République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE MOHAMED KHIDER BISKRA
FACULTE DES SCIENCES ET DES SCIENCES DE L'INGENIEUR
DEPARTEMENT DE MATHEMATIQUES

MEMOIRE

Présenté en vue de l'obtention du diplôme de

Magister en Mathématiques

Par

BERKANE Hassiba

Thème

Méthodes du Bootstrap pour les queues de distributions

Option

Analyse & Modèles Aléatoires

Soutenu publiquement le : 17/05/2005

Devant le jury :

Président :	**B. MEZERDI**	**PR.**	**U.M.K. Biskra.**
Rapporteur :	**A. NECIR**	**M.C.**	**U.M.K. Biskra.**
Examinateur :	**K. BOUKHETALA**	**PR.**	**U.S.T.H.B. Alger.**
Examinateur :	**S. BAHLALI**	**DR. M.A.C.**	**U.M.K. Biskra.**

Remerciements

Le moment est venu d'exprimer toute ma reconnaissance à mon encadreur, ***Dr.*** *A. Necir,* Maître de Conférences à l'université de Biskra, de m'avoir proposé un sujet très intéressant, et qui m'a fait découvrir la statistique, pour son enthousiasme et son soutien sans faille durant la réalisation de ce mémoire. Il a toujours été disponible pour me prodiguer ses conseils.

J'exprime ma profonde gratitude au ***Pr.*** *B. Mezerdi* qui m'a fait l'honneur de présider le jury de ce mémoire. Je lui suis très reconnaissante pour l'attention qu'il a porté à ce travail.

J'adresse mes sincères remerciements au ***Dr.*** *S. Bahlali* et au ***Pr.*** *K. Boukhetala* pour l'intérêt qu'ils ont manifesté a ce mémoire en acceptant de l'examiner.

L'occasion m'est donnée ici de remercier ma sœur *Houda,* ***M***^{elles}***.*** *R. Fouzia,* et *G. Soufia* pour toute l'aide qu'elles m'ont apporté. Je souhaite remercier aussi ***M***^{er}*. D. Yahia,* pour leur aide.

J'exprime ma gratitude à ma famille qui m'a toujours soutenue et encouragée dans la voie que je m'étais fixée. Je remercie particulièrement mes parents qui m'ont stimulée et encouragé pendant mes études.

Mes vifs remerciements vont également à tous mes enseignants en graduation, et en post graduation, notamment, ***M***^{r}***.*** *B. Labed,* ***Dr.*** *L. Melkmi,* ***Dr.*** *A. Belaggoun.* J'adresse un amical remerciement le plus sincère à touts mes collègues pour leur sympathie et leur soutien.

Table des matières

0.1 Notations

v.a.r.	Variable aléatoire réelle.
$\theta,\ T_n$	Valeur de population, fonction des données.
Θ	Espace des paramètres.
$\Im$	Espace des fonctions de distributions.
$\xrightarrow{\mathcal{D}}$	Convergence en distribution.
$\xrightarrow{p}$	Convergence en probabilité.
$\xrightarrow{p.s}$	Convergence presque sûre.
$\overset{\mathcal{D}}{=}$	Egalité en distribution.
F	Fonction de répartition.
F^{-1}	Fonction des quantiles de la queue.
P_n, F_n	Loi de probabilité, fonction de la distribution empirique.
F_n^{-1}	Fonction des quantiles du queue empirique.
$G_{\gamma,\sigma}$	Fonction de répartition de la loi des valeurs extrêmes.
i.i.d.	Indépendante et identiquement distribuée.
$X^* = (X_1^*, ..., X_n^*)$	Échantillon Bootstrap.
FDC	Fonction de distribution cumulative.
FDE	Fonction de distribution empirique.
Md_n	Médiane empirique.
PR	Probabilité de rejeter.
$J(T_n)$	Estimateur du jackknife de T_n.
IC	Intervalle de confiance.
CBa	Correction pour le biais et accélération.
EQM	Erreur quadratique moyenne.
$G_{\gamma,\sigma},\ DPG$	Distribution de Paréto Généralisée.
γ	Index des valeurs extrêmes
$\{\mathcal{B}(t); 0 \leq t \leq 1\}$	Pont Brownien.
$\{W(t);\ 0 \leq t < \infty\}$	Processus de Wiener.

Liste des figures

Liste des tableaux

0.2 Introduction

Le terme de rééchantillonnage, ou en anglais, "*bootstrap*", qui évoque l'action de "se hisser en tirant sur ses propres lacets", désigne un ensemble des méthodes qui consistent à faire de l'inférence statistique sur des "nouveaux" échantillons tirés à partir d'un échantillon initial. Disposant d'un échantillon destiné à donner une certaine information sur une population, on tire au sort, parmi la sous-population réduite à cet échantillon, un nouvel échantillon de même taille n, et on répète cette opération B fois, où B est grand. On analyse ensuite les nouvelles observations ainsi obtenues pour affiner l'inférence faite sur les observations initiales. A priori, on peut avoir des doutes sur l'efficacité d'une telle méthode et penser qu'il n'y a aucune amélioration à espérer en rééchantillonnant à partir du même échantillon. En effet, aucune information supplémentaire ne peut être espérée, toute l'information étant contenue dans l'échantillon initial. Cependant, comme on va le voir, ce rééchantillonnage, s'il ne rajoute aucune information, permet, dans certains cas, d'extraire de l'échantillon de base l'information souhaitée.

Le *bootstrap* est une technique de rééchantillonnage introduite par *Efron (1979)* [15] permettant de simuler la distribution d'un estimateur quelconque pour en apprécier le biais, la variance donc le risque quadratique ou encore pour en estimer un intervalle de confiance même si la loi théorique est inconnue, les méthodes de bootstrap sont mises en œuvre afin d'améliorer la précision des estimations statistiques.

Cette méthode est employée pour analyser la variabilité de paramètres statistiques en produisant des intervalles de confiance de ces paramètres. Elle est particulièrement utile quand les hypothèses de base sont irréalistes, ou les distributions des paramètres sont complexes ou inconnues (cas de la plupart des méthodes multidimensionnelles).

Le *bootstrap* n'est rien d'autre qu'une technique de simulation particulière, fondée sur la distribution empirique de l'échantillon de base. *Efron et Tibshirani (1993)* [17] réservent le nom de *bootstrap non-paramétrique* à ce type de simulation, et qualifie de *bootstrap paramétrique* les simulations qui mettent en jeu une distribution théorique et des paramètres calculés à partir de l'échantillon *(simulations classiques)*. On note que le Jackknife est déterministe et fait intervenir de façon symétrique l'échantillon (sans nécessiter de tirages pseudo-aléatoires), contrairement au *bootstrap*.

Le *bootstrap* fait l'objet de beaucoup des recherches dans les statistiques depuis son introduction par *Efron (1979)* [15]. Les résultats de cette recherche sont synthétisés dans les ouvrages par *Beran* et *Ducharme (1991)* [3], *Davison* et *Hinkley (1997)* [9], *Efron* et *Tibshirani (1993)* [17], *Hall (1992a)* [25], *Mammen (1992)* [36], et *Hall (1994)* [27], *Horowitz (1997-2000-2002)* [29 – 30], *Maddala* et *Jeong (1996)* [35] et *Vinod (1993)* [42]*...etc..*

Les travaux sur le *bootstrap* et la modélisation montrent qu'il convient parfois d'adapter la méthode aux modèles, de manière à pouvoir obtenir de bonnes propriétés. De nombreuses adaptations du *bootstrap* ont été proposées dans la littérature pour tenir compte des spécificités de certains modèles statistiques. La plupart de ces méthodes sont des cas particuliers du *bootstrap* généralisé introduit par *Lo (1991)* [34] et *Mason* et *Newton (1992)* [38], qui remet par ailleurs en cause l'intuition initiale du *bootstrap* d'efron.

Ce mémoire est réparti en quatre chapitres.

On présente dans le premier chapitre de ce travail quelques définitions, propriétés de bases en statistiques paramétrique et non paramétrique et aussi les théories permettent d'appliquer le *bootstrap* en général.

Au deuxième chapitre, on s'intéresse aux principes du bootstrap et l'algorithme générale de cette méthode, référée à *Efron et Tibshirani (1993)*. Nous étudions la validité du bootstrap d'abord dans le cas général, nous décrivons les théorèmes centrale limite et son erreur d'approximation liés, qui seront utilisé pour arriver à une bonne approximation de distribution bootstrapée. *Beran et Ducharme (1991)* [3] ont montrés la consistance du bootstrap et *Mammen (1992)* [36] a donné des conditions nécessaires et suffisantes pour réaliser la consistante appliquée aux fonctionnels linéaires.

Nous présentons dans le dernier chapitre le mécanisme du *bootstrap* dans la résolution des problèmes d'inférence statistique et sur l'estimation des paramètres. En outre, on expose les méthodes de rééchantillonnage et leurs apports sur l'estimation de l'erreur -standard et le biais d'un estimateur. Nous donnons aussi quelques méthodes de détermination des limites de confiance d'un paramètre estimé, et l'utilisation du bootstrap sur les modèles de régression.

Au $4^{ème}$ chapitre nous appliquons la méthode du *bootstrap* aux queues de distributions, aux valeurs extrêmes, et en particulier dans l'estimation de l'index de Pareto.

Nous achevons le présent mémoire par une étude simulatrice d'un test d'ajustement des modèles de type Pareto. Nos considérations sont basées sur la statistique-test introduite récemment par *Necir et Boukhetala (2005)* [40].

Chapitre 1

NOTIONS DE STATISTIQUE

1.1 Théorie de l'estimation

1.1.1 Modèles Statistiques

Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^n$ défini sur un espace de probabilité (Ω, τ, P). La loi de X n'est que partiellement connue. On suppose que cette loi appartient à une classe donnée $\{P_\theta\,;\,\theta \in \Theta\}$ de lois de probabilités, i.e. il existe $\theta_0 \in \Theta$ tel que $X \sim P_{\theta_0}$. Le problème d'estimation est typiquement la détermination de ce paramètre à partir de l'observation de X. Posons $\chi = X(\Omega) \subset \mathbb{R}^n$, χ est l'espace des observations. P_θ est une loi sur χ munit de sa tribu borélienne.

Définition 1.1.1 : Un modèle statistique est un triplet $(\chi, \mathcal{F}, \{P_\theta; \theta \in \Theta\})$ où χ est une partie de $\mathbb{R}^n$ représentant l'espace des observations ; $\mathcal{F}$ est la tribu borélienne sur χ; $\{P_\theta; \theta \in \Theta\}$ est une famille de lois de probabilité définies sur $(\chi, \mathcal{F})$.

Lorsqu'il n'existe aucune ambiguïté sur l'espace des observations, on appellera modèle statistique la famille $\{P_\theta; \theta \in \Theta\}$. Si on note $p(x;\theta)$ la densité de P_θ, la famille de densités $\{p(x;\theta)\,;\theta \in \Theta\}$ sera aussi appelée modèle statistique. On notera E_θ (resp E) l'espérance associée à P_θ (resp E).

D'une façon générale :

$$\chi = \mathbb{R}^n,\ \mathcal{F} = \mathcal{B}(\mathbb{R}^n),$$

où

$$\Theta \in \mathbb{R}^n \rightarrow \text{ modèle paramétrique,}$$

et

$$\Theta \text{ plus gros } \rightarrow \text{ modèle non paramétrique.}$$

On dit que le problème est *paramétrique* si la famille de départ est paramétrée par un vecteur de dimension fini ; dans ce cas sa résolution se résume à l'identification de paramètre.

Le problème est *semi- paramétrique* s'il s'agit d'estimer un vecteur de dimension finie, mais que la famille de départ n'est pas paramétrée.

Le problème est *non- paramétrique* s'il s'agit d'estimer un vecteur de dimension infini.

X est une variable aléatoire définie sur $(\Omega, \mathcal{F}, P)$, à valeurs dans $\mathbb{R}^n$. Donc si I est une partie de $\mathbb{R}^n$, on a :

$$P(X \in I) = P_\theta(I).$$

Pour des raisons de commodité on utilisera souvent la notation suivante

$$P_\theta(X \in I).$$

Cette notation, quoique formelle, présente l'intérêt de faire apparaître que l'on considère "la probabilité que l'observation X soit dans I, sous l'hypothèse que la loi de X est P_θ". De même si φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$, on a :

$$E\varphi(X) = E_\theta\varphi.$$

Dans la suite on utilisera aussi la notation $E_\theta\varphi(X)$.

On aura souvent à manipuler des suites $X_1, ..., X_n$ de variables indépendantes suivent toutes une même loi.

Définition 1.1.2 : On appelle $n-$*échantillon* d'une loi F_θ, de densité $f(.;\theta)$, sur un espace A toute suite $X_1, ..., X_n$ de v.a indépendantes de même loi F_θ.

On considère un *n-échantillon* d'une loi de densité $f(.;\theta)$ sur $\mathcal{A} \in \mathbb{R}$ dépendant d'un paramètre inconnu $\theta \in \Theta$. Le modèle statistique associé à cette situation est le suivant :

$$\chi = \mathcal{A}^n,\ \mathcal{F} = \text{boréliens de } \chi,\ P_\theta = \text{loi de densité } p(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta),$$

où $x = (x_1, ..., x_n) \in \chi$.

1.1.2 Estimateur

On se donne un modèle statistique $(\chi, \mathcal{F}, \{P_\theta; \theta \in \Theta\})$. On cherche à estimer θ le paramètre inconnu.

Définition 1.1.3 : On appelle estimateur toute fonction mesurable T de χ à valeur dans Θ (de l'espace des observations dans l'espace des paramètres). On parlera d'estimateur de θ.

Définition 1.1.4 : Une suite d'estimateurs $\hat{\theta}_n = \hat{\theta}_n(y_1, ..., y_n)$ est dite *consistante* si $\hat{\theta}_n \to \theta$ en probabilité, et *fortement consistance* si la convergence a lieu presque sûrement.

Cette notion d'estimateur est très large. On verra dans les paragraphes suivants deux grandes classes d'estimateurs (ou plutôt de méthodes d'estimation), les estimateurs paramétriques et les estimateurs non paramétriques.

1.2 Méthodes paramétriques

On se donne un modèle $(\chi, \mathcal{F}, \{P_\theta; \theta \in \Theta\})$, et on suppose que Θ est une partie de $\mathbb{R}^d$. On note $p(x;\, \theta)$ la densité associée à la loi P_θ.

1.2.1 Biais et risque

Définition 1.2.1 : Le biais de l'estimateur T est la fonction

$$\theta \to b(\theta, T) \overset{\mathcal{D}}{=} E_\theta T - \theta \in \mathbb{R}^d.$$

Le *risque* de l'estimateur T est la fonction

$$\theta \to R(\theta, T) \overset{\mathcal{D}}{=} E_\theta |T - \theta|^2 .$$

Le biais définit donc *l'erreur moyenne* d'estimation. Le risque donne un critère de qualité de l'estimateur T. Un estimateur T est dite *estimateur sans biais (SB)* si

$$b(\theta, T) = 0, \quad \forall \theta \in \Theta.$$

Etant donnés deux estimateurs T_1 et T_2 de θ, T_1 sera dit *préférable* à T_2 si

$$R(T_1, \theta) \leq R(T_2, \theta), \quad \forall \theta \in \Theta.$$

Définition 1.2.2 : Un estimateur T^* est dit minimax s'il minimise $\sup_{\theta \in \Theta} R(T, \theta)$;

$$\sup_{\theta \in \Theta} R(T^*,\ \theta) = \inf \sup_{\theta \in \Theta} R(T, \theta).$$

Estimation par le maximum de vraisemblance

Définition 1.2.3 : Soit $(\Omega, \mathcal{A}, \mathbb{P}_\theta ; \theta \in \Theta)$ un modèle dominé par μ. On note

$$p(\theta, \omega) = \frac{dP_\theta}{d\mu}(\omega).$$

On appelle *estimateur du maximum de vraisemblance* le point $\hat{\theta}(\omega)$ tel que :

(i) $p(\hat{\theta}(\omega), \omega) \geq p(\theta, \omega), \quad \forall\ \theta \in \Theta.$

(ii) le *maximum* existe et est unique, avec $\hat{\theta} = \arg\max p(\theta, \omega)$.

Définition 1.2.4 : On appelle *fonction de vraisemblance*, calculée à partir d'un échantillon $(X_1, ..., X_n)$, la fonction $\mathcal{L}$ donnée par l'équation

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(\theta, x_i).$$

Estimateurs bayésien et "MAP"

On considère un modèle bayésien, i.e. un modèle statistique $(\chi\ ,\ \mathcal{F},\ \{P_\theta\ ; \theta \in \Theta\})$ et une densité $p_{\underline{\theta}}(.)$ de la loi d'une variable aléatoire $\underline{\theta}$ définie sur Θ. Etant donné T un estimateur de θ, on définit le *risque bayésien* associé à la loi *a* priori $p_{\underline{\theta}}$

$$\tilde{R}(T) = \begin{cases} \int_\Theta R(\theta, T)\, p_{\underline{\theta}}(\theta) d\theta, & \text{(si } p_{\underline{\theta}} \text{ est continue)}, \\ \sum_{\theta \in \Theta} R(\theta, T)\, p_{\underline{\theta}}(\theta), & \text{(si } p_{\underline{\theta}} \text{ est discrète)}. \end{cases}$$

Définition 1.2.5 : Soit $\tilde{\theta}$ un estimateur du paramètre θ. Si pour tout estimateur T de θ ,on a $\tilde{R}\left(\tilde{\theta}\right) \leq \tilde{R}(T)$, alors $\tilde{\theta}$ est appelé *estimateur bayésien* associé à la densité a priori $p_{\underline{\theta}}(.)$, s'écrit explicitement en fonction de $P(.; \theta)$, la densité de P_θ, de la manière suivante

$$\tilde{\theta}(x) = \begin{cases} \dfrac{\int_\Theta \theta p(x; \theta)\, p_{\underline{\theta}}(\theta) d\theta}{\int_\Theta p(x; \theta)\, p_{\underline{\theta}}(\theta) d\theta}, & \text{(si } p_{\underline{\theta}} \text{ est continue)}, \\ \dfrac{\sum_{\theta \in \Theta} \theta p(x; \theta)\, p_{\underline{\theta}}(\theta)}{\sum_{\theta \in \Theta} p(x; \theta)\, p_{\underline{\theta}}(\theta)}, & \text{(si } p_{\underline{\theta}} \text{ est discrète)}. \end{cases}$$

Remarque 1.2.1 : Soit $p(x)$ la densité de la loi d'une v.a X à valeurs dans $\mathbb{R}$. Quel est le point x_0 de $\mathbb{R}$ qui peut le mieux représenter X ? Il existe deux approches. Premièrement, on peut dire que ce point est l'espérance de X, i.e. $x_0 = \int x p(x)\, dx$.

Cette approche est justifiée par le fait que x_0 ainsi défini minimise la fonction $x \to E(X - x)^2$ (*l'erreur quadratique*). Deuxièmes, on peut choisir un point lel que la probabilité que X tombe dans un voisinage de ce point soit plus grande possible. Cela revient à choisir x_0 qui maximise la fonction $x \to P(x - \alpha \leq X \leq x + \alpha)$ (α petit). Donc x_0 est le point qui maximise $p(x)$, x_0 est appelé le mode de $p(x)$. Cette

seconde approche suppose que $p(x)$ admette un maximum unique. Dans le cas d'une densité gaussienne, le mode et l'espérance coïncident.

D'après cette remarque, on peut choisir comme estimateur celui qui maximise la densité a posteriori :

$$p_{\underline{\theta}|X}(\theta|X) = \begin{cases} \dfrac{p(x;\theta)\, p_{\underline{\theta}}(\theta)}{\int_\Theta p(x;\theta)\, p_{\underline{\theta}}(\theta) d\theta}, & \text{(si } p_{\underline{\theta}} \text{ est continue)}, \\ \dfrac{p(x;\theta)\, p_{\underline{\theta}}(\theta)}{\sum_{\theta\in\Theta} p(x;\theta)\, p_{\underline{\theta}}(\theta)}, & \text{(si } p_{\underline{\theta}} \text{ est discrète)}. \end{cases}$$

Définition 1.2.6 : L'estimateur "MAP" (maximum a posteriori) de θ pour la densité a prioiri $p_{\underline{\theta}}(\theta)$, noté $\tilde{\theta}_{MAP}$, est défini par :

$$\tilde{\theta}_{MAP}(x) \in \ Arg\max_{\theta\in\Theta} p_{\underline{\theta}|X}(\theta|X) = Arg\max_{\theta\in\Theta} p(x;\theta)\, p_{\underline{\theta}}(\theta), \quad \forall x \in \mathcal{X}.$$

Supposons la densité $p(x;\theta)$ unimodale, c'est -à- dire ne possédant "qu'une seule bosse". Prenons pour $p_{\underline{\theta}}(\theta)$ une densité uniforme sur un intervalle $[-K, K]$ (K grand). Il est alors facile de vérifier que l'estimateur MAP et l'EMV coïncident. En effet

$$\begin{aligned} Arg\max_{\theta\in\Theta} p_{\underline{\theta}|X}(\theta|X) &= Arg\max_{\theta\in\Theta} \frac{1}{c} p(x;\theta)\, p_{\underline{\theta}}(\theta) = Arg \max_{-K\le\theta\le K} \frac{1}{2Kc} p(x;\theta) \\ &= Arg \max_{-K\le\theta\le K} p(x;\theta) = Arg \max_{\theta\in\Theta} p(x;\theta)\ . \end{aligned}$$

Cette dernière égalité est dûe au fait que, pour K suffisamment grand, le maximum en θ de $p(x;\theta)$ est atteint sur l'intervalle $[-K, K]$.

Théorie asymptotique

Soit $X = (X_1, X_2, \ldots, X_n \ldots)$ un échantillon de taille infinie d'une loi F_θ de densité $f(.;\theta)$ sur $\mathcal{A} \subset \mathbb{R}^k$. On suppose que θ est unidimensionnel ($\Theta \subset \mathbb{R}$). Le modèle statistique correspondant est le suivant :

$$\chi = \mathcal{A}^\infty, \quad p(x;\theta) = \prod_{i=1}^{\infty} f(x_i;\theta), \quad x = (x_1, ..., x_n ...) \in \chi.$$

On dispose d'une suite $\{T_n\}_{n\in\mathbb{N}}$ d'estimateurs de θ tels que, pour tout n, $T_n = T_n(X_1, X_2, .., X_n...)$, (i.e. T_n ne dépend que de $(X_1, X_2, .., X_n...)$. Il s'agit d'étudier le comportement des estimateurs T_n lorsque $n \to \infty$.

Définition 1.2.7 : On dira que la suite d'estimateurs $\{T_n\}$ de θ est consistante si pour tout $\theta \in \Theta$,

$$T_n = T_n(X_1, X_2, .., X_n...) \underset{n\to+\infty}{\to} \theta\,, \quad P_\theta - p.s^1$$

Théorème 1.2.1 : *Supposons la densité $f(.;\theta)$ régulière par rapport à θ. On considère l'EMV $\hat{\theta}_n$ défini à partie de $X_1, ..., X_n$, i.e. $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ est solution de l'équation de la vraisemblance*

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(x_i;\theta) = 0.$$

La suite est consistante

$$\hat{\theta}_n \underset{n\to+\infty}{\to} \theta, \quad P_\theta - p.s, \quad (\forall\theta \in \Theta)\,.$$

Définition 1.2.8 : Soit $\{T_n\}$ une suite d'estimateurs de θ telle que :

$$loi\left(\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)}\right) \underset{n\to+\infty}{\to} N(0,1)\,,$$

pour une suite $(\mu_n(\theta), \sigma_n(\theta)) \in \mathbb{R}\times\mathbb{R}$ donnée. Une telle suite d'estimateurs est dite

asymptotiquement normale. $\mu_n(\theta)$ est appelée la moyenne asymptotique et $\sigma_n(\theta)^2$ la variance asymptotique.

La quantité suivante $\dfrac{\mu_n(\theta) - \theta}{\sigma_n(\theta)}$ définit le biais asymptotique. La suite $\{T_n\}$ est dite asymptotiquement sans biais si le biais asymptotique tend vers 0 lorsque $n \to +\infty$.

1.2.2 Intervalles de confiance

Construire un intervalle de confiance pour un paramètre consiste à trouver des bornes qui soient telles que l'on puisse connaître la probabilité pour que le paramètre

[1] $\zeta_n \underset{n\to\infty}{\to} \zeta$ $P_\theta - p.s.$ signifie que $P_\theta\left(\zeta_n \underset{n\to\infty}{\to} \zeta\right) = 1$.

(inconnu) appartienne à l'intervalle formé par ces deux bornes. Pour une certaine probabilité, notée $1-\alpha$ et appelée niveau de confiance.

