commune f et d'une fonction de répartition F continue, tel que : $F(x) = P(X \leq x)$, pour tout $x \in \mathbb{R}$. Les variables aléatoires

$$X_{1,n} \leqslant X_{2,n} \leqslant ... \leqslant X_{n,n}, \tag{1.1}$$

désignent les statistiques d'ordre de l'échantillon.

Pour $1 \leqslant k \leqslant n$, la $(v.a)$ $X_{k,n}$ est appelée la $k^{ième}$ statistique d'ordre (ou statistique d'ordre du rang k)

La statistique du premier ordre (c'est-à-dire le minimum) et la statistique d'ordre n

(c-à-d le maximum) sont données par :

$$X_{1,n} := \min(X_1,\ X_2, ...,\ X_n) = \min_{1 \leqslant k \leqslant n} X_k,$$
$$X_{n,n} := \max(X_1,\ X_2, ...,\ X_n) = \max_{1 \leqslant k \leqslant n} X_k.$$

Proposition 1.1.1 *La relation entre les statistiques d'ordre extrême, est la suivante :*

$$\min(X_1,\ X_2, ...,\ X_n) = -\max(-X_1,\ -X_2, ...,\ -X_n).$$

La distance entre les statistiques d'ordre extrême (c-à-d le minimum et le maximum), est appelée déviation extrême ou l'étendue de l'échantillon, notée W_n donnée par : $W_n = X_{n,n} - X_{1,n}$.

Définition 1.1.1 *La médiane de l'échantillon divise l'échantillon aléatoire en deux moitiés, une qui contient des échantillons avec des valeurs inférieures, et l'autre qui contient des échantillons avec des valeurs plus élevées. C'est comme la statistique d'ordre intermédiaire/central, elle est mathématiquement définie comme :*

$$\textit{médiane}\ \{X_1,\ X_2, ...,\ X_n\} = \begin{cases} X_{(k),n}, & \textit{si } n \textit{ est impair } \ k = \frac{n+1}{2}, \\ \dfrac{X_{(k),n} + X_{(k+1),n}}{2}, & \textit{si } n \textit{ est pair } \ k = \frac{n}{2}. \end{cases}$$

De plus amples détails pour les propriétés générales des statistiques d'ordre peuvent être trouvés dans les références [1], [2], [9], [11], [7].

1.2 Distributions des statistiques d'ordre

Nous présentons dans cette partie quelques formules importantes pour la fonction de répartition et la densité des statistiques d'ordre.

Proposition 1.2.1 *Soit $X_1,\ X_2, ...,\ X_n$ un échantillon de n $(v.a)$ i.i.d de densité*

commune f et de fonction de répartition F continue, alors la fonction de densité de probabilité de $X_{k,n}$ est donnée par suit

$$f_{X_{k,n}}(x) = f_{k,n}(x) = \frac{n!}{(k-1)!(n-k)!}\left[F(x)\right]^{k-1}\left[1-F(x)\right]^{n-k} f(x), \quad x \in \mathbb{R}. \quad (1.2)$$

1.2.1 Densité Conjointe de n statistiques d'ordre

Lemme 1.2.1 *La densité conjointe de la statistique d'ordre $(X_{1,n},\ X_{2,n}, ...,\ X_{n,n})$ est donnée par :*

$$f_{(X_{1,n},\ X_{2,n},...,\ X_{n,n})}(x_1,\ x_2, ...,\ x_n) = \begin{cases} n!\prod_{k=1}^{n} f(x_k), & si \quad -\infty < x_1 < x_2 < ... < x_n < +\infty, \\ 0, & sinon. \end{cases} \quad (1.3)$$

Preuve. *Voir Arnold et al. (1992) ([1], chapitre 2).*

1.2.2 Distribution de la k-ième statistique d'ordre

Proposition 1.2.2 *Soit $X_1,\ X_2, ...,\ X_n$ un échantillon de n $(v.a)$ i.i.d de densité commune f et de fonction de répartition F continue. Alors, la f.d.c de $X_{k,n}$ peut être obtenue en intégrant la f.d.p de $X_{k,n}$, en (1.2) :*

$$\begin{aligned} F_{_{k,n}}(x) &= F_{X_{k,n}}(x) = P(X_{k,n} \leqslant x) \\ &= P(au\ moins\ k\ de\ X_1,\ X_2, ...,\ X_n,\ sont\ inferieurs\ ou\ égale\ à\ x) \\ &= \sum_{r=k}^{n} P(exactement\ r\ de\ X_1,\ X_2, ...,\ X_n,\ sont\ inferieurs\ ou\ égale\ à\ x) \\ &= \sum_{r=k}^{n} \binom{n}{r}\left[F(x)\right]^{r}\left[1-F(x)\right]^{n-r}, \quad x \in \mathbb{R}, \quad 1 \leq k \leq n. \end{aligned} \quad (1.4)$$

En utilisant la relation :

$$\sum_{r=k}^{n} \binom{n}{r} p^r (1-p)^{n-r} = \int_0^p \frac{n!}{(k-1)!(n-k)!} t^{k-1}(1-t)^{n-k} dt, \quad 0 < p < 1.$$

Ce qui se prouve facilement par intégration répétée par parties, on peut écrire la f.d.c de $X_{k,n}$, de manière équivalente à

$$\begin{aligned} F_{k,n}(x) &= \int_0^{F(x)} \frac{n!}{(k-1)!(n-k)!} t^{k-1}(1-t)^{n-k} dt = \int_0^{F(x)} \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} t^{k-1}(1-t)^{n-k} dt \\ &= \frac{1}{B(k, n-k+1)} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt = I_{F(x)}(k, n-k+1), \quad x \in \mathbb{R}, \quad 1 \leq k \leq n, \end{aligned}$$

où

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, \quad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad \Gamma(k) = (k-1)!, \quad \textit{pour } k = 1,\ 2, ...$$

$B(\cdot,\cdot)$ et $\Gamma(\cdot)$ désignent respectivement la fonction bêta et la fonction gamma complète définies par

$$B(p,q) := \int_0^1 t^{p-1}(1-t)^{q-1} dt, \ \Gamma(p) := \int_0^\infty e^{-t} t^{p-1} dt, \ p, q > 0.$$

1.2.3 Distribution du minimum et de maximum

Proposition 1.2.3 *Soit X_1, $X_2, ...$, X_n un échantillon de n $(v.a)$ i.i.d de densité commune f et de fonction de répartition F continue. Alors, les densités des statistiques d'ordre du minimum et du maximum sont respectivement :*

$$\begin{aligned} f_{1,n}(x) &= f_{X_{1,n}}(x) = n\left[1 - F(x)\right]^{n-1} f(x), \\ f_{n,n}(x) &= f_{Xn,n}(x) = n\left[F(x)\right]^{n-1} f(x). \end{aligned}$$

Les fonctions de distribution du minimum et du maximum (par l'intégration des f.d.p dans les deux formules précédentes)

$$F_{1,n}(x) = F_{X_{1,n}}(x) = 1 - [1 - F(x)]^n , \tag{1.5}$$

$$F_{n,n}(x) = F_{X_{n,n}}(x) = [F(x)]^n . \tag{1.6}$$

Preuve. En utilisant la propriété d'indépendance des variables aléatoires X_1, X_2, ..., X_n, et puisque nos $(v.a)$ sont distribuées de manière identique, nous en déduisons que :

$$F_{1,n}(x) = P(X_{1,n} \leq x) = 1 - P(X_{1,n} > x) = 1 - P\left\{\bigcap_{k=1}^{n} X_k > x\right\} = 1 - [1 - F(x)]^n .$$

$$F_{n,n}(x) = P(X_{n,n} \leq x) = P\left\{\bigcap_{k=1}^{n} X_k \leq x\right\} = \prod_{k=1}^{n} P(X_k \leq x) = [F(x)]^n .$$

■

1.2.4 Distribution conjointe de deux statistiques d'ordre

Lemme 1.2.2 *La fonction de densité de probabilité conjointe peut nous aider à mieux comprendre la relation entre deux $(v.a)$ (statistiques à deux ordres dans notre cas).*

La densité conjointe de $(X_{i,n} \leq X_{j,n})$ avec $1 \leq i < j \leq n$ et $-\infty < x < y < +\infty$ est donnée par l'équation suivante :

$$\begin{aligned} f_{i,j:n}(x,y) = {} & \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} [F(x)]^{i-1} f(x) \\ & \times [F(y) - F(x)]^{j-i-1} f(y) [1 - F(y)]^{n-j} , \end{aligned} \tag{1.7}$$

où $x, y \in \mathbb{R}$. Pour la fonction de distribution conjointe de $(X_{i,n} \leq X_{j,n})$, on a deux cas :

1^{er} *cas* $x < y$:

$$F_{i,j:n}(x,y) = \sum_{s=j}^{n} \sum_{r=i}^{s} \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} [F(x)]^{r-1} f(x) \times [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s}, \quad x, y \in \mathbb{R}. \tag{1.8}$$

Preuve. Nous avons

$$\begin{aligned} F_{i,j:n}(x,y) &= P(X_{i,n} \leq x, X_{j,n} \leq y) \\ &= P(\text{au moins } i \text{ de } X_1,\ X_2, ...,\ X_n, \text{ sont inferieurs ou égale à } x \text{ et} \\ &\text{au moins } j \text{ de } X_1,\ X_2, ...,\ X_n, \text{ sont inferieurs où égale à } y) \\ &= \sum_{s=j}^{n} \sum_{r=i}^{s} P(\text{exactement } r \text{ de } X_1,\ X_2, ...,\ X_n, \text{ sont inferieurs ou égale à } x \text{ et} \\ &\text{exactement } s \text{ de } X_1,\ X_2, ...,\ X_n, \text{ sont inferieurs ou égale à } y) \\ &= \sum_{s=j}^{n} \sum_{r=i}^{s} \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} [F(x)]^{r-1} f(x) \\ &\times [F(y) - F(x)]^{s-r-1} f(y) [1 - F(x)]^{n-s}, \quad x, y \in \mathbb{R}. \end{aligned}$$

$2^{ème}$ *cas* $x \geq y$: il est claire que

$$F_{i,j:n}(x,y) = P(X_{i,n} \leq x, X_{j,n} \leq y) = P(X_{j,n} \leq y) = F_{X_{j,n}}(y). \tag{1.9}$$

Définition 1.2.1 *(Fonction de distribution empirique) Soit* $X_1, X_2, ..., X_n$ *une suite de* $(v.a)$ *i.i.d définies sur le même espace de probabilité, d'une fonction de répartition* F *tell que* $\{F(x) = P(X \leq x),\ x \in \mathbb{R}\}$. *La fonction de répartition empirique notée*

F_n, est définie par :

$$F_n(x) := \frac{1}{n}\sum_{k=1}^{n} I_{\{X_k \leq x\}} := \begin{cases} 0 & si \quad x < X_{1,n}, \\ \frac{k-1}{n} & si \quad X_{k-1,n} \leq x < X_{k,n}, \quad 2 \leq k \leq n, \\ 1 & si \quad x \geq X_{n,n}. \end{cases} \tag{1.10}$$

$$E\left[(F_n(x))\right] = F(x), \quad Var\left[(F_n(x))\right] = \frac{F(x)\left(1-F(x)\right)}{n},$$

où $I_{\{X_k \leq x\}}$ est la fonction indicatrice définie par :

$$I_{\{X_k \leq x\}} := \begin{cases} 1 & si \quad X_k \leq x, \\ 0 & sinon. \end{cases}$$

Définition 1.2.2 *(Fonction quantile et quantile de queue) La fonction quantile de la fonction de distribution F est la fonction inverse généralisée de F définie par :*

$$Q(s) = F^{\leftarrow}(s) = \inf\{x \in \mathbb{R} : F(x) \geq s\}, \quad 0 < s < 1.$$

Dans la théorie des extrêmes, une fonction notée U et appelée fonction quantile de queue, est définie par :

$$U(t) = Q(1 - 1/t) = (1/\bar{F})^{\leftarrow}(t), \quad 1 < t < \infty.$$

où $\{\bar{F}(x) = P(X > x) = 1 - F(x),\, x \in \mathbb{R}\}$ est la fonction de survie.

Définition 1.2.3 *(Fonctions empiriques de quantile et de quantile de queue) La fonction quantile empirique de l'échantillon $(X_1, X_2, ..., X_n)$ est définie par :*

$$Q_n(s) = F_n^{\leftarrow}(s) = \inf\{x \in \mathbb{R} : F_n(x) \geq s\}, \quad 0 < s < 1.$$

La fonction empirique de quantile de queue correspondante est : $U_n(t) = Q_n(1-1/t)$, $1 < t < \infty$. *On peut être exprimée comme une fonction simple des statistiques d'ordre relatives à l'échantillon* $(X_1, X_2, ..., X_n)$ *et nous avons :*

$$Q_n(s) := \begin{cases} X_{k,n}, & si \quad \frac{(k-1)}{n} < s \leq \frac{k}{n}, \\ X_{[np]+1,n}, & si \quad 0 < s \leq 1. \end{cases}$$

Notons que pour $0 < p < 1$, $X_{[np]+1,n}$ *est le quantile empirique de l'ordre* p, *défini par :* $x_p = F^{\leftarrow}(p)$.

Définition 1.2.4 *(Convergence presque sûre des statistiques d'ordre) soit* $0 < p < 1$, *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*

$$1 \leq k(n) \leq n \quad et \quad k(n)/n \longrightarrow p \quad quand \quad n \longrightarrow \infty.$$

Alors la suite des quantiles empiriques $\left(X_{k(n),n}\right)_{n\geq 1}$ *converge presque sûrement vers* x_p.

Lemme 1.2.3 *(Transformation quantile) soit* $X_1,\ X_2, ...,\ X_n$ *des* $(v.a)$ *indépendantes et de fonction de répartition* F. *Soit* $U_1, ..., U_n$ *des* $(v.a)$ *indépendantes de loi uniforme standard, alors :*

(a) Pour toute fonction de distribution F, on a $X_{i,n} \stackrel{d}{=} F^{\leftarrow}(U_{i,n})$, $i = 1, 2, ..., n$.

(b) Lorsque F est continue, on a $F(X_{i,n}) \stackrel{d}{=} U_{i,n}$, $i = 1, 2, ..., n$.

1.3 Représentations des statistiques d'ordre

Nous démontrons quelques relations importantes, qui nous permettent d'exprimer des statistiques d'ordre, en termes de sommes ou produits de $(v.a)$ indépendantes.

Soient $U_{1,n} \leq U_{2,n} \leq ... \leq U_{n,n}$ les statistiques d'ordre associées à un échantillon U_1, $U_2, ..., U_n$ uniformément distribué sur $[0,1]$ et $Z_{1,n} \leq Z_{2,n} \leq ... \leq Z_{n,n}$ les statistiques d'ordre associées à un échantillon Z_1, $Z_2, ..., Z_n$ qui suit la loi exponentielle standard. Nous avons

$$U_{k,n} \stackrel{d}{=} W_k^{1/k} W_{k+1}^{1/k+1} ... W_n^{1/n} \stackrel{d}{=} \frac{\nu_1 + ... + \nu_k}{\nu_1 + ... + \nu_{n+1}},$$

et

$$Z_{k,n} \stackrel{d}{=} \frac{\nu_1}{n} + \frac{\nu_2}{n-1} + ... + \frac{\nu_k}{n-k+1}, \quad k = 1,\ 2, ...,\ n,$$

où W_1, W_2, ...et ν_1, $\nu_2, ...$ sont deux suites de $(v.a)$ $i.i.d$, uniformément distribué sur $[0,1]$ (dans le premier cas) et la distribution exponentielle standard $\mathcal{E}(1)$ (dans le deuxième cas).

Dans le cas général, lorsque F est une $f.d.c$, nous introduisons la fonction inverse par : $G(s) = \inf \{x : F(x) \geq s\}$, $0 < s < 1$, et les statistiques d'ordre exprès $X_{k,n}$ via les statistiques d'ordre uniformes ou exponentiels comme suit :

$$X_{k,n} \stackrel{d}{=} G(U_{k,n}) \stackrel{d}{=} G(1 - \exp(-Z_{k,n})).$$

En raison de ces relations, nous avons les représentations utiles suivantes pour les statistiques d'ordre dans le cas général :

$$\begin{aligned} X_{k,n} &\stackrel{d}{=} G(\frac{\nu_1 + ... + \nu_k}{\nu_1 + ... + \nu_{n+1}}) \\ &\stackrel{d}{=} G(1 - \exp(-(\frac{\nu_1}{n} + \frac{\nu_2}{n-1} + ... + \frac{\nu_k}{n-k+1}))), \end{aligned}$$

où $k = 1,\ 2, ...,\ n$.