Si $X_1, ..., X_n$ est un échantillon aléatoire issu d'une loi normale avec moyenne μ (inconnue) et variance σ^2 (connue) telle que

$$\bar{X} = \frac{1}{n}\left(X_1 + ... + X_n\right), \quad se(\bar{X}) = \frac{\sigma}{\sqrt{n}} \text{ (erreur standard).}$$

Alors un intervalle de confiance (IC) de niveaux $1-\alpha$ aura la forme

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],$$

où $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ est la valeur telle que, pour $Z \sim N(0,1)$

$$P\left(Z > z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} = P\left(Z < -z_{\frac{\alpha}{2}}\right).$$

Si n est assez grand et $X_1, ..., X_n$ i.i.d avec $\mathbb{E}(x) = \mu$, $\mathbb{V}(x) = \sigma^2 < \infty$, alors

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n}) \quad \text{ou} \quad \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \approx N(0,1).$$

On obtient donc la même forme d'IC mais c'est un IC approximatif. Si σ^2 n'est pas connue, on peut l'estimer avec S^2. Autrement dit, on va utiliser

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \quad \text{au lieu de} \quad \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}.$$

Cette statistique est distribuée suivant la loi de *student* de paramètre $n-1$ qu'on appelle des degrés de liberté. L'intervalle de confiance que l'on en déduit a une forme très similaire au cas précédent, à savoir

$$\left[\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right],$$

où la valeur $t_{\frac{\alpha}{2},n-1}$ est obtenue dans la table de la loi student. Mais le cas non normal avec n grand et variance inconnue en utilisant le *TCL*, on a :

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0,1).$$

On en déduit l'IC *approximatif*

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right].$$

1.2.3 Test d'hypothèses

Dans un problème d'estimation on recherche la vraie valeur d'un paramètre θ, et la réponse est un point de l'espace des paramètres Θ. Dans un problème d'intervalle de confiance, on recherche un intervalle contenant la vraie valeur du paramètre et la réponse est un intervalle de Θ. Dans un problème de test, il n'existe que deux réponses possibles : "*Oui*" ou "*non*".

On se donne un modèle statistique $\{P_\theta; \theta \in \Theta\}$, le problème de test se formule ainsi : étant donnée une partie Θ_0 de Θ, la vraie valeur de θ est-elle dans Θ_0 ou non ?

Hypothèses simples

Définition 1.2.9 : Un problème de test d'hypothèses est la donnée d'un modèle statistique $\{P_\theta; \theta \in \Theta\}$ et d'une partie Θ_0 de Θ. L'hypothèse H_0 : "$\theta \in \Theta_0$" est appelée *hypothèse nulle*. On pose $\Theta_1 = \Theta \diagdown \Theta_0$, L'hypothèse H_1 : "$\theta \in \Theta_1$" est appelée *hypothèse alternative.*

Définition 1.2.10 : Soit $\{P_\theta; \theta \in \Theta\}$ un modèle statistique et Θ_0 un ensemble de test. Soit D une partie de l'espace des observations χ.

La *règle de décision* ρ_D - ou test d'hypothèse -, fondée sur l'ensemble D est

$$x \in \chi \to \rho_D(x) = \begin{cases} 0, & \text{i.e accepter } H_0 \quad \text{si } x \notin D, \\ 1, & \text{i.e rejeter } H_0 \quad \text{si } x \in D. \end{cases} \tag{1.1}$$

D est appelé *région* de rejet de l'hypothèse H_0 (D sera parfois aussi appelé test).

Définition 1.2.11 : On appelle *hypothèse de base* l'hypothèse dont le rejet à tort a les conséquences les plus graves. C'est habituellement H_0. On appelle *risque de première espèce* du test ρ la probabilité de rejeter à tort l'hypothèse de base, soit $\mathbb{P}_\theta$ $(\rho = 0)$, $\theta \in \Theta_1$.

Ainsi on appelle *risque de seconde espèce* du test ρ la probabilité de rejeter à tort l'hypothèse alternative, soit $\mathbb{P}_\theta$ $(\rho = 1)$, $\quad \theta \in \Theta_0$.

Définition 1.2.12 : On dit que le test ρ est *exactement de niveau* α, $\alpha \in [0, 1]$, ssi $\forall \theta \in \Theta, \mathbb{P}_\theta$ $(\rho = 1) \leq \alpha$. Pour tester θ_0 contre θ_1 (*i.e.* $\mathbb{E}_\theta(\rho) \leq \alpha, \quad \forall \theta \in \Theta_0$) est dit *sans biais* si

$$\mathbb{E}_\theta(\rho) \geq \alpha, \quad \forall \theta \in \Theta_1.$$

Puissance

Définition 1.2.13 : On appelle puissance du test ϕ la quantité $\mathbb{E}_\theta(\phi)\,;\ \theta \in \Theta_1$. Un test ϕ est dit *uniformément le plus puissant (UPP)* de niveau α

$$\begin{cases} \phi \text{ est de niveau } \alpha. \\ \\ \forall \phi^{'} \text{ test de niveau } \alpha, \quad \mathbb{E}_\theta(\phi) \geq \mathbb{E}_\theta(\phi^{'}); \quad \forall \theta \in \Theta_1. \end{cases} \tag{1.2}$$

Rechercher un test UPP de niveau α implique que les hypothèses H_0 et H_1 ne sont pas considérées de la même manière. Donc rechercher un test UPP de niveau α revient à

$$\begin{cases} \text{choisir parmi tous les tests tels que } P\ (\textit{erreur de type } I) \leq \alpha. \\ \\ \text{celui qui minimise } P\ (\textit{erreur de type } II). \end{cases}$$

Test de deux hypothèses simples

Dans la plupart des problèmes il n'existe pas de test *UPP*. Une exception est le cas où l'espace des paramètres est réduit à deux points

$$\Theta \overset{\mathcal{D}}{=} \{\theta_0, \theta_1\} \text{ et } \begin{cases} H_0 : \text{“}\theta = \theta_0\text{”} \\ \\ H_1 : \text{“}\theta = \theta_1\text{”} \end{cases} \tag{1.3}$$

Dans ce cas nous allons voir qu'il existe un test *UPP.*

Théorème 1.2.2 : *(lemme de Neyman- Pearson) On considère un modèle statistique* $\{P_\theta; \theta \in \Theta\}$ *avec* $\Theta = \{\theta_0, \theta_1\}$, P_θ *de densité* $P(x, \theta)$ $(x \in \chi)$. *On veut tester l'hypothèse*

$$H_0 : \text{“}\theta = \theta_0\text{”}.$$

On définit la région de rejet

$$D \overset{\mathcal{D}}{=} \left\{ x \in \chi;\ \frac{P(x, \theta_0)}{P(x, \theta_1)} \leq c \right\}$$

où c *est choisi tel que* $P_{\theta_0}(D) = \alpha$. *Alors, le test* ρ_D *de région de rejet* D, *est un test*

UPP de niveau α.

Hypothèses multiples

Un test est dit d'hypothèses multiples si

a) Tests unilatères (one-tailed tests) : $\Theta_0 = \{\theta \leq \theta_0\}$, il s'agit de tester $\Theta_0 = \{\theta \leq \theta_0\}$ (resp $\Theta_0 = \{\theta \geq \theta_0\}$ contre $\Theta_1 = \{\theta > \theta_0\}$ (resp $\Theta_1 = \{\theta < \theta_0\}$.

b) Tests bilatères (two-tailed tests) : Soient $\theta_1 < \theta_2$, il s'agit de tester l'une des trois hypothèses suivantes :

- $\theta \in [\theta_1, \theta_2]$ contre $\theta \notin [\theta_1, \theta_2]$.
- $\theta \notin [\theta_1, \theta_2]$ contre $\theta \in [\theta_1, \theta_2]$.
- $\theta = \theta_0$ contre $\theta \neq \theta_0$.

Probabilité critique et règle de décision associée

Définition 1.2.14 : Soit t une réalisation de la statistique de test T. La probabilité critique mesure la probabilité d'obtenir t ou une valeur encore plus espacée de θ_0 si H_0 est vraie. C'est une mesure de l'accord entre l'hypothèse testée et le résultat obtenu. Plus elle est proche de 0, plus forte est la contradiction entre H_0 et le résultat obtenu. La contradiction au sens logique du terme correspond à une valeur nulle de la probabilité critique.

- Cas d'un test unilatéral : la région de rejet est de la forme $R = \{T \geq l\}$. On appelle *probabilité critique* et on note P_c, $P_c(t) = \mathbb{P}(T \geq t \mid \theta = \theta_0)$.
- Cas d'un test bilatéral : la région de rejet est de la forme $R = \{|T| \geq l\}$. On appelle *probabilité critique* $P_c(t) = \mathbb{P}(|T| \geq t \mid \theta = \theta_0)$.

On a dans les deux cas la propriété suivante, qui permet de procéder à la décision d'acceptation ou de rejet au vu de la probabilité critique :

$$P_c(t) < \alpha \Leftrightarrow t \in R,$$

où R est la région de rejet d'un test de niveau α.

1.3 Méthodes non-paramétriques

1.3.1 Estimation non-paramétrique

Supposons que l'on veuille des estimateurs utilisables pour un grand nombre de lois possibles. Le premier problème non trivial est de déterminer ce que l'on veut estimer.

Dans un cas paramétrique ce problème est simple : la loi de l'observation appartient à une classe $\{P_\theta; \theta \in \Theta\}$ et on veut estimer le paramètre θ. Dans un cas *non-paramétrique* on cherche estimer :

1. La loi de l'observation P_X,
2. Les moments de P_X,
3. Les quantiles[2] de P_X.

On se place dans un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ et on considèr $X = X_1, ..., X_n$ un n-échantillon d'une loi P sur $\mathbb{R}$. La loi de l'observation est donc $P_X = P^n$. On note F la fonction de répartition associée.

1.3.2 Modèle d'échantillonnage d'une loi sur $\mathbb{R}$

Définition 1.3.1 : La loi empirique de l'échantillon $X = X_1, ..., X_n$ est la loi sur $\mathbb{R}$ définie par[3] :

$$\hat{P}_n(I) = \frac{1}{n} \# \{X_i \in I\} = \frac{1}{n} \sum_{i=1}^{n} \delta_{Xi}, \quad \forall I \text{ intervalle de } \mathbb{R}. \tag{1.4}$$

La fonction de *répartition empirique* est définie par :

$$\hat{F}_n(x) = \frac{1}{n} \# \{X_i \leq x\} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)} = \hat{P}_n \left( ]-\infty, x] \right), \quad \forall x \in \mathbb{R}. \tag{1.5}$$

$\hat{F}_n$ est la fonction de répartition de la loi empirique $\hat{P}_n$.

[2] Soit P une loi sur $\mathbb{R}$ et $\alpha \in [0,1]$, le quantile de cette loi est le nombre $x_\alpha \in [-\infty, +\infty]$ tel que $P(]-\infty, +\infty]) = \alpha$.

[3] $\# \{X_i \in I\}$ est le nombre de i tels que $X_i \in I$.

Ainsi $\hat{P}_n$ peut aussi être définie comme la loi discrète qui associe la probabilité $1/n$ à chacun des points X_i (cette dernière définition se généralise au cas multidimensionnel). On peut aussi remarquer que $\hat{P}_n$ dépend de l'observation $\{X_i\}$, donc $\hat{P}_n$ est aléatoire.

La loi empirique $\hat{P}_n$ est un estimateur de la loi P. En effet, soit I un intervalle de $\mathbb{R}$, si on veut estimer $P(I) = \mathbb{P}(X_i \in I)$ la probabilité pour que X_i appartienne à I, il suffit de calculer

$$\hat{P}_n(I) = \frac{1}{n} \# \{X_i \in I\} = \frac{1}{n}. \tag{1.6}$$

Soit $f : \mathbb{R} \to \mathbb{R}$, si on veut estimer $Ef = \mathbb{E}f(X_i)$ ($\mathbb{E}$ désigne l'espérance associée à $\mathbb{P}$) on calcule

$$\hat{E}_n f \overset{\mathcal{D}}{=} \frac{1}{n} \sum_{i=1}^{n} f(x_i), \tag{1.7}$$

$\hat{E}_n$ est l'espérance associée à $\hat{P}_n$. Pour estimer les moments $\mathbb{E}(X_i)^k$ on utilise les moments empiriques :

Définition 1.3.2 : les moments empiriques d'un n-échantillon $X = X_1, ..., X_n$ sont par définition les moments de la loi empirique

$$\hat{\mu}_n^k \overset{\mathcal{D}}{=} \frac{X_1^k + ... + X_n^k}{n}, \quad (k \in \mathbb{N}). \tag{1.8}$$

La variance empirique est la variance de la loi empirique

$$\hat{\sigma}_n^2 \overset{\mathcal{D}}{=} \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n^1)^2. \tag{1.9}$$

Pour estimer les quantiles de la loi P on introduit les définitions suivantes : soit $X_{(1)}$ le plus petit élément parmi $\{X_1, ..., X_n\}$, i.e. $X_{(1)} = \min\{X_i\}$, $X_{(2)}$ le suivant, ainsi de suite

Définition 1.3.3 : On définit

$$X_{(1)} \overset{\mathcal{D}}{=} \min_{i=1...n} \{X_i\}, \qquad X_{(j)} \overset{\mathcal{D}}{=} \min_{i=1...n} \{X_i; X_i \geq X_{(j-1)}\}, \tag{1.10}$$

i.e. $X_{(j)}$ est le $j^{ième}$ plus petit élément parmi $\{X_1, ..., X_n\}$. $X_{(j)}$ est appelée la $j^{ième}$ statistique d'ordre de l'échantillon $X = X_1, ..., X_n$.

Définition 1.3.4 : On appelle *quantile empirique* d'ordre p

$$\hat{z}_{n,p} = \inf\left\{x : \hat{F}_n(x) \geq p\right\}, \tag{1.11}$$

et on appelle *estimateur par quantile*

$$Q(\theta) = \psi\left(z_{p_1}(\theta), ..., z_{p_r}(\theta)\right),$$

où ψ fonction continue et $z_{p_i}(\theta)$ quantile d'ordre p_i de $\mathbb{P}_\theta$. Dans ce cas on,

$$\hat{Q}(\theta) = \psi\left(\hat{z}_{n,p_1}(\theta), ..., \hat{z}_{n,p_r}(\theta)\right).$$

Théorème 1.3.1 : *Soit* $0 < p < 1$ *tel que* $z_p(\theta)$, *est l'unique solution de* $F(-x) \leq p \leq F(x)$. *Alors*

$$\hat{z}_{n,p} \to z_p \quad p.s. \tag{1.12}$$

Plus précisément,

$$\mathbb{P}\left(|\hat{z}_{n,p} - z_p| > \varepsilon\right) \leq 2e^{-2n\delta_\varepsilon^2}, \tag{1.13}$$

avec $\delta_\varepsilon = \min\left\{F(z_p + \varepsilon) - p,\ p - f(z - p - \varepsilon)\right\}$.

Pour estimer x_α le quantile d'ordre α de la loi P on utilise le quantile empirique défini par

Définition 1.3.5 : Soit $\alpha \in [0.1]$, le quantile empirique d'ordre α de l'échantillon $X = X_1, ..., X_n$ est défini par :

$$\hat{x}_\alpha \overset{\mathcal{D}}{=} X_{([n\alpha]+1)}, \tag{1.14}$$

où $[n\alpha]$ désigne la partie entière de $n\alpha$.

Les propriétés de ces estimateurs sont rassemblées dans les deux résultats suivants :

Théorème 1.3.2 :

(i) *Pour tout* I *intervalle de* $\mathbb{R}$

$$\mathbb{E}\hat{P}_n(I) = P(I), \quad \mathbb{E}\left(\hat{P}_n(I) - P(I)\right)^2 = \frac{1}{n}P(I)\left(1 - P(I)\right).$$

(ii) *Pour tout* $x \in \mathbb{R}$

$$\mathbb{E}\hat{F}_n(x) = F(x), \quad \mathbb{E}\left(\hat{F}_n(x) - F(x)\right)^2 = \frac{1}{n}F(x)\left(1 - F(x)\right).$$

(iii) *Pour tout* $f : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}\left(\hat{E}_n f\right) = Ef, \quad \mathbb{E}\left(\hat{E}_n f - Ef\right)^2 = \frac{1}{n} Var\left(f(x_i)\right).$$

(iv) *Soit* $p(x)$ *la densité de la loi* P*, alors pour tout* $\alpha \in [0,1]$ *et pour* n *grand*

$$\mathbb{E}\hat{x}_\alpha \simeq x_\alpha, \quad \mathbb{E}\left(\hat{x}_\alpha - x_\alpha\right)^2 \simeq \frac{1}{n} \frac{\alpha\left(1-\alpha\right)}{p(x_\alpha)}.$$

Théorème 1.3.3 : *Pour tout* I *intervalle de* $\mathbb{R}$, $x \in \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ *et* $\alpha \in [0,1]$

$$\begin{aligned} &\hat{P}_n\left(I\right) \underset{n\to\infty}{\to} P(I), \quad \mathbb{P}-p.s, \quad \hat{F}_n\left(x\right) \underset{n\to\infty}{\to} F(x), \quad p.s, \\ &\hat{E}_n f \underset{n\to\infty}{\to} Ef, \quad \mathbb{P}-p.s, \quad \hat{x}_\alpha \underset{n\to\infty}{\to} x_\alpha; \quad p.s. \end{aligned} \tag{1.15}$$

Théorème 1.3.4 : (Théorème de Glivenko-Cantelli)

$$\sup_{x\in\mathbb{R}} \left|F_n\left(x\right) - F(x)\right| \underset{n\to\infty}{\to} 0, \quad p.s. \tag{1.16}$$

1.3.3 Tests non-paramétriques

Dans un cadre paramétrique $(\Omega, \mathcal{F}, \mathbb{P}, \{P_\theta; \theta \in \Theta\})$ les problèmes de test portent toutjours sur le paramètre θ. Il existe pourtant des questions que l'on peut se poser sur un échantillon qui ne demandent pas de spécifier un modèle paramétrique, comme par exemple

(1) L'échantillon $X_1, ..., X_n$ provient-il d'une loi P donnée ?

(2) Les échantillons $X_1, ..., X_n$ et $Y_1, ..., Y_n$ proviennent-ils d'une même loi ?

(3) Les échantillons $X_1, ..., X_n$ et $Y_1, ..., Y_n$ sont-ils indépendants ?

Dans les paragraphes suivants nous allons présenter des outils statistiques permettant de répondre à ces questions sans avoir à préciser un modèle paramétrique.

Test d'adéquation

On dispose d'un n-échantillon $X = X_1, ..., X_n$, i.e. les X_i sont des variables i.i.d. définies sur un espace $(\Omega, \mathcal{F}, \mathbb{P})$. On se donne une loi de probabilité P sur $\mathbb{R}$ et on s'intéresse au problème de test de l'hypothèse

$$H : \quad "X_1, ..., X_n \text{ est un échantillon de la loi } P\ ".$$

Adéquation a une loi donnée

Il s'agit du cas où $\mathcal{F}_0 = \{F_0\}$, F_0 étant une f.r. donnée. On veut savoir si l'échantillon observé peut être considéré comme issu de cette loi, d'où le problème de test suivant :

$$\begin{cases} H_0 : & F = F_0 \\ H_1 : & F \neq F_0. \end{cases}$$

Test de Kolmogorov-Smirnov

Il est basé sur la fonction de répartition empirique F_n et sa distance $d_k(F_n, F_0)$ à la f.r.F_0. Pour tester l'hypothèse H, une méthode simple consiste à comparer F_0 et F_n.

Définition 1.3.6 : Le test d'adéquation de *Kolmogorov-Smirnov* consiste à accepter l'hypothèse H si

$$\max_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \leq c,$$

pour un seuil c convenablement choisi. Posons

$$d = \max_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

On montre que $K_n = \sqrt{n} d_k(F_n, F_0)$ admet une loi limite de fonction de répartition H définie par :

$$H(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2} \mathbf{1}_{\mathbb{R}_+^*}(x).$$

Pour un test unilatéral avec une alternative de la forme :

$$H_1^{'}: \quad F < F_0 \qquad \text{où} \qquad H_1^{''}: \quad F > F_0,$$

on utilise les statistiques :

$$K_n^+ = \sqrt{n} d^+ \left(F_n, F_0\right) \qquad \text{où} \qquad K_n^- = \sqrt{n} d^+ \left(F_0, F_n\right),$$

la distance d^+ entre deux fonctions f et g étant définie par :

$$d^+ \left(f, g\right) = \sup_{x \in \mathbb{R}} \left[f\left(x\right) - g(x)\right].$$

Remarque 1.3.1 : Pour le calcul pratique de ces statistique, on utilise les formules :

$$d_k \left(F_n, F_0\right) = \max_{1 \leq i \leq n} \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right|, \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\}. \tag{1.17}$$

Test de Cramér-von Mises

On utilise la statistique :

$$W_n^2 = n \int_{\mathbb{R}} \left[F_n\left(x\right) - F_0\left(x\right)\right]^2 w\left[F_0\left(x\right)\right] dF_0\left(x\right), \tag{1.18}$$

où w est une fonction de poids positive définie sur $[0,1]$ choisie selon certains critères relatifs à la puissance du test. Si F_0 est continue, la loi de W_n^2 est indépendante de F_0 car :

$$\begin{aligned} W_n^2 &= n \int_{\mathbb{R}} \left[\frac{1}{n} \sum_{j=1}^n \mathbf{1}_{\mathbb{R}}\left(x - X_j\right) - F_0\left(x\right)\right]^2 w\left[F_0\left(x\right)\right] dF_0\left(x\right) \\ &= n \int_0^1 \left[\frac{1}{n} \sum_{j=1}^n \mathbf{1}_{\mathbb{R}}\left(t - F_0\left(X_j\right)\right) - t\right]^2 w\left(t\right) dt \quad p.s. \end{aligned} \tag{1.19}$$

et $F_0\left(X_j\right)$ suit une loi $\mathcal{U}\left(0,1\right)$. Ce test est convergent et W_n^2 admet une loi limite.

Pour $w\left(t\right) = 1$, on obtient :

$$W_n^2 = \sum_{i=1}^n \left[F_0\left(X_i\right) - \frac{2i-1}{2n}\right]^2 + \frac{1}{12n}.$$

Chapitre 2

PROCEDURE DU BOOTSTRAP

2.1 Principe du Bootstrap

Le premier cadre du bootstrap est celui de la statistique classique : l'échantillon est constitué de n variables aléatoires $(X_1, \ldots, X_n)$ i.i.d de fonction de répartition F est supposée inconnue. On cherche à estimer la loi de $T_n = T(X_1, \ldots, X_n)$, T_n étant un estimateur d'une grandeur θ telle que :

$$\theta := T(F);$$

qui permet d'estimer la variance, le biais et les intervalles de confiance.... L'idée pour cela est de substituer F par la fonction de répartition empirique F_n obtenue à partir de l'échantillon :

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X_k \leq x}, \quad x \in \mathbb{R}$$

Définition 2.1.1 : On appelle estimateur *plug-in* d'un paramètre de F, l'estimateur obtenu en remplaçant la loi F par la loi empirique :

$$\hat{\theta} = t\left(\hat{F}\right).$$

Ce principe de substitution (ou plug-in) est justifié asymptotiquement puisqu'on a presque sûrement, d'après le théorème *Glivenko-Cantelli.*

On étudie les propriétés de T_n par l'intermédiaire de $T_n^* = T(X_1^*, ..., X_n^*)$ où les X_k^* sont des variables i.i.d de fonction de répartition F_n.

Soit $X^* = (X_1^*, ..., X_n^*)$ échantillons bootstrap. On note indifféremment :

$$P\left(X_j^* = X_i \mid X\right) = 1/n, \quad 1 \leq i,\ j \leq n,$$

ou

$$P\left(X_j^* = X_i \mid F_n\right) = 1/n, \quad 1 \leq i,\ j \leq n,$$

car dès qu'on connaît X, on peut déduire F_n et réciproquement.

L'espérance de T_n :

$$E\left[T_n\right] = \int ... \int T\left(x_1, ..., x_n\right) dF\left(x_1\right) ... dF\left(x_n\right),$$

est ainsi estimé par :

$$E\left[T_n^*\right] = \int ... \int T\left(x_1, ..., x_n\right) dF_n\left(x_1\right) ... dF_n\left(x_n\right).$$

Cela s'écrit également :

$$\begin{aligned} E\left[T_n^*\right] &= \frac{1}{n^n} \int ... \int T\left(x_1, ..., x_n\right) \sum_{l_1=1}^{n} \delta_{X_{l_1}}\left(x_1\right) ... \sum_{l_n=1}^{n} \delta_{X_{l_n}}\left(x_n\right) dx_1 ... dx_n \\ &= \frac{1}{n^n} \sum_{l_1=1}^{n} ... \sum_{l_n=1}^{n} T\left(X_{l_1}, ..., X_{l_n}\right). \end{aligned}$$

Il s'agit tout simplement de la moyenne de T sur l'ensemble des échantillons de taille n tirés avec remise à partir de l'échantillon initial $(X_1, ..., X_n)$. Sur ce même principe, théoriquement on peut fournir une estimation de la loi de T_n.