Au début, nous montrerons à quel point les résultats pour $U_{k,n}$ et $Z_{k,n}$ peuvent être réécrits pour les statistiques d'ordre à partir d'une distribution arbitraire. Pour toute

$f.d$ F nous déterminons la fonction inverse

$$G(s) = \inf \{x : F(x) \geq s\}, \quad 0 < s < 1. \tag{1.11}$$

Exemple 1.3.1 *Soit $F(x)$ une $f.d.c$ continue, d'une $(v.a)$ X. Montrer que dans ce cas $F(G(x)) = x$, pour $0 < x < 1$, et $Y = F(X)$ a la distribution uniforme sur l'intervalle $[0, 1]$.*

Remarque 1.3.1 *Dans l'exemple précédent nous avons prouvé, en particulier, que la fonction inverse $G(s)$ fournit l'égalité $F(G(s)) = s$, pour $0 < s < 1$, si F est une $f.d.c$ continue. En outre, il est facile de voir que la double égalité*

$$G(F(s)) = s, \quad -\infty < s < \infty,$$

elle est vraie pour tous les s, où $F(s)$ augmente fortement.

Remarque 1.3.2 *Pour toute variable aléatoire avec une $f.d.c$ F continue, nous avons l'égalité :*

$$F(X) \overset{d}{=} U, \tag{1.12}$$

où $Y \overset{d}{=} Z$ indique que les variables aléatoires (ou vecteurs aléatoires) Y et Z ont la même distribution, U en (1.12) étant une variable aléatoire, qui a la distribution uniforme sur $[0, 1]$. En effet, (1.12) échoue si F a des points de saut, depuis, les valeurs de $F(X)$, contrairement à U, ne couvrent pas tous les intervalle $[0, 1]$.

Remarque 1.3.3 *On a la relation $X \overset{d}{=} G(U)$, où G est l'inverse de $f.d.c$ F, s'applique pour toute variable aléatoire, tandis que la double égalité : $F(X) \overset{d}{=} U$, est valide pour les variables aléatoires avec des fonctions de distribution continue.*

Une preuve détaillée est donnée dans Ahsanullah et al. ([7], chapitre 4).

1.4 Des relations générales sur les moments des statistiques d'ordre

Proposition 1.4.1 *Soit $X_{k,n}$ la $k^{ième}$ Statistique d'ordre associée à l'échantillon de taille n de densité f et de fonction de distribution F continue. Alors le $m^{ième}$ $(m = 1, 2, ...)$ moment de $k^{ième}$ $(k = 1, 2, ...)$ statistique d'ordre est :*

$$\begin{aligned}\mu_{k,n}^{(m)} &= E(X_{k,n}^m) = \int_{-\infty}^{+\infty} x^m f_{X_{k,n}}(x)dx < \infty \\ &= \frac{n}{(k-1)!\,(n-k)!} \int_{-\infty}^{+\infty} x^m \left[F(x)\right]^{k-1} \left[1-F(x)\right]^{n-k} f(x)dx. \end{aligned} \tag{1.13}$$

L'égalité en (1.13) peut être exprimée comme :

$$\mu_{k,n}^{(m)} = \frac{n}{(k-1)!\,(n-k)!} \int_0^1 \left(G(u)\right)^m (u)^{k-1} (1-u)^{n-k}\, du,$$

où $G(u) = F_X^{\leftarrow}(u)$. Nous indiquerons $\mu_{k,n}^{(1)}$ par $\mu_{k,n}$ pour plus de commodité,à partir les deux premiers moments, on peut déterminer la variance de $X_{k,n}$ par :

$$\sigma_{k,k,n} = \sigma_{k,n}^{(2)} = Var(X_{k,n}) = \mu_{k,n}^{(2)} - \mu_{k,n}^2, \quad 1 \le k \le n. \tag{1.14}$$

1.4.1 Existence des moments des statistique d'ordre

Théorème 1.4.1 *Soit X_1 , $X_2, ..., X_n$ un échantillon de n $(v.a)$ et d'une distribution F continue, et $X_{1,n}$, $X_{2,n}, ..., X_{n,n}$ les statistiques d'ordre associées. Soit m un entier strictement positif. Si X possède un moment d'ordre m, alors, pour tout k appartenant à $\{1,\ 2, ...,\ n\}$, la $k^{ième}$ statistique d'ordre $X_{k,n}$ possède un moment d'ordre m.*

Preuve. *On a la distribution de $X_{k,n}$ est donnée par :*

$$F_{X_{k,n}}(t) = \frac{n!}{(k-1)!\,(n-k)!} \int_{-\infty}^{t} [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)dx.$$

Nous avons alors

$$E(X_{k,n}^m) = \frac{n!}{(k-1)!\,(n-k)!} \int_{-\infty}^{+\infty} x^m [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)dx,$$

et donc

$$\begin{aligned} E(\left|X_{k,n}^m\right|) &\leq \frac{n!}{(k-1)!\,(n-k)!} \int_{-\infty}^{+\infty} |x|^m [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)dx \\ &\leq \frac{n!}{(k-1)!\,(n-k)!} \int_{-\infty}^{+\infty} |x|^m f(x)dx \\ &= \frac{n!}{(k-1)!\,(n-k)!} E(|X|^m) < \infty, \end{aligned}$$

ce qui démontre le théorème. ■

Remarque 1.4.1 *Le théorème ci-dessus montre que l'existence du moment d'ordre m de X entraîne l'existence du moment d'ordre m de $X_{k,n}$.*

Remarque 1.4.2

$$\{E(X^m) \text{ existe } \Longrightarrow \forall k \in \{1,\ 2, ...,\ n\},\ E(X_{k,n}^m) \text{ existe}\}, \tag{1.15}$$

la réciproque n'est pas vraie

$$\forall k \in \{1,\ 2, ...,\ n\},\ E(X_{k,n}^m) \text{ existe} \not\Rightarrow E(X^m) \text{ existe}. \tag{1.16}$$

Théorème 1.4.2 *Soit $X_1,\ X_2, ...,\ X_n$ un échantillon de n $(v.a)$ et d'une distribution F continue, et $X_{1,n},\ X_{2,n}, ...,\ X_{n,n}$ les statistiques d'ordre associées. Si $E(X)$*

existe, alors, pour tous couples (r,s) *tels que* $1 \leq r < s \leq n$, *et* m_r, $m_s = 1, 2, ...,$ $E(X_{r,n}X_{s,n})$ *existe.*

Preuve. Nous avons vu, dans la $2^{ème}$ section, que la loi du couple $(X_{r,n}X_{s,n})$ est donnée par :

$$F_{X_{r,n},X_{s,n}}(u,v) = \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!}\int\int_{\{x<y,x\leq u,y\leq v\}}[F(x)]^{r-1}f(x) \times [F(y)-F(x)]^{s-r-1}f(y)\,[1-F(x)]^{n-s}\,dxdy,$$

et,donc

$$\begin{aligned}\mu_{r,s:n}^{(m_r,m_s)} &= E(X_{r,n}^{m_r}X_{s,n}^{m_s}) = \int\int_{-\infty<x<y<+\infty} x^{m_r}y^{m_s}f_{r,s:n}(x,y)dxdy \\ &= \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!}\int\int_{-\infty<x<y<+\infty}[F(x)]^{r-1}f(x) \\ &\times [F(y)-F(x)]^{s-r-1}f(y)\,[1-F(y)]^{n-s}\,dxdy. \qquad (1.17)\end{aligned}$$

Une autrefois, et pour plus de commodité,nous utiliserons $\mu_{r,s:n}$ au lieu de $\mu_{r,s:n}^{(1,1)}$, la covariance de $_{X_{r,n}}$ et $_{X_{s,n}}$ peut alors être déterminée par :

$$\sigma_{r,s:n} = Cov(X_{r,n},X_{s,n}) = \mu_{r,s:n} - \mu_{r:n}\mu_{s:n}, \quad 1 \leq r < s \leq n. \qquad (1.18)$$

■

1.5 Moments des statisques d'ordre de l'uniforme et l'exponentielles

Dans la $3^{ème}$ section nous avons prouvé quelques représentations pour des statistiques d'ordre uniformes et exponentielles, qui nous permettent d'exprimer ces statis-

tiques d'ordre via des sommes ou des produits de variables aléatoires indépendantes. En vertu des expressions correspondantes, on peut facilement trouver les moments simples et conjoints de statistiques d'ordre exponentielles et uniformes.

1.5.1 Statistiques d'ordre uniformes

En effet, dans le cas de la distribution uniforme standard, on peut utiliser l'expression (1.13) pour trouver les moments des statistiques d'ordre de $X_{k,n}$. En fait, dans ce cas pour tout $\alpha > -k$ nous obtenons le résultat suivant :

$$\begin{aligned} E(U_{k,n})^{\alpha} &= \frac{n!}{(k-1)!\,(n-k)!}\int_0^1 x^{\alpha}x^{k-1}(1-x)^{n-k}dx \\ &= \frac{n!}{(k-1)!\,(n-k)!}B(\alpha+k, n-k+1) \\ &= \frac{n!\Gamma(\alpha+k)\Gamma(n-k+1)}{(k-1)!\,(n-k)!\Gamma(n-\alpha+1)} \qquad (1.19) \\ &= \frac{n!\Gamma(\alpha+k)}{(k-1)!\Gamma(n-\alpha+1)}, \qquad (1.20) \end{aligned}$$

où $B(a,b)$ et $\Gamma(s)$ désignent respectivement la fonction bêta et la fonction gamma. Si α est un entier, alors la relation on (1.20) est simplifié, comme suit :

$$\begin{aligned} E(U_{k,n}) &= \frac{n!\Gamma(k+1)}{(k-1)!\Gamma(n+2)} = \frac{n!k!}{(k-1)!(n+1)!} \\ &= \frac{k}{n+1}, \quad 1 \le k \le n, \qquad (1.21) \end{aligned}$$

de même manière

$$E(U_{k,n})^2 = \frac{k(k+1)}{(n+1)\,(n+2)}, \quad 1 \le k \le n, \qquad (1.22)$$

en général, pour $r = 1, 2, ...$, nous avons :

$$E(U_{k,n})^r = \frac{k(k+1)...(k+r-1)}{(n+1)(n+2)...(n+r)}, \quad 1 \leq k \leq n,$$

il résulte de (1.21) et (1.22) que :

$$\sigma_{k,k,n} = \sigma^{(2)}_{k,n} = Var(U_{k,n}) = \mu^{(2)}_{k,n} - \mu^2_{k,n} = \frac{k(n-k+1)}{(n+1)^2(n+2)}, \quad 1 \leq k \leq n. \tag{1.23}$$

Considérons les deux statistiques d'ordre uniformes $U_{r,n}$ et $U_{s,n}$, alors :

$$E(U_{r,n}U_{s,n}) = \frac{r(s+1)}{(n+1)(n+2)}, \quad 1 \leq r < s \leq n. \tag{1.24}$$

À partir de (1.21), (1.23) et (1.24), nous obtenons l'expression suivante pour la covariance entre les statistiques d'ordre de loi uniforme :

$$\begin{aligned} Cov(U_{r,n}, U_{s,n}) &= E(U_{r,n}U_{s,n}) - E(U_{r,n})E(U_{s,n}) \\ &= \frac{r(n-s+1)}{(n+1)^2(n+2)}, \quad r \leq s, \end{aligned} \tag{1.25}$$

Remarquons que l'expression $Cov(U_{r,n}, U_{s,n})$ est celle de $Var(U_{k,n})$ si $s = r$.

1.5.2 Statistiques d'ordre exponentielles

Soient $Z_{1,n}$, $Z_{2,n}, ...$, $Z_{n,n}$, $n = 1, 2, ...$, les statistiques d'ordre correspondant à la distribution exponentielle standard avec $f.d.c : \{F(x) = 1 - \exp(-x),\ x > 0\}$, pour

obtenir la formule $E(Z_{k,n})^{\alpha}$, $k = 1, 2, ..., n$, il faut calculer les intégrales

$$\begin{aligned}
&\frac{n!}{(k-1)!\,(n-k)!}\int_0^{\infty} x^{\alpha}\left(F\left(x\right)\right)^{k-1}\left(1-F\left(x\right)\right)^{n-k}f(x)dx \\
&= \frac{n!}{(k-1)!\,(n-k)!}\int_0^{\infty} x^{\alpha}\left(1-\exp(-x)\right)^{k-1}\exp(-(n-k+1))dx \\
&= \frac{n!}{(k-1)!\,(n-k)!}\sum\nolimits_{r=0}^{k-1}(-1)^{r}\binom{k-1}{r}\int_0^{\infty} x^{\alpha}\exp(-x(n-k+r+1))dx,
\end{aligned}$$

depuit

$$\begin{aligned}
\int_0^{\infty} x^{\alpha}\exp(-x(n-k+r+1))dx &= (n-k+r+1)^{-(\alpha+1)}\int_0^{\infty} u^{\alpha}\exp(-u)du \\
&= \frac{\Gamma(\alpha+1)}{(n-k+r+1)^{(\alpha+1)}},
\end{aligned}$$

on obtient que

$$E(Z_{k,n})^{\alpha} = \frac{n!}{(k-1)!\,(n-k)!}\sum\nolimits_{r=0}^{k-1}(-1)^{r}\binom{k-1}{r}\frac{\Gamma(\alpha+1)}{(n-k+r+1)^{(\alpha+1)}}. \qquad (1.26)$$

Par exemple, si $k = 1$, alors

$$E(Z_{1,n})^{\alpha} = n\frac{\Gamma(\alpha+1)}{(n)^{(\alpha+1)}} = \frac{\Gamma(\alpha+1)}{n^{\alpha}}, \quad \alpha > -1,$$

pour $k = 2$ et $\alpha > -1$ on a : $E(Z_{2,n})^{\alpha} = n\,(n-1)\,\Gamma(\alpha+1)\left\{(n-1)^{-(\alpha+1)} - n^{-(\alpha+1)}\right\}$.

Exemple 1.5.1 *Rappelons que les statistiques d'ordre exponentiel sont exprimées en termes de somme de* $(v.a)$ *indépendantes :*

$$(Z_{1,n},\ Z_{2,n}, ...,\ Z_{n,n}) \overset{d}{=} \left(\frac{\nu_1}{n}, \frac{\nu_1}{n} + \frac{\nu_2}{n-1}, ...\frac{\nu_1}{n} + \frac{\nu_2}{n-1} + ... + \frac{\nu_{n-1}}{2} + \nu_n\right),$$

où $\nu_1,\ \nu_2, ...,$ *sont des* $(v.a)$ *exponentielles indépendantes* $\mathcal{E}(1)$, *immédiatement on*

obtient que :

$$E(Z_{k,n}) = E\left(\frac{\nu_1}{n} + \frac{\nu_2}{n-1} + ... + \frac{\nu_k}{n-k+1}\right) = \sum_{r=1}^{k} \frac{1}{n-r+1}, \tag{1.27}$$

$$Var(Z_{k,n}) = \sum_{r=1}^{k} Var\left(\frac{\nu_r}{n-r+1}\right) = \sum_{r=1}^{k} \frac{1}{(n-r+1)^2}, \tag{1.28}$$

Pour autant que $E(\nu) = Var(\nu) = 1$ *si* ν *a la distribution exponentielle standard. Il découle de (1.27) et (1.28) que*

$$E(Z_{k,n})^2 = \sum_{r=1}^{k} \frac{1}{(n-r+1)^2} + \left(\sum_{r=1}^{k} \frac{1}{n-r+1}\right)^2, \tag{1.29}$$

en comparant (1.27) et (1.28) avec (1.26) $(sous\ \alpha = 1\ et\ \alpha = 2)$*, on obtient le résultat suivant :*

$$\frac{n!}{(k-1)!(n-k)!} \sum_{r=0}^{k-1} (-1)^r \frac{\binom{k-1}{r}}{(n-k+r+1)^2} = \sum_{r=1}^{k} \frac{1}{n-r+1},$$
$$\frac{2(n!)}{(k-1)!(n-k)!} \sum_{r=0}^{k-1} (-1)^r \frac{\binom{k-1}{r}}{(n-k+r+1)^3} = \sum_{r=1}^{k} \frac{1}{(n-r+1)^2} + \left(\sum_{r=1}^{k} \frac{1}{n-r+1}\right)^2.$$

Remarque 1.5.1 *Il est intéressant de voir que* $E(Z_{1,n}) = \frac{1}{n}$*, et* $Var(Z_{1,n}) = \frac{1}{n^2}$ *tendent vers le zéro, quand* $n \longrightarrow \infty$*, alors que*

$$E(Z_{n,n}) = \sum_{r=1}^{k} \frac{1}{n-r+1} = \sum_{r=1}^{k} \frac{1}{r} \sim \log n \longrightarrow \infty, \quad n \longrightarrow \infty.$$
$$Var(Z_{n,n}) = \sum_{r=1}^{k} \frac{1}{r^2} \longrightarrow \frac{\pi^2}{6}, \quad n \longrightarrow \infty.$$

Il est possible de trouver des relations utiles pour les moments de l'exponentielle statistique d'ordre par exemple :

$$E(Z_{k,n}^r) = E(Z_{k-1,n}^r) + \frac{r}{n-k+1} E(Z_{k,n}^{r-1}), \quad r \geq 1, \quad 2 \leq k \leq n.$$

Pour plus de détails, voir Ahsanullah et al. ([7], chapitre 8).