La formule précédente montre qu'il est nécessaire en général de calculer n^n termes (C_{2n-1}^n en réalité si T est symétrique), ce qui est en pratique impossible dès que n dépasse la dizaine. Pour s'en sortir, on a recours à des simulations dites de *Monte-Carlo*. On effectue un grand nombre de tirages avec remise à partir de l'échantillon initial, et on calcule leur estimateurs. Au prix de cette double approximation (estimation de F par F_n, puis approximation du calcul analytique par simulation), on dispose alors d'un estimateur pour l'ensemble des caractéristiques de T_n.

Par exemple, on estime $E[T_n]$ par la moyenne empirique sur l'ensemble des tirages :

$$E_B\left[T_n^*\right] = \frac{1}{B} \sum_{k=1}^{B} T_n^{*(k)},$$

où B est le nombre de tirages avec remise effectués et $T_n^{*(k)}$ la statistique obtenue au k-ième tirage. De même, la fonction de répartition G de T_n est estimée par :

$$G_B^*\left(x\right) = \frac{1}{B} \sum_{k=1}^{B} \mathbf{1}_{T_n^{*(k)} \leq x}.$$

L'avantage de cette méthode est donc qu'il est possible, moyennant un jeu d'hypothèses très faibles (et en particulier, aucune paramétrisation de la loi des X_k), de fournir une estimation convergente de la loi de T_n. Cela nous permet d'obtenir des intervalles de confiance sur θ sans postulat de normalité.

L'idée du Bootstrap

1. **Problèmes paramétriques et non paramétriques :**

   La loi F_n associée à l'échantillon peut être, comme dans l'exemple ci-dessus de l'estimation d'une moyenne, la loi empirique. C'est le cas lorsqu'on a affaire à un problème non paramétrique. Mais la loi F_n peut être une loi issue d'un modèle paramétrique : les paramètres sont estimés en employant le modèle, en principe par le maximum de vraisemblance, et alors F_n est la loi appartenant au modèle, dont les paramètres sont ceux estimés à partir de l'échantillon.

2. **Paramètres et fonctionnelles :**

   On a vu que pour estimer une fonctionnelle $T(F)$ de la distribution inconnue F on remplaçait F, dans le cas non paramétrique, par la loi empirique F_n associée à l'échantillon. Mais si par exemple on veut estimer un paramètre comme le centre de symétrie d'une loi symétrique, ce centre de symétrie peut correspondre à plusieurs fonctionnelles différentes : la moyenne, la médiane de F et beaucoup d'autres encore ; par exemple les moyennes α tronquées. Ces dernières sont obtenues en prenant la moyenne des observations qui restent lorsqu'on a ôté les plus grandes et les plus petites, en proportion α. Il est donc nécessaire de dire précisément quelle est la fonctionnelle que l'on veut estimer.

3. **Approximation d'une statistique bootstrap :**

   Il faut bien distinguer deux éléments différents dans les méthodes bootstrap :

- Le principe consiste à remplacer la loi initiale inconnue par une loi associée à l'échantillon observé, et toutes les lois dérivées nécessaires. Les paramètres d'intérêt sont ainsi remplacés par une statistique "bootstrap", en principe complètement calculables.

- Le calcul proprement dit de la statistique bootstrap : bien que la statistique bootstrap soit en principe complètement calculable, souvent son calcul effectif serait trop long. Il s'agit en général d'espérances fondées sur la loi F_n et des dérivées de cette loi. Aussi, *Efron* a-t-il suggéré de le faire par une méthode de type *Monte-Carlo* qui consiste à rééchantillonner à partir de l'échantillon initial, obtenant des échantillons de même taille n. Si le nombre des rééchantillonnages est assez grand, on aura une bonne approximation de l'espérance cherchée à cause de la loi des grands nombres.

Maintenant nous récrivons la procédure de bootstrap non paramétrique ci-dessus dans les étapes suivantes. Qui a référée à *Efron et Tibshirani (1993)* [17] pour des discussions détaillées. Considérer le cas où un échantillon aléatoire de la taille n est tiré d'une distribution non spécifiée de probabilité F.

Algorithme du bootstrap non-paramétrique

L'algorithme du bootstrap peut être résumé comme suit :

1. Construire une distribution empirique de probabilité F_n, de l'échantillon en plaçant une probabilité de $1/n$ à chaque point, $x_1, \ldots, x_n$ de l'échantillon. C'est la fonction de distribution empirique de l'échantillon, qui est l'estimation *non paramétrique* du maximum de vraisemblance de la distribution de population.

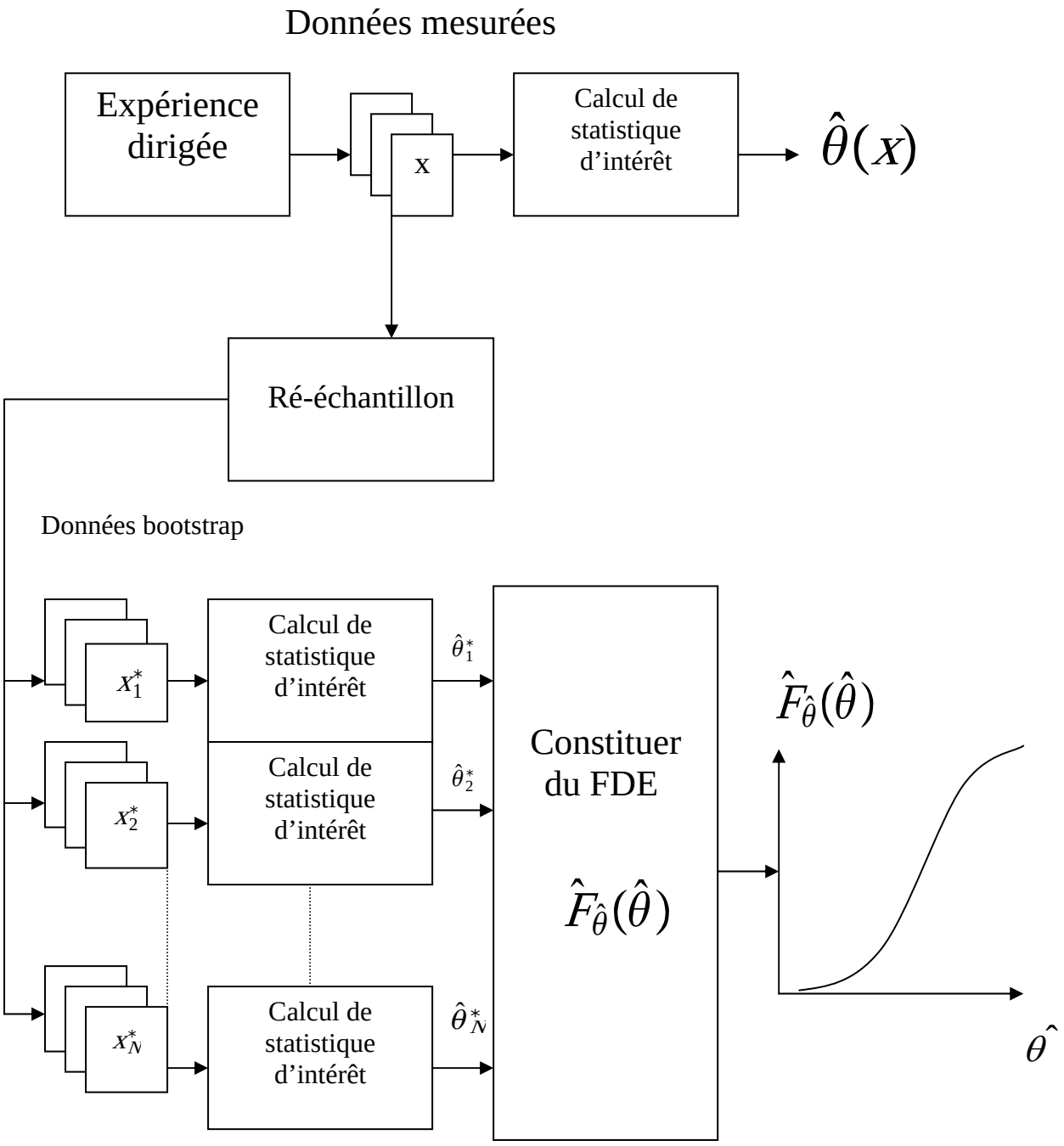
2. De la fonction de distribution empirique F_n, tirer un échantillon aléatoire de taille n avec remplaçement.

3. Calculer la statistique d'intérêt T_n pour ce rééchantillon, en replaçant T_n^*.

4. Répéter les étapes 2 et 3 B fois, où B est un nombre grand, afin de créer B rééchantillon. La taille particulière de B dépendent des testes à courir sur les données. Typiquement, B est au moins égal à 1000 où une estimation d'intervalle de confiance autour de T_n est exigée.

5. Construire l'histogramme de fréquence relative du B nombre de T_n^* en plaçant une probabilité de $1/B$ à chaque point, T_n^{*1}, T_n^{*2}, ..., T_n^{*B}, la distribution obtenue est l'estimation bootstrapée de la distribution d'échantillon T_n. Cette distribution peut être maintenant utilisée pour faire des inférences sur le paramètre θ, qui doit être estimé par T_n.

Procédure du bootstrap



2.2 Validité du Bootstrap

Nous étudions maintenant la validité du bootstrap dans le cas générale et son erreur d'approximation liée par le théorème de centrale limite, utilisés pour arriver à une bonne approximation de distribution bootstrapée.

2.2.1 Théorie asymptotique du bootstrap

Soient $\{X_i : i = 1, \ldots, n\}$ un échantillon des v.a de taille n et d'une distribution de probabilité de fonction de distribution cumulative (FDC) F_0.

Définition 2.2.1 : Si $\Im$ est une famille de dimension finie indexée par le paramètre θ et θ_0 valeur de population , alors $F_0 \in \Im$. On note $F_0(x, \theta_0)$ pour $P(X \leq x)$ et $F(x, \theta)$ pour un membre général de la famille paramétrique.

Définition 2.2.2 : Soit $T_n = T_n(X_1, \ldots, X_n)$ une statistique. Notons :

$$H_n(\tau, F_0) \equiv P(T_n \leq x),$$

appelé FDC exact, d'échantillon finie de T_n, $H_n(., F)$ est la FDC exacte de T_n quand les données sont échantillonnées par la distribution de F. On utilise, $H_n(\tau, F)$ est une fonction différente de τ pour des distributions différentes de F.

Remarques 2.2.1 : On approxime H_n si $H_n(., F)$ ne dépend pas de F, dans ce cas T_n est centré. Le bootstrap est une méthode pour l'estimation de $H_n(., F_0)$ ou les particularités de $H_n(., F_0)$ comme son quantiles quand F_0 est inconnu.

Définition 2.2.3 : Une statistique *asymptotiquement centrale*, signifie que sa distribution asymptotique est indépendante des paramètres de population inconnus. Notons $H_\infty(., F_0)$ la distribution asymptotique de T_n et $H_\infty(., F)$ la FDC asymptotique de T_n quand les données sont échantillonnées dont FDC est F. Si T_n est asymptotiquement central, on a :

$$H_\infty(., F) \equiv H_\infty(.);$$

indépendante de F. Donc, si n est suffisamment grand, $H_n(., F_0)$ peut être estimé par $H_\infty(.)$ sans avoir F_0.

Définition 2.2.4 : Si un estimateur est *asymptotiquement normalement distribué*, alors sa distribution asymptotique dépend de deux paramètres inconnus, la moyenne et la variance, qui peut être souvent estimées sans difficulté, la distribution

normal avec estimées ses paramètres sont utilisés pour rapprocher l'inconnu $H_n(.,F_0)$ si n est suffisamment grand.

Définition 2.2.5 : Le bootstrap fournit une *approximation alternative* à la distribution d'échantillon de $T_n(X_1,\ldots,X_n)$ statistique finie. Tandis que des approximations asymptotiques de premier ordre remplaçent la fonction inconnu de distribution H_n avec la fonction connue $H_\infty(.)$, le bootstrap remplace la fonction de distribution F_0 inconnu avec un estimateur connu. Notons F_n l' estimateur de F_0 ; alors il existe deux possibilités du choix de F_n par :

1) La fonction de distribution empirique (FDE) des données :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(X_i \leq x\right), \quad -\infty < x < +\infty.$$

Il suit du théorème *Glivenko-Cantelli* que : $F_n\left(x\right) \overset{p.s.}{\to} F_0\left(x\right)$, quand $n \to \infty$

2) Un estimateur paramétrique de F_0; Supposons que $F_0\left(.\right) = F(.,\theta_0)$ dans le cas de dimension finie de θ_0 est estimé successivement par θ_n.

Si $F(.,\theta)$ est une fonction continue de θ au voisinage de θ_0, alors

$$F(x,\theta_n) \to F(x,\theta_0), \quad \text{quand } n \to \infty$$

pour tout x, telle que

$$\theta_n \overset{p.s.}{\to} \theta_0.$$

Indépendant du choix de F_n, l'estimateur de bootstrap de $H_n(.,F_0)$ est $H_n(.,F_n)$.

On peut pas utiliser, $H_n(.,F_n)$ dans l'estimation analytique, cependant elle est estimée par la simulation de Monte-Carlo où les échantillons aléatoires représentent F_n.

Procédure de Monte-Carlo pour l'estimation de Bootstrap de $\mathbf{H_n(\tau,F_0)}$

Pas 1 : construire un échantillon bootstrap $\{X_i^* : i = 1 : n\}$ de taille n, en échantillonnant la distribution correspondante à F_n aléatoirement.

Si F_n est l'*FDE* d'ensemble des données d'estimation, donc l'échantillon bootstrap peut être obtenu en échantillonnant les données d'estimation aléatoirement avec remplaçement.

Pas 2 : Calculer $T_n^* \equiv T_n(X_1^*,\ldots,X_n^*)$.

Pas 3 : Utiliser les résultats répétitions des *pas 1* et *2* pour calculer la probabilité empirique de l'événement $T_n^* \leq \tau$.

Brown (1999) [5] et *Hall (1992)* [26] discutent les méthodes de simulation qui prennent avantage de techniques pour réduction de variation d'échantillons dans la simulation de *Monte-Carlo*.

Remarque 2.2.3 : Puisque F_n et F_0 sont des fonctions différentes, $H_n(., F_n)$ et $H_n(., F_0)$ sont aussi des fonctions différentes à moins que T_n n'est pas central. Donc, l'estimateur de bootstrap $H_n(., F_n)$ est seulement une approximation exacte de l'échantillon fini T_n de $H_n(., F_0)$.

2.2.2 Théorème centrale limite

Théorème 2.2.1 (*Lindeberg -Levy*) **:** *Soit* $\{X_i\}$ *une suite de variables aléatoires i.i.d, de moyenne* μ *et de variance finie* σ^2. *Alors* :

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

Le théorème précédent peut être généralisé aux variables aléatoires indépendantes, qui ne sont pas nécessairement identiquement distribuées. Cela comme suit :

Théorème 2.2.2 (*Lindeberg -Feller*) **:** *Soit* $\{X_i\}$ *une suite de variables aléatoires indépendantes, avec moyennes* $\{\mu_i\}$, *variances* $\{\sigma_i^2\}$ *finies et les fonctions de distribution* $\{F_i\}$. *On suppose que* $B_n^2 = \sum_{i=1}^{n} \sigma_i^2$ *satisfait*

$$\frac{\sigma_n^2}{B_n^2} \to 0, B_n \to \infty, \quad \text{quand } n \to \infty.$$

Alors, $n^{-1}\sum_{i=1}^{n} X_i$ *suit une loi normale* $\mathcal{N}\left(n^{-1}\sum_{i=1}^{n}\mu_i, n^{-2}B_n^2\right)$ *si et seulement si la condition de Lindeberg est satisfaite*

$$\lim_{n\to\infty} B_n^{-2}\sum_{i=1}^{n}\int_{|t-\mu_i|<\varepsilon B_n} (t-\mu_i)^2\, dF_i(t) = 0, \quad \forall \varepsilon > 0.$$

Dans les théorèmes précédents on associe la normalité asymptotique pour un ordre des sommes $\sum_{i=1}^{n} X_i$ constituées par un ordre unique $X_1, ..., X_n$ de v.a. Plus généralement, on considère un double tableau de v.a.

$$\begin{bmatrix} X_{11,} & X_{12,} & . & . & . & X_{2K_1,} \\ X_{21,} & X_{22,} & . & . & . & X_{2K_{2,}} \\ . & . & & . & & . \\ . & . & & . & & . \\ X_{n1,} & X_{n1,} & . & . & . & X_{2K_n,} \\ . & . & & . & & . \\ . & . & & . & & . \\ . & . & & . & & . \end{bmatrix}$$

On suppose K_n des variables aléatoires qui tends vers l'infini pour chaque $n > 1$. Dans le cas ou $K_n = n$ le tableau est "triangulaire".

Notons : F_{nj} la fonction de distribution du X_{nj}. En outre

$$\mu_{nj} = EX_{nj},$$

$$A_n = E\sum_{j=1}^{K_n} X_{nj} = \sum_{j=1}^{K_n} \mu_{nj},$$

$$B_n^2 = Var\left(\sum_{j=1}^{K_n} X_{nj}\right).$$

Théorème 2.2.3 (*Lindeberg-Feller*) **:** *Soit* $\{X_{nj} : 1 \leq j \leq K_n; n = 1, 2, ...\}$ *est un double tableau de variables aléatoires indépendantes dans les colonne alors la condition est "asymptotique uniforme négligeable".*

$$\max_{1\leq j\leq K_n} P\left(\left|X_{nj} - \mu_{nj}\right| > \tau B_n\right) \to 0, \quad n \to \infty, \quad \text{pour } \tau > 0,$$

avec $\sum_{j=1}^{K_n} X_{nj}$ la condition asymptotique normale de $\mathrm{AN}(A_n, B_n^2)$; dite l'ensemble prise si et seulement si la condition de *Lindeberg*

$$B_n^{-2}\sum_{j=1}^{n}\int_{|t-\mu_i|>\varepsilon B_n} (t-\mu_i)^2\, dF_i(t) \to 0, \quad n \to \infty, \quad \text{pour } \tau > 0,$$

est satisfaite.

On remarque que l'indépendance est assumée seulement dans les colonnes, qui sont déjà arbitrairement dépendantes.

Corollaire 2.2.1 : Supposons que pour $\upsilon > 2$,

$$\sum_{j=1}^{n} E\left|X_{nj} - \mu_{nj}\right|^{\upsilon} = 0\left(B_n^{\upsilon}\right),\ n \to \infty.$$

Alors

$$\sum_{j=1}^{K_n} X_{nj} \quad \text{est} \quad AN\left(A_n, B_n^2\right).$$

Propriétés asymptotiques du bootstrap

Soit un rééchantillonnage de taille $N = n$ et soit $T^* = T(\mathcal{E}^*, F_n)$ où T vaut successivement

$$T_1 = \bar{X}_n - \mathbb{E}_P(X), \qquad T_2 = \frac{\bar{X}_n - \mathbb{E}_P(X)}{\sigma_p(X)}, \quad T_3 = F_n^{-1}(x) - F^{-1}(x).$$

Théorème 2.2.4 :

(i) *Si* $E(|X|^2) < \infty$, *alors*

$$A_n = \sup_{t\in\mathbb{R}} |\mathbb{P}\left(\sqrt{n}T_1 \leq t\right) - \mathbb{P}\left(\sqrt{n}T_1^* \leq t\right)| \overset{p.s}{\to} 0 \quad (n \to \infty) \text{ avec}$$

$$T_1^* = \bar{X}_n^* - \bar{X}_n.$$

La vitesse de convergence est donnée par :

$$\limsup_{n\to\infty} \frac{\sqrt{n}}{\sqrt{\log\left(\log n\right)}} A_n = c_1,$$

c_1 *est constante.*

(ii) *Si* $E(|X|^3) < \infty$, *alors*

$$B_n = \sup_{t\in\mathbb{R}} |\mathbb{P}\left(\sqrt{n}T_2 \leq t\right) - \mathbb{P}\left(\sqrt{n}T_2^* \leq t\right)| \overset{p.s.}{\to} 0 \quad (n \to \infty)\,; \text{ avec}$$

$$T_2^* = \frac{\bar{X}_n^* - \bar{X}_n}{S_n'}.$$

L*a vitesse de convergence de* B_n *est supérieure à* $1/\sqrt{n}$, *au sens où*

$$\limsup_{n\to\infty} \sqrt{n}B_n \leq c_2,$$

avec c_2 *constante.*

(iii) *Si* F'' *existe au voisinage de* $F^{-1}(t)$ *et si* $F'(F^{-1}(t)) > 0$, *alors*

$$C_n = \sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}T_3 \leq t\right) - \mathbb{P}\left(\sqrt{n}T_3^* \leq t\right)\right| \overset{p.s.}{\to} 0 \quad (n \to \infty);$$

avec

$$\lim_{n\to\infty} \sup \frac{\sqrt[4]{n}}{\sqrt{\log(\log n)}} C_n = c_3(x, F).$$

Théorème 2.2.5 : *Soit* $\mathcal{E} = (X_i)_{i=1,\ldots,N}$ *une suite de variable i.i.d. de loi* $\mathbb{P}$, *d'espérance* m *et de variance* σ^2. *Soit* $\mathcal{E}^* = (X_i^*)_{i=1,\ldots,N}$ *un échantillon bootstrap extrait de* $\mathcal{E}$. *Conditionnellement à* $\mathcal{E}$, *pour* $N \to \infty$ *et* $n \to \infty$,

(i) $\sqrt{N}\left(\bar{X}_n^* - \bar{X}_n\right) \overset{\mathcal{L}}{\to} \mathcal{N}(0, \sigma)$;

(ii) $\mathbb{P}\left(|S_N^* - \sigma| > \epsilon\right) \overset{p.s.}{\to} 0 \quad (n \to \infty)$ *avec* $S_N^* = \sqrt{\frac{1}{N}\sum_{i=1}^N \left(X_i - \bar{X}_n^*\right)^2}$.

Théorème 2.2.6 : *On suppose que la fonction de répartition* F *d'une v.a.* X *possède une unique médiane* Md *et une dérivée* f *positive et continue sur un voisinage de* Md. *Alors* :

$$\sqrt{n}\left(Md_n^* - Md_n\right) \overset{\mathcal{L}}{\to} \mathcal{N}\left(0, \frac{1}{2f(Md)}\right), \quad (n \to \infty).$$

où Md_n la médiane empirique et Md_n^* la médiane de l'échantillon bootstrap de taille n.

2.3 Consistance du bootstrap

Supposons que F_n est un estimateur consistant de F_0. Cela définit que pour tout x dans le support de X,

$$F_n(x) \overset{p.s.}{\to} F_0(x), \quad \text{quand } n \to \infty$$

Si F_0 est une fonction continue, donc par suite du théorème de *Polya* on a :

$$F_n \overset{p.p.s}{\to} F_0.$$

Ainsi, F_n et F_0 sont uniformément près si n est grand ; de plus on considère $H_n(\tau, F)$ comme une fonctionnelle de F continue dans un sens approprié, on peut tendre $H_n(\tau, F_n)$ vers $H_n(\tau, F_0)$ quand n est grand.

D'autre part, si n est grand, $H_n(., F_0)$ est uniformément près de la distribution asymptotique $H_\infty(., F_0)$, si $H_\infty(., F_0)$ est continu. Alors on suggère que l'estimateur de bootstrap $H_n(., F_0)$ et la distribution asymptotique $H_\infty(., F_0)$ doivent être uniformément près si n réalise des conditions.

Définition 2.3.1 : Supposons P_n la distribution de probabilité jointe d'échantillon $\{X_i : i = 1, \ldots, n\}$. L'estimateur de bootstrap $H_n(., F_n)$ est consistant si pour tout $\varepsilon > 0$ et $F_0 \in \Im$, on a :

$$\lim_{n\to\infty} P_n \left[\sup_{\tau} |H_n(\tau, F_n) - H_\infty(\tau, F_0)| > \varepsilon \right] = 0.$$

Le théorème fondamental pour la compréhension du bootstrap de *Beran* et *Ducharme (1991)* [3] donne des conditions sous lequelles l'estimateur de bootstrap est consistant.

Théorème 2.3.1 [3] **:** *Soit ρ une distance sur l'espace $\Im$. $H_n(., F_n)$ est consistant si pour tout $\varepsilon > 0$ et $F_0 \in \Im$*

(i) $\lim_{n\to\infty} P_n \left[\rho \left(F_n, F_0 \right) > \varepsilon \right] = 0$;

(ii) *$H_\infty(\tau, F)$ est une fonction continue de τ pour tout $F \in \Im$;*

(iii) *Pour tout τ et toute suite $\{G_n\} \in \Im$ tel que $\lim_{n\to\infty} \rho \left(G_n, F_0 \right) = 0$,*

$$H_\infty(\tau, G_n) \to H_\infty(\tau, F_0).$$

Nous allons données un exemple dont les conditions du *Théorème (2.3.1)* sont satisfaites :

Exemple 2.3.1 (distribution de la moyenne d'échantillon**) :** Supposons que $\bar{X}$ est la moyenne d'échantillon aléatoire $\{X_i : i = 1, \ldots, n\}$. On définit

$$T_n = n^{1/2} \left(\bar{X} - \mu \right),$$

où $\mu = E(X)$. On a :

$$H_n(\tau, F_0) = P_n \left[n^{1/2} \left(\bar{X} - \mu \right) \leq \tau \right].$$

Pour estimer $H_n(\tau, F_0)$, le bootstrap de T_n est :

$$T_n^* = n^{1/2}\left(\bar{X^*} - \bar{X}\right),$$

où X^* désigne la moyenne d'échantillon aléatoire de taille n de F_n, l'estimateur bootstrap de $H_n(\tau, F_0)$ est

$$H_n(\tau, F_n) = P_n^*\left[n^{1/2}\left(\bar{X^*} - \bar{X}\right) \leq \tau\right],$$

où P_n^* est la distribution de probabilité citée par le processus d'échantillons bootstrap. $H_n(\tau, F_n)$ satisfait les conditions du *théorème (2.3.1)* donc le théorème de *Glivenko-Cantelli est consistant,* et la loi forte des grands nombres satisfait que la condition (i) du théorème *(2.3.1).*

Le théorème de *Lindeberg-Levy* implique que T_n est asymptotiquement normalement distribué, l'FDC normale est continue, donc les conditions (ii) et (iii) sont réalisées.■

Le théorème de *Mammen (1992)* [35] donne des conditions nécessaires et suffisantes pour réaliser la consistance du bootstrap fonctionnelles des distributions linéaires de F_0 quand F_n est l*'FDE.*

Théorème 2.3.2 [35] **:** *Soit* $\{X_i : i = 1, \ldots, n\}$ *un échantillon aléatoire de population. Pour une suite des fonctions* g_n*, et des suites de nombres* t_n *et* σ_n*, on définit*

$$\bar{g_n} = n^{-1}\sum_{i=1}^{n} g_n(X_i) \quad et \quad T_n = \left(\bar{g_n} - t_n\right)/\sigma_n.$$

Pour l'échantillon bootstrap $\{X_i^* : i = 1, \ldots, n\}$*, on a :*

$$\bar{g_n^*} = n^{-1}\sum_{i=1}^{n} g_n(X_i^*), \quad T_n^* = \left(\bar{g_n^*} - \bar{g_n}\right)/\sigma_n,$$

et

$$H_n(\tau) = P(T_n \leq \tau), \quad H_n^*(\tau) = P^*(T_n^* \leq \tau),$$

où P^* *est la distribution de probabilité citée par l'échantillons bootstrap. Alors* $H_n^*(.)$ *estime consistant* H_n *si et seulement si* $T_n \xrightarrow{\mathcal{D}} N(0,1)$.■

En effet, si $E[g_n(X)]$ et $Var[g_n(X)]$ existent pour tout n, la condition de normalité asymptotique du *théorème (2.3.2)* est donnée par :

$$t_n = E\left(\bar{g_n}\right), \sigma_n^2 = Var\left(\bar{g_n}\right) \quad \text{ou} \quad \sigma_n^2 = n^{-2}\sum_{i=1}^{n}\left[g_n\left(X_i\right) - \bar{g_n}\right]^2.$$

L'application directe du *théorème (2.3.2)* donne la consistance de l'estimateur du bootstrap de distribution du concentré de la moyenne d'échantillon normalisé dans l'exemple.

Le bootstrap n'a pas besoin d'être consistant si les conditions du *théorème (2.3.1)* ne sont pas satisfaites, il est inconsistant si la condition de normalité asymptotique du théorème *(2.3.2)* n'est pas satisfaite. Particulièrement le bootstrap est inconsistant si F_0 est un point de discontinuité de la fonction de distribution asymptotique $H_\infty(\tau, .)$, ou un point de super efficacité.

Les exemples suivants illustrent des conditions sous lesquelles le bootstrap est inconsistant. *Donald* et *Paarsch (1996)* [13] décrivent les applications économétriques qui sont des particulièrement semblable à certains exemples, quoique la consistance du bootstrap dans applications n'est pas examinées.

Exemple 2.3.2 (Distributions a queue lourde**) :** Soient F_0 la distribution de Cauchy standard , $\{X_i\}$ l'ensemble d'échantillons de cette distribution et $T_n = \bar{X}$ sa moyenne. Alors T_n est la distribution de Cauchy standard. L'échantillon bootstrap de T_n est

$$T_n^* = \overline{X}^* - m_n,$$

où X^* est la moyenne d'échantillon bootstrap et m_n le médiane des données. La condition de normalité asymptotique du *Théorème (2.3.2)* n'est pas satisfaite et l'estimateur de bootstrap de la distribution de T_n est inconsistant.

Athreya (1987) [2] et *Hall (1990)* [24] fournissent la nouvelle discussion du comportement du bootstrap avec distributions a queue lourde.■

L'exemple suivant est dû à *Bickel et Freedman (1981)* [4].

Exemple 2.3.3 (Distribution du maximum d'un échantillon) **:**

Soit $\{X_i : i = 1, \ldots, n\}$ un échantillon aléatoire d'une distribution avec FDC absolument continue F_0 et à support $[0, \theta_0]$ et $\theta_n = Max(X_1, \ldots, X_n)$, on définit $T_n = n\left(\theta_n - \theta_0\right)$ avec F_n l'FDE d'échantillon . Le bootstrap de T_n est

$$T_n^* := n\left(\theta_n^* - \theta_n\right),$$

où θ_n^* est le maximum d'échantillon bootstrap $\{X_i^*\}$. Le bootstrap ne fait pas une estimation

$$H_n(-\tau, F_0) = P_n(T_n \leq -\tau) \quad (\tau \geq 0).$$

On observons que

$$P_n^*\left(T_n^* = 0\right) = 1 - (1 - 1/n)^n \to 1 - e^{-1}, \quad \text{quand } n \to \infty.$$

On montre facilement que la distribution asymptotique de T_n est

$$H_\infty\left(-\tau, F_0\right) = 1 - \exp\left[-\tau f\left(\theta_0\right)\right],$$

où $f(x) = dF(x)/dx$ est la fonction de densité de probabilité de X. Donc

$$P(T_n = 0) \to 0,$$

l'estimateur de bootstrap de $H_n\left(., F_0\right)$ est inconsistant.■

2.3.1 Distribution de la Statistique

Soit $\{X_i : i = 1, \ldots, n\}$ un échantillon aléatoire d'une distribution de probabilité dont *FDC* est F_0, et $T_n = T_n(X_1, ..., X_n)$ est une statistique. On note

$$H_n(\tau, F_0) = P(T_n \leq \tau),$$

appelée l'échantillon exacte fini d'une *FDC* de T_n puisque $H_n(\tau, F_0)$ ne peut pas être calculée analytiquement à moins que T_n n'est pas central.

L'objectif c'est d'obtenir une approximation à $H_n(\tau, F_0)$ qui est applicable quand $T_n(X_1..., X_n)$ n'est pas central.

Pour obtenir des approximations utiles à $H_n(\tau, F_0)$, il est nécessaire de faire certaines suppositions sur la forme de la fonction $T_n(X_1..., X_n)$.

Supposons que $T_n(X_1..., X_n)$ est une fonction lisse d'échantillons de moments de X. On a :

$$T_n = n^{1/2}\left[G\left(\bar{Z}_1, ..., \bar{Z}_J\right) - G\left(\mu_{Z_1}, ..., \mu_{Z_J}\right)\right],$$

où la fonction estimée de scalaire G est lisse dans un sens qui est défini précisément ci-dessous,

$$\bar{Z}_j = n^{-1}\sum_{i=1}^{n} Z_J\left(X_i\right), \quad \text{pour tout } \; j = 1, \ldots, J,$$

et

$$\mu_{Z_J} = E\left(Z_J\right).$$

Maintenant on retourne au problème d'approximation de $H_n\left(\tau, F_0\right)$ pour obtenir cette approximation, écrivons

$$G\left(\bar{Z}_1, ..., \bar{Z}_J\right) = G\left(\bar{Z}\right),$$

où $\bar{Z} = \left(\bar{Z}_1, ..., \bar{Z}_J\right)'$.

On définit

$$\mu_Z = E\left(\bar{Z}\right), \quad \partial G\left(Z\right)/\partial Z, \text{ et } \quad \Omega = E\left[\left(\bar{Z} - \mu_Z\right)\left(\bar{Z} - \mu_Z\right)'\right],$$

chaque fois que ces quantités existent. Supposons que :

(K1) $T_n = n^{1/2}\left[G\left(\bar{Z}\right) - G\left(\mu_Z\right)\right]$, où $G(z)$ est six fois continuellement partiellement différentiable pour toutes composantes de z au voisinage de μ_Z.

(K2) $\partial G\left(\mu_Z\right) \neq 0$.

(K3) La valeur attendue du produit de touts les 16 composantes de Z existes.

Sous la supposition (K1), (K2) et (K3) une approximation de série *Taylor* donne

$$n^{1/2}\left[G\left(\bar{Z}\right) - G\left(\mu_Z\right)\right] = \partial G\left(\mu_Z\right)' n^{1/2}\left(\bar{Z} - \mu_Z\right) + o_p\left(1\right). \tag{2.1}$$

Par application du théorème *Lindeberg-Levy* de limite central sur (2.1) on a :

$$n^{1/2}\left[G\left(\bar{Z}\right) - G\left(\mu_Z\right)\right] \xrightarrow{d} N\left(0, V\right),$$

où $V = \partial G\left(\mu_Z\right)' \Omega\, G\left(\mu_Z\right)$. Ainsi, la *FDC* asymptotique de T_n est

$$H_\infty\left(\tau, F_0\right) = \Phi\left(\tau/V^{1/2}\right),$$

où Φ est *FDC* de la loi normal standard. De plus, on applique le théorème *Berry-Esséen,* on résulte :

$$\sup_\tau \left|H_n\left(\tau, F_0\right) - H_\infty\left(\tau, F_0\right)\right| = O\left(n^{-1/2}\right).$$

Ainsi, sous les suppositions (K1), (K2) et (K3) du fonction modèle lisse, les approximations asymptotiques de premier ordre à la distribution d'échantillon finie de T_n fait une erreur de taille $O(n^{-1/2})$.

En effet, il est possible de prouver fortement le résultat :

$$\sup_{\tau} |H_n(\tau, F_n) - H_\infty(\tau, F_0)| \overset{p.s.}{\to} 0.$$

Ce résultat assure que le bootstrap donne une bonne approximation à la distribution asymptotique de T_n si n est suffisamment grand. cependant, ça ne dit rien, de l'exactitude de $H_n(., F_n)$ comme une approximation d'échantillon finie exacte d'une distribution $H_n(., F_0)$.

Le théorème suivant, qui est prouvé par *Hall (1992 a)* [25], donne un résultat essentiel.

Théorème 2.3.3 : *Soient les conditions vérifiées (K1),(K2) et (K3), on suppose que*

$$\lim \sup_{\|t\| \to \infty} \left| E\left[\exp\left(tt'Z\right)\right]\right| < 1, \tag{2.2}$$

où $t = \sqrt{-1}$. *Alors*

$$H_n(\tau, F_0) = H_\infty(\tau, F_0) + \frac{1}{n^{1/2}} g_1(\tau, F_0) + \frac{1}{n} g_2(\tau, F_0) + \frac{1}{n^{3/2}} g_3(\tau, F_0) + O\left(n^{-2}\right), \tag{2.3}$$

uniformément sur τ *et*

$$H_n(\tau, F_n) = H_\infty(\tau, F_n) + \frac{1}{n^{1/2}} g_1(\tau, F_n) + \frac{1}{n} g_2(\tau, F_n) + \frac{1}{n^{3/2}} g_3(\tau, F_n) + O\left(n^{-2}\right), \tag{2.4}$$

uniformément presque sûrement sur τ. De plus, g_2 et g_3 sont des fonctions différentiables dans leur premier argument, g_2 est une fonction impaire, différentiable en son premier argument et H_∞, g_1, g_2 et g_3 sont les fonctions continues de leurs deuxièmes relativement à la norme de l'espace des fonctions de distribution.

Si T_n est asymptotiquement central, donc H_∞ est la fonction de distribution normale standard. Autrement, $H_\infty(\tau, F_0)$ est la fonction de distribution $N(0, V)$ et $H_\infty(., F_n)$ est la fonction de distribution $N(0, V_n)$, où V_n est une quantité obtenue de V en remplaçant des espérances de population et moments par l'espérances et les moments relatives à F_n.

La condition (2.2) est appelée la condition de *Cramér*. Elle est satisfaite si le vecteur aléatoire Z a densité de probabilité respectivement a la mesure de *Lebesgue*.

Il est maintenant possible d'évaluer l'exactitude de l'estimateur de bootstrap $H_n(\tau, F_n)$ comme une approximation d'échantillon finie de $H_n(\tau, F_0)$. Il suit de (2.3) et (2.4) que

$$\begin{aligned} H(\tau, F_n) - H_n(\tau, F_0) &= [H_\infty(\tau, F_n) - H_\infty(\tau, F_0)] + \frac{1}{n}[g_2(\tau, F_n) - g_2(\tau, F_0)] \\ &+ \frac{1}{n^{1/2}}[g_1(\tau, F_n) - g_1(\tau, F_0)] + O\left(n^{-3/2}\right) \quad p.s. \end{aligned} \tag{2.5}$$

Le terme principal sur le côté droit de (2.5) est $[H_\infty(\tau, F_n) - H_\infty(\tau, F_0)]$. La taille de ce terme est $O(n^{-1/2})$ uniformément presque sûrement sur τ parce que $F_n - F_0 = O(n^{-1/2})$, uniformément presque sûrement sur le support de F_0. Ainsi, le bootstrap fait une erreur de taille $O(n^{-1/2})$ presque sûrement, la même de erreur faite en de taille premier ordre d'approximation asymptotique.

En termes de taux de convergence à zéro de l'erreur d'approximation, le bootstrap à la même exactitude que les approximations asymptotiques de premier ordre. Dans ce sens, rien n'est perdu dans les termes d'exactitude en utilisant le bootstrap au lieu les approximations de premier ordre, mais encore rien n'est obtenu non plus.

Supposons maintenant que T_n est asymptotiquement central. Alors la distribution asymptotique de T_n est indépendante de F_0 et $H_\infty(\tau, F_n) = H_\infty(\tau, F_0)$ pour tout τ.

Les équations (2.3) et (2.4) sont données par

$$\begin{aligned} H_n(\tau, F_n) - H_n(\tau, F_0) = \; & \frac{1}{n^{1/2}}[g_1(\tau, F_n) - g_1(\tau, F_0)] \\ & + \frac{1}{n}[g_2(\tau, F_n) - g_2(\tau, F_0)] + O\left(n^{-3/2}\right) \quad p.s. \end{aligned} \tag{2.6}$$

Le terme principal sur le côté droit de (2.6) est

$$n^{-1/2}[g_1(\tau, F_n) - g_1(\tau, F_0)].$$

Cela suite de la continuité de $g_{1,}$ en ce qui concerne son deuxième argument ; ce terme est à la taille $O(n^{-1})$ uniformément presque sûrement sur τ.

Maintenant le bootstrap fait une erreur est d'ordre $O(n^{-1})$, plus petit que l'erreur fait par les approximations asymptotiques de premier ordre quand $n \to \infty$. Ainsi, le bootstrap est plus précis que la théorie asymptotique de premier ordre pour l'estimation de la distribution d'un lisse asymptotiquement central.

Si T_n est asymptotiquement central, donc l'exactitude de bootstrap est même plus grande pour l'estimation de la fonction de distribution symétrique

$$P(|T_n| \leq \tau) = H_n(\tau, F_0) - H_n(-\tau, F_0).$$

Soit Φ la fonction de distribution normale standard, alors, il suit de (2.3) et la symétrie de g_1, $g_2$et g_3, que

$$\begin{aligned} &H_n(\tau, F_0) - H_n(-\tau, F_0) = [H_\infty(\tau, F_0) - H_\infty(-\tau, F_0)] + \frac{2}{n}[g_2(\tau, F_0)] \\ &+O(n^{-2}) = 2\Phi(\tau) - 1 + \frac{2}{n} g_2(\tau, F_0) + O(n^{-2}) \quad p.s. \end{aligned} \tag{2.7}$$

De la même façon il suit de (2.4) que

$$\begin{aligned} &H_n(\tau, F_n) - H_n(-\tau, F_n) = [H_\infty(\tau, F_n) - H_\infty(-\tau, F_n)] + \frac{2}{n}[g_2(\tau, F_n)] \\ &+O(n^{-2}) = 2\Phi(\tau) - 1 + \frac{2}{n} g_2(\tau, F_n) + O(n^{-2}) \quad p.s. \end{aligned} \tag{2.8}$$

Le reste des termes dans (2.7) et (2.8) est d'ordre $O(n^{-2})$ mais pas $O(n^{-3/2})$ parce que $O(n^{-3/2})$ est un terme d'une *expansion Edgeworth*, $n^{-3/2} g_3(\tau, F)$ même pour g_1. On utilise le fait que $F_n - F_0 = O(n^{-1/2})$, pour obtenir

$$\begin{aligned} &[H_n(\tau, F_n) - H_n(-\tau, F_n)] - [H_n(\tau, F_0) - H_n(-\tau, F_0)] = \\ &\frac{2}{n}[g_2(\tau, F_n) - g_2(\tau, F_0)] + O(n^{-2}) = O\left(n^{-3/2}\right) \quad p.s. \end{aligned} \tag{2.9}$$

Remarque 2.3.1 : si T_n est asymptotiquement central. Ainsi, l'erreur fait par l'approximation de bootstrap à la fonction de distribution symétrique $P(|T_n| \leq \tau)$ est d'ordre $O\left(n^{-3/2}\right)$ comparé à l'erreur $O(n^{-1})$ fait par les approximations asymptotiques de premier ordre.

Définition 2.3.2 : L'erreur de l'approximation de bootstrap à la fonction de *distribution unilatérale* est

$$H_n(\tau, F_n) - H_n(\tau, F_0) = O\left(n^{-1}\right) \quad p.s. \tag{2.10}$$

L'erreur dans l'approximation de bootstrap à un fonction de *distribution symétrique* est

$$[H_n(\tau, F_n) - H_n(-\tau, F_n)] - [H_n(\tau, F_0) - H_n(-\tau, F_0)] = O\left(n^{-3/2}\right) \quad p.s. \tag{2.11}$$

Remarque 2.3.2 : Au contraire, les erreurs faits par les approximations asymptotique de premier ordre sont $O(n^{-1/2})$ et $O(n^{-1})$, respectivement, pour la distribution unilatérale et la fonction symétrique.

Chapitre 3

UTILISATION DU BOOTSTRAP POUR LES PROBLEMES STATISTIQUES

Dans ce chapitre, on s'intéresse à l'utilisation du bootstrap pour le calcul de l'erreur -standard et du biais d'un estimateur et pour la détermination des limites de confiance d'un paramètre estimé ainsi que l'application des méthodes de bootstrap sur les modèles de régression. Les différentes méthodes exposées sont illustrées par un exemple.

3.1 Méthodes de rééchantillonnage

3.1.1 Bootstrap des individus

On considère un échantillon de n observations $(x_1, x_2, ..., x_i, ..., x_n)$, prélevé de manière aléatoire et simple dans une population. Ces observations peuvent concerner une seule variable, ou, au contraire, être relatives à plusieurs variables. Dans ce cas, les x_i représentent des vecteurs de dimension p, p étant le nombre de variables. Afin de ne pas alourdir les notations, nous ne distinguerons pas ces deux situations et, de manière plus condensée, nous désignerons l'échantillon initial par le symbole x, qu'il s'agisse d'un vecteur ou d'une matrice.

Le principe de la méthode du *bootstrap* est de prélever une série d'échantillons aléatoires et simples avec remise de n observations dans l'échantillon initial, considéré comme une population. Ces échantillons successifs seront notés :

$$(x_1^*, x_2^*, ..., x_k^*, ..., x_B^*),$$

où B étant le nombre de rééchantillonnages effectués.

Exemple 3.1 : À titre d'illustration, nous considérons le problème de l'estimation de diverses caractéristiques de la population, à partir d'un échantillon aléatoire et simple de 9 observations.

La deuxième colonne du tableau 1, notée x, donne les premières observations de l'échantillon. Les colonnes suivantes donnent les premières et les dernières observations de 1000 échantillons de 9 observations prélevés dans l'échantillon initial et notés x_1^*, x_2^*, ..., x_{1000}^*.

TAB.3.1 : Échantillon initial et résultats de 1000 rééchantillonnages (données partielles).

Obs	x	x_1^*	x_2^*	x_3^*	...	x_{999}^*	x_{1000}^*
1	52	46	50	104	...	52	31
2	10	40	27	46	...	27	104
3	40	27	31	52	...	104	146
...	...	...	...	...	...	...	...
9	46	31	46	10	...	52	104

Pour l'ensemble des B échantillons obtenus par bootstrap, les observations x_i n'apparaissent pas en nombre égal et on peut définir les proportions d'apparition P_i^* de chacune des observations, P_i^* étant égal au nombre de fois que l'observation x_i a été prélevée pour l'ensemble des B échantillons, divisé par le nombre total de prélèvements, qui est égal à nB.

3.2 Erreur standard et biais d'un paramètre

3.2.1 Estimation de l'erreur-standard

Soit un paramètre θ de la population et soit :

$$\hat{\theta} = f(x_{1,}..., x_n) = f(x),$$

une estimation de ce paramètre, obtenue à partir des données de l'échantillon initial x. Chaque échantillon obtenu par rééchantillonnage permet de calculer une répétition du bootstrap de l'estimation $\hat{\theta}$:

$$\hat{\theta}_k^* = f(x_k^*); \quad (k = 1, ..., B),$$

la fonction f étant la même que celle utilisée pour la définition de $\hat{\theta}$.

Supposons qu'on s'intéresse à la moyenne, à la médiane et à la variance et qu'on se propose d'estimer ces trois paramètres à partir de l'échantillon x. Si on utilise les estimateurs classiques, le paramètre $\hat{\theta}$ s'écrit, successivement pour les trois paramètres considérés :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \tilde{x} = \frac{1}{2}\left(x_{([\frac{n+1}{2}])} + x_{([\frac{n+1}{2}]+1)}\right),$$

$$\text{et} \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

où $x_{(\left[\frac{n+1}{2}\right])}$ et $x_{(\left[\frac{n+1}{2}\right]+1)}$ étant les observations de rangs $\left[\frac{n+1}{2}\right]$ et $\left[\frac{n+1}{2}\right]+1$ de l'échantillon initial. Les valeurs numériques pour ces trois estimations sont données dans la première partie du tableau 2, sur la linge intitulée $\hat{\theta}$ d'après l'algorithme de l'erreur-standard.

TAB.3.2 : Paramètres estimés pour l'échantillon initial $\left(\hat{\theta}\right)$ et pour les trois premiers échantillons obtenus par rééchantillonnage $\left(\hat{\theta}_1^*, \hat{\theta}_2^* \text{ et } \hat{\theta}_3^*\right)$; moyennes $\left(\hat{\theta}^*\right)$ et écarts-types $\left(\hat{\sigma}_{\hat{\theta}^*}\right)$ des paramètres estimés pour 1000 rééchantillonnages.

Paramètre	Moyenne	Médiane	Variance
$\hat{\theta}$	56.22	46	1799.19
$\hat{\theta}_1^*$	45.56	50	1779.44
$\hat{\theta}_2^*$	62.44	50	2348.75
$\hat{\theta}_3^*$	79.44	31	2662.86
...	...	...	...
$\hat{\theta}^*$	56.10	45.53	7670.13
$\hat{\sigma}_{\hat{\theta}^*}$	12.2579	7.112	8.089

Les calculs des trois paramètres peuvent être répétés pour les échantillons, x_1^*, x_2^* et x_3^*. Les résultats obtenus sont repris dans la seconde partie du *tableau* 2, sur les lignes intitulées $\hat{\theta}_1^*, \hat{\theta}_2^*$ et $\hat{\theta}_3^*$. Cette seconde partie du tableau peut évidemment être complétée au fur et à mesure des rééchantillonnages fournissant x_4^*, $x_5^*, ..., x_B^*$.