1.6 Comportement asymptotique des statistiques d'ordre

Nous commençons à étudier différents théorèmes limites pour les statistiques d'ordre. Dans cette section, nous considérons les distributions asymptotiques pour les statistiques dites d'ordre modérées et intermédiaires.

Il s'avère que celui qui veut étudier des distributions asymptotiques des statistiques d'ordre $X_{k,n}$ convenablement normalisées et centrées doit distinguer trois options différentes. Puisque nous considérons le cas où n (la taille de l'échantillon) tend vers l'infini, il est naturel que $k = k(n)$ soit une fonction de n. Les statistiques d'ordre $X_{k(n),n}$, sont dites extrêmes si $k = k(n)$ ou $n - k(n) + 1$ est fixe, quand $n \to \infty$, si

$$0 < \liminf_{n\to\infty} \frac{k(n)}{n} \leq \limsup_{n\to\infty} \frac{k(n)}{n} < 1,$$

alors les statistiques d'ordre $X_{k(n),n}$ sont dites modérées. Enfin, le cas où $k(n) \to \infty$, $\frac{k(n)}{n} \to 0$ ou $n - k(n) \to \infty$, $\frac{k(n)}{n} \to 1$, correspond aux statistiques dites d'ordre intermédiaire.

La possibilité d'exprimer les statistiques d'ordre uniformes et exponentielles via des sommes de termes indépendants nous permet d'étuliser des théorèmes limites pour des sommes de $(v.a)$ indépendantes. En fait, nous devons connaître le théorème de *Lyapunov* suivant.

Soient X_1, $X_2, ...,$ X_n des $(v.a)$ indépendantes d'espérances $\alpha_k = E(X_k)$ variances σ_k^2 et des moments finis $\gamma_k = E\left|X_k - \alpha_k\right|^3$, $k = 1, 2, ...$ indiquent :

$$S_n = \sum_{k=1}^{n} \left(X_k - \alpha_k\right), \quad B_n^2 = var\left(S_n\right) = \sum_{k=1}^{n} \sigma_k^2.$$

et

$$L_n = \sum\nolimits_{k=1}^{n} \frac{\gamma_k}{B_n^3} \tag{1.30}$$

Si le rapport de *lyapunov* (1.30) converge vers zéro, quand $n \to \infty$, alors :

$$\sup \left| P\left\{ S_n / B_n < x \right\} - \Phi(x) \right| \to 0, \tag{1.31}$$

où $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{t^2}{2}) dt$ est le $d.f$ de la distribution normale standard.

Exemple 1.6.1 *Considérons une suite de statistiques d'ordre exponentielles* $Z_{k(n),n}$, $n = 1,\ 2, ...,$ *où* $k(n) \to \infty$ *et lim* sup $k(n)/n < 1$, *quand* $n \to \infty$. *Comme nous le savons depuis la* $3^{\text{ème}}$ *section, la relation suivante est valable pour les statistiques d'ordre exponentielles :*

$$Z_{k,n} \overset{d}{=} \frac{\nu_1}{n} + \frac{\nu_2}{n-1} + ... + \frac{\nu_k}{n-k+1}, \quad k = 1,\ 2, ...,\ n,$$

où $\nu_1,\ \nu_2, ...$*est une suite de* $(v.a)$ *i.i.d, ayant la distribution exponentielle standard* $\mathcal{E}(1)$. *Vérifions si le théorème de lyapunov peut être utilisé pour des termes indépendants* $X_k = \frac{\nu_k}{n-k+1}$, *on obtient facilement que :*

$$\alpha_k = E\left(X_k\right) = \frac{1}{n-k+1}, \quad \sigma_k^2 = Var\left(X_k\right) = \frac{1}{(n-k+1)^2},$$

et $\gamma_k = \frac{1}{(n-k+1)^3}$, *où :*

$$\gamma = E\left|\nu - 1\right|^3 = \int_0^1 (1-x)^3 e^{(-x)} dx + \int_1^\infty (x-1)^3 e^{(-x)} dx = \frac{12}{e} - 2.$$

Dans notre cas

$$B_{k,n}^2 = Var\left(Z_{k,n}\right) = \sum_{r=1}^{k} \sigma_r^2 = \sum_{r=1}^{k} \frac{1}{(n-k+1)^2} = \sum_{r=n-k+1}^{n} \frac{1}{r^2},$$
$$\Gamma_{k,n} = \sum_{r=1}^{k} E\left|X_r - \alpha_r\right|^3 = \gamma \sum_{r=1}^{k} \frac{1}{(n-r+1)^3} = \gamma \sum_{r=n-k+1}^{n} \frac{1}{r^3}.$$

La restriction lim sup $k(n)/n < 1$, *quand* $n \to \infty$, *signifie qu'il existe un tel* p, $0 < p < 1$, *que* $k \leq pn$, *pour tout* n *suffisamment grand, pour de telles valeurs* n *on a évidemment les inégalités suivantes :*

$$B_{k,n}^2 \geq \tfrac{k}{n^2}, \quad \Gamma_{k,n} \leq \tfrac{k\gamma}{(n(1-p))^3}, \quad L_{k,n} = \tfrac{\Gamma_{k,n}}{B_{k,n}^3} \leq \tfrac{\gamma}{(1-p)^3 k^{1/2}},$$

donc, si lim sup $k(n)/n < 1$ *et* $k(n) \to \infty$, *quand* $n \to \infty$, *alors le rapport de lyapunov* $L_{k(n),n}$, *tend vers zéro et ceci fournit la normalité asymptotique de* $\left(Z_{k(n),n} - A_{k,n}\right)/B_{k,n}$ *où* $\alpha_{k,n} = \sum_{r=1}^{k} \frac{1}{n-r+1}$.

Remarque 1.6.1 *Il est facile de voir que l'exemple précédent couvre le cas des quantiles empirique* $Z_{[pn]+1,n}$, $0 < p < 1$, *de plus si* $k(n) \sim pn$, $0 < p < 1$, *quand* $n \to \infty$, *alors :*

$$\alpha_{k,n} = \sum_{r=1}^{k} \frac{1}{n-r+1} = \sum_{r=n-k+1}^{n} \frac{1}{r} \sim -\log(1-p), \quad n \to \infty,$$
$$B_{k,n}^2 = \sum_{r=n-k+1}^{n} \frac{1}{r^2} \sim \int_{n-k+1}^{n} x^{-2} dx \sim \frac{p}{(1-p)^n}, \quad quand \;\; n \to \infty.$$

Tous ces propos avec quelques arguments supplémentaires nous permettent de montrer que pour tout $0 < p < 1$, *quand* $n \to \infty$,

$$\sup_x \left| P\left\{ \left(Z_{[pn]+1,n} + \log(1-p)\right) \left(n(1-p)/p\right)^{1/2} \right\} < x - \Phi(x) \right| \to 0.$$

En particulier, la relation suivante est valable pour la médiane exponentielle de

l'échantillon

$$\sup_x \left| P\left\{ n^{1/2} \left(Z_{[n/2]+1,n} - \log 2 \right) \right\} < x - \Phi(x) \right| \to 0, \quad \textit{quand} \quad n \to \infty.$$

Remarque 1.6.2 *Des arguments plus compliqués et détaillés permettent de prouver que les statistiques d'ordre exponentielles convenablement centrées et normalisées* $Z_{k(n),n}$ *ont la distribution asymptotiquement normale si* $\min\{k(n), n - k(n) + 1\} \to \infty$, *quand* $n \to \infty$.

Chapitre 2

Estimation

2.1 Estimation optimale

2.1.1 Propriétés des estimateurs basés sur les statistiques d'ordre

Nous étudions certaines propriétés des estimateurs basés sur les statistiques d'ordre. Nous donnons des exemples des statistiques suffisantes. Nous considérons également des estimateurs sans biais de certains paramètres inconnus ainsi que des estimateurs avec des erreurs quadratiques moyennes minimales.

Les statistiques d'ordre jouent un rôle très important dans différents problèmes statistiques. Très souvent ils sont utilisés comme les meilleurs (dans un certain sens) ou simplement comme estimateurs pratiques des paramètres inconnus . Ci-dessous, nous étudierons différentes propriétés de divers estimateurs basés sur les statistiques d'ordre.

Nous commençons de la définition de statistiques suffisantes.

Définition 2.1.1 *Supposons que* X_1, $X_2, ...,$ X_n *sont* n *observations d'une distri-*

bution avec f.d.c $F_\theta(x)$, ici θ est un paramètre inconnu dont les valeurs possibles forment un ensemble Θ. soit :

$$T = T(X_1,\ X_2, ...,\ X_n),$$

une fonction connue de X_1, $X_2, ..., X_n$. Alors, T est appelée une statistique suffisante pour θ si la distribution conditionnelle de X_1, $X_2, ...,$ X_n donnée $T = t$ ne dépend pas de $\theta \in \Theta$.

Remarque 2.1.1 *La statistique T et le paramètre θ en définition 2.1.1 peuvent être des vecteurs.*

Exemple 2.1.1 *Soient X_1, $X_2, ..., X_n$ n observations d'une distribution absolument continue avec f.d.c $F_\theta(x)$, $\theta \in \Theta$. Considérer le vecteur des statistiques d'ordre $T = (X_{1,n},\ X_{2,n}, ...,\ X_{n,n})$. Alors T est suffisant pour $\theta \in \Theta$. En fait, que $f_\theta(x)$ soit le f.d.p de $f_\theta(x)$. Il est facile de vérifier que :*

$$P\{X_1 = \beta_1,\ X_2 = \beta_2, ...,\ X_n = \beta_n \mid X_{1,n} = x_1,\ X_{2,n} = x_2, ...,\ X_{n,n} = x_n\} = \frac{1}{n!}. \tag{2.1}$$

Pour tous les $n!$ permutations $(\beta_1, ..., \beta_n)$ de valeurs choisies arbitrairement x_1, $x_2, ...,$ x_n, de sorte que $x_1 \leq x_2 \leq ... \leq x_n$ et $f_\theta(x_k) > 0$, pour tout $k = 1, 2, ..., n$ et $\theta \in \Theta$. On voit que la relation de (2.1) ne dépend pas de θ. Ainsi, T est une statistique suffisante pour $\theta \in \Theta$.

Remarque 2.1.2 *Il existe le critère de factorisation pour vérifier si certaines statistiques $T = T(X_1,\ X_2, ...,\ X_n)$ est suffisant pour un paramètre $\theta \in \Theta$. Nous le formulons pour des distributions absolument continues sous la forme suivante.*

Théorème 2.1.1 *Soient X_1, $X_2, ..., X_n$ n observations indépendantes d'une distribution avec une fonction de densité $f_\theta(x)$, $\theta \in \Theta$. Alors $T = T(X_1,\ X_2, ...,\ X_n)$, est*

une statistique suffisante pour θ, si la f.d.p conjointe de X_1, $X_2, ..., X_n$ est de la forme

$$f_\theta(x_1,\ x_2, ...,\ x_n) = f_\theta(x_1)f_\theta(x_2)...f_\theta(x_n),$$

peut s'exprimer comme suit :

$$f_\theta(x_1,\ x_2, ...,\ x_n) = h(x_1,\ x_2, ...,\ x_n)g\left(T(x_1,\ x_2, ...,\ x_n), \theta\right), \tag{2.2}$$

où h est une fonction non négative de x_1, $x_2, ..., x_n$ seulement et ne dépend pas de θ, et g est une fonction non négative de θ et $T(x_1, x_2, ..., x_n)$ uniquement.

Exemple 2.1.2 *Soient X_1, $X_2, ..., X_n$, n observations d'une distribution $U[0, \theta]$ uniforme de f.d.p $f_\theta(x)$, ce qui équivaut à $1/\theta$, si $0 < x < \theta$, où $\theta > 0$ est un paramètre inconnu, et il est null sinon . Nous allons montrer que*

$$T(X_1,\ X_2, ...,\ X_n) = \max\{X_1,\ X_2, ...,\ X_n\} = X_{n,n}.$$

Est une statistique suffisante pour le paramètre θ, on peut écrire la fonction de densité de probabilité jointe des observations X_1, $X_2, ..., X_n$, comme suit :

$$f_\theta(x_1,\ x_2, ...,\ x_n) = \frac{1}{\theta^n}\prod_{i=1}^{n} R(x_i)C(x_i, \theta), \tag{2.3}$$

où

$$C(x,\theta) = \begin{cases} 1, & si \quad x < \theta, \\ 0, & sinon. \end{cases} \quad et \quad R(x) = \begin{cases} 0, & si \quad x \leq 0, \\ 1, & si \quad x > 0. \end{cases}$$

On remarque que $\prod_{i=1}^{n} C(x_i, \theta)$, coïncide avec la fonction $C\{\max(x_1, x_2, ..., x_n), \theta\}$, qui peut s'écrire aussi $C\{T(x_1, x_2, ..., x_n), \theta\}$, ainsi dans notre cas, la relation (2.2) est

vraie avec les fonctions

$$h(x_1,\ x_2, ...,\ x_n) = \prod_{i=1}^{n} R(x_i),$$

qui ne dépend pas de θ, *et* $g(x, \theta) = \frac{C(x,\theta)}{\theta^n}$, *donc*

$$T(X_1,\ X_2, ...,\ X_n) = \max\{X_1,\ X_2, ...,\ X_n\} = X_{n,n}, \tag{2.4}$$

est une statistique suffisante pour le paramètre θ.

Définition 2.1.2 *Si une statistique* $T = T(X_1,\ X_2, ...,\ X_n)$, *est telle que* $E(T) = \theta$, *pour tout* $\theta \in \Theta$, *alors* T *est appelé un estimateur sans biais de* θ.

2.1.2 Estimateurs linéaires sans biais de variance minimale

Nous donnons des définitions et des exemples des meilleurs estimateurs non biaisés (au sens de la variance minimale) des paramètres de position et d'échelle, basés sur des combinaisons linéaires des statistiques d'ordre.

Dans la section précédente, nous avons introduit la définition d'un estimateur sans biais de variance minimale. Nous allons maintenant présenter les estimateurs linéaires sans biais de variance minimale $(MVLUE)\,s$ (en anglais Minimum Variance Linear Unbiased Estimators) qui sont exprimés sous forme de combinaisons linéaires des statistiques d'ordre. Commençons par les $(MVLUE)\,s$ des paramètres de position et d'échelle.