Disposant des B répétitions, on peut déterminer la moyenne :

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_k^*,$$

et l'écart-type des $\hat{\theta}_k^*$:

$$\hat{\sigma}_{\hat{\theta}^*} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_k^* - \hat{\theta}^*\right)^2}.$$

Définition 3.2.1 : On appelle estimation bootstrap de l'écart-type $\hat{\sigma}_F\left(\hat{\theta}\right)$ de $\hat{\theta}$, son estimation plug-in : $\sigma_{\hat{F}}\left(\hat{\theta}\right)$.

Une approximation de l'estimateur bootstrap de l'écart-type de $\hat{\theta}$ est obtenue par une simulation (*Monte-carlo*) décrite dans l'algorithme ci-dessous.

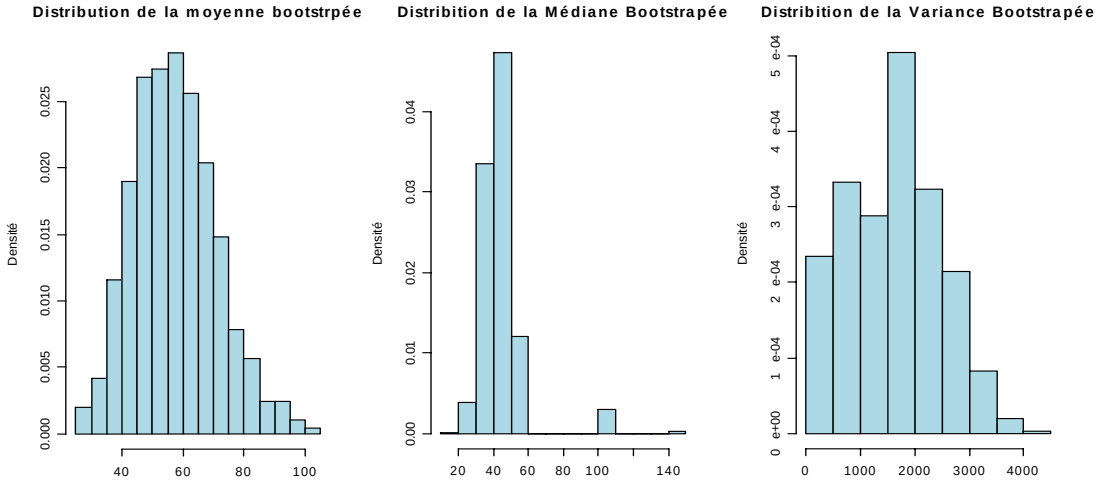


FIG. 3.1. Histogrammes de fréquences de tableau 2, B=1000.

Pour un paramètre θ est un échantillon x donnés, on note $\hat{\theta} = s(x)$ l'estimation obtenue sur cet échantillon. Une réplication bootstrap de $\hat{\theta}$ est donnée par : $\hat{\theta}^* = s(x^*)$.

ALGORITHME : Estimation bootstrap de l'écart-type

1. Tirer B échantillons bootstrap $x^{*1}, x^{*2}, ..., x^{*B}$ par tirage avec remise dans x.

2. Calculer la copie bootstrap $\hat{\theta}^*(b) = s(x^{*b}); \quad b = 1, 2, ..., B$.

3. Calculer l'ecart-type de l'échantillon ainsi construit :

$$\hat{\sigma}_B^2 = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(.) \right)^2,$$

avec

$$\hat{\theta}^*(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b).$$

où $\hat{\sigma}_B$ est l'approximation bootstrap de l'estimation plug-in recherchée de l'ecart type de $\hat{\theta}$.

L'écart-type $\hat{\sigma}_{\hat{\theta}^*}$ est une estimation de l'erreur-standard de l'estimateur du paramètre θ. Pour les situations où on dispose d'un estimateur de cette erreur-standard,

et pour autant que les conditions d'application soient remplies, on peut montrer que l'écart-type des $\hat{\theta}_k^*$ tend vers le résultat analytique, lorsque B tend vers l'infini.

Ainsi, pour la moyenne d'un échantillon aléatoire et simple, on sait que l'erreur-standard de la moyenne est égale à $\sigma/\sqrt{n}$. Si B tend vers l'infini, l'écart-type $\hat{\sigma}_{\hat{\theta}^*}$ tend vers $\hat{\sigma}_{plug}/\sqrt{n}$, avec :

$$\hat{\sigma}_{plug} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

L'estimation $\hat{\sigma}_{plug}$ de l'écart-type de la population donnée ci-dessus est appelée estimation par insertion. Pour un estimateur par insertion, la formule conduisant à l'estimation est la même que celle utilisée pour la définition du paramètre de la population. En effet, pour une population finie de taille N et de moyenne m_X on a :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)^2}.$$

En d'autres mots, pour un estimateur par insertion, on considère l'échantillon comme une population particulière et on utilise la formule relative au paramètre de la population.

Nombre de réplications bootstrap nécessaires

D'une manière générale, lorsque B tend vers l'infini, la valeur $\hat{\sigma}_{\hat{\theta}^*}$ tend vers une valeur fixée qui correspond à l'estimation de l'erreur-standard du bootstrap idéal. Efron et Tibshirani (1993) [17] proposent les règles empiriques suivantes pour le choix de B :

- un nombre réduit de répétitions ($B = 25$, par exemple) permet d'obtenir une première information et $B = 50$ est généralement suffisant pour avoir une bonne estimation de l'erreur-standard ;
- il est très rare que plus de 200 répétitions soient nécessaires pour estimer une erreur-standard.

TAB.3.3 : Erreur standard de la moyenne ($SE(m)$) de tableau **1**.

B	25	50	100	200	400	1000	3200	10000	15000
$SE(m)$	13.28	15.16	12.84	11.55	13.83	13.30	12.86	13.23	13.28

Remarque 3.2.1 : On peut noter que le choix de B n'est pas fonction de la taille n de l'échantillon.

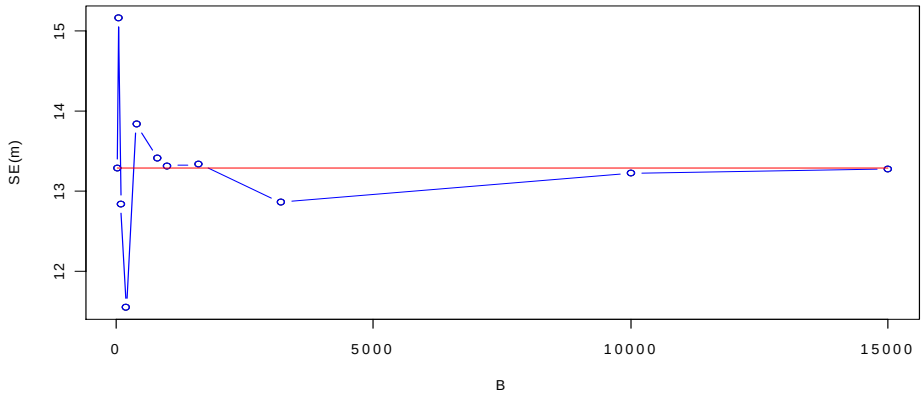


FIG. 3.2. Nombre de réplications bootstrap nécessaires.

3.2.2 Estimation du biais

L'un des emplois les plus fréquents du bootstrap est d'éliminer le biais d'un estimateur de la manière suivante : Soit T un estimateur de θ, son biais est :

$$b(T) = E(T|F) - \theta.$$

On estime ce biais par :

$$b^*(T) = E(T^*|X) - T,$$

où T^* est calculé sur un échantillon bootstrap X^* issu de l'échantillon initial X.

L'estimateur T est ensuite "corrigé de son biais" et donc remplacé par :

$$T - b^*(T) = 2T - E(T^*|X).$$

Comme $T - b(T)$ est sans biais pour θ, $T - b^*(T)$ sera presque sans biais. Prenons un exemple. On peut utiliser le bootstrap pour réduire ce biais. Supposons que l'on veuille estimer :

$$\theta(F) = \left[\int x dF(x)\right]^r,$$

à partir d'un échantillon auquel est associée la fonction de répartition empirique F_0 telle que $F_0(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq x\}}$. On choisit l'estimateur

$$\hat{\theta}(F) = \theta(F_0) = \left[\int x dF(x)\right]^r.$$

Quel est son biais ?

Par définition

$$Biais = E\{\theta(F_0) - \theta(F_1) \mid F_0\}.$$

Comme on ne connaît pas F, on utilise le principe du bootstrap en remplaçant dans cette équation F par F_0 et F_0 par F_1, où F_1 est la loi associée à un n échantillon d'une variable de loi F_0 :

$$\widehat{Biais} = E\{\theta(F_0) - \theta(F_1) \mid F_0\}.$$

Donc l'estimateur sans biais de θ s'obtient en retranchant à $\theta(F_0)$ cet estimateur de son biais, soit :

$$\text{Estimateur sans biais de } \theta = \theta(F_0) - \widehat{Biais}.$$

Pour obtenir un estimateur sans biais, on doit donc ajouter t à $\theta(F_0)$ où t est défini par

$$E(\theta(F_0) - \theta(F) + t) = 0.$$

On a donc remplacé l'équation initiale qui donne la correction t que l'on devrait faire pour supprimer le biais de l'estimateur $\widehat{\theta(F_0)}$ par une équation bootstrap qui donne une correction t^*, en principe calculable, et dont on espère qu'elle est une bonne estimation de t.

On remarque que t est un paramètre qui dépend de F alors que t^* est une statistique dépendant de F_0. De cette équation se déduit la correction

$$t^* = \theta(F_0) - E(\theta(F_1) \mid F_0).$$

On doit donc calculer la quantité :

$$E(\theta(F_1) \mid F_0),$$

l'estimateur sans biais est alors égal à :

$$\theta(F_0) + t^* = 2\theta(F_0) - E(\theta(F_1) \mid F_0).$$

Procédure de Monte-Carlo pour bootstrap du biais :

Soit x un échantillon et θ un paramètre.

B1. Pour $b = 1 : B$

-utiliser les données d'estimation pour calculer θ.

-Sélectionner l'échantillon bootstrap $x^{*b} = \left\{x_1^{*b}, x_2^{*b}, ..., x_n^{*b}\right\}$ par tirage avec remise dans x.

B2. Estimer sur cet échantillon la réplication bootstrap de $\hat{\theta}$

$$\hat{\theta}^*(b) = s(x^*).$$

B3. Approcher $E_{\hat{F}}[s(x^*)]$ par $\hat{\theta}^*(.) = \frac{1}{B}\sum_{k=1}^{B}\left(\hat{\theta}_b^*\right)$

-L'approximation bootstrap du biais est

$$\widehat{Biais}_B\left(\hat{\theta}\right) = \hat{\theta}^*(.) - \hat{\theta}.$$

3.2.3 Estimations par le jackknife

La méthode dite du ≪ j*ackknife* ≫, nom anglais d'un couteau scout à plusieurs lames prêt a être utilisé dans un grand nombre de situations, a été introduite par M. Quenouille (1949, 1956).[1] L'objectif initial du jackknife est de réduire le biais d'un estimateur.

Une autre forme de rééchantillonnage permettant d'estimer l'erreur standard et le biais d'un paramètre est la technique du jackknife. On calcule n fois la valeur de paramètre à partir d'un échantillon de $(n-1)$ observations, chacune des observations étant éliminée à son tour de l'échantillon. En désignant par $\hat{\theta}_{(-i)}$ l'estimation obtenue après élimination de l'observation i, on peut estimer l'erreur-standard du paramètre $\hat{\theta}$ par la relation suivante (*Efron, Tibshirani (1993)* [17] *; Dagnelie, (1998)* [10]) *:*

$$\hat{\sigma}_{\hat{\theta}_J} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta}_J)^2},$$

avec

$$\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)},$$

[1] La dénomination ≪ jackknife ≫ est due à J.TAKEY(1958)

et le biais vaut :

$$biais_J(\hat{\theta}) = (n-1)(\hat{\theta}_J - \hat{\theta}).$$

Définitions

Soient n réalisations indépendantes $(X_1, ..., X_n)$ d'une variable X de loi $\mathbb{P}_\theta$ dépendant d'un paramètre réel θ, on possède un estimateur T_n biaisé de θ :

$$\mathbb{E}(T_n) = \theta + B(n, \theta).$$

On note ε_i le sous-échantillon $(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$ obtenu à partir de l'échantillon initial en supprimant la i^e observation, cela revient à dire que l'on fait un sondage dans l'échantillon de base en tirant $n-1$ observations sans remise.

T^i_{n-1} désigne la statistique fondée sur ε_i selon la même règle de décision que celle de T_n.

Définition 3.2.2 : On *appelle pseudo-valeur* d'ordre i de T_n la statistique

$$J_{\ i}\ (T) = nT_n - (n-1)T^i_{n-1}.$$

Définition 3.2.3 : On appelle *jackknife* de T_n la statistique $J(T_n)$ moyenne des pseudo-valeurs :

$$\begin{aligned} J(T_n) &= \frac{1}{n}\sum_{i=1}^{n} J_i(T_n) = nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T^i_{n-1} \\ &= T_n - \frac{n-1}{n}\sum_{i=1}^{n}(T^i_{n-1} - T_n), \end{aligned}$$

où $J(T_n)$ est appelé estimateur du jackknife de T_n, ou *jackknifé* de T_n.

Remarque 3.2.2 : Si T_n est sans biais, alors $J(T_n)$ l'est aussi.

Calculons la variance empirique $S^2_{PJ}(T)$ des n pseudo-valeurs $J_i(T)$:

$$\begin{aligned} S^2_{PJ}(T) &= \frac{1}{n-1}\sum_{i=1}^{n}(J_i(T_n) - J(T_n))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\frac{n-1}{n}\sum_{i=1}^{n} T^i_{n-1} - (n-1)T^i_{n-1})^2. \end{aligned}$$

D'où

$$S^2_{PJ}(T) = (n-1)\sum_{i=1}^{n}(T^i_{n-1} - \bar{T}^i_{n-1})^2,$$

avec

$$\bar{T}^i_{n-1} = \frac{1}{n}\sum_{i=1}^{n} T^i_{n-1}.$$

D'autre part

$$\mathbb{V}(J(T_n)) = \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbb{V}(J_i\ (T_n)) + \sum_{i\neq j}\mathbb{C}ov(J_i\ (T_n), J_j\ (T_n))\right].$$

Les $J_i(T_n)$ peuvent être considérés comme i.i.d ; sous cette conjecture

$$\mathbb{V}(J(T_n)) = \frac{1}{n}\mathbb{V}(J_1(T_n)).$$

On peut estimer $\mathbb{V}(J(T_n))$ par $S^2_{PJ}(T)/n$, c'est-à-dire :

$$\begin{aligned}\hat{\mathbb{V}}(J(T_n)) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(J_i(T_n) - J(T_n))^2 \\ &= \frac{(n-1)}{n}\sum_{i=1}^{n}(T^i_{n-1} - \bar{T}^i_{n-1})^2.\end{aligned}$$

Par la suite, on notera

$$JV(T_n) = S^2_{PJ}(T)/n;$$

où $S^2_{PJ}(T_n)/n$ est souvent utilisé pour estimer $V(T_n)$.

Lorsque la taille n de l'échantillon est inférieure au nombre de répétitions utilisé dans la méthode du bootstrap, soit le plus souvent entre 50 et 200, le calcul de l'erreur-standard et du biais est plus rapide par le jackknife. Par contre, la méthode du jackknife, qui peut être considérée comme une approximation du bootstrap, donne, de manière générale, de moins bonnes estimations.

De plus, la méthode du jackknife peut donner des résultats aberrants lorsque la statistique $\hat{\theta}$ n'est pas une fonction continue des observations x_i, c'est-à-dire une fonction variant de manière régulière lorsque les données changent. La médiane est un cas typique de fonction non continue.

3.3 Bootstrap pour les tests d'hypothèse

Soit T_n une statistique pour tester une hypothèse H_0 de la population échantillonnée. Supposons que sous H_0, T_n est asymptotiquement central et satisfait les suppositions (K1), (K2) et (K3) et (2.2).Ce test rejete H_0 à un niveau α si

$$|T_n| > z_{n,\alpha/2},$$

où $z_{n,\alpha/2}$ l'échantillon fini exacte, α le niveau de la valeur critique $(1-\alpha/2)$ quantile de la distribution de T_n. La valeur critique réalise l'équation

$$H_n\left(z_{n,\alpha/2}, F_0\right) - H_n\left(-z_{n,\alpha/2}, F_0\right) = 1-\alpha. \tag{3.1}$$

À moins que T_n est exactement central, cependant, l'équation (3.1) ne peut pas être résolue par application ; parce que F_0 est inconnu. Donc la valeur critique exacte, d'échantillon finie ne peut pas être obtenue par application si T_n n'est pas central. Des approximations asymptotiques de premier ordre obtiennent une version possible de (3.1) en remplaçant H_n par H_∞. Ainsi la valeur critique asymptotique $z_{\infty,\alpha/2}$,résoud

$$H_\infty\left(z_{\infty,\alpha/2}, F_0\right) - H_\infty\left(-z_{\infty,\alpha/2}, F_0\right) = 1-\alpha. \tag{3.2}$$

Puisque H_∞ est la distribution normale standard quand T_n est asymptotiquement central, alors $z_{\infty,\alpha/2}$ peut obtenir des tableaux de quantiles normale standard, composés par (2.7), (3.1) et (3.2) par suite on a :

$$\left[H_\infty\left(z_{n,\alpha/2}, F_0\right) - H_\infty\left(-z_{n,\alpha/2}, F_0\right)\right] - \left[H_n\left(z_{\infty,\alpha/2}, F_0\right) - H_n\left(-z_{\infty,\alpha/2}, F_0\right)\right]$$

$$= O\left(n^{-1}\right),$$

qui implique que

$$z_{n,\alpha/2} - z_{\infty,\alpha/2} = O\left(n^{-1}\right),$$

ainsi, la valeur critique asymptotique s'approche de la valeur critique exacte d'échantillon finie avec l'erreur dont la taille est $O\left(n^{-1}\right)$.

Le bootstrap obtient une version faisable de (3.1) en remplaçant F_0 par $F_{n.}$ Ainsi, le bootstrap de la valeur critique $z^*_{n,\alpha/2}$ résoudre

$$H_n\left(z^*_{n,\alpha/2}, F_n\right) - H_n\left(-z^*_{n,\alpha/2}, F_n\right) = 1-\alpha. \tag{3.3}$$

L'équation (3.3) ne peut pas être résolue analytiquement, mais $z^*_{n,\alpha/2}$ peut être estimé par simulation de *Monte-Carlo*. On suppose souvent dans les applications que T_n est asymptotiquement normale, l'estimateur de Studend d'un paramètre θ dont la valeur sous H_0 est θ_0. C'est-à-dire.

$$T_n = \frac{n^{1/2}\left(\theta_n - \theta_0\right)}{s_n},$$

où θ_n est l'estimateur de θ, $n^{1/2}\left(\theta_n - \theta_0\right) \xrightarrow{d} N(0, \sigma^2)$ sous H_0 et s_n^2 est un estimateur consistant de σ^2. Alors la procédure de *Monte-Carlo* pour l'estimation $z^*_{n,\alpha/2}$ est comme suit :

Procédure de Monte-Carlo pour calculer le bootstrap de valeur critique

T1. Utiliser les données d'estimation pour calculer θ_n.

T2. Construire un échantillon bootstrap de taille n en échantillonnant la distribution correspondant à F_n.

-Calculer θ^*_n et s^*_n d'échantillon bootstrap, la version de bootstrap de T_n est

$$T^*_n = n^{1/2}(\theta^*_n - \theta_0)/s^*_n.$$

T3. Utiliser les résultats de répétitions de $T2$ pour calculer la distribution empirique de $\left|T^*_n\right|$.

L'estimation de valeur critique du bootstrap $z^*_{n,\alpha/2}$ est comme un estimateur de la valeur exacte d'échantillon fini de valeur critique $z_{n,\alpha/2}$, en (2.7) et (3.1) pour obtenir :

$$2\Phi\left(z_{n,\alpha/2}\right) - 1 + \frac{2}{n} g_2\left(z_{n,\alpha/2}, F_0\right) = 1 - \alpha + O(n^{-2}) \quad p.s. \tag{3.4}$$

Combinant de la même façon (2.8) et (3.1) pour obtenir :

$$2\Phi\left(z^*_{n,\alpha/2}\right) - 1 + \frac{2}{n} g_2\left(z^*_{n,\alpha/2}, F_n\right) = 1 - \alpha + O(n^{-2}) \quad p.s. \tag{3.5}$$

Alors, d'après Hall (1992 a) nous avons :

$$z_{n,\alpha/2} = z_{\infty,\alpha/2} - \frac{1}{n}\frac{g_2\left(z_{\infty,\alpha/2}, F_0\right)}{\phi\left(z_{\infty,\alpha/2}\right)} + O(n^{-2}), \tag{3.6}$$

où ϕ est la fonction de densité normale standard et

$$z^*_{n,\alpha/2} = z_{\infty,\alpha/2} - \frac{1}{n}\frac{g_2\left(z_{\infty,\alpha/2}, F_n\right)}{\phi\left(z_{\infty,\alpha/2}\right)} + O(n^{-2}) \quad p.s. \tag{3.7}$$

Il suit de (3.6) et (3.7) que

$$z^*_{n,\alpha/2} = z_{\infty,\alpha/2} + O\left(n^{-3/2}\right) \quad p.s. \tag{3.8}$$

Considérons maintenant la probabilité de rejet du test basé sur T_n (PR) quand H_0 est vrai, alors PR est

$$P(|T_n| > z_{\infty,\alpha/2}) = \alpha,$$

telle que l'asymptotique de la valeur critique PR est

$$P(|T_n| > z_{\infty,\alpha/2}) = 1 - \left[H_n\left(z_{\infty,\alpha/2}, F_0\right) - H_n\left(-z_{\infty,\alpha/2}, F_0\right)\right] = \alpha + O\left(n^{-1}\right), \tag{3.9}$$

où $\tau = z_{\infty,\alpha/2}$ dans (2.7). Ainsi, avec la valeur critique asymptotique PR vrai et nominal diffère de $O(n^{-1})$.

Considérons maintenant PR avec le bootstrap de valeur critique

$$P(|T_n| > z^*_{n,\alpha/2}),$$

parce que $z^*_{n,\alpha/2}$ est une variable aléatoire ;

$$P(|T_n| > z^*_{n,\alpha/2}) \neq 1 - \left[H_n\left(z^*_{n,\alpha/2}, F_0\right) - H_n\left(-z^*_{n,\alpha/2}, F_0\right)\right].$$

Donc

$$P(|T_n| > z^*_{n,\alpha/2}) = \alpha + O(n^{-2}). \tag{3.10}$$

Les hypothèses générales et la statistique de test sont données dans *Beran (1988)* [3] , *Hall (1986, 1988)* [21 − 22 − 23].

3.4 Intervalle de Confiance et Bootstrap

Soit θ un paramètre de population de valeur vraie mais inconnue θ_0 et θ_n est l'estimateur $n^{1/2}$ consistant asymptotiquement normal de θ, s_n est un estimateur consistant de l'écart-type de la distribution asymptotique de $n^{1/2}(\theta_n - \theta_0)$.

L'intervalle théorique asymptotique pour θ_0 est

$$\theta_n - z_{\infty,\alpha/2} s_n/n^{1/2} \leq \theta_0 \leq \theta_n + z_{\infty,\alpha/2} s_n/n^{1/2}.$$

On définit :

$$T_n = n^{1/2}(\theta_n - \theta_0)/s_n.$$

Alors la probabilité d'intervalle de confiance asymptotique est

$$P(|T_n| \leq z_{\infty,\alpha/2}\,).$$

Il suit de (3.9) que la différence entre la vraie probabilité d'intervalle et le nominal probabilité de $1-\alpha$, est $O(n^{-1})$. Si T_n satisfait le *Théorème (2.3.1)*, donc l'intervalle de confiance de bootstrap de valeur critique est :

$$\theta_n - z^*_{\infty,\alpha/2}\, s_n/n^{1/2} \leq \theta_0 \leq \theta_n + z^*_{\infty,\alpha/2}\, s_n/n^{1/2}. \tag{3.11}$$

La probabilité de cet intervalle est :

$$P(|T_n| \leq z^*_{\infty,\alpha/2}\,). \tag{3.12}$$

D'après (3.10), on a :

$$P(|T_n| \leq z^*_{\infty,\alpha/2}\,) = 1 - \alpha + O(n^{-2}).$$

3.4.1 Méthode de l'erreur-standard

Une première solution consiste à définir l'intervalle de confiance par la méthode de l'erreur-standard :

$$\hat{\theta} \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}^*},$$

où $z_{1-\alpha/2}$ étant le pourcentile $(1-\alpha/2)$ de la distribution normale réduite et $(1-\alpha)$ étant le degré de confiance retenu.

Pour que cette approche soit satisfaisante, il faut que la distribution d'échantillonnage du paramètre étudié soit approximativement normale, que l'estimateur soit non biaisé, et que $\hat{\sigma}_{\hat{\theta}^*}$ soit une bonne estimation de l'erreur-standard de la distribution du paramètre.

Le fait que ces conditions soient remplies ou non dépend des circonstances. La condition de normalité peut être vérifiée à partir de la distribution des $\hat{\theta}^*_k$ et il peut être utile éventuellement d'effectuer une transformation de manière à rendre la distribution plus proche de la normale. Le biais de l'estimateur peut être estimé, comme nous l'avons vu au paragraphe (3.2), mais sa prise en compte risque d'augmenter la variance de l'estimateur. Enfin, la qualité de l'estimation de l'erreur-standard est liée au nombre de répétitions B.

3.4.2 Méthode des pourcentiles simples

Dans la méthode des pourcentiles simples, les limites de confiance sont données par les pourcentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des $\hat{\theta}^*_k$. Nous les notons $\hat{\theta}^*_{[\alpha/2]}$ et $\hat{\theta}^*_{[1-\alpha/2]}$.

Contrairement à la méthode de l'erreur-standard, la distribution d'échantillonnage du paramètre étudié ne doit pas être normale pour que la méthode des pourcentiles soit satisfaisante. Par contre, le nombre de rééchantillonnages B doit être plus élevé que dans le cas de la méthode de l'erreur-standard, car il faut un plus grand nombre d'observations pour estimer, avec une précision suffisante, un pourcentile que pour estimer un écart-type. B sera par exemple de 1.000.

Pour 1.000 rééchantillonnages et pour un degré de confiance de 95%, les pourcentiles 0.025 et 0.975 correspondent approximativement à l'observation de rang 25 et à l'observation de rang 975, la valeur exacte pouvant dépendre de l'algorithme utilisé pour le calcul de ces pourcentiles.

Il faut noter aussi qu'une procédure de calcul un peu différente a été proposée par *Hall (1992)* [26] et est décrite par *Manly (1997)* [37]. La méthode consiste à calculer les écarts :

$$\hat{e}^*_k = \hat{\theta}^*_k - \hat{\theta},$$

et à déterminer les pourcentiles $\alpha/2$ et $1-\alpha/2$, notés $\hat{e}^*_{[\alpha/2]}$ et $\hat{e}^*_{[1-\alpha/2]}$ de cette distribution. Les limites de confiance sont alors données par les relations :

$$\hat{\theta} - \hat{e}^*_{[1-\alpha/2]} \qquad \text{et} \qquad \hat{\theta} - \hat{e}^*_{[\alpha/2]}.$$

3.4.3 Méthode des pourcentiles corrigés pour le biais

On détermine d'abord la proportion p de valeurs $\hat{\theta}_k^*$ inférieures à $\hat{\theta}$ et on calcule le pourcentile z_p relatif à la distribution normale réduite.

Soit α_1 et α_2 les valeurs de la fonction de répartition de la normale réduite aux points z_1 et z_2 :

$$\alpha_1 = \Phi(z_1) \qquad \text{et} \qquad \alpha_2 = \Phi(z_2),$$

avec $z_1 = 2z_{_P} + z_{\alpha/2}$ et $z_2 = 2z_{_P} + z_{1-\alpha/2}$.

Les limites de confiance déterminées par la méthode des pourcentiles corrigés pour le biais sont alors les pourcentiles $\hat{\theta}^*_{[\alpha_1]}$ et $\hat{\theta}^*_{[\alpha_2]}$ de la distribution des $\hat{\theta}_k^*$.

Des informations concernant l'origine de cette correction sont données dans *Efron et Tibshirani (1993)* [17] et dans *Chermick (1999)* [6].

On remarque que si $P = 0.5$, c'est-à-dire si $\hat{\theta}$ est la médiane de la distribution des $\hat{\theta}_k^*$, il n'y a pas de correction pour le biais, puisque $z_P = 0$, et on retrouve la méthode précédente. Si P est inférieur à 0.5, les limites de confiance correspondent à des pourcentiles inférieurs respectivement à $\alpha/2$ et $1 - \alpha/2$. Au contraire si P est supérieur à 0.5, les limites correspondent à des pourcentiles supérieurs à $\alpha/2$ et $1 - \alpha/2$.

3.4.4 Méthode des pourcentiles avec correction pour le biais et accélération

La méthode précédente, qui prend en compte le biais, peut être étendue de manière à tenir compte d'un éventuel changement de l'erreur-standard de $\hat{\theta}$ lorsque θ varie. Elle porte alors le nom de méthode des pourcentiles avec correction pour le biais et accélération (CBa). Une justification de cette méthode est donnée par *Efron* et *Tibshirani (1993)* [17].

Les limites de confiance sont les pourcentiles $\hat{\theta}^*_{[\alpha_1]}$ et $\hat{\theta}^*_{[\alpha_2]}$ de la distribution des $\hat{\theta}_k^*$, α_1 et α_2 étant les valeurs de la fonction de répartition de la variable normale réduite aux points z_1 et z_2 définis de la manière suivante :

$$z_1 = z_P + \left(z_P + z_{\alpha/2}\right) / \left[1 - a\left(z_P + z_{\alpha/2}\right)\right]$$

$$z_2 = z_P + \left(z_P + z_{1-\alpha/2}\right) / \left[1 - a\left(z_P + z_{1-\alpha/2}\right)\right].$$

Dans ces relations, z_P est défini comme précédemment et la constante a est appelée accélération, car elle est liée au taux de variation de l'erreur standard de $\hat{\theta}$ lorsque

le paramètre θ varie. Cette constante peut être estimée de différentes manières. Une solution consiste à utiliser la technique du *jackknife*. On obtient alors le paramètre a par la relation suivante :

$$a = \sum_{i=1}^{n} \left(\hat{\theta}_j - \hat{\theta}_{(-i)}\right)^3 /6. \left[\sum_{i=1}^{n} \left(\hat{\theta}_j - \hat{\theta}_{(-i)}\right)^2\right]^{3/2}.$$

Dans cette relation, $\hat{\theta}_{(-i)}$ est l'estimation du paramètre θ obtenue à partir de l'échantillon initial, dont on a enlevé la $i^{ème}$ observation et $\hat{\theta}_j$ est la moyenne des n valeurs $\hat{\theta}_{(-i)}$.

On peut constater que si $a = 0$, on retrouve la méthode des pourcentiles corrigés pour le biais. La prise en compte de l'accélération constitue donc bien une extension de la méthode précédente.

On constate que pour la médiane les résultats sont identiques à ceux obtenus par la méthode précédente, le paramètre a étant nul, du fait de la parfaite symétrie de la distribution des $\tilde{x}_{(-i)}$.

3.4.5 Méthode du bootstrap-t

l'idée mise en application dans le bootstrap-t est de définir une statistique dont la distribution ne soit pas fonction de la valeur réelle et inconnue du paramètre θ. La statistique T :

$$T = \frac{\hat{\theta} - \theta}{\sigma_\theta},$$

peut remplir ce rôle. Il s'agit alors d'approcher la distribution théorique de T par rééchantillonnage. Dans ce but, on calcule :

$$t_k^* = \frac{\hat{\theta}_k^* - \hat{\theta}}{\hat{\sigma}\left(\hat{\theta}_k^*\right)},$$

où $\hat{\sigma}\left(\hat{\theta}_k^*\right)$ étant l'erreur-standard de $\hat{\theta}_k^*$, qui dépend donc de l'échantillon k. Elle peut

être calculée par une formule théorique, lorsqu'une telle formule est disponible, ou à partir du rééchantillonnage de l'échantillon x_k^* utilisé pour calculer $\hat{\theta}_k^*$. Il s'agit alors d'un *bootstrap* à deux niveaux, puisque chaque échantillon x_k^* fait lui-même l'objet d'un rééchantillonnage, permettant de calculer l'erreur-standard.

Une autre possibilité pour estimer l'erreur-standard est le recours à la méthode du *jackknife*. Pour l'échantillon x_k^*, on détermine les n valeurs $\hat{\theta}^*_{(-i)}$, en éliminant à tour de rôle chacune des observations, et on calcule l'erreur-standard du paramètre par la relation suivante :

$$\hat{\sigma}\left(\hat{\theta}_k^*\right) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}^*_{(-i)} - \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^*_{(-i)}\right)^2}.$$

Lorsque l'effectif n de l'échantillon est inférieur au nombre de répétitions au second niveau, noté B', la méthode du jackknife exige moins de calculs, mais l'estimation est généralement moins bonne que l'estimation par *bootstrap*, du moins lorsque B' est suffisamment grand (de 50 à 200, par exemple).

Disposant de la distribution des t_k^*, on en détermine les pourcentiles $\alpha/2$ et $1-\alpha/2$ notés $t^*_{[\alpha/2]}$ et $t^*_{[1-\alpha/2]}$, et on obtient les limites de confiance du paramètre θ par les relations.

$$\begin{aligned}\hat{\theta}_1 &= \hat{\theta} - t^*_{[1-\alpha/2]}\hat{\sigma}\left(\hat{\theta}_k^*\right),\\ \hat{\theta}_2 &= \hat{\theta} - t^*_{[\alpha/2]}\hat{\sigma}\left(\hat{\theta}_k^*\right).\end{aligned}$$

Choix d'une méthode

Dans les paragraphes précédents, diverses méthodes de calcul de l'intervalle de confiance ont été décrites.

TAB.3.4 : Caractéristiques des méthodes de calcul de l'intervalle de confiance d'un paramètre (**1.** méthode de l'erreur-standard ; **2.** méthode des pourcentiles simples **3.** méthode des CBa ; **4.** méthode du bootstrap-t).

Méthodes	1	2	3	4
Nombre de rééchantillonnages	100	1.000	1.000	1.000×100
Respect du domaine	non	oui	oui	non
Ordre de précision	$n^{-1/2}$	$n^{-1/2}$	n^{-1}	n^{-1}

Le tableau 4 reprend quelques caractéristiques de ces méthodes. La méthode du pourcentile avec correction du biais n'a pas été reprise car il s'agit d'un cas particulier de la méthode avec correction du biais et accélération.

La première ligne du tableau 4 concerne l'ordre de grandeur du nombre d'échantillons qu'il faut prélever pour le calcul des limites de confiance. La méthode de l'erreur standard est la plus rapide ($B = 100$, par exemple) alors que la méthode du

bootstrap-t est la plus coûteuse, puisqu'elle fait appel au bootstrap à deux niveaux : par exemple, $B = 1.000$ répétitions au premier niveau, chacune de celles-ci faisant l'objet de $B' = 100$ répétitions pour l'estimation de l'erreur standard, sauf si n est plus petit que 100 et qu'on utilise la méthode du *jackknife*, qui est cependant en général moins bonne.

La deuxième ligne du tableau 4 signale si l'intervalle de confiance respecte le domaine dans lequel doit se trouver le paramètre θ. Ainsi, par exemple, les méthodes qui ne respectent pas le domaine (méthode de l'erreur standard et méthode du bootstrap-t) peuvent conduire à des limites de confiance qui ne sont pas dans le domaine $(-1, 1)$ pour un coefficient de corrélation, alors que les autres méthodes ne donneront jamais des limites de confiance situées en dehors de ce domaine.

Enfin, la troisième ligne a trait à la proportion d'intervalles de confiance corrects. Idéalement, pour un degré de confiance égal à 0.95 par exemple, la probabilité que la limite inférieure de l'intervalle de confiance soit supérieure à θ doit être égale à 0.025. De même, la probabilité que la limite supérieure de l'intervalles de confiance soit inférieure à θ doit être égale à 0.025.

Plutôt que de considérer les deux limites simultanément, on peut s'intéresser à une seule limite.

Soit $\theta_{[\alpha]}$ la limite calculée, telle que, idéalement :

$$P\left(\theta < \theta_{[\alpha]}\right) = \alpha.$$

On peut montre que les méthodes de l'erreur standard et des pourcentiles simples conduisent à des probabilités, en général à :

$$P\left(\theta < \theta_{[\alpha]}\right) = \alpha + O(n^{-1/2}).$$

Tandis que les méthodes du bootstrap-t et des pourcentiles avec correction du biais et accélération sont du type :

$$P\left(\theta < \theta_{[\alpha]}\right) = \alpha + O(n^{-1}).$$

Les termes $O(n^{-1/2})$ et $O(n^{-1})$ indiquent respectivement l'ordre de grandeur de l'erreur. Ces relations montrent que les deux dernières méthodes citées sont préférables de ce point de vue.

La comparaison des différentes méthodes proposées montre que la méthode des pourcentiles corrigés pou le biais et accélération offre, dans l'ensemble, le plus d'avantages et est, de ce fait, préconisée par Efron et Tibshirani (1993) [17].

Exemple 3.2 : -Distribution du Bootstrap loi normale-

Soit $X_1, ..., X_{200} \overset{i.i.d}{\sim} N(\theta, 1)$, pour $R = 1500$ réplications bootstrap de la moyenne.

TAB.3.5 : Erreurs standards, biais et l'intervalles de confiance de loi normale Bootstrappée :

Statistique du Bootstrap	Original	Biais	Erreur-standard
	t_1^* 1.545965	0.001086968	0.1197003
IC asymptotique	Niveau	Normale	Student
	95%	$(-0.0173, 0.2187)$	$(-0.1249, -0.1250)$
IC Bootstrap	Niveau	Normale	Basic
	95%	$(-0.0378, 0.2305)$	$(-0.0463, 0.2248)$
		Percentile	CBa
		$(-0.0333, 0.2379)$	$(-0.0301, 0.2448)$

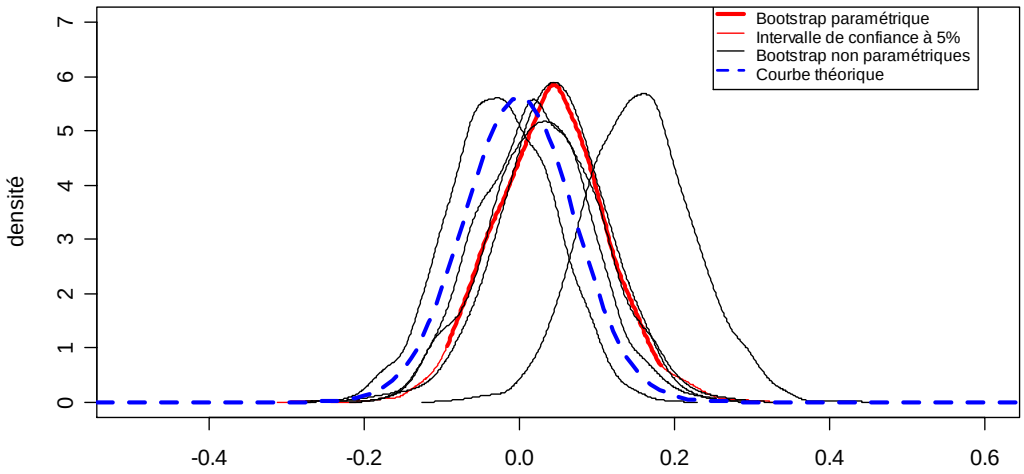


FIG. 3.3. Pertinence du bootstrap.

3.5 Techniques du bootstrap pour les modèles de régression

L'utilisation du bootstrap pour les modèles de régression a initialement été abordée par *Freedman* (1981) [18]. Jeong et *Maddala* (1993) [35], *Vinod* (1993) [42] offrent des synthèses des nombreux développements et applications des techniques de bootstrap dans le domaine de l'économétrie qui est ensuite apparu. *Horowitz* (1997) [30] s'intéresse aux performances théoriques et numériques du bootstrap en économétrie.

3.5.1 Bootstrap des Résidus

Le modèle de régression linéaire multiple est noté :

$$Y = X\beta + u, \tag{3.13}$$

où Y est un vecteur $(n, 1)$, X une matrice (n, p), β le vecteur des coefficients à estimer $(p, 1)$ et u le vecteur des erreurs aléatoires $(n, 1)$. Un rang d'observations i $(i = 1, \ldots, n)$ de la matrice X, correspondant à une ligne, est noté $X_i(1, p)$ les paramètres estimés par la méthode des moindres carrés ordinaires (MCO) $\hat{\beta}$ et les résidus $\hat{u}$ sont définis comme :

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y, \quad \hat{u} = Y - X\hat{\beta}.$$

Le modèle théorique bootstrap est le suivant :

$$Y^* = X\hat{\beta} + u^*, \tag{3.14}$$

où u^* est un terme aléatoire issu des résidus $\hat{u}$ de la régression initiale. L'application de la procédure bootstrap consiste à répéter B fois les étapes suivantes :

1. à chaque itération $b(b = 1 : B)$, un échantillon $\{y_i^*\}_{i=1}^n$ de dimension $(n, 1)$, est constitué à partir du modèle bootstrap (3.14) ;

2. les résidus MCO étant plus petits que les erreurs qu'ils estiment, le terme aléatoire du modèle théorique bootstrap est construit à partir des résidus transformés suivants, qui sont de même norme que les termes erreurs u_i :

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{1 - h_i}} - \frac{1}{n}\sum_{s=1}^{n} \frac{\hat{u}_s}{\sqrt{1 - h_s}},$$

où h_i est l'élément diagonal (i, i) de la matrice $X(X^T X)^{-1} X^T$.

En effet, les erreurs u et les résidus $\hat{u}$ sont liés par la relation

$$\hat{u} = (1 - X(X^T X)^{-1} X^T)u.$$

Le modèle théorique bootstrap s'exprime donc comme :

$$y_i^* (b) = X_i \hat{\beta} + \tilde{u}_i^*, (b), \quad i = 1, ..., n \tag{3.15}$$

où $\tilde{u}_i^* (b)$ est rééchantillonné à partir des $\tilde{u}_i$.

Soit la variable aléatoire z_j définie comme :

$$z_j = \frac{\hat{\beta}_j - \beta_j}{s\left(\hat{\beta}_j\right)},$$

où $s\left(\hat{\beta}_j\right)$ désigne l'écart type estimé du coefficient. L'intervalle de confiance standard de β_j découle de l'hypothèse selon laquelle z_j est distribuée selon une loi de Student à $n - p$ degrés de liberté. Ainsi, pour un niveau de confiance $1 - 2\alpha$, cet intervalle de confiance prend la forme suivante :

$$\left[\hat{\beta}_j - s\left(\hat{\beta}_j\right) t_{(1-\alpha),n-p}, \hat{\beta}_j - s\left(\hat{\beta}_j\right) t_{(\alpha),n-p}\right] \tag{3.16}$$

où t est la valeur des quantiles α et $1 - \alpha$ de la distribution de Student à $n - p$ degrés de liberté.

Les intervalles de confiance bootstrap sont construits à partir des deux approches pourcentile et pourcentile$-t$. Pour un niveau $1 - 2\alpha$, l'intervalle de confiance pourcentile pour le paramètre β_j est donné par :

$$\left[\hat{\beta}_j^* - (\alpha B), \hat{\beta}_j^* (1 - \alpha) B\right] \tag{3.17}$$

où $\hat{\beta}_j^* (\alpha B)$ représente la $\alpha B - ième$ valeur (respectivement $\hat{\beta}_j^* ((1 - \alpha) B)$ la $(1 - \alpha B) - ième$ valeur) de la liste ordonnée des B réplications bootstrap. Les valeurs seuils sont donc choisies telles que $\alpha\%$ des réplications ont fourni des $\hat{\beta}_j^*$ plus petits (grands) que la borne inférieure (supérieure) de l'intervalle de confiance percentile.

La procédure bootstrap pourcentile-t consiste à estimer la fonction de répartition de z_j directement à partir des données.

Cela revient à construire une table statistique à partir de la fonction de répartition empirique des B réplications bootstrap z_j^*. Cette table est nommée table bootstrap. Les z_j^* sont définies comme :

$$z_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{s^*\left(\hat{\beta}_j^*\right)}. \tag{3.18}$$

Notons, en comparaison avec la méthode percentile, un calcul supplémentaire dans cette approche. En effet, pour chacune des réplications bootstrap, il est nécessaire de calculer l'écart type estimé bootstrap $s^*\left(\hat{\beta}_j^*\right)$.

Soit $\hat{F}_{z_j^*}$ la fonction de répartition empirique des z_j^*. Le fractile à $\alpha\%$, $\hat{F}_{z_j^*}^{-1}(\alpha)$ est estimé par la valeur $\hat{t}^{(\alpha)}$ telle que :

$$\#\left\{z_j^*(b) \leq \hat{t}^{(\alpha)}\right\}/B = \alpha.$$

Finalement, l'intervalle de confiance percentile-t pour β_j s'écrit :

$$\left[\hat{\beta}_j - s\left(\hat{\beta}_j\right)\hat{t}^{(1-\alpha)}, \hat{\beta}_j - s\left(\hat{\beta}_j\right)\hat{t}^{(\alpha)}\right]. \tag{3.19}$$

Ainsi, l'intervalle de confiance pourcentile-t est l'analogue bootstrap de l'intervalle de confiance standard.

En résumé, l'intervalle de confiance pourcentile-t substitue, aux valeurs critiques de la loi de Student utilisées dans l'intervalle standard, les valeurs seuils de la table bootstrap. Notons que ces dernières peuvent être très différentes. Cette différence est d'autant plus importante que la distribution (inconnue) des erreurs est éloignée de la loi normale. De plus, nous remarquons que les valeurs des quantiles α et $1-\alpha$ de la distribution de Student, symétriques par nature, entraînent directement la symétrie de l'intervalle de confiance standard autour de l'estimation $\hat{\beta}_j$. Par opposition, les valeurs $\hat{t}^{(\alpha)}$ et $\hat{t}^{(1-\alpha)}$ de la table bootstrap peuvent être asymétriques et permettent alors des intervalles de confiance asymétriques autour de $\hat{\beta}_j$. Cette prise en compte d'une possible asymétrie constitue un avantage important des intervalles de confiance bootstrap.

3.5.2 Les intervalles de prédiction bootstrap

Pour un nouveau rang f d'observation des variables explicatives X_f, la prédiction de coût $\hat{y}_f$ est calculée à partir du modèle de régression :

$$\hat{y}_f = X_f \hat{\beta}.$$

L'intervalle de prédiction standard découle, comme les intervalles de confiance des coefficients de la régression, de l'hypothèse de normalité des erreurs. Ainsi, pour un niveau de confiance $1 - 2\alpha$, cet intervalle de prédiction standard s'écrit

$$\left[\hat{y}_f - s_f.t_{(1-\alpha),n-p}, \hat{y}_f - s_f.t_{(\alpha),n-p}\right], \tag{3.20}$$

où, en notant s l'écart type résiduel, s_f est l'écart type estimé de l'erreur de prédiction, défini comme :

$$s_f = s\sqrt{1 + X_f(X^T X)^{-1} X_f^T}.$$

L'utilisation du bootstrap, pour préciser les intervalles de prédiction, conduit à étudier la distribution de l'erreur de prédiction. Aussi, afin de conserver le même processus générateur de données (PGD) pour les estimations des coefficients et des prédictions, les intervalles de prédiction bootstrap sont obtenues avec la procédure du bootstrap des résidus. De manière similaire à la construction des intervalles de confiance, il existe deux principales méthodes de construction des intervalles de prédiction bootstrap : l'approche pourcentile et l'approche pourcentile-t.

L'intervalle de prédiction pourcentile

La méthode pourcentile consiste à utiliser l'approximation bootstrap de la distribution de l'erreur de prédiction :

$$e_f = \hat{y}_f - y_f,$$

pour construire un intervalle de prédiction de y_f.

Les réplications bootstrap de la valeur y_f^*, pour le nouveau rang d'observations X_f, sont générées suivant le même modèle (3.15) :

$$y_f^* = X_f \hat{\beta} + \tilde{u}_f^*. \tag{3.21}$$

Le terme d'erreur $\tilde{u}_f^*$ est issu, comme les $\tilde{u}^*$, d'un tirage avec remise dans la distribution empirique des résidus transformés.

Pour chacune des B réplications bootstrap, nous calculons l'estimateur bootstrap. Ainsi, la prévision et l'erreur de prédiction bootstrap s'écrivent respectivement :

$$\hat{y}_f^* = X_f \hat{\beta}^* (b), \quad e_f^* (b) = \hat{y}_f^* (b) - y_f^* (b). \tag{3.22}$$

En utilisant l'équation (3.20), nous pouvons réécrire l'erreur de prédiction bootstrap comme :

$$e_f^* = \hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*. \tag{3.23}$$

Cette dernière dépend donc, par nature, de la prédiction MCO initiale $\hat{y}_f$.

Les B réplications bootstrap de l'erreur de prédiction fournissent la distribution empirique de e_f^* : G^*. Les quantiles de cette distribution empirique, notés $G^{*-1}(1-\alpha)$ et $G^{*-1}(\alpha)$, sont alors utilisés pour construire un intervalle de prédiction bootstrap.