Supposons que X a une fonction de distribution absolument continue de la forme :

$$F(\frac{x-\mu}{\sigma}), \quad -\infty < \mu < +\infty, \ \ \sigma > 0. \tag{2.5}$$

Supposons en outre que :

$$E(X_{r,n}) = \mu + \alpha_r \sigma, \quad r = 1, 2, ..., n.$$
$$Var(X_{r,n}) = V_{rr}\sigma^2, \quad r = 1, 2, ..., n.$$
$$Cov(X_{r,n}, X_{s,n}) = Cov(X_{s,n}, X_{r,n}) = V_{rs}\sigma^2, \quad 1 \leq r < s \leq n.$$

Soit $\mathbf{X}' = (X_{1,n},\, X_{2,n}, ...,\, X_{n,n})$, on peut écrire

$$E(\mathbf{X}) = \mu \mathbf{1} + \sigma\alpha, \tag{2.6}$$

où $\mathbf{1} = (1,\, 1, ...,\, 1)'$, $\alpha = (\alpha_1,\, \alpha_2, ...,\, \alpha_n)'$, et $Var(\mathbf{X}) = \sigma^2 \sum$ avec $\sum$ est une matrice avec des éléments V_{rs}, $1 \leq r,\, s \leq n$. Ensuite, les $(MVLUE)\, s$ des paramètres de position et d'échelle μ et σ sont :

$$\hat{\mu} = \frac{1}{\Delta}\left\{\alpha' \sum\nolimits^{-1} \alpha 1' \sum\nolimits^{-1} - \alpha' \sum\nolimits^{-1} 1 \alpha' \sum\nolimits^{-1}\right\} X, \tag{2.7}$$
$$\hat{\sigma} = \frac{1}{\Delta}\left\{1' \sum\nolimits^{-1} 1 \alpha' \sum\nolimits^{-1} - 1' \sum\nolimits^{-1} \alpha 1' \sum\nolimits^{-1}\right\} X, \tag{2.8}$$

où

$$\Delta = (\alpha' \sum\nolimits^{-1} \alpha)(1' \sum\nolimits^{-1} 1) - (\alpha' \sum\nolimits^{-1} 1)^2. \tag{2.9}$$

La variance et la covariance de ces estimateurs sont données comme suit :

$$Var(\hat{\mu}) = \frac{\sigma^2(\alpha' \sum\nolimits^{-1} \alpha)}{\Delta}, \tag{2.10}$$
$$Var(\hat{\sigma}) = \frac{\sigma^2(1' \sum\nolimits^{-1} 1)}{\Delta}, \tag{2.11}$$
$$Cov(\hat{\mu}, \hat{\sigma}) = -\frac{\sigma^2(\alpha' \sum\nolimits^{-1} 1)}{\Delta}. \tag{2.12}$$

Notons que pour toute distribution symétrique :

$$\alpha_j = -\alpha_{n-j+1}, 1'\sum^{-1}\alpha = \alpha'\sum^{-1}1 = 0, \quad \text{et} \quad \Delta = (\alpha'\sum^{-1}\alpha)(1'\sum^{-1}1).$$

Par conséquent, les meilleures estimations linéaires sans biais de μ et σ pour le cas symétrique sont :

$$\hat{\mu}^* = \frac{1'\sum^{-1}X}{1'\sum^{-1}1}, \tag{2.13}$$

$$\hat{\sigma}^* = \frac{\alpha'\sum^{-1}X}{\alpha'\sum^{-1}\alpha}. \tag{2.14}$$

Et la covariance correspondante des estimateurs est nulle et leurs variances sont données comme :

$$Var(\hat{\mu}^*) = \frac{\sigma^2}{1'\sum^{-1}1}, \tag{2.15}$$

$$Var(\hat{\sigma}^*) = \frac{\sigma^2}{\alpha'\sum^{-1}\alpha}. \tag{2.16}$$

Nous pouvons utiliser les formules ci-dessus pour obtenir les $(MVLUE)\,s$ des paramètres de position et d'échelle pour toute distribution numériquement à condition que les variances des statistiques d'ordre existent. Pour certaines distributions des $(MVLUE)\,s$ des paramètres de position et d'échelle peuvent être exprimées sous forme simplifiée.

Le lemma suivant (voir Garybill (1983), p. 198) sera utile pour trouver l'inverse de la matrice de covariance.

Lemme 2.1.1 *Soit* $\sum = (\sigma_{r,s})$ *une matrice* $n \times n$ *avec des éléments, qui satisfont la relation*

$$\sigma_{r,s} = \sigma_{s,r} = c_r d_s, \quad 1 \leq r,\ s \leq n,$$

pour certains positifs $c_1, ..., c_n$ *et* $d_1, ..., d_n$, *alors son inverse* $\sum^{-1} = (\sigma^{r,s})$, *a des éléments donnés comme suit :*

$$\begin{aligned}
\sigma^{1,1} &= \frac{c_2}{c_1\left(c_2 d_1 - c_1 d_2\right)}, \\
\sigma^{n,n} &= \frac{d_{n-1}}{d_n\left(c_n d_{n-1} - c_{n-1} d_n\right)}, \\
\sigma^{k+1,k} &= \sigma^{k,k+1} = -\frac{1}{c_{k+1} d_k - c_k d_{k+1}}, \\
\sigma^{k,k} &= \frac{c_{k+1} d_{k-1} - c_{k-1} d_{k+1}}{\left(c_k d_{k-1} - c_{k-1} d_k\right)\left(c_{k+1} d_k - c_k d_{k+1}\right)}, \quad k = 2, ..., \ n-1,
\end{aligned}$$

avec $\sigma^{i,j} = 0, \ \ si \ |i-j| > 1$.

2.1.3 Estimateurs linéaires sans biais de variance minimale et prédicteurs basés sur des échantillons censurés

Nous considérons le cas où certaines observations les plus petites et les plus grandes sont manquantes. Dans cette situation, nous construisons les estimateurs linéaires sans biais de la variance minimale pour les paramètres d'échelle. Nous discutons également du problème de la recherche du meilleur prédicteur linéaire sans biais (au sens de la variance minimale) de la statistique d'ordre $X_{s,n}$, basé sur des observations données $X_{1,n}$, $X_{2,n}, ..., X_{n,n}$, où $1 \leq r \leq s \leq n$.

Dans le chapitre précédent, nous avons examiné les $(MVLUE)$ s dans le cas où toutes les n observations sont disponibles pour un statisticien. Supposons maintenant que le plus petit r_1 et le plus grand r_2 de ces observations sont perdus et que nous pouvons traiter les statistiques d'ordre $X_{r_1+1,n} \leq ... \leq X_{n-r_2,n}$.

Nous examinerons ici les estimateurs linéaires sans biais de variance minimale des paramètres de position et d'échelle basée sur les éléments donnés de la série variationnelle.

Supposons que X a une fonction de distribution absolument continue de la forme :

$$F(\frac{x-\mu}{\sigma}), \quad -\infty < \mu < +\infty, \quad \sigma > 0.$$

Supposons en outre que :

$$\begin{aligned} E(X_{r,n}) &= \mu + \alpha_r \sigma, \\ Var(X_{r,n}) &= V_{rr}\sigma^2, \quad r_1 + 1 \leq r \leq n - r_2, \\ Cov(X_{r,n}, X_{s,n}) &= Cov(X_{s,n}, X_{r,n}) = V_{rs}\sigma^2, \quad r_1 + 1 \leq r, \ s \leq n - r_2. \end{aligned}$$

Soit

$$\mathbf{X}' = (X_{r_1+1,n}, \ X_{2,n}, ..., X_{n-r_2,n}), \tag{2.17}$$

on peut écrire

$$E(\mathbf{X}) = \mu\mathbf{1} + \sigma\alpha,$$

où $\mathbf{1} = (1, \ 1, ..., \ 1)'$, $\alpha = (\alpha_{r_1+1}, \ \alpha_2, ..., \ \alpha_{n-r_2})'$ et $Var(\mathbf{X}) = \sigma^2 \sum$, où $\sum$ est une matrice $(n-r_2-r_1)\times(n-r_2-r_1)$ avec des éléments V_{rs}, $r_1 < r$, $s \leq n-r_2$. Ensuite, pour obtenir les $(MVLUE)$ s des paramètres de position et d'échelle μ et σ basés sur les statistiques d'ordre

$$\mathbf{X}' = (X_{r_1+1,n}, ..., \ X_{n-r_2,n}).$$

Un statisticien doit utiliser les formules suivantes, qui sont similaires à celles données au section précédente :

$$\hat{\mu}^* = \frac{1}{\Delta}\left\{\alpha'\sum^{-1}\alpha 1'\sum^{-1} - \alpha'\sum^{-1}1\alpha'\sum^{-1}\right\}X, \tag{2.18}$$

$$\hat{\sigma}^* = \frac{1}{\Delta}\left\{1'\sum^{-1}1\alpha'\sum^{-1} - 1'\sum^{-1}\alpha 1'\sum^{-1}\right\}X, \tag{2.19}$$

$$\Delta = (\alpha'\sum^{-1}\alpha)(1'\sum^{-1}1) - (\alpha'\sum^{-1}1)^2. \tag{2.20}$$

La variance et la covariance de ces estimateurs sont données comme suit :

$$Var(\hat{\mu}^*) = \frac{\sigma^2(\alpha' \sum^{-1} \alpha)}{\Delta}. \tag{2.21}$$

$$Var(\hat{\sigma}^*) = \frac{\sigma^2(1' \sum^{-1} 1)}{\Delta}. \tag{2.22}$$

$$Cov(\hat{\mu}^*, \hat{\sigma}^*) = -\frac{\sigma^2(\alpha' \sum^{-1} 1)}{\Delta}. \tag{2.23}$$

2.1.4 Estimation des paramètres basée sur les quantiles

Nous continuons à rechercher les meilleurs estimateurs non biaisés (au sens de la variance minimale) des paramètres de position et d'échelle. Dans ce chapitre, les $(MVLUE)\, s$, basés sur certaines statistiques d'ordre fixe, sont discutés.

Nous rappelons tout d'abord les définitions d'un quantile d'ordre λ et d'un quantile empirique d'ordre λ, $0 < \lambda < 1$, données au chapitre 1. Une valeur x_λ est appelée un quantile d'ordre λ, $0 < \lambda < 1$, si

$$P(X < x_\lambda) < \lambda < P(X \leq x_{\lambda'}).$$

Si X a un $f.d.c$ F continue, alors toute solution x_λ de l'équation $F(x_\lambda) = \lambda$, est un quantile d'ordre λ.

Suivant les définitions données au chapitre 1, nous déterminons le quantile empirique d'ordre λ, $0 < \lambda < 1$, comme $X_{[n\lambda]+1,n}$. Pour des raisons de simplicité, au lieu de $X_{[n\lambda]+1,n}$, nous utiliserons la notation x^s_λ.

Considérons un ensemble de nombres réels $0 < \lambda_1 < \lambda_2 < ... < \lambda_k < 1$, et que $x^s_{\lambda_1}$, $x^s_{\lambda_2}, ..., x^s_{\lambda_k}$.

Sont les quantiles correspondants de l'échantillon. *Mosteller* (1946) a montré que la

distribution conjointe des quantiles normalisés des échantillons

$$\sqrt{n}\left(x_{\lambda_1}^s - x_{\lambda_1}\right),\ \sqrt{n}\left(x_{\lambda_2}^s - x_{\lambda_2}\right), ...,\ \sqrt{n}\left(x_{\lambda_k}^s - x_{\lambda_k}\right),$$

correspondant à une distribution dont la $f.d.p$ est $f(x)$ tend, lorsque $n \to \infty$, vers une distribution normale à k-dimensions. à moyenne nulle et à matrice de covariance :

$$\begin{pmatrix} \frac{\lambda_1(1-\lambda_1)}{\left(f(x_{\lambda_1})\right)^2} & \frac{\lambda_1(1-\lambda_2)}{f(x_{\lambda_1})f(x_{\lambda_2})} & \cdots & \frac{\lambda_1(1-\lambda_k)}{f(x_{\lambda_1})f(x_{\lambda_k})} \\ \frac{\lambda_1(1-\lambda_2)}{f(x_{\lambda_1})f(x_{\lambda_2})} & \frac{\lambda_2(1-\lambda_2)}{\left(f(x_{\lambda_2})\right)^2} & \cdots & \frac{\lambda_2(1-\lambda_k)}{f(x_{\lambda_2})f(x_{\lambda_k})} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\lambda_1(1-\lambda_k)}{f(x_{\lambda_1})f(x_{\lambda_k})} & \cdots & \cdots & \frac{\lambda_k(1-\lambda_k)}{\left(f(x_{\lambda_k})\right)^2} \end{pmatrix} \tag{2.24}$$

Lorsque la taille n d'un échantillon est grande, un statisticien peut avoir des problèmes techniques pour construire les estimateurs appropriés basés sur un grand nombre de statistique d'ordre .*Ogawa* (1951) a suggéré de simplifier le problème en considérant les $(MVLUE)\,s$ des paramètres de position et d'échelle, qui n'utilisent qu'un nombre fixe k, disons $k = 2$ ou $k = 3$, des statistiques d'ordre. Cette simplification est d'un grand intérêt, si n est suffisamment grand, et alors il vaut mieux résoudre le problème en termes de k quantiles empiriques.

L'énoncé du problème est le suivant. Nous avons des observations, correspondant à une fonction de répartition $F((x-\mu)/\sigma)$. Ici $F(x)$ est connue et a la $f.d.p$ $f(x)$, μ et σ sont les paramètres de position et d'échelle, l'un ou les deux paramètres sont inconnus. Nous voulons construire le $(MVLUE)$ de(s) paramètre(s) inconnu(s) basé sur un ensemble fixe

$$\left(x_{\lambda_1}^s,\ x_{\lambda_2}^s, ...,\ x_{\lambda_k}^s\right),$$

des quantiles empiriques,où $0 < \lambda_1 < \lambda_2 < ... < \lambda_k < 1$.

Pour simplifier nos calculs nous supposons que $n \longrightarrow \infty$ et au lieu de la distribution

conjointe des quantiles empiriques, on peut utiliser son expression asymptotique, donnée ci-dessus. *Ogawa* a examiné certaines situations importantes.

1ercas. $(MVLUE)$ du paramètre de position μ lorsque le paramètre d'échelle σ est connu, soit $\hat{\mu}_q$, la $(MVLUE)$ de μ basée sur des combinaisons linéaires de k quantiles empiriques fixes $\left(x^s_{\lambda_1},\ x^s_{\lambda_2},...,\ x^s_{\lambda_k}\right)$.

$$\hat{\mu}_q = \frac{T}{K_1} - \frac{K_3}{K_1}\sigma, \tag{2.25}$$

et la variance de $\hat{\mu}_q$, se comporte $Var(\hat{\mu}_q) \sim \frac{\sigma^2}{nK_1}$, quand $n \longrightarrow \infty$, où :

$$\begin{aligned} T &= T\left(x^s_{\lambda_1},\ x^s_{\lambda_2},...,\ x^s_{\lambda_k}\right) \\ &= \sum\nolimits_{i=1}^{k+1} \frac{\left(f(x_{\lambda_i}) - f(x_{\lambda_{i-1}})\right)\left(f(x_{\lambda_i})x^s_{\lambda_i} - f(x_{\lambda_{i-1}})x^s_{\lambda_{i-1}}\right)}{\lambda_i - \lambda_{i-1}}, \\ \lambda_0 &= 0,\ \lambda_{k+1} = 1,\ f(x_{\lambda_0}) = 0,\ f(x_{\lambda_{k+1}}) = 0. \end{aligned}$$

$$K_1 = \sum\nolimits_{i=1}^{k+1} \frac{\left(f(x_{\lambda_i}) - f(x_{\lambda_{i-1}})\right)^2}{\lambda_i - \lambda_{i-1}}. \tag{2.26}$$

$$K_3 = \sum\nolimits_{i=1}^{k+1} \frac{\left(f(x_{\lambda_i}) - f(x_{\lambda_{i-1}})\right)\left(f(x_{\lambda_i})x_{\lambda_i} - f(x_{\lambda_{i-1}})x_{\lambda_{i-1}}\right)}{\lambda_i - \lambda_{i-1}}. \tag{2.27}$$

2èmecas. $(MVLUE)$ du paramètre d'échelle σ, lorsque le paramètre de position μ est connu, soit $\hat{\sigma}_q$ le $(MVLUE)$ de σ, il se trouve que :

$$\hat{\sigma}_q = \frac{S}{K_2} - \frac{K_3}{K_2}\sigma, \quad Var(\hat{\sigma}_q) \sim \frac{\sigma^2}{nK_2}, \quad n \longrightarrow \infty, \tag{2.28}$$

où

$$S = S\left(x^s_{\lambda_1},\ x^s_{\lambda_2}, ...,\ x^s_{\lambda_k}\right)$$
$$= \sum_{i=1}^{k+1} \frac{\left(f(x_{\lambda_i})x_{\lambda_i} - f(x_{\lambda_{i-1}})x_{\lambda_{i-1}}\right)\left(f(x_{\lambda_i})x^s_{\lambda_i} - f(x_{\lambda_{i-1}})x^s_{\lambda_{i-1}}\right)}{\lambda_i - \lambda_{i-1}},$$

$$K_2 = \sum_{i=1}^{k+1} \frac{\left(f(x_{\lambda_i})x_{\lambda_i} - f(x_{\lambda_{i-1}})x_{\lambda_{i-1}}\right)^2}{\lambda_i - \lambda_{i-1}}, \tag{2.29}$$

et K_3 est donnée en (2.27).

3èmecas. $(MVLUE)$ s des paramètres de position et d'échelle, lorsque μ et σ sont inconnus, soient $\hat{\mu}_{q0}$ et $\hat{\sigma}_{q0}$ sont les $(MVLUE)$ s de μ et σ, basés sur des combinaisons linéaires de k quantiles empiriques fixes $x^s_{\lambda_1}$, $x^s_{\lambda_2}$, ..., $x^s_{\lambda_k}$, respectivement, alors

$$\hat{\mu}_{q0} = \frac{K_2T - K_3S}{\Delta}, \quad \text{et} \quad \hat{\sigma}_{q0} = \frac{-K_3T - K_1S}{\Delta},$$

où $\Delta = K_1K_2 - K_3^2$, et les expressions de

$$T = T\left(x^s_{\lambda_1},\ x^s_{\lambda_2}, ...,\ x^s_{\lambda_k}\right),$$
$$S = S\left(x^s_{\lambda_1},\ x^s_{\lambda_2}, ...,\ x^s_{\lambda_k}\right).$$

K_1, K_2 et K_3 sont données ci-dessus.

Les variances et les covariances des estimations $\hat{\mu}_{q0}$ et $\hat{\sigma}_{q0}$ se comporte comme suit :

$$Var(\hat{\mu}_{q0}) \sim \frac{K_2}{n\Delta}\sigma^2, \quad \text{et} \quad Var(\hat{\sigma}_{q0}) \sim \frac{K_1}{n\Delta},$$

et

$$Cov(\hat{\mu}_{q0}, \hat{\sigma}_{q0}) \sim -\frac{K_3}{n\Delta}\sigma^2, \quad \text{quand} \quad n \longrightarrow \infty. \tag{2.30}$$

Si la $f.d.p$ $F(x)$ est symétrique par rapport à zéro et l'emplacement de $0 < \lambda_1 <$

$\lambda_2 < ... < \lambda_k < 1$, est symétrique par rapport à $1/2$, c-à-d :

$$\lambda_j + \lambda_{k-j+1} = 1, \quad j = 1,\ 2, ...,\ k.$$

Alors $K_3 = 0$ et $Cov(\hat{\mu}_{q0}, \hat{\sigma}_{q0}) = 0$, de plus, $\hat{\mu}_{q0}$ et $\hat{\sigma}_{q0}$ dans ce cas coïncidé avec $\hat{\mu}_{q0}$ du 1^{er} cas et $\hat{\sigma}_{q0}$ de $2^{ème}$ cas respectivement.

2.2 Théorie des valeurs extrêmes

La théorie des valeurs extrêmes (EVT, en anglais Extreme Value Theory) a été développée pour l'estimation de probabilités d'occurrences d'évènements extrêmes, qui sont souvent rares. Elle permet d'extrapoler le comportement de la queue de distribution à partir des plus grandes valeurs observées (les observations extrêmes de l'échantillon).

Cette section réservée pour les valeurs extrêmes et les queues de distributions. Nous énonçons les principaux résultats concernant les distributions limites des plus grandes observations d'un échantillon ainsi que les domaines d'attraction.

De plus amples détails pour la théorie des valeurs extrêmes peuvent être trouvés dans les références : de Haan, Ferreria [6], Embrechts et al. [3] et Reiss, Thomas (1997) [10].

En règle générale, les résultats de minima peuvent être déduits des résultats correspondant aux maxima par la relation triviale

$$\min(X_1,\ X_2, ...,\ X_n) = -\max(-X_1,\ -X_2, ...,\ -X_n).$$

2.2.1 Quelques résultats fondamentaux de la théorie des valeurs extrêmes

Pour commencer notre étude et les explications de la théorie des valeurs extrêmes, il faut avoir un grand bagage, alors notre point de départ sera les statistiques d'ordre (déjà évoquées au premier chapitre).

Théorèmes limites

L'objectif principal de la théorie des valeurs extrêmes (TVE) est de trouver les lois limites possibles pour suites des maximums $\{max(X_1,\ X_2, ...,\ X_n),\ n \in \mathbb{N}^*\}$ si on connaît la loi exacte de la $v.a\ X$.

Loi des grands nombres Essentiellement, la loi des grands nombres (LGN) indique que lorsque l'on fait un tirage aléatoire dans une série de grandes tailles, plus on augmente la taille de l'échantillon, plus les caractéristiques statistiques du tirage (l'échantillon) se rapprochent des caractéristiques statistiques de la population. Elle se subdivise elle-même en deux catégories :

a Loi faibles des grands nombres

Théorème 2.2.1 *Soit X_1, X_2, ..., X_n une suite de $(v.a)$ i.i.d, de carrée intégrable et définie sur le même espace de probabilité (Ω, F, P) de variance finie σ^2 et d'espérance finie μ, on pose :*

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \quad alors \quad \bar{X}_n \xrightarrow{p} \mu. \tag{2.31}$$

$i.e\ \ \forall \varepsilon > 0 \ \ \lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) = 0.$

b Loi forte des grands nombres

Considérons n variables aléatoires $i.i.d$ $(X_1,\ X_2, ...,\ X_n)$ d'espérance finie μ ($i.e$ $E(|X|) < \infty$), on pose

$$\bar{X_n} = \frac{X_1 + X_2 + ... + X_n}{n}, \quad \text{alors } \bar{X_n} \overset{p.s}{\to} \mu. \tag{2.32}$$

La loi forte des grands nombres précise que $\bar{X_n}$, converge vers μ " presque sûrement ", c-à-d que :

$i.e\ P(\lim_{n\to\infty} \bar{X_n} = \mu) = 1$

Théorème de la limite centrale

Théorème 2.2.2 *(TCL) Si $X_1,\ X_2, ...,\ X_n$ est une suite de variables aléatoires définies sur le même espace de probabilité de variance σ^2 finie et de moyenne μ, on pose $S_n = X_1 + X_2 + ... + X_n$, alors :*

$$S_n \overset{loi}{\to} \mathcal{N}(n\mu, \sigma\sqrt{n}) \quad i.e\ \tfrac{S_n - n\mu}{\sigma\sqrt{n}} \overset{loi}{\to} \mathcal{N}(0,1) \quad quand\ n \to \infty. \tag{2.33}$$

Remarque 2.2.1 *Pour plus de détails sur les théorèmes (LGN) et (TCL) en peut cité le livre Saporta. [12].*

Théorème 2.2.3 *(Fonction survie ou fonction de queue) On appelle fonction de survie ou fonction de queue de la variable aléatoire X, la fonction $\bar{F} : \mathbb{R} \longrightarrow [0,1]$ définie par*

$$\bar{F}(x) = P(X > x) = 1 - F(x), \quad x \in \mathbb{R}. \tag{2.34}$$

Théorème 2.2.4 *(Glivenko-Cantelli, 1933) Soit $(X_n,\ n \in \mathbb{N})$ une suite des $(v.a)$ réelles indépendantes et de même loi de F, avec la fonction de répartition empirique F_n, alors :*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p.s} 0, \quad quand\ n \to \infty. \tag{2.35}$$

Définition 2.2.1 *(Point terminal) Le point terminal d'une fonction F est défini par*

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}.$$

Définition 2.2.2 *(Point extrême) on note par x_F (resp x_F^*) le point extrême supérieur (resp inférieur) de la distribution F (i.e : la plus grande valeur possible pour $X_{k,n}$ peut prendre la valeur $+\infty$ (resp $-\infty$)) au sens où :*

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\} \leq \infty, \tag{2.36}$$

$$x_F^* = \inf\{x \in \mathbb{R} : F(x) > 0\}. \tag{2.37}$$

Définition 2.2.3 *(Fonction de Von-Mises) F est une fonction de Von-Mises (avec la fonction auxiliaire σ), s'il existe un certain $z \leq x_F$, tel que : $\forall x \in]z, x_F]$*

$$\bar{F}(x) = c\exp(-\int_z^x \frac{dt}{\sigma(t)}), \tag{2.38}$$

où $c > 0$, et σ est une fonction positive absolument continue (par rapport la mesure de Lebesgue) avec la densité σ' vérifiant $\lim_{x \longrightarrow x_F} \sigma'(x) = 0$.

2.2.2 Lois limites des valeurs extrêmes

Comportement asymptotique des extrêmes

Nous considérons n variables aléatoires $X_1, X_2, ..., X_n$ *i.i.d* de fonction de répartition F : une manière simple d'étudier le comportement des événements extrêmes est de considérer la variable aléatoire :

$$X_{n,n} = \max(-X_1, -X_2, ..., -X_n),$$

la loi exacte de $X_{n,n}$ est :

$$F_{X_{n,n}}(x) = [F(x)]^n . \tag{2.39}$$

La formule (2.39) présente un intérêt très limité. De plus, la loi d'une variable aléatoire parente X est rarement connue avec précision et même si la loi de cette variable parente X est connue avec exactitude, la loi du terme maximum n'est pas toujours facilement calculable, *i.e* : La difficulté provient du fait que l'on ne connaît pas en général la fonction de répartition F, en plus la loi du maximum quand n tend vers l'infini est une loi dégénérée.

$$\lim_{n\to+\infty} F_{X_{n,n}}(x) \begin{cases} 0 & si \quad x < x_F, \\ 1 & si \quad x \geq x_F. \end{cases}$$

Pour cette raison, on essaie de trouver des constantes (a_n) et (b_n) $n \in \mathbb{N}^*$ de telle sorte que la loi du maximum normalisée, *i.e*. la loi de $(X_{n,n} - b_n)/a_n$ ne le soit plus, noton :

$$Y_n = \frac{X_{n,n} - b_n}{a_n}, \quad \text{avec} \quad a_n > 0 \quad \text{et} \quad b_n \in \mathbb{R}. \tag{2.40}$$
$$F_{Y_n}(x) = P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n).$$

On étudie $\lim\limits_{n\longrightarrow\infty} F_{Y_n}(x) = \lim\limits_{n\longrightarrow\infty} F^n(a_n x + b_n)$.

Définition 2.2.4 *On appelle loi asymptotique des extrêmes, la loi de la (v.a) Y, telle que*

$$Y_n = \frac{X_{n,n} - b_n}{a_n} \xrightarrow{L} Y, \quad quand \quad n \longrightarrow \infty. \tag{2.41}$$

Fisher et Tippett (1928), Gnedenko (1943) et de Hann (1970) *ont montré que les seules distributions limites non dégénérée H possibles sont les distributions de valeurs extrêmes. Le théorème suivant donne une condition nécessaire et suffisante pour l'existence d'une loi limite non dégénérée pour le maximum.*

Théorème 2.2.5 *(Fisher et Tippett(1928), Gnedenko(1943))*

Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de variables aléatoires $i.i.d$ de fonction de répartition F. S'il existe deux suites normalisantes réelles $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$, telles que

$$\lim_{n\longrightarrow\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = H(x). \tag{2.42}$$

Pour tout x, où H est une fonction de distribution non dégénérée. Alors H est du même type que l'une des fonctions suivantes :

1. **Loi de Gumbel**

$$H_0(x) = \Lambda(x) = \exp(\exp(-x)), \quad \forall x \in \mathbb{R}. \tag{2.43}$$

2. **Loi de Fréchet**

$$H_\alpha(x) = \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), & x > 0 \end{cases} \quad \text{avec} \quad \alpha > 0 \tag{2.44}$$

3. **Loi de Weibull**

$$H_\alpha(x) = \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha), & x \leq 0 \\ 1, & x > 0 \end{cases} \quad \text{avec} \quad \alpha < 0 \tag{2.45}$$

Ces trois Λ, Φ_α et Ψ_α sont appelées" **les distributions des valeurs extrêmes standard**" et les variables aléatoires correspondantes sont "**les variables aléatoires extrémales**".

– Ce théorème présente un intérêt important, car si l'ensemble des distributions est grand alors, l'ensemble des distributions des valeurs extrêmes est très petit. Ce théorème n'est valable que si les suites (a_n) et (b_n) existent et admettent des

limites.

– Les suites de normalisation (a_n) et (b_n) ne sont pas uniques.

– Ce théorème est un résultat important car il n'est pas nécessaire de faire d'hypothèses paramétriques sur la loi des X_i.

Proposition 2.2.1 *(Relation entre Λ et Φ_α et Ψ_α)*

Soit Y une variable aléatoire positive $(Y > 0)$ alors les affirmations suivantes sont équivalentes :

$$(Y \backsim \Phi_\alpha) \Longleftrightarrow (\ln Y^\alpha \backsim \Lambda) \Longleftrightarrow \left(-Y^{-1} \backsim \Psi_\alpha\right).$$

Preuve. On peut trouver la démonstration de ce théorème dans le livre Embrechts et al. (1997) [3].

2.2.3 Loi généralisée des valeurs extrêmes (GEV)

Grâce aux travaux de *Von Mises* (1936) et de *Jenkinson* (1955), on a une forme unifiée de la fonction de répartition de la loi des valeurs extrêmes à un facteur d'échelle et de position près.

Représentation de Jenkinson-Von Mises

Il est possible de rassembler les trois familles des lois des valeurs extrêmes en une seule famille paramétrique $(H\gamma(x),\, \gamma \in \mathbb{R})$ dite famille des lois des valeurs extrêmes généralisées notées $GEVD$ (Generalized Extreme Value Distribution). Elle est paramétrée par une seule variable $\gamma \in \mathbb{R}$, mais toujours à un facteur de changement d'échelle et de translation prés. La fonction de répartition est pour $\gamma \in \mathbb{R}$, et pour tout x, tel que $1 + \gamma\,(x) > 0$.

$$H_\gamma(x) := \begin{cases} \exp\left\{-\left[1+\gamma\,(x)\right]^{-1/\gamma}\right\}, & \gamma \neq 0, \quad 1+\gamma\,(x) > 0, \\ \exp(-\exp(-x)), & \gamma = 0, \quad x \in \mathbb{R}. \end{cases} \tag{2.46}$$

La forme la plus générale de la distribution des valeurs extrêmes est :

$$H_{\mu,\sigma,\gamma}(x) := \begin{cases} \exp\left\{-\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\gamma}\right\}, & \gamma \neq 0, \quad 1+\gamma\left(\frac{x-\mu}{\sigma}\right) > 0, \\ \exp(-\exp(-\frac{x-\mu}{\sigma})), & \gamma = 0, \quad x \in \mathbb{R}, \end{cases} \tag{2.47}$$

avec $\mu \in \mathbb{R}$ et $\sigma > 0$.

La fonction de répartition H_γ est appelée loi des valeurs extrêmes. μ est un paramètre de localisation, il est directement lié à la valeur la plus probable de la loi, il indique donc approximativement où se trouve le coeur de la distribution. σ est un paramètre de dispersion (paramètre d'échelle), il indique l'étalement des extrêmes. γ est l'indice de queue (paramètre de forme).

Proposition 2.2.2 *Suivant le signe du paramètre de forme, on définit trois types de GEV*

$$H_\gamma(x) = \begin{cases} \Lambda(x) & \gamma = 0 \quad x \in \mathbb{R}, \\ \Phi_{1/\gamma}(x) & \gamma > 0 \quad x > \mathbb{R}, \\ \Psi_{-1/\gamma}(x) & \gamma < 0 \quad x < \mathbb{R}. \end{cases}$$

On peut facilement montrer que la fonction de densité correspondante à $H_{\gamma,\mu,\sigma}$ pour ($\gamma \in \mathbb{R}$ et $1+\gamma\left(\frac{x-\mu}{\sigma}\right) > 0$) est :

$$h_{\gamma,\mu,\sigma}(x) = \begin{cases} \frac{1}{\sigma}\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-(1+\gamma)/\gamma} H_{\mu,\sigma,\gamma}(x), & \gamma \neq 0, \quad 1+\gamma\left(\frac{x-\mu}{\sigma}\right) > 0, \\ \frac{1}{\sigma}\left[\exp(-\left(\frac{x-\mu}{\sigma}\right) - \exp(-\frac{x-\mu}{\sigma}))\right], & \gamma = 0, \quad x \in \mathbb{R}. \end{cases} \tag{2.48}$$

Loi max-stable

Définition 2.2.5 *(Loi max-stable) une loi H non dégénérée est dite max-stable s'il existe des constantes $a_n \in \mathbb{R}_+^*$ et $b_n \in \mathbb{R}$, telles que*

$$H^n\left(a_n x + b_n\right) = H(x), \quad n \in \mathbb{N}^* \quad et \quad \forall x \in \mathbb{R}. \tag{2.49}$$

Exemple 2.2.1 *La loi de Weibull* (Ψ_α) *est une loi max-stable, en effet :*

$$\Psi_\alpha^n(n^{\frac{-1}{\alpha}}x) = \exp\left(-(n^{\frac{-1}{\alpha}}x)^\alpha\right)^n = \exp\left(-(n^{-1}x^\alpha)n\right)$$
$$= \exp\left(-(-x^\alpha)\right) = \Psi_\alpha(x).$$

2.2.4 Domaines d'attraction

L'analyse des valeurs extrêmes repose principalement sur les distributions limites des extrêmes et leurs domaines d'attraction. Ces distributions apparaissent comme les seules distributions limites possibles du maximum d'un échantillon de $(v.a)$ $(i.i.d)$ X_1, $X_2, ..., X_n$ de loi F.