Un intervalle de prédiction pourcentile est finalement de la forme suivante :

$$\left[\hat{y}_f - G^{*-1}(1-\alpha), \hat{y}_f - G^{*-1}(\alpha)\right]. \tag{3.24}$$

L'intervalle de prédiction pourcentile-t

De manière identique à l'intervalle de confiance, la construction de l'intervalle de prédiction avec la méthode pourcentile-t implique le calcul, pour chaque échantillon bootstrap, de l'estimateur bootstrap de l'écart type. Ainsi, pour établir des intervalles de prédiction pourcentile-t, l'estimateur bootstrap de l'écart type de prédiction est nécessaire, pour chacune des réplications. Il s'écrit :

$$s_f^* = s^*.\sqrt{(1-h_f)} \tag{3.25}$$

où s^* est l'estimateur bootstrap de l'écart type des termes erreurs et où

$$h_f = X_f(X^TX)^{-1}X_f^T.$$

La procédure percentile-t consiste à construire les statistiques z_f^*, telles que :

$$z_f^* = \frac{e_f^*}{s_f^*} = \frac{\hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*}{s_f^*}. \tag{3.26}$$

La distribution bootstrap de z_f^* définit l'intervalle de prédiction bootstrap pourcentile-t. Les quantiles $z_{f(1-\alpha)}^*$ et $z_{f(\alpha)}^*$ remplacent ainsi les valeurs critiques de la distribution de Student, prises en compte dans l'intervalle de prédiction standard (3.20).

Un intervalle de prédiction pourcentile-t s'écrit donc :

$$\left[\hat{y}_f - s_f.z_{f(1-\alpha)}^*, \hat{y}_f - s_f.z_{f(\alpha)}^*\right]. \tag{3.27}$$

Notons que, comme pour l'intervalle de confiance des coefficients, le quantile $1-\alpha$ de la distribution de z_f^* définit la borne inférieure de l'intervalle de prédiction et inversement pour le quantile α.

Une distribution symétrique de z_f^* implique donc la symétrie de l'intervalle de prédiction pourcentile-t. Cependant, dans le cas contraire, l'asymétrie est retranscrite de manière inversée pour ce dernier. Par exemple, si z_f^* possède une queue de distribution plus longue vers la droite, les quantiles $z_{f(1-\alpha)}^*$ et $z_{f(\alpha)}^*$ sont décalés vers les valeurs élevées des erreurs de prédiction bootstrap, comparativement aux quantiles correspondants d'une distribution symétrique. L'intervalle de prédiction pourcentile-t résultant est donc décalé vers la gauche, asymétrique autour de la valeur prédite MCO. Ainsi, sa construction implique une sorte de (correction automatique du biais) et permet l'acceptation, pour un niveau de confiance donné, de valeurs prédites plus faibles que l'intervalle de prédiction standard, symétrique.

3.5.3 Bootstrap par paires

Cette seconde approche bootstrap des modèles de régression consiste à rééchantillonner directement dans les données d'origine, à partir des paires (y_i, X_i). Notons cependant que le retirage simultané de (y_i, X_i) introduit une corrélation entre les régression et les erreurs du processus générateur de données (PGD) bootstrap.

L'application de la procédure bootstrap par paires consiste à répéter B fois les étapes suivantes :

1. à chaque itération b ($b = 1 : B$), le vecteur Y^* et la matrice des variables explicatives X^* sont construits, en effectuant n tirages aléatoires avec remise[2] de paires (y_i, X_i) dans l'échantillon d'origine. Ainsi, si le terme erreur u_i associé à X_i a une grande variance, la relation sera préservée dans l'échantillon bootstrap ;

2. une estimation par MCO des coefficients du modèle de régression bootstrap est ensuite réalisée :

$$\hat{\beta}^*(b) = \left(X^{*^T}(b)\, X^*(b)\right)^{-1} X^{*^T}(b)\, Y^*(b).$$

Notons, qu'à la différence de la procédure bootstrap des résidus, la matrice des variables explicatives $X^*(b)$ est différente à chaque itération b.

Les B réplications $\hat{\beta}^*$ fournissent alors la fonction de répartition empirique bootstrap. Ainsi, les B réplications bootstrap indépendantes, obtenues suivant les procédures de bootstrap présentées ci-dessus, fournissent un échantillon aléatoire des $\hat{\beta}^*$

[2]Notons que Freedman (1981) envisage des échantillons bootstrap de taille m différente de n.

qui est utilisé pour estimer la distribution bootstrap de $\hat{\beta}^*$. Cette dernière permet alors la construction des intervalles de confiance bootstrap des paramètres du modèle de régression.

Exemple 3.3 : La distribution des coefficients d'une Régression linéaire

Données brutes : Les données que l'on possède nous renseignent sur la consommation d'essence en litres (cs) d'une voiture suivant la distance (Km) parcourue par cette dernière. On a un échantillon de 10 observations.

Méthodes : On réalise une régression simple avec comme variable expliquée (observée) la consommation (cs) et comme variable explicative (prédicteur) (Km) la distance.

Le modèle : $y = X\beta + e, \quad y_i = ax_i + b + e_i \quad \forall i \in \{1, ..., n\}$.

Le modèle théorique bootstrap est : $y^* = X\hat{\beta} + e^*$

TAB.3.6 : Estimation, erreurs standards, biais et intervalles de confiance de modèle linéaire Bootstrappé :

Données	Obs	1	2	3	4	5	6	7	8	9	10
	Km	5	20	25	15	300	140	550	25	60	650
	cs	0.1	1.1	1.7	0.9	20.7	10.1	38.2	2.3	4.8	45.8

Résidus	Min	1Q	Médiane	3Q	Max
	−0.3555	−0.3062	−0.1555	0.2934	0.5484

		Estim.	Erreurs standards	valeur t	$Pr(>\|t\|)$
Coefficients	Intercept	0.0574285	0.1543543	0.372	0.72
	Km	0.0699026	0.0005325	131.262	$1.27e-14$

IC asympt		Normale	Student
	95%	$(-4.3254, 2.8364)$	$(-4.4475, 3.0122)$

Bootstrap		Original	Biais	Erreurs standards
	t_1^*	−0.738095	−0.16581440	1.85985246
	t_2^*	14.298973	0.02001199	0.09662666

$B = 99$	Normale	Basic	Student
	$(-4.5170, 3.0262)$	$(-4.6309, 3.1399)$	$(-4.6757, 3.1392)$
IC à 95%	Percentile	CBa	
	$(-4.6161, 3.1547)$	$(-5.2172, 3.0862)$	

- Ce modèle a un coefficient de qualité bootstrapée $R^2 = 0,9997$ très proche de 1 c'est à dire que la variation de (cs) est expliquée à 99.97% par la distance parcourue (Km), c'est donc un trés bon ajustement des données bootstrapées.

- L'estimation de la variance $\hat{\sigma}^2$ bootstrapée obtenue est faible donc tous les points sont très proches de la droite de régression.

- Les erreurs standards sont très faibles. La P-Value (probabilité critique) du test sur le coefficient correspondant au (Km) étant inférieure à 5% il est il est possible de dire que les Km ont significativement un effet sur la consommation.

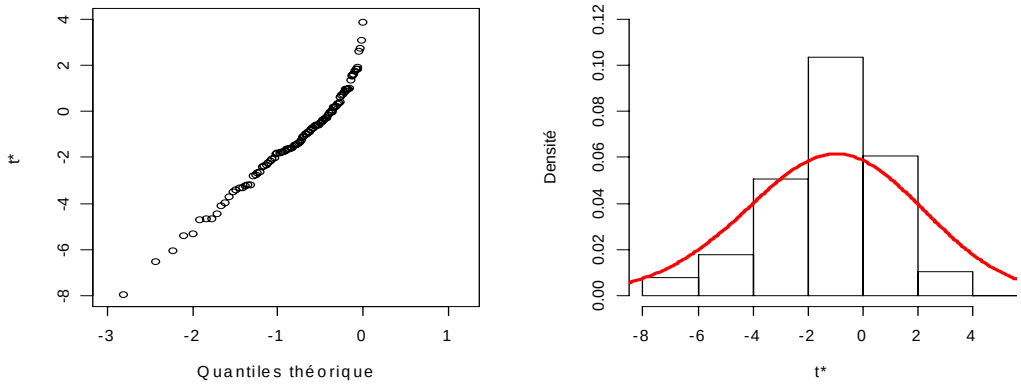


FIG. 3.4. Résidus de régression linéaire non paramétrique bootstrappée

Exemple 3.4 : -Précision des prédictions -

Calcul de l'erreur quadratique moyenne pour une régression classique avec bootstrap paramétrique.

Soit $x_1, ..., x_{200} \overset{i.i.d}{\sim} N(\theta, 1)$, pour $B = 999$ réplications bootstrap.

Le modèle

$$y = sin(x_1) + cos(x_2) - x_1x_2 + u$$

Le modèle théorique bootstrap est :

$$y^* = sin(x_1^*) + cos(x_2^*) - x_1^*x_2^* + u^*.$$

Le résultat est une liste dont essentiellement deux composantes nous intéresseront : t, qui contient toutes les estimations de la statistique à étudier et t_0 qui contient l'estimation pour tout l'échantillon.

TAB. 3.7 : Biais, variance, erreurs quadratique du modèle linéaire Bootstrappé :

t	Biais	Variance	EQM
$[1,]$	-0.0193045633	2.193542	2.182947
$[2,]$	-0.0226059880	2.442703	2.431001
$[3,]$	-0.0404100680	2.942414	2.929335
...	...	...	...
$[998,]$	-0.2876407286	2.143876	2.215894
$[999,]$	0.2305238002	3.100343	3.137983

	Biais	Variance	EQM
t_0	0.0009416182	2.3026197828	2.3140859706

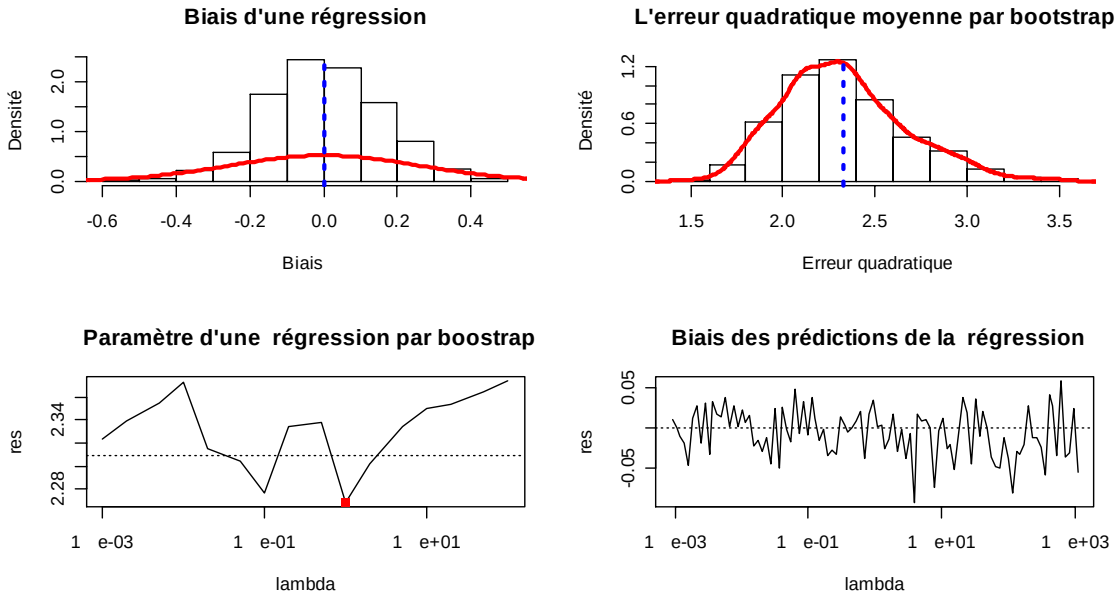


FIG. 3.5. Graphes des résidus de régression bootstrappée

Chapitre 4

BOOTSTRAP POUR LES VALEURS EXTRÊMES

Dans ce chapitre nous présentons quelques méthodes d'estimation des quantiles extrêmes et nous appliquons la méthode du *bootstrap* aux queues de distributions, aux valeurs extrêmes, et en particulier dans l'estimation de l'index de Pareto.

4.1 Méthodes classiques d'estimation des quantiles extrêmes

Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires réelles i.i.d. de distribution F inconnue

$$F(x) = P(X_1 \leq x), \quad x \in \mathbb{R}.$$

Pour chaque entier $n \geq 1$, on note par $X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}$ la statistique d'ordre associée à $X_1, X_2, ..., X_n$; avec :

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i<x\}}, \quad x \in \mathbb{R} \\ &= \frac{i}{n}, \quad \text{pour} \quad x \in]X_{i,n}, X_{i+1,n}], \end{aligned}$$

où $0 \leq i \leq n$, $X_{0,n} = -\infty$, $X_{n+1,n} = +\infty$, la fonction de distribution empirique basée sur les n premières variables aléatoires. On définit la fonction des quantiles Q par :

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

où F^{-1} est l'inverse généralisée de la fonction de distribution F. La fonction des quantiles empirique Q_n est définie pour $n \geq 1$ par :

$$Q_n(s) = \inf\{x \in \mathbb{R},\ F_n(x) \geq s\} = \begin{cases} X_{0,n} & \text{pour } s < 0, \\ X_{i,n} & \text{pour } \frac{i-1}{n} < s \leq \frac{i}{n}, \quad 1 \leq i \leq n-1, \\ X_{n,n} & \text{pour } s > 1. \end{cases}$$

On note par x_p les quantiles extrêmes définis par :

$$x_p = F^{-1}(1-p) = Q(1-p), \quad \text{quand } p \to 0.$$

On définit la fonction des quantiles du queue U par :

$$U(s) := Q\left(1 - \frac{1}{s}\right), \quad \text{quand } s \to \infty,$$

notons que

$$x_p = U\left(\frac{1}{p}\right), \quad \text{quand } p \to 0.$$

Estimation du p-quantile

Soit x_p le quantile de la loi de X_i et $\hat{x}_p$ son estimateur donnée par :

$$\hat{x}_p = \begin{cases} X_{(k+1)} & \text{si } \frac{k}{n} < p < \frac{k+1}{n}, \\ \dfrac{X_{(k)} + X_{(k+1)}}{2} & \text{si } p = \frac{k}{n}. \end{cases}$$

4.1.1 Méthode des valeurs extrêmes

Théorème 4.1.1 : *On suppose une distribution F* qui *satisfait la condition suivante : il existe deux suites de nombres réels (a_n) et (b_n), $n \in \mathbb{N}$, avec $a_n > 0$ et $b_n \in \mathbb{R}$, tels que :*

$$\forall x \in \mathbb{R} \quad , \lim_{n\to\infty} F^n(a_n x + b_n) = \lim_{n\to\infty} P\left\{\frac{X_{n,n} - b_n}{a_n} \leq x\right\} = G_\gamma(x),$$

où G_γ est la fonction de répartition de la loi des valeurs extrêmes :

$$G_\gamma(x) := \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right) & \text{pour tout } x \text{ tel que } 1+\gamma x > 0 \text{ si } \gamma \neq 0, \\ \exp(-e^{-x}) & \text{pour tout } x \in \mathbb{R}\,; \quad \text{si } \gamma = 0. \end{cases}$$

Lorsque $\gamma \neq 0$, $1 + \gamma x \leq 0$, alors $G_\gamma(x) = 0$.

On dit que la fonction de répartition F est dans le domaine d'attraction de Fréchet, de Gumbel ou de Weibull respectivement si $\gamma > 0$, $\gamma = 0$ ou $\gamma < 0$.

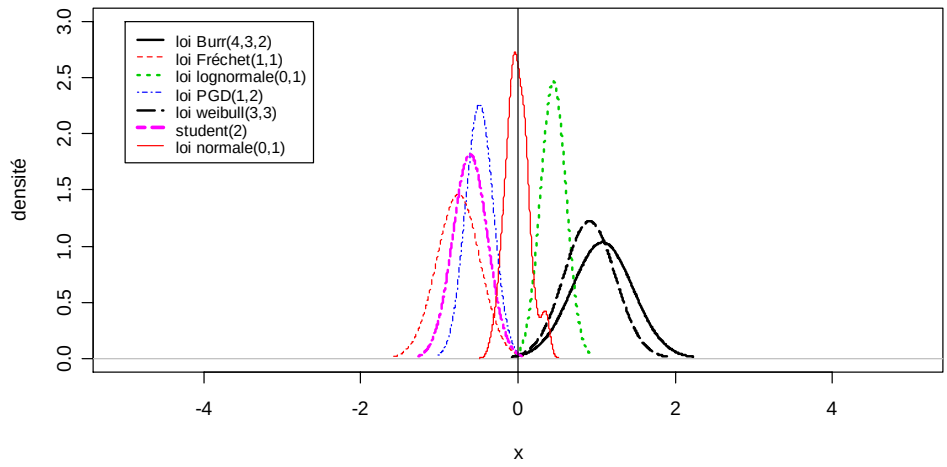


FIG. 4.1. Densités des lois extrêmes bootstrapés.

4.1.2 Méthode des excès

Soit u un réel suffisamment grand et inférieur au point limite supérieur[1] de F $(u < w(F))$, appelé seuil. La méthode des excès s'appuie sur une approximation de la loi des excès au dessus du seuil u de la v.a.r X. On note par F_u la distribution des excès au-delà du seuil u définie par :

$$F_u(y) := P(X - u \leq y \mid X > u) = \begin{cases} \dfrac{F(u+y) - F(u)}{1 - F(u)} & \text{si} \quad y \geq 0, \\ 0 & \text{si} \quad y < 0. \end{cases}$$

La loi asymptotique des excès est donnée par le théorème suivant :

Théorème 4.1.2 (Pickands (1975)) : *On Considère la probabilité conditionnelle de X ne dépasse pas $t + x$ sachant que X dépasse t, i.e., on considère*

$$F_t(x) = \frac{F(t+x) - F(t)}{1 - F(t)}, \quad x > 0.$$

[1] On appelle point limite supérieur d'une fonction de répartition F le réel (fini ou infini) défini par $w(F) = \sup\{x : F(x) < 1\}$.

Alors pour toute fonction de distribution continue F : $F \in D(G_\gamma)$ *pour* $\gamma \in \mathbb{R}$ *si et seulement si, pour une fonction positive* $\sigma(.)$

$$\lim_{t \uparrow x_F^*} \sup_{0<x\leq x_F^*-t} \left|F_t(x) - G_{\gamma,\sigma(t)}(x)\right| = 0,$$

où $G_{\gamma,\sigma}(x)$ *est la distribution de Pareto Généralisée* (DPG), *définie pour* $x > 0$ *et* $\sigma > 0$ *par :*

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \gamma\dfrac{x}{\sigma}\right)^{-1/\gamma} & \text{pour } \gamma \neq 0, \quad 1 + \gamma x/\sigma > 0, \\ 1 - e^{-x/\sigma} & \text{pour } \gamma = 0. \end{cases}$$

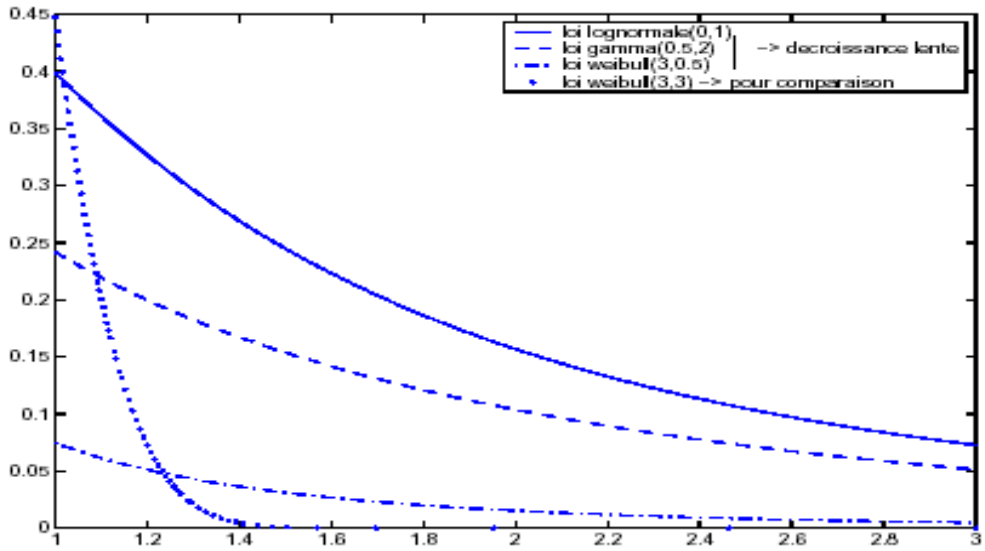


FIG. 4.2. Exemples de queues de distributions lourdes $\mathcal{LN}(0,1), Gamma(0.5,2)$, discontinu $\mathcal{W}(3,0.5)$ et loi $\mathcal{W}(3,3)$.

Théorème 4.1.3 : *Soit* F *une fonction de distribution et* $Q(s)$ *la fonction des quantiles. Alors* $F \in D(G_\gamma)$ *si et seulement si, il existe une fonction positive* $a(\cdot)$ *telle que pour tout* $x > 0$ on a :

$$\lim_{s\downarrow 0} \frac{Q(1-sx) - Q(1-s)}{a(s)} = \begin{cases} \dfrac{x^{-\gamma}-1}{\gamma}, & \gamma \neq 0, \\ -\log x, & \gamma = 0. \end{cases}$$

Si on définit une fonction $U(\cdot)$ *par :*

$$U(x) := \left(\frac{1}{1-F}\right)^{-1}(x), \quad 0 < x < \infty.$$

Alors $F \in D(G_\gamma)$ *si et seulement si pour tout* $x > 0$,

$$\lim_{t \to \infty} \frac{U(tx) - U(t)}{\widetilde{a}(t)} = \begin{cases} \dfrac{x^\gamma - 1}{\gamma}, & \gamma \neq 0 \\ \log x, & \gamma = 0 \end{cases}$$

avec $\widetilde{a}(t) = a(1/t)$.

4.1.3 Méthode des quantiles

Soit F_n la fonction de distribution empirique et soit U_n la fonction des quantiles empiriques des queues définie par :

$$U_n(s) := Q_n\left(1 - \frac{1}{s}\right),$$

notons :

$$F_n(X_{k,n}) = \frac{k}{n}, \quad k = 1, 2, ..., n,$$

et

$$X_{n-k+1,n} = U\left(\frac{n}{k}\right), \quad k = 1, 2, ..., n.$$

En remplaçant U par son équivalent empirique U_n, pour x, $y > 0$, $y \neq 1$, alors

$$\lim_{t \to \infty} \frac{U(tx) - U(t)}{U(ty) - U(t)} = \frac{x^\gamma - 1}{y^\gamma - 1}.$$

Dans le paragraphe suivant on présente un résultat sur la transformation des quantiles, qui est nécessaire dans la simulation d'échantillon et par suite dans l'estimation des p-quantiles.

Transformation des quantiles

Soit $X_1, ..., X_n$ une suite de variables aléatoires i.i.d. de distribution F, et $U_1, ..., U_n$ une suite de variables aléatoires i.i.d. uniformément distribuées sur $[0, 1]$, notant par $U_{(1)}, ..., U_{(n)}$ la statistique d'ordre correspondante, alors

- $F^{-1}(U_1) \overset{\mathcal{D}}{=} X_1$.
- Pour tout $n \in \mathbb{N}$, $(X_{(1)},...,X_{(n)}) \overset{\mathcal{D}}{=} (F^{-1}(U_{(1)}),...,F^{-1}(U_{(n)}))$.
- La variable aléatoire $F(X_1)$ a une distribution uniforme sur $[0,1]$ si et seulement si F est une fonction continue.

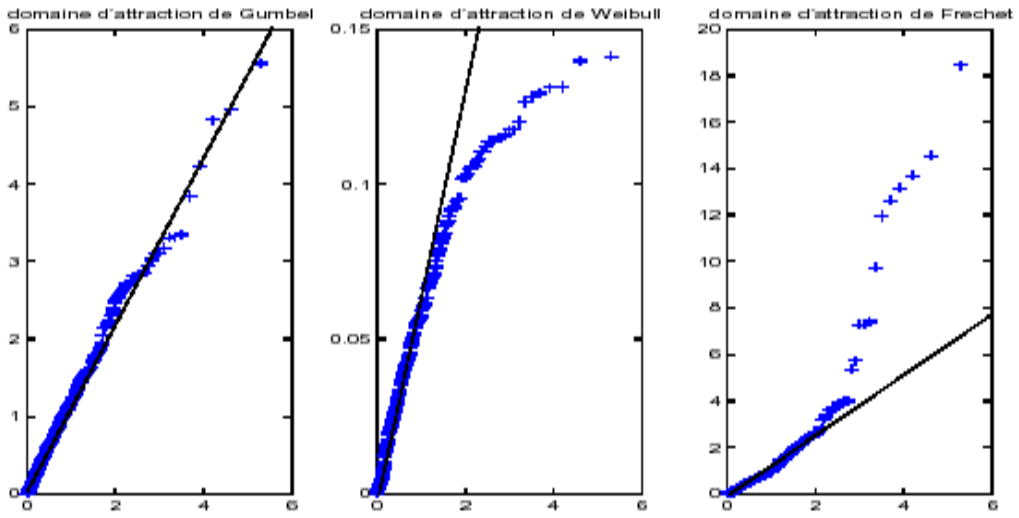


FIG. 4.3. qqplot dans le DA Gumbel (loi normale), DA Weibull (loi beta) et DA Fréchet (loi de Student) pour des echantillons de taille $n = 1000$ et $K_n = 200$.