La recherche du domaine d'attraction est équivalente à la réponse à la question suivante :

Etant donnée une distribution de valeurs extrêmes H, sous quelles conditions la fonction de distribution F des maxima $X_{n,n}$ normalisés converge-t-elle faiblement vers H ?

Définition 2.2.6 *(Domaine d'Attraction) On dit qu'une distribution* F *appartient au domaine d'attraction de* H*, et en note* $F \in DA(H_\gamma)$ *si la distribution du maximum renormalisée converge vers* H*. Autrement dit , s'il existe des constantes réelles* $a_n > 0$ *et* $b_n \in \mathbb{R}$*, telles qu'en tout point de continuité* x *de* H :

$$\lim_{n\longrightarrow\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\longrightarrow\infty} F^n(a_n x + b_n) = H_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}), \quad \forall x \in \mathbb{R}, \tag{2.50}$$

avec $1 + \gamma x > 0$.

Proposition 2.2.3 *(caractérisation de* $DA(H_\gamma)$*) On a* $F \in DA(H_\gamma)$ *si et seulement*

si : $\forall x \in \mathbb{R}$

$$\lim_{n\longrightarrow\infty} n\,[1 - F(a_n x + b_n)] = \lim_{n\longrightarrow\infty} n\,\bar{F}(a_n x + b_n) = -\log H_\gamma(x). \tag{2.51}$$

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* $(a_n^{-1}\,(X_{n,n} - b_n)\,,\ n \geq 1)$ *vers une variable aléatoire de fonction de répartition* H_γ.

La loi limite du maximum dépend donc du seul paramètre γ appelé l'indice des valeurs extrêmes. Selon le signe de γ, on définit trois types de domaines d'attraction :

1. Si $\gamma > 0$, on dit que F appartient au domaine d'attraction de *Fréchet* et on notera $F \in DA(\Phi_\gamma)$, ce domaine d'attraction regroupe les distributions à queue lourde (Cauchy, Pareto, Student, etc).
2. Si $\gamma = 0$, on dit que F appartient au domaine d'attraction de *Gumbel* et on notera $F \in DA(\Lambda_\gamma)$, ce domaine d'attraction est celui des distributions à queues légères, c'est -à-dire qui ont une fonction de servie à décroissance exponentielle (Exponentielle, Normale, Gamma, Weibull, etc).
3. Si $\gamma < 0$, on dit que F appartient au domaine d'attraction de $Weibull$ et on notera $F \in DA(\Psi_\gamma)$,nous trouvons des distributions dont le point terminal est fini (Uniforme, Beta, etc).

Le paramètre γ est un paramètre de forme. Il contrôle la forme de la queue de distribution.

Exemple 2.2.2 *Supposons donner, n variables aléatoires* $X_1,\ X_2, ...,\ X_n$ *i.i.d de fonction de répartition commune F, soit X suit la loi exponentielle standard du paramètre* $1\ (X \sim \mathcal{E}(1))\,,\ c\ à\ d : \{F(x) = 1 - \exp(-x),\ x \geq 0\}$, *et on pose* $a_n = 1$ *et*

$b_n = \ln n$

$$\lim_{n\longrightarrow\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\longrightarrow\infty} P\left(\frac{X_{n,n} - \ln n}{1} \leq x\right) = \lim_{n\longrightarrow\infty} F(x + \ln n)^n$$
$$= \lim_{n\longrightarrow\infty} [1 - \exp(-x + \ln n)]^n = \lim_{n\longrightarrow\infty} \left[1 - \frac{\exp(-x)}{n}\right]^n$$
$$= \exp(-\exp(-x)) = \Lambda(x),$$

car $\lim_{n\longrightarrow\infty} \left[1 - \frac{x}{n}\right]^n = \exp(-x)$, *alors la distribution exponentielle appartient au domaine d'attraction de* ***Gumbel.***

Caractérisation des domaines d'attraction

Nous donnons dans cette partie la caractérisation des trois domaines d'attraction, Fréchet, Weibull et Gumbel.

D'abord, nous donnons la définition d'une fonction à variation régulière parce qu'elle est présente dans la caractérisation des domaines d'attraction.

Notion de fonction à variation régulière La notion de fonction à variation régulière est très utilisée dans le contexte de la caractérisation des domaines d'attraction dans la théorie des valeurs extrêmes. Nous présentons ici quelques résultats principaux, pour plus de détails voir Bingham et al. (1987).

Définition 2.2.7 *(Fonction à variation régulière) Une fonction mesurable* $U : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, *est dite à variations régulières d'indice* $\alpha \in \mathbb{R}$ *à l'infini si pour tout* $t > 0$:

$$\lim_{x\longrightarrow\infty} \frac{U(tx)}{U(x)} = t^\alpha. \tag{2.52}$$

On notera dans la suite RV_α *l'ensemble des fonctions à variations régulières d'indice* α.

Définition 2.2.8 *Si une fonction mesurable* $L:\ \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, *est à variations régulières d'indice* 0 $(L \in RV_0)$, *on dit que* L *est à variations lentes à* ∞ :

$$\lim_{x\longrightarrow\infty} \frac{L(tx)}{L(x)} = 1, \quad t > 0. \tag{2.53}$$

Une fonction à variation régulière d'indice $\alpha \in R$ *peut toujours s'écrit sous la forme*

$$U(x) = x^{\alpha} L(x), \quad L \in RV_0.$$

Le résultat ci-dessous fournit une représentation des fonctions à variations régulières.

Proposition 2.2.4 *(Représentation de Karamata Resnick (1987)) Toute fonction à variation lente* L *à l'infini s'écrit sous la forme*

$$L(x) = c(x) \exp\left(-\int_1^x \frac{\Delta t}{t} dt\right), \tag{2.54}$$

où $c(\cdot) > 0$ *et* $\Delta(\cdot)$ *sont deux fonctions mesurables, telles que*

$$\lim_{x\longrightarrow\infty} c(x) = c_0 \in]0, +\infty[\quad et \quad \lim_{x\longrightarrow\infty} \Delta(x) = 0.$$

Si la fonction $c(\cdot)$ *est constante, alors on dit que* L *est normalisée. La relation (2.54) implique que si* L *est normalisée alors* L *est dérivable, de dérivée* L' *avec pour tout* $x > 0$:

$$L'(x) = \frac{L(x)\Delta(x)}{x},$$

en particulier, on a

$$\lim_{x\longrightarrow\infty} \frac{xL'(x)}{L(x)} = 0.$$

Définition 2.2.9 *(Convergence uniforme locale) Si* $U \in RV_\alpha$, $\alpha \in \mathbb{R}$ *alors pour tout*

intervalle $i = [a, b] \subset \mathbb{R}$, $0 < a < b < \infty$, *on a :*

$$\lim_{x\to\infty} \sup_{t\in I} \left| \frac{U(tx)}{U(x)} - t^{\alpha} \right| \longrightarrow 0, \quad \textit{quand } n \to \infty. \tag{2.55}$$

Si $\alpha < 0$, *le résultat de convergence uniforme est vrai pour des intervalles* I *de la forme* $[a, \infty[$, $a > 0$.

Si $\alpha > 0$ et si U est bornée sur l'intervalle $]0, b]$, $b > 0$, alors la convergence uniforme est vraie sur les intervalles $]0, b]$, $b > 0$.

Une propriété intéressante des fonctions à variations régulières est la conservation des équivalents à l'infini.

Proposition 2.2.5 *(Resnick,1987) Si* $U \in RV_{\alpha}$ *avec* $\alpha > 0$, *alors l'inverse généralisée de* U *est à variations régulières d'indice* $1/\alpha$ $\left(U^{\leftarrow}(\cdot) \in RV_{1/\alpha}\right)$.

Remarque 2.2.2 *Si* $U \in RV_{\alpha}$ *avec* $\alpha < 0$, *alors la fonction* $U^{\leftarrow}(1/\cdot)$ *est à variations régulières d'indice* $-1/\alpha$. *En effet, si* $U \in RV_{\alpha}$ *alors* $1/U \in RV_{-\alpha}$. *Donc, d'après la proposition* 2.2.5, $(1/U)^{\leftarrow} \in RV_{-1/\alpha}$. *Il reste à remarquer que* $(1/U(\cdot))^{\leftarrow} = U^{\leftarrow}(1/\cdot)$.

Domaine d'attraction de Fréchet

Le résultat ci-dessous assure que toute fonction appartenant au domaine d'attraction de Fréchet est une fonction à variations régulières.

Théorème 2.2.6 *Une fonction de répartition* F *appartient au domaine d'attraction de Fréchet* $(F \in \mathcal{DA}(\Phi_{\gamma})$ *avec un indice des valeurs extrêmes* $\gamma > 0)$ *si et seulement si la fonction de survie* $\bar{F} \in \mathcal{RV}_{-1/\gamma}$.

Des suites possibles de normalisation ((a_n) et (b_n), $\forall n > 0$), sont données par :

$$a_n = F^{\leftarrow}(1 - \tfrac{1}{n}), \quad \text{et} \quad b_n = 0.$$

Autrement dit, une fonction de répartition F appartenant au domaine d'attraction de Fréchet s'écrit sous la forme :

$$F(x) = 1 - x^{-1/\gamma} L(x), \quad L \in RV_0.$$

Remarque 2.2.3 *Toutes les fonctions de répartition du domaine d'attraction de Fréchet ont un point terminal infini* $(x_F = +\infty)$. *En effet, par définition, une fonction à variations régulières ne peut pas être nulle à partir d'un certain rang.*

Domaine d'attraction de Weibull

Le résultat suivant montre que l'on passe du domaine d'attraction de Fréchet à celui de Weibull par un simple changement de variable dans la fonction de répartition.

Théorème 2.2.7 *Une fonction de répartition* F *appartient au domaine d'attraction de Weibull* $(F \in \mathcal{DA}(\Psi_\gamma)$ *avec un indice des valeurs extrêmes* $\gamma < 0)$ *si et seulement si son point terminal* x_F *est fini* $(x_F < +\infty)$ *et si* $1 - F^*$ *est une fonction à variation régulière d'indice* $1/\gamma$ *avec :*

$$F^*(x) \begin{cases} 0, & si \quad x \leq 0, \\ F(x_F - 1/x), & si \quad x > 0. \end{cases}$$

Des suites possibles de normalisation $((a_n)$ *et* $(b_n), \ \forall n > 0)$, *sont données par :*

$$a_n = x_{F-} F^{\leftarrow}(1 - \tfrac{1}{n}), \quad et \quad b_n = x_F.$$

Ainsi, une fonction de répartition F *du domaine d'attraction de Weibull s'écrit pour* $x \leq x_F$:

$$F(x) = 1 - (x_F - x)^{1/\gamma} \left[L\left(x_F - x\right)^{-1} \right], \quad L \in RV_0.$$

Domaine d'attraction de Gumbel

La caractérisation des fonctions de répartition du domaine d'attraction de Gumbel est plus complexe. Une distribution F appartient au $DA(\Lambda_\gamma)$ si F est une fonction de Von Mises.

Théorème 2.2.8 *Une fonction de répartition F appartient au domaine d'attraction de Gumbel $F \in DA(\Lambda_\gamma)$ si et seulement s'il existe $z < x_F \leq \infty$, tel que*

$$\bar{F}(x) = c(x) \exp\left(-\int_z^x \frac{g(t)}{a(t)} dt\right), \quad z < x < x_F,$$

où c et g sont deux fonctions mesurables telles que $c(x) \longrightarrow c > 0$ et $g(x) \longrightarrow 1$ lorsque $x \longrightarrow x_F$, et α une fonction positive et dérivable, de dérivée a' telle que $a'(x) \longrightarrow 0$ lorsque $x \longrightarrow x_F$, alors F est une fonction de Von Mises et α sa fonction auxiliaire.

Des suites possibles de normalisation ((a_n) et (b_n), $\forall n > 0$), sont données par :

$$a_n = a(b_n), \quad \text{et} \quad b_n = F^{\leftarrow}(1 - \tfrac{1}{n}).$$

Le domaine d'attraction de Gumbel regroupe une grande diversité des lois comptant parmi elles la plupart des lois usuelles (loi normale, exponentielle, gamma, log-normale). Cette famille étant difficile à étudier dans toute sa généralité, de nombreux auteurs se sont concentrés sur une sous-famille : les lois à queue de type Weibull.

2.2.5 Loi généralisée de pareto (GPD)

Distribution des excès

On cherche à partir de la loi F de X à définir une loi conditionnelle F_μ par rapport au seuil u pour les variables aléatoires dépassant ce seuil.

Nous supposons une suite d'observations X_1, X_2, ..., X_n *i.i.d.* de fonction de répartition F et x_F un point terminal, alors pour un seuil $u < x_F$ fixé, on déduit la variable $X_i - u$, $i = 1, ..., n$, l'excès au-dessus du seuil u, la fonction de répartition des excès est :

$$F_\mu(y) = P(X - u \leq y / X > u) = \frac{F(y+u)-F(u)}{1-F(u)}, \quad \text{pour} \quad 0 \leq y \leq x_F - u. \tag{2.56}$$

Distribution de pareto généralisé (GPD)

La fonction de distribution de Pareto généralisée notée GPD (en anglais Generalized Pareto Distribution), pour $\gamma \in \mathbb{R}$ et $\beta > 0$, est définie par :

$$G_{\gamma,\beta}(x) := \begin{cases} 1 - \left[1 + \gamma \frac{x}{\beta}\right]^{-1/\gamma}, & \gamma \neq 0, \\ 1 - \exp(-\frac{x}{\beta}), & \gamma = 0, \end{cases} \tag{2.57}$$

cette distribution est définie pour :

$$\begin{cases} x \geq 0, & si \quad \gamma \geq 0, \\ 0 \leq x \leq -\frac{\beta}{\gamma}, & si \quad \gamma < 0, \end{cases}$$

et sa densité s'écrit alors :

$$g_{\gamma,\beta}(x) := \begin{cases} \frac{1}{\beta}\left[1 + \gamma \frac{x}{\beta}\right]^{-(1/\gamma)-1}, & \gamma \neq 0, \\ \frac{1}{\beta}\exp(-\frac{x}{\beta}), & \gamma = 0. \end{cases} \tag{2.58}$$

Le GPD avec deux paramètres regroupe trois distributions selon les valeurs du paramètre de forme. Lorsque $\gamma > 0$, c'est la loi Pareto, lorsque $\gamma < 0$, nous avons la loi Bêta et $\gamma = 0$ donne la loi exponentielle.

Pour plus de détails sur le GPD voir Embrechts et al. [3].

2.2.6 Théorème de Balkema-Haan et Pickands

Le théorème suivant faire le lien entre le comportement asymptotique de la distribution des excès et la loi de Pareto généralisée.

Théorème 2.2.9 *(de Balkema-Haan(1974), Pickands(1975)) Soit F_μ la distribution des excès. Si $F \in DA(H_\gamma)$, la GPD est la distribution limitée de la distribution des excès lorsque le seuil μ tend vers x_F, alors pour $\beta(u) > 0$ et $\gamma \in \mathbb{R}$:*

$$\lim_{\mu \to x_F} \sup_{0<x<x_F-\mu} \left| F_\mu(x) - G_{\gamma,\beta(u)}(x) \right| = 0, \tag{2.59}$$

où $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ est le point terminal de F et $G_{\gamma,\beta(u)}$ est la fonction de répartition de la loi de Pareto généralisée.

Chapitre 3

Tests statistiques

Un test statistique (test d'hypothèse) est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations. Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes. On distinguera deux classes de tests :

1. **Test paramétrique :** est un test pour lequel on fait une hypothèse sur la forme des données sous H_0 (Normale, Poisson, ...). Les hypothèses du test concernant alors les paramètres gouvernant cette loi, par exemple (*Test T, Test Z, ANOVA...*)
2. **Test non paramétrique :** est un test ne nécessitant pas d'hypothèse sur la forme des données. Les données sont alors remplacées par des statistiques ne dépendant pas des moyennes/variances des données initiales (tables de contingence, statistique d'ordre ...), par exemple (*test de Kolmogorov-Smirnov, test de Wilcoxon, test de Shapiro, ...*)

Les tests paramétriques, quand leurs conditions sont remplies, sont les plus puissants que les tests non paramétriques. Les tests non paramétriques s'emploient lorsque les conditions d'application des autres méthodes ne sont pas satisfaites, même après d'éventuelle transformation de variables. Ils peuvent s'utiliser même pour des échantillons de taille très faible.

On fonction de l'hypothèse testée, plusieurs types de tests peuvent être réalisés :

- Le test de conformité.
- Le test d'ajustement ou d'adéquation.
- Le test d'homogénéité ou de comparaison.
- Le test d'indépendance ou d'association.

Différentes étapes doivent être survies pour tester une hypothèse :

1. définir l'hypothèse nulle (notée H_0) à contrôler,
2. choisir un test statistique ou une statistique pour contrôler H_0,
3. définir la distribution de la statistique sous l'hypothèse "H_0 est réalisée",
4. définir le niveau de signification du test α et la région critique associée,
5. calculer, à partir des données fournies par l'échantillon, la valeur de la statistique,
6. prendre une décision consernant l'hypothèse posée.

3.1 Test de normalité

En statistiques, les tests de normalité permettent de vérifier si des données réelles suivent une loi normale ou non. Les tests de normalité sont des cas particuliers de tests d'adéquation (ou tests d'ajustement, tests permettant de comparer des distributions), appliqués à une loi normale.

Le test de normalité est vraiment un test d'hypothèse. L'hypothèse nulle (H_0) est

que vos données ne sont pas différentes de la normale, votre hypothèse alternative (H_1) est que vos données sont différentes de la normale. Indépendamment du test de normalité statistique que vous utilisez, vous prendrez votre décision de rejeter ou non la valeur nulle en fonction de votre $valeur - P$.

Il existe un certain nombre de tests statistiques que vous pouvez utiliser (test Anderson-Darling, test de Shapiro-Wilk, test Kolmogorov-Smirnov, Lilliefors, D'Agostino's K-Squared, Chen-Shapiro). Les hypothèses nulles et alternatives sont les mêmes que celles décrites ci-dessus.

Le modèle paramétrique spécifié par l'hypothèse H_0 est donc ici

$$\begin{cases} \text{Il existe } (\mu, \sigma^2) \text{ appartenant à } \mathbb{R} \times \mathbb{R}_+^* \\ \text{tel que } X_1,\ X_2, ...,\ X_n \text{ est une échantillon de la loi } \mathcal{N}(\mu, \sigma^2). \end{cases}$$

La normalité est une condition indispensable à vérifier pour la réalisation des tests paramétriques en statistiques.

Deux méthodes sont à retenir :

1. **La méthode graphique :** examen visuel de la représentation graphique (l'histogramme, QQ-plot ou la boîte à moustache). Facile mais subjective.
2. **La méthode théorique (les tests de normalités)** : Le test de Kolmogrov-Smirnov, Lilliefors et test de Shapiro-Wilk largement utilisés, et les plus répandus.

Dans ce chapitre, nous allons mettre l'accent sur le test de Shapiro-Wilk (recommandé pour les échantillons de petite taille).

3.1.1 Test de normalité de Shapiro-Wilk

Il existe plusieurs tests qui permettent de tester l'hypothèse nulle (H_0)selon laquelle un échantillon X_1, $X_2, ..., X_n$ serait issu d'une population normalement distribuée. Celui-ci dit Shapiro-Wilk est dû à Samuel Sanford Shapiro et Martin Bradbury Wilk (1965). D'après la littérature ce serait le plus puissant devant d'autres tests de normalité (notamment celui de Kolmogorov et Smirnov appliqué à la loi Normale).

Le test de Shapiro-Wilk est basé sur le statistique W. En comparaison avec les autres tests, il est particulièrement puissant pour les petits effectifs ($n \leq 50$). La statistique W peut donc être interprétée comme le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générés à partir de la loi normale et les quantiles empiriques obtenus à partir des données. Plus W est élevé, plus la compatibilité avec la loi normale est crédible.

Le test d'adéquation de S-W est utilisé pour déterminer si un échantillon aléatoire X_1, $X_2, ..., X_n$ est tiré d'une distribution de probabilité gaussienne normale avec la moyenne et la variance réelles, μ et σ^2, respectivement, c-à-d $X \sim \mathcal{N}(\mu, \sigma^2)$. Ainsi, nous souhaitons tester l'hypothèse suivante :

$$\begin{cases} H_0 : \text{l'échantillon aléatoire a été tiré d'une population normale } \mathcal{N}(\mu, \sigma^2) \\ H_1 : \text{l'échantillon aléatoire ne suit pas } \mathcal{N}(\mu, \sigma^2) \end{cases}$$

Pour tester cette hypothèse, nous utilisons la statistique du test de Shapiro-Wilk, qui est donnée par :

$$W = \frac{\left[\sum_{i=1}^{[\frac{n}{2}]} (\alpha_{i,n} * d_i)\right]^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{B^2}{S^2}, \tag{3.1}$$

d'où : $\frac{n\alpha_{1,n}^2}{n-1} \leq W \leq 1$, avec $n > 1$.

La région critique du test W est définie par :

$$R.C : W < W_{crit}.$$

Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk (Table L2, [11], p 354).

Voici les différentes étapes pour calculer la valeur W du test de Shapiro-Wilk à partir d'un échantillon avec n observations $X_1, X_2, ..., X_n$

1. On trie les observations dans l'ordre croissant $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$.
2. On calcule la moyenne de ces observations $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
3. On calcule S^2 qui est la somme des écarts à la moyenne $S^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.
4. On calcule B^2 (un autre estimateur de la variance des données)
   - On calcule d : $d_i := X_{n-i+1:n} - X_{i:n}$.
   - $[\frac{n}{2}]$ est la partie entière du rapport $\frac{n}{2}$. L'avantage de cette formule, c'est quelle réduit, en moitié, le calcul de la somme du numérateur B^2.
   - $B^2 = \left[\sum_{i=1}^{[\frac{n}{2}]}(\alpha_{i,n} * d_i)\right]^2$, où $\alpha_{i,n}$ sont des constantes générées à partir de la moyenne et de la matrice de variance co-variance des quantiles d'un échantillon de taille n suivant la loi normale.

     $$(\alpha_{i,n})_1^n = \frac{m^t V^{-1}}{\sqrt{(m^t V^{-1})\,(V^{-1} m)}}.$$

     Ces constantes sont fournies dans des tables spécifiques (Table L1, [11], p352-353).
5. On calcule : $W = \frac{B^2}{S^2}$.
6. On compare la valeur de W observée avec la valeur de W_{crit} donnée dans la table de Shapiro-Wilk (Table L2, [11], p 354) Si $W_{obs} < W_{crit}$, on rejette l'H_0, ce qui indique la non normalité des données.

Remarque 3.1.1 *Il existe une commande dans R qui permet de calculer directement la valeur de W_{obs} et la probabilité de dépassement $(valeur - P)$ associée à cette valeur : shapiro.test(X)*

- si la $valeur - P$ est inférieure à un niveau α choisi (par exemple 0.05), alors l'hypothèse nulle est rejetée (*i.e.* il est improbable d'obtenir de telles données en supposant qu'elles soient normalement distribuées).
- si la $valeur - P$ est supérieure au niveau α choisi (par exemple 0.05), alors on ne doit pas rejeter l'hypothèse nulle. La valeur de la $valeur - P$ alors obtenue ne présuppose en rien de la nature de la distribution des données.

Pour $n = 2$, la normalité ne peut jamais être rejetée, donc le test n'est utile que pour $n \geq 3$.

Pour des valeurs de W suffisamment proches de 1 (dépendant de n) l'hypothèse de normalité ne sera pas rejetée. Pour les plus petits W sera rejetée.

Exemple : soit le tableau suivant correspondant aux résultats de mesures d'un alésage (en mm)

1	2	3	4	5	6	7	8	9	10
12.124	12.232	12.327	12.242	12.466	12.215	12.026	12.359	12.215	12.387

1. Classement des valeurs de mesure par ordre croissant :

12.026	12.124	12.215	12.215	12.230	12.242	12.327	12.359	12.387	12.466

2. Calculer la moyenne empirique observée : $\overline{x} = 12.259$.

3. Calculer la valeur observée de S^2 :

$$s^2 = \sum_{i=1}^{10} (x_i - \bar{x})^2 \simeq 0.154.$$

4. Calculer les différences respectives :

$$d_1 = 12.466 - 12.026 = 0.440$$

$$d_2 = 12.387 - 12.124 = 0.263$$

$$d_3 = 12.359 - 12.215 = 0.144$$

$$d_4 = 12.327 - 12.215 = 0.112$$

$$d_5 = 12.242 - 12.230 = 0.012$$

5. A chacune de ces différences, on affecte les coefficients $\alpha_{i,n}$, donnés par la Table L1.

$$0.440 \times 0.5739 = 0.2525$$

$$0.263 \times 0.3291 = 0.0865$$

$$0.144 \times 0.2141 = 0.0308$$

$$0.112 \times 0.1224 = 0.0137$$

$$0.012 \times 0.0399 = 0.0005$$

6. Calculer la valeur observée de B^2 :

$$b^2 = \left[\sum_{i=1}^{5} \alpha_{i,n} \left(x_{n-i+1} - x_i\right)\right]^2 = 0.384.$$

7. Calculer la valeur observée rapport du rapport $W := B^2/S^2$:

$$w = (0.384)^2 / 0.1514 \simeq 0.974.$$

8. Pour une risque , le seuil critique lue dans la Table L2 :

$$w_{crit} = 0.842.$$

Puis on compare entre w et w_{crit}, comme $0.974 > 0.842$ alors on accepte la normalité des données.

Vérifions nos résultats en utilisant le langage R :

x<-c(12.124,12.232,12.327,12.242,12.466,12.215,12.026,12.359,12.215,12.387)

shapiro.test(x)

Après l'exécution on obtient :

```
Shapiro-Wilk normality test

data : x

W = 0.97471, p-value = 0.9308
```

La **p-value=**$0.9308 > 0.05$, donc en effet des données sont ajustées par la loi normale.

Chapitre 4

Simulation

1-Loi exponentielle

Soit $(X_1,\ X_2, ...,\ X_n)$ une suite des variables aléatoires de la loi exponentielle $\mathcal{E}(2)$, d'une fonction de distribution $\{F(x) = 1 - \exp(-2x),\ x \geq 0\}$.

Code R :

```
m=100
M=rep(NA,m)
n=1000
###génération du max pour un échantillon exp de paramètre lamda
for(i in 1 :m){
L=2
X=rexp(n,rate=L)
M[i]=max(X)
}
#calcul des coefficients de normalisations
#le max normalisé
```

```
Yn=L*M-log(n)

library(evd)

Tn=rgumbel(m)

#test de kolmogrove égalité de deux lois

ks.test(Yn,Tn)

##########################

hist(Yn,freq=F,ylim=c(0,0.4),col="yellow",ylab = "densités",

xlab ="Valeurs du maximum pour n=1000",main = "Histogramme -densité-Gumbel"

)

lines(density(Yn),col="red",pch=19)

curve(dgumbel(x),col="green",pch=20,add=TRUE)
```

La valeur du test K-S et la p-value selon la valeur de n :

Two-sample Kolmogorov-Smirnov test

data : Yn and Tn

pour n=100 :

D = 0.15, p-value = 0.2106

pour n=1000 :

D = 0.17, p-value = 0.1111

pour n=10000 :

D = 0.15, p-value = 0.2106

pour n=100000 :

D = 0.1, p-value = 0.6994

alternative hypothesis : two-sided

Commentaire

On remarque que la $valeur - P$ est supérieure au risque $\alpha := 0.05$, ce qui implique que Y_n suit la loi Gumbel.

Illustration graphique :

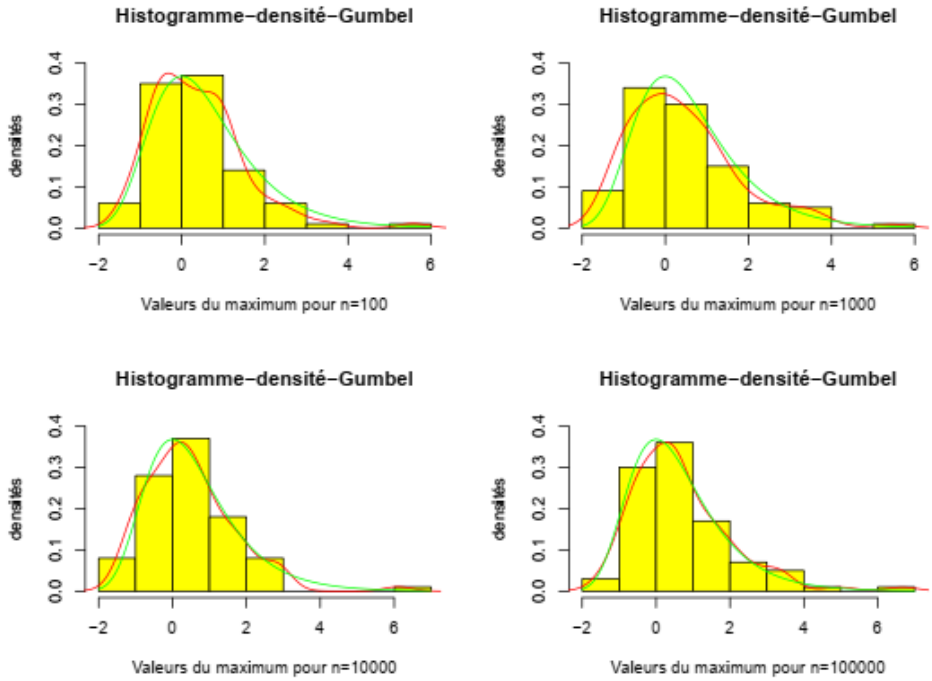

FIG. 4.1 – Ajustement de la loi du maximum avec celle du Gumbel.

2-Loi de pareto

Soit $(X_1,\ X_2, ...,\ X_n)$ une suite des variables aléatoires de la loi de Pareto, d'une fonction de distribution $\{F(x) = 1 - (x)^{-\alpha},\ \alpha > 0,\ x \geq 1\}$.

Code R :

```
m=100

M=rep(NA,m)

###génération du max pour échantillon uniforme de paramètre teta

for(i in 1 :m){

n=1000
```

```
a=3
U=runif(n)
X=(1-U)^(-1/a)
M[i]=max(X)
}
#calcul des coefficients de normalisations
#le max normalisé
Yn=M/n^(1/a)
library(evd)
Tn=rfrechet(m,loc=0,scale=1,shape=a)
#test de kolmogrove égalite de deux lois
ks.test(Yn,Tn)
##########################
hist(Yn,freq=F,ylim=c(0,1.2),col="yellow",ylab = "densités",
xlab ="Valeurs du maximum pour n=1000",
main = "Histogramme-densité-Fréchet" )
lines(density(Yn),col="red",pch=19)
curve(dfrechet(x,loc=0,scale=1,shape=a),col="green",pch=20,add=TRUE)
```

La valeur du test K-S et la p-value selon la valeur de n :

Two-sample Kolmogorov-Smirnov test

data : Yn and Tn

pour n=100 :

D = 0.13, p-value = 0.3667

pour n=1000 :

D = 0.08, p-value = 0.9062

pour n=10000 :

D = 0.12, p-value = 0.4676

pour n=100000 :

D = 0.15, p-value = 0.2106

alternative hypothesis : two-sided

Commentaire :

On remarque que la $valeur - P$ est supérieure au risque $\alpha := 0.05$, ce qui implique que Y_n suit la loi Fréchet.

Illustration graphique :

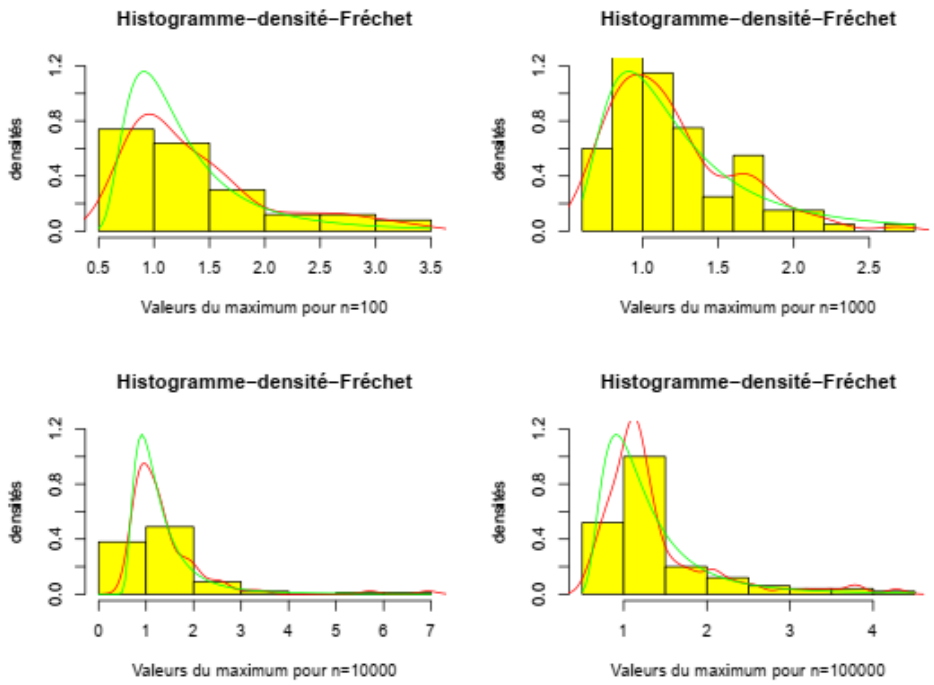

FIG. 4.2 – Ajustement de la loi du maximum avec celle du Fréchet.

3-Loi uniforme

Soit $(X_1,\ X_2, ...,\ X_n)$ une suite des variables aléatoires de la loi uniforme sur l'inter-

valle $[0, 1]$.

Code R :

```
m=100
M=rep(NA,m)
###génération du max pour un échantillon uniforme de paramètre teta
for(i in 1 :m){
n=1000
X=runif(n)
M[i]=max(X)
}
#calcul des coefficients de normalisations
an=1/n ;bn=1
#le max normalisé
Yn=n*(M-1)
###génération d'un échantillon de weibull
library(stats)
Tn=log(runif(m))
#test de kolmogrove égalité de deux lois
ks.test(Yn,Tn)
##########################
hist(Yn,freq=F,ylim=c(0,0.8),xlim = c(0,-5.5),col="yellow",ylab = "densités",
xlab ="Valeurs du maximum pour n=1000",main = "Histogramme-densité-weibull"
)
lines(density(Yn),col="red",pch=19)
```

curve(exp(x),col="green",pch=20,add=TRUE)

La valeur du test K-S et la p-value selon la valeur de n :

Two-sample Kolmogorov-Smirnov test

data : Yn and Tn

pour n=100 :

D = 0.09, p-value = 0.8127

pour n=1000 :

D = 0.11, p-value = 0.5806

pour n=10000 :

D = 0.12, p-value = 0.4676

pour n=100000 :

D = 0.13, p-value = 0.3667

alternative hypothesis : two-sided

Commentaire :

On remarque que la $valeur - P$ est supérieure au risque $\alpha := 0.05$, ce qui implique que Y_n suit la loi Weibull.

Illustration graphique :

3-Loi normale

Soit $(X_1,\ X_2, ...,\ X_n)$ une suite des variables aléatoires de la loi normale centrée réduite, d'une fonction de densité $\{f(x) = 1/\sqrt{2\pi}\exp(-x^2/2)\}$.

Code R :

m=100

M=rep(NA,m)

###génération du max pour un échantillon normale de paramètres u et sd

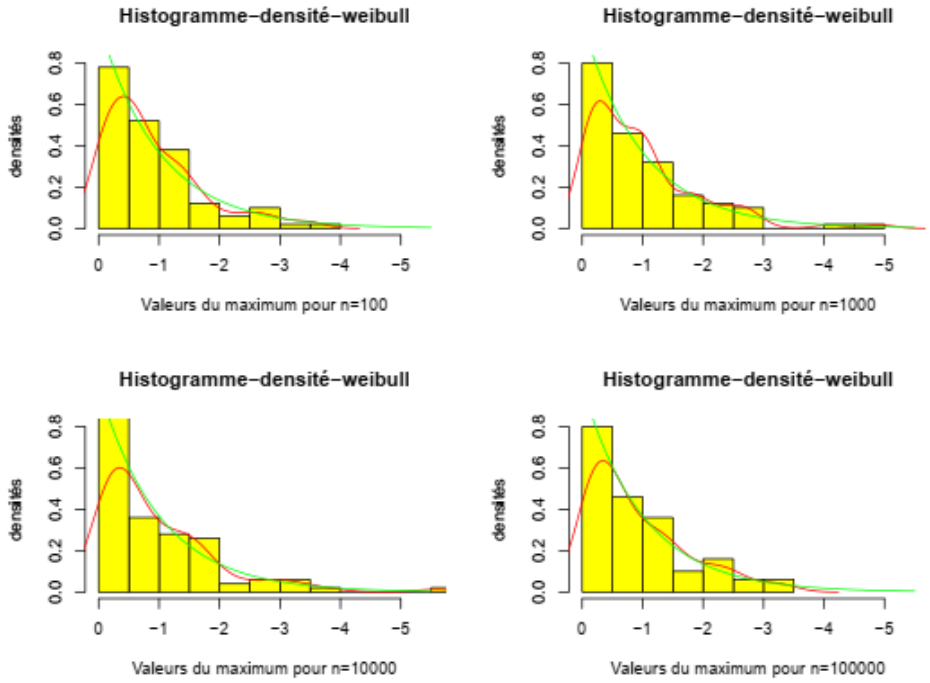

FIG. 4.3 – Ajustement de la loi du maximum avec celle du Weibull.

```
for(i in 1 :m){
s=0
k=1
n=1000
X=rnorm(n,mean=s,sd=k)
M[i]=max(X)
}
#calcule des coefficients de normalisations
bn=sqrt(2*log(n))-(log(4*pi)+log(log(n)))/(2*sqrt(2*log(n)))
an=1/sqrt(2*log(n))
#le max normalisé
Yn=(M-bn)/an
```

```
library(evd)
Tn=rgumbel(m)
#test de kolmogorov égalite des deux lois
ks.test(Yn,Tn)
##########################
hist(Yn,probability = TRUE,ylim=c(0,0.5),col="yellow",ylab = "densités",
xlab ="Valeurs du maximum pour n=1000",main = "Histogramme-densité-Gumbel"
)
lines(density(Yn),col="red",pch=19)
curve(dgumbel,col="green",pch=20,add=TRUE)
```

La valeur du test K-S et la p-value selon la valeur de n :

Two-sample Kolmogorov-Smirnov test

data : Yn and Tn

pour n=100 :

D = 0.16, p-value = 0.1545

pour n=1000 :

D = 0.13, p-value = 0.3667

pour n=10000 :

D = 0.12, p-value = 0.4676

pour n=100000 :

D = 0.19, p-value = 0.0541

alternative hypothesis : two-sided

Commentaire :

On remarque que la $valeur - P$ est supérieure au risque $\alpha := 0.05$, ce qui implique

que Y_n suit la loi Gumbel

Illustration graphique :

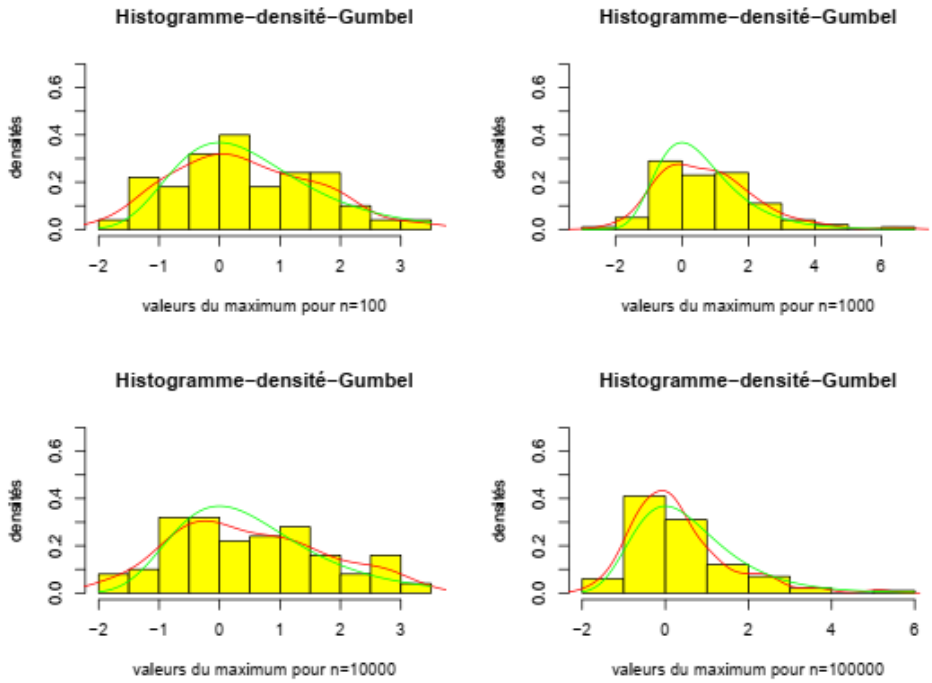

FIG. 4.4 – Ajustement de la loi du maximum avec celle du Gumbel.

Remarque 4.0.2 *on a utilisé la commande* ***par(mfrow=c(2,2)),*** *pour avoir les quatres graphes sur la même figure.*

Conclusion

Les statistiques d'ordre jouent un rôle important dans plusieurs domaines de la statistique, à savoir l'estimation paramétrique, et non paramétrique, les tests statistiques, la théorie des valeurs extrêmes. Comme application de celles-ci nous citons l'estimateur de l'index des valeurs extrêmes (Hill, 1965), l'estimateur linéaire sans biais de variance minimale ($MVLUE$), test de Shapiro-Wilk (1965).

Bibliographie

[1] Arnold, B.C, Balakrishan, N., Nagaraja, H.N. (1992). A First Course in Order Statistics. Wiely, New York.

[2] David, H.A, Nagaraja, H.N (2003). Order Statistics Third Edition. Wiely

[3] Embrechts, P. Klüppelberg, C. Mikosch, T. (1997). Modelling Extremal Events, in : Applications in Mathematics. Springer-Verlag, New York. vol 33.

[4] Fisher, R.A., Tippet, L.H.C., (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. Proceedings of the Cambridge Philosophical Society, 24, 180-190.

[5] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Annales de Mathématiques 44, 423 453.

[6] Laurens de Haan, Ferreira, A. (2006). Extreme Value Theorie An Introduction. Springer.

[7] Mohammad, A.Valery, B. N. et Mohammad, S. (2013). An Introduction to Order Statistics. Atlantis studies in Probability and Statistics, 3. Atlantis press, Paris.

[8] Michel Lejeune, Statistique La théorie et ses applications Deuxième édition. (2010). Springer.

[9] N. Balakrishnan, C.R. Rao. (1998). Order Statistics : Applications, Handbook of Statistics 17.

[10] Reiss, R.D., Thomas, M. (1997). Statistical analysis of extreme values with applications to insurance,finance, hydrology and other fields Birkhäuser, Basel.

[11] Philippe Capéraà, Bernard Van Cutsem. (1988). Méthodes et Modèles en Statistiques Non Paramétrique

exposé fondamentale.

[12] Saporta, G. (1990). Probabilité, analyse des données et statistique. Editions Technip, Paris.

[13] http ://www.biostat.ulg.ac.be/pages/Site_r/Normalite.html,Vérifier la normalité des données.

[14] https ://www.sciencedirect.com/topics/psychology/shapiro-wilk-test, Shapiro-WilkTest.

[15] https ://math.mit.edu/~rmd/465/shapiro.pdf,The Shapiro-Wilk And Related Tests For Normality.

[16] https ://www.nrc.gov/docs/ML1714/ML17143A100.pdf,Power comparisons of Shapiro-Wilk,Kolmogorov-Smirnov, Lilliefors and Anderson-Darlingtests.

[17] https ://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf, wikistat.

[18] http ://irma.math.unistra.fr/~gardes/Poly_extreme.pdf.

[19] http ://bu.univ-ouargla.dz/master/pdf/negais-chaima.pdf.

[20] https ://cprp.sti-beziers.fr/wp-content/uploads/2017/10/carte-de-controle-Xbar-R.pdf.

Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$(X_1, X_2, ..., X_n)$	:	Échantillons de taille n de $(v.a)'s$.
$X_{1,n}, X_{2,n}, ..., X_{n,n}$	:	Statistiques d'ordre associées à $(X_1, X_2, ..., X_n)$.
$:=$	:	Égalité par définition.
S_n	:	Somme arithmétique.
$E(X),\ \mu$	:	Espérance de X.
$V(X),\ \sigma^2$	:	Variance de X.
$X_{i,n}$	:	La $i-ième$ statistique d'ordre.
$\mathbb{R}$	:	Ensemble des nombres réels.
$\mathbb{N}$	:	Ensemble des nombres entiers naturels.
$\Lambda(x)$	:	Distribution des valeurs extrêmes de Gumbel.
$\Phi_\alpha(x)$	:	Distribution des valeurs extrêmes de Fréchet.
$\Psi_\alpha(x)$	:	Distribution des valeurs extrêmes de Weibull.
W	:	La statistique de Shapiro-Wilk.
Θ	:	Espace des paramètres.
al	:	Autres.
$\mathbf{X}'$	:	Le vecteur transposé de X.

$\xrightarrow{p.s}$	:	Convergence en prèsque sûre.
$\xrightarrow{p}$	:	Convergence en probabilité.
$\xrightarrow{d}$	:	Convergence en distribution.
$F^{\leftarrow}(.)$	:	L'inverse généralisée de F.
$F(.)$	:	Fonction de répartition.
$\overline{F}(.)$	:	Fonction de survie.
$DA(.)$	:	Domaine d'attraction.
$F_n(.)$	:	Fonction de répartition empirique.
f	:	Densité de probabilité d'une variable aléatoire.
$f.d.p$	:	Fonction de densité de probabilité.
$f.d.c$	:	Fonction de distribution cumulative.
$v.a$	:	Variale aléatoire "au singulier comme au pluriel".
$i.e$	:	En d'autre terme.
$i.i.d$	:	Indépendante et identiquement distribuée.
$\sum$	:	Matrice variance-covariance.
$\sum^{-1}$	:	Matrice inverse de $\sum$.
x_F	:	Le point limite de F.
σ	:	Paramètre d'échelle.
μ	:	Paramètre de position.
γ	:	Paramètre de forme.
u	:	Seuil.
H	:	Fonction de répatition de la $(v.a)$ Y.
$\stackrel{d}{=}$	:	Égalité en distribution.
L	:	Fonction à variation lente.
θ	:	Vecteur colonne des paramètres.
$\binom{n}{m}$	:	La combinaison de m objets parmi n objets sans remise.

Résumé

Ce mémoire est un aperçu général sur les statistiques d'ordre associées à un échantillon et à ces applications. On s'intéresse aux caractéristiques de ces dernières pour des populations continues. Nous présentons leurs distributions, leurs moments d'ordre supérieurs, ces estimations optimales et le comportement des valeurs extrêmes. En outre, on définit les tests statistiques associés à savoir le test de Shapiro-Wilk. A la fin des simulations illustrons notre travail sont données.

Abstract

This memory is a general over view of order statistics pertaining to a given sample and its applications. We focus on the characteristics of these statistics for continuous populations. We present their distributions, their higher order moments, their optimal estimates and the behavior of extreme values. Furthermore, we define the respective statistical tests, such as the Shapiro-Wilk test. We illustrate our work by simulations.

ملخص

هذه المذكرة هي لمحة عامة عن الإحصاءات المرتبة المرتبطة بالعينة و تطبيقاتها. عملنا يتمحور في دراسة خصائص هذه الأخيرة للمتغيرات العشوائية المستمرة. نتكلم هنا عن توزيعاتها الاحتمالية و العزوم و التقديرات المثالية و سلوك القيم القصوى .إضافة إلى ذلك نعرف الاختبارات المتعلقة بها كاختبار Shapiro-Wilk . ننهي عملنا بدراسة محاكاة توضح النقاط المذكورة اعلاه .