Estimateur de Hill

Cet estimateur est basé sur la condition que la queue de la distribution F est telle que $:1 - F(x) \sim cx^{-1/\gamma}$ quand $x \to \infty$ pour $\gamma > 0$ et $c > 0$. Donc l'estimateur de Hill (1975) [29] est applicable seulement dans le cas où l'index des valeurs extrêmes est positif et la fonction de distribution F est a queue lourde [1] .

On définit l'estimateur de Hill de l'index de la queue γ par:

$$\widehat{\gamma}_n^{(H)}(k_n) = k_n^{-1} \sum_{i=1}^{k_n} \log\left(X_{n-i+1,n}/X_{n-k_n,n}\right), \tag{4.1}$$

[1] On dit que la fonction de distribution F est a queue lourde si la fonction du queue $\overline{F} := 1-F$ est a variation régulière en $+\infty$ d'index $-1/\gamma$, $\overline{F}(x) = x^{-1/\gamma}L(x)$,où γ est l'index des valeurs extrêmes et L est une fonction à variation lente.

où k_n est une suite des entiers positifs satisfaisant les conditions suivantes :

$$1 \leq k_n < n, \quad k_n \to \infty \quad \text{et} \quad k_n/n \to 0 \quad \text{quand} \quad n \to \infty.$$

Notons de plus, que cet estimateur peut être représenté en terme des espacements des logarithmes des observations :

$$\widehat{\gamma}_n^{(H)}(k_n) := k_n^{-1} \sum_{i=1}^{k_n} i \left(\log X_{n-i+1,n} - \log X_{n-i,n} \right).$$

On suppose que $1 - F$ est de type Paréto avec l'index $\gamma > 0$, tel que

$$1 - F(x) = x^{-\gamma} \tilde{L}(x), \quad x \to \infty,$$

où $\tilde{L}$ est une fonction a variation lente [2] au voisinage de l'infini ;

$$F^{-1}(1 - x) = x^{-1/\gamma} \tilde{L}(x), \quad x \to \infty.$$

Pour tout n fixé, on définit l'estimateur de Hill bootstrapé par :

$$1/\widehat{\gamma}_n^{*(H)}(k_n) := k_n^{-1} \textstyle\sum_{i=1}^{k_n} \log X^*_{n-i+1,n} - \log X^*_{n-i,n}. \tag{4.2}$$

Nous avons simulé un échantillon de taille $n = 1500$ de la loi de Student : $t(\nu)$ à $\nu = 2$ degré de liberté. Pour la vraie valeur $\gamma = 1/\nu$ et $\rho = -2/\nu$, la figure (4.4) présente les valeurs moyennes bootstrapées obtenues pour les estimateurs de Hill en fonction de $\alpha = theta$, $theta = \frac{1}{m}$ à $\frac{m}{m}$ avec $m = 1000$ et en fonction de kn; $kn = k_n = [n^\alpha]$ varié de 1 à n et $B = 100$ réplications bootstrap.

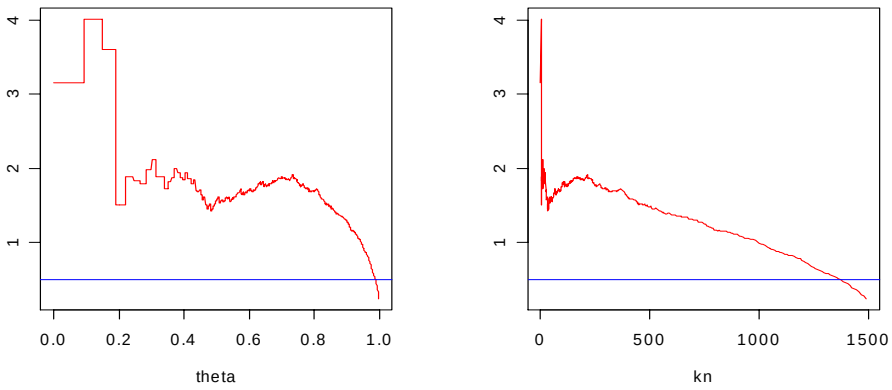


FIG. 4.4. Estimateur de Hill Bootstrapé pour $n = 1500$, $\gamma = 1/2$, $B = 100$.

[2] On dit que l est à variation lente au voisinage de ∞ si $\lim_{t\to\infty} (l(tx)/l(t)) = 1, \ x > 0$.

Bootstrap pondéré

Soit $X_n = (X_1, ..., X_n)$ n variables aléatoires i.i.d. de la loi de probabilité P, supposée inconnue, on définit la mesure empirique par $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, où les δ_{X_i} sont des masses de dirac en X_i.

Conditionnellement à $(X_1, ..., X_n)$ une suite d'échantillons bootstrap de taille k_n, lorsque n varie, est un tableau triangulaire $X_k^{(n)} = \left(X_1^{(n)}, ..., X_n^{(n)}\right)$ de v.a et i.i.d. de loi de probabilité P_n.

Soit $W^{(n)} := \left\{W_i^{(n)} : 1 \leq i \leq n\right\}$ des variables aléatoires échangeables (dont la loi jointe est invariante par permutation) conditionnellement à l'échantillon $(X_1, ..., X_n)$ et telles que

$$\sum_{i=1}^{n} W_i^{(n)} = k_n.$$

Ce système de poids peut aussi s'interpréter comme un plan de rééchantillonnage (*voir Efron (1979)* [15]). La probabilité bootstrap au sens de *Mason-Newton* [38] est donnée par

$$P_{W,k_n}^{(n)} := k_n^{-1} \sum_{i=1}^{n} W_i^{(n)} \delta_{X_i},$$

pondération des masses de dirac par des poids aléatoires.

Lorsque les prennent des valeurs entières, $P_{W,k_n}^{(n)}$ est en fait, l'analogue de la probabilité empirique de l'échantillon bootstrap

$$P_{k_n}^{(n)} := k_n^{-1} \sum_{i=1}^{n} \delta_{X_i^{(n)}},$$

dans lequel X_i apparaîtrait $W_i^{(n)}$ fois. Le bootstrap d'Efron consistant à effectuer un tirage avec remise dans les $(X_i)_{1\leq i\leq n}$ s'obtient en choisissant des poids $\left(W_i^{(n)}\right)_{1\leq i\leq n}$ de loi multinomiale $Mult\,(k_n, 1/n)$; il n'est donc qu'un cas particulier de cette méthode.

4.2 Test d'adéquation pour les queues de distributions

Cependant, lorsque l'on s'intéresse à l'estimation des événements rares, les tests usuels (par exemple Anderson- Darling ou Cramér- von Mises) ne permettent pas de déceler une mauvaise estimation de la queue de distribution, c'esp pour ça, on propose un test par la distance $\mathbb{L}^2$-Wasserstein entre deux fonctions des quantiles, qui permet, dans le cadre restreint du domaine d'attraction où $\gamma > 0$ et pour la classe de distributions a queues lourdes.

4.2.1 Test d'adéquation via la distance L^2-Wasserstein pondérés

Soit $\Im$ un ensemble de fonctions de distributions G tels que

$$\int_0^1 \left(G^{-1}(s)\right)^2 ds < \infty, \tag{4.3}$$

où $G^{-1}(t) := \inf\{x : g(x) \geq t\}$, $0 < t < 1$, désigne l'inverse généralisé ou la fonction des quantiles. La distance $\mathbb{L}^2$-Wasserstein entre deux fonctions de distributions G_1 et G_2 (voir, *Bickel* et *Freedman (1981)* [4] ou *Shorack* et *Wellner (1986), page 63)* [41], est donnée par :

$$d^2(G_1, G_2) := \int_0^1 \left\{G_1^{-1}(s) - G_2^{-1}(s)\right\}^2 ds.$$

Pour la classe de distributions a queues lourdes (par exemple, Pareto, log-normal, log-gamma, distributions de Cauchy) la condition (4.3) n'est pas vérifiée, dans ce cas, nous proposons une autre version pour $d^2(G_1, G_2)$ comme suit :

$$d_w^2(G_1, G_2) := \int_0^1 w(s) \left\{G_1^{-1}(1-s) - G_2^{-1}(1-s)\right\}^2 ds,$$

pour tout G_1 et $G_2 \in \Im_w$, où $\Im_w$ est l'ensemble des fonctions de distributions G tels que

$$\int_0^1 w(s) \left(G^{-1}(s)\right)^2 ds < \infty,$$

avec $w(.)$ est une fonction de poids souhaitable sur $(0, 1)$ tels que

$$\int_0^1 w(s)\, ds = 1.$$

$\mathbb{L}^2$-Wasserstein pour la statistique des queues pondérées

Soit $X_1, X_2, ...$une suite des variables aléatoires i.i.d. d'une fonction de distribution commune $F(x) = P(X \leq x)$, $x \in \mathbb{R}$. Nous supposons, que $1 - F$ est de type Pareto avec l'index $\gamma > 0$, tel que

$$1 - F(x) = x^{-1/\gamma} \tilde{L}(x), \quad x \to \infty, \tag{4.4}$$

où $\tilde{L}$ est une fonction a variation lente au voisinage de l'infini.

Pour vérifier la validité de (4.4), nous proposons la distance $d_w^2(F_n, F)$, pour une fonction de poids

$$w_{\delta,k}(s) := \begin{cases} (1+2\delta)(k/n)^{-(2\delta+1)} s^{2\delta}, & 0 \le s \le k/n, \\ 0 & ailleurs, \end{cases}$$

où $\delta > 0$ est une constante et $k := k_n$, $n \ge 1$ est une suite des entiers positifs satisfaisant les conditions $1 \le k_n < n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$.

Il est clair que $\int_0^1 w_{\delta,k}(s)ds = 1$, pour tout $\delta > 0$. La distance correspondante est

$$d_{\delta,k}^2(F_n, F) := \int_0^{k/n} \frac{(2\delta+1)s^{2\delta}}{(k/n)^{(2\delta+1)}} \left\{F_n^{-1}(1-s) - F^{-1}(1-s)\right\}^2 ds.$$

Ceci peut être réécrit, pour $\delta > 0$

$$d_{\delta,k}^2(F_n, F) := (2\delta+1)\int_0^1 s^{2\delta} \left\{F_n^{-1}(1-ks/n) - F^{-1}(1-ks/n)\right\}^2 ds.$$

Finalement, on suppose que la fonction de la queue $\bar{F} = 1 - F$ est à variation régulière au second ordre dont le premier indice est $-1/\gamma$ et le second étant $\rho \le 0$. Par conséquent, il existe une fonction $A^*(t) \to 0$ quand $t \to 0$ et qui a un signe constant :

$$\lim_{t\to\infty} \frac{1}{A^*(t)} \left\{ \frac{1-F(tx)}{1-F(t)} - x^{-1/\gamma} \right\} = x^{-1/\gamma} \frac{x^\rho - 1}{\rho}, \quad \text{pour tout } x > 0. \tag{4.5}$$

Si $\rho = 0$, on remplace $\dfrac{x^\rho - 1}{\rho}$ par $\log x$. Par simplicité, on pose

$$A(t) := A^*\left[F^{-1}(1-1/t)\right], \quad t > 1,$$

où $F^{-1}(1-1/t)$ est la fonction des quantiles de la queue. Notons que $|A|$ est à variation régulière à l'infini d'indice $\rho\gamma$, avec $A(t) \to 0$, quand $n \to \infty$. (voir, par exemple, de Haan et Stadtmüler, pour plus de détails sur les propriétés de la fonction $A(.)$ et de la variation régulière du deuxième ordre).

Théorème 4.2.1 [40] **:** *Supposons que* (4.5) *est satisfaite avec* $\gamma > 0$. *Soit* $k = k_n$ *telle que* $k \to \infty$, $k/n \to 0$, $kA(n/k) \to 0$ *quand* $n \to \infty$. *Alors pour tout,* $\delta > \gamma$, *nous avons*

$$\frac{k}{\gamma^2 (2\delta + 1) \left[F^{-1}(1 - k/n)\right]^2} d^2_{\delta,k}(F_n, F) \overset{\mathcal{D}}{=} \int_0^1 s^{2(\delta-\gamma-1)} W^2(s)\, ds + o_p(1),$$

quand $n \to \infty$, *où* W *est le processus de Wiener standard.*

Preuve

Soit $\xi_1, \xi_2, ...,$ une suite de v.a indépendantes et uniformément distribuées sur $(0, 1)$. Pour tout entier $n \geq 1$, la fonction des quantiles uniforme est définie par

$$\mathbb{V}_n(t) = \xi_{i,n}, \qquad \text{pour } (i-1)/n < t \leq i/n, \quad i = 1, ..., n,$$

avec $\mathbb{V}_n(0) = \xi_{1,n}$, où $\xi_{1,n} \leq ... \leq \xi_{n,n}$ désigne les statistiques d'ordre basées sur $\xi_1, ..., \xi_n$. Supposons sans perte de généralité que les variables aléatoires $(X_n)_{n\geq 1}$ sont définies sur un espace de la probabilité $(\Omega, \mathcal{A}, P)$ sur lequel emporte la suite $(\xi_n)_{n\geq 1}$ tel que $X_n = F^{-1}(\xi_n)$, pour $n = 1, 2, ...,$ et par conséquent, $X_{i,n} = F^{-1}\left(\xi_{i,n}\right)$ pour tout $1 \leq i \leq n$ et $n \geq 1$. Observons que pour tout entier $n \geq 1$, nous avons $\mathbb{V}_n(1 - i/n) = \xi_{n-i+1,n}$, et pour tout $0 < s < 1$

$$F^{-1}(1 - \mathbb{V}_n(1 - i/n)) \overset{\mathcal{D}}{=} Q(\mathbb{V}_n(1 - i/n)) \overset{\mathcal{D}}{=} X_{n-i+1,n}, \quad i = 1, 2, ...$$

Pour tout $\delta > 0$

$$\begin{aligned} d^2_{\delta,k}(F_n, F) &:= \int_0^{k/n} \frac{(2\delta + 1)\, s^{2\delta}}{(k/n)^{(2\delta+1)}} \left\{F_n^{-1}(1 - s) - F^{-1}(1 - s)\right\}^2 ds \\ &= (2\delta + 1) \int_0^1 s^{2\delta} \left\{Q_n(1 - ks/n) - Q(1 - ks/n)\right\}^2 ds. \end{aligned}$$

On utilise les techniques utilisées par *Necir et Boukhetala (2004)* [39] ou *Greoneboon et al. (2003)* (Lemme 4.1), nous écrivons

$$\begin{aligned} \frac{k d^2_{\delta,k}(F_n, F)}{Q^2(1 - k/n)} &= (2\delta + 1)\, k \int_0^1 s^{2\delta} \frac{Q^2(1 - ks/n)}{Q^2(1 - k/n)} \\ &\quad \times \left[\frac{Q(\mathbb{V}_n(1 - ks/n))}{Q(1 - ks/n)} - 1\right]^2 ds, \\ &= (2\delta + 1)\left\{S_n + T_n\right\}, \end{aligned}$$

où

$$S_n := k\int_0^{1/k} s^{2\delta}\frac{Q^2\left(1-ks/n\right)}{Q^2\left(1-k/n\right)}\left[\frac{Q_n\left(1-ks/n\right)}{Q\left(1-ks/n\right)}-1\right]^2 ds,$$

et

$$T_n := k\int_{1/k}^1 s^{2\delta}\frac{Q^2\left(1-ks/n\right)}{Q^2\left(1-k/n\right)}\left[\frac{Q_n\left(1-ks/n\right)}{Q\left(1-ks/n\right)}-1\right]^2 ds.$$

Ensuite, on prouve que $S_n = o_p(1)$, quand $n \to \infty$. C'est clair que (4.**3**) implique que

$$Q\left(1-s\right) = s^{-\gamma}L(s), \quad s \downarrow 0$$

où L est une fonction à variation lentement au voisinage de zéro. D'après la représentation de Karamata (voir, e. g., *Seneta (1976)*), nous avons

$$L(u) = \exp\left(\int_u^1 \frac{b\left(s\right)}{s}ds\right),$$

au voisinage de zéro, où $b\left(\cdot\right)$ est une fonction réelle telle que $b\left(s\right) \to 0$ quand $s \downarrow 0$, alors, il est facile de vérifier que, pour chaque $\varepsilon > 0$

$$\frac{Q\left(1-s\right)}{Q\left(1-k/n\right)} = s^{-\gamma-\varepsilon}\left(\frac{k}{n}\right)^{\gamma-\varepsilon}, \quad \text{quand} \quad n \to \infty.$$

D'autre part, on a :

$$Q_n\left(1-s\right) \overset{\mathcal{D}}{=} Q\left(1-\xi_{1,n}\right), \quad \text{pour tout } 0 < s \leq 1/n,$$

ensuite, pour tout $\varepsilon > 0$ et pour tout $0 < s \leq 1/n$,

$$\frac{Q_n\left(1-s\right)}{Q\left(1-s\right)} \overset{\mathcal{D}}{=} \left(\frac{\left(\xi_{1,n}\right)}{s}\right)^{-\gamma\pm\varepsilon}, \quad \text{quand } n \to \infty.$$

Donc, pour chaque $\varepsilon > 0$

$$\begin{aligned} S_n &= (k/n)^{-2\delta}\int_0^{1/n} s^{2\delta}\frac{Q^2\left(1-s\right)}{Q^2\left(1-k/n\right)}\left[\frac{Q_n\left(1-s\right)}{Q\left(1-s\right)}-1\right]^2 ds \\ &\overset{\mathcal{D}}{=} (k/n)^{2(\gamma-\delta-\varepsilon)}\int_0^{1/n} s^{2(\delta-\gamma-\varepsilon)}\left[\left(\frac{\left(\xi_{1,n}\right)}{s}\right)^{\gamma\pm\varepsilon}-1\right]^2 ds \\ &= S_{n1}+S_{n2}+S_{n3}, \end{aligned}$$

où

$$S_{n1} := (k/n)^{2(\gamma-\delta-\varepsilon)} \int_0^{1/n} s^{2(\delta-\gamma-\varepsilon)} \left(\frac{\left(\xi_{1,n}\right)}{s} \right)^{2(\gamma\pm\varepsilon)} ds,$$

$$S_{n2} := -2\,(k/n)^{2(\gamma-\delta-\varepsilon)} \int_0^{1/n} s^{2(\delta-\gamma-\varepsilon)} \left(\frac{\left(\xi_{1,n}\right)}{s} \right)^{\gamma\pm\varepsilon} ds,$$

et

$$S_{n3} := (k/n)^{2(\gamma-\delta-\varepsilon)} \int_0^{1/n} s^{2(\delta-\gamma-\varepsilon)} ds.$$

Rappelons que $n\left(\xi_{1,n}\right) \xrightarrow{p} 1$ quand $n \to \infty$. Par suit, pour tout n suffisamment grand nous obtenons

$$\begin{aligned} S_{n1} &= O_p(1)\,(k)^{2(\gamma-\delta-\varepsilon)}\, n^{-1}, \\ S_{n2} &= O_p(1)\,(k)^{2(\gamma-\delta-\varepsilon)}\, n^{-1}, \end{aligned}$$

et

$$S_{n3} = \frac{1}{2\,(\delta-\gamma-\varepsilon)+1} k^{2(\gamma-\delta-\varepsilon)}.$$

Maintenant il est clair que :

$$S_{n1} + S_{n2} + S_{n3,} = o_p(1), \text{ quand } n \to \infty,\ \delta > \gamma > 0.$$

Après nous nous servons de l'expression suivante pour obtenir le comportement asymptotique de T_n. Pour tous réels a, b, c et d, nous avons :

$$a^2\,(b-1)^2 = [(a-d)+d]^2\,[(b-c)+(c-1)]^2 = \sum\nolimits_{i=1}^{9} p_i$$

où

$$\begin{aligned} p_1 &:= (a-d)^2\,(b-c)^2, & p_2 &:= 2\,(a-d)^2\,(b-c)\,(c-1), \\ p_3 &:= (a-d)^2\,(c-1)^2, & p_4 &:= 2d\,(a-d)\,(b-c)^2, \\ p_5 &:= 4d\,(a-d)\,(b-c)\,(c-1), & p_6 &:= 2d\,(a-d)\,(c-1)^2, \\ p_7 &:= d^2\,(b-c)^2, & p_8 &:= 2d^2\,(b-c)\,(c-1), \end{aligned}$$

et

$$p_9 := d^2 (c-1)^2 .$$

Cela, nous permet de réécrire a, b, c et d de la manière suivante :

$$a := \frac{Q(1-ks/n)}{Q(1-k/n)}, \qquad b := \frac{Q_n(1-ks/n)}{Q(1-ks/n)},$$

$$c := \left(\frac{1-\mathbb{V}_n(1-ks/n)}{ks/n}\right)^{-\gamma}, \quad d := s^{-\gamma},$$

nous obtenons

$$T_n = \sum\nolimits_{i=1}^{9} T_{ni},$$

où

$$T_{n1} := k\int_{1/k}^{1} s^{2\delta} p_1 ds, \quad T_{n2} := 2k\int_{1/k}^{1} s^{2\delta} p_2 ds,$$

$$T_{n3} := k\int_{1/k}^{1} s^{2\delta} p_3 ds, \quad T_{n4} := 2k\int_{1/k}^{1} s^{2\delta} p_4 ds,$$

$$T_{n5} := 4k\int_{1/k}^{1} s^{2\delta} p_5 ds, \quad T_{n6} := 2k\int_{1/k}^{1} s^{2\delta} p_6 ds,$$

$$T_{n7} := k\int_{1/k}^{1} s^{2\delta} p_7 ds, \quad T_{n8} := 2k\int_{1/k}^{1} s^{2\delta} p_8 ds,$$

et

$$\begin{aligned} T_{n9} &:= k\int_{1/k}^{1} s^{2\delta} p_9 ds = k\int_{1/k}^{1} s^{2\delta} d^2 (c-1)^2 ds \\ &:= k\int_{1/k}^{1} s^{2(\delta-\gamma)} \left\{\left(\frac{1-(\mathbb{V}_n(1-ks/n))}{ks/n}\right)^{-\gamma} - 1\right\}^2 ds. \end{aligned}$$

Puisque $kA(k/n) \to 0$, il est facile de vérifier que $T_{ni} = o(1)$, $i = 1, ...8$, quand $n \to \infty$, et

$$\begin{aligned} T_{n9} &= k(k/n)^{2\gamma} \int_{1/k}^{1} s^{2\delta} \left\{(1-(\mathbb{V}_n(1-ks/n)))^{-\gamma} - (ks/n)^{-\gamma}\right\}^2 ds \\ &= k(k/n)^{2\gamma-2\delta-1} \int_{1/n}^{k/n} s^{2\delta} \left\{(1-(\mathbb{V}_n(1-s)))^{-\gamma} - s^{-\gamma}\right\}^2 ds. \end{aligned}$$

En utilisant le théorème de la valeur moyenne, nous obtenons

$$T_{n9} = \gamma^2 k \left(k/n\right)^{2\gamma-2\delta-1} \int_{1/n}^{k/n} s^{2\delta} \left(s + \xi_{s,n}\right)^{-2(\gamma+1)} \left\{1 - \mathbb{V}_n\left(1-s\right) - s\right\}^2 ds,$$

où $\left|\xi_{s,n}\right| \leq \left|\mathbb{V}_n\left(s\right) - s\right|$.

Soit $\theta_n := k^{1/2}/n$ une suite des nombres réels positifs tel que $1/n < \theta_n < k/n$. Alors T_{n9} peut être décomposé comme suit

$$T_{n9} = \tilde{T}_{n9} + T^*_{n9},$$

où

$$\tilde{T}_{n9} := \gamma^2 k \left(k/n\right)^{2\gamma-2\delta-1} \int_{1/n}^{\theta_n} s^{2\delta} \left(s + \xi_{s,n}\right)^{-2(\gamma+1)} \left\{1 - \mathbb{V}_n\left(1-s\right) - s\right\}^2 ds,$$

et

$$T^*_{n9} := \gamma^2 k \left(k/n\right)^{2\gamma-2\delta-1} \int_{\theta_n}^{k/n} s^{2\delta} \left(s + \xi_{s,n}\right)^{-2(\gamma+1)} \left\{1 - \mathbb{V}_n\left(1-s\right) - s\right\}^2 ds.$$

Observons que

$$\tilde{T}_{n9} = \gamma^2 k \left(k/n\right)^{2(\gamma-\delta)-1} \int_{1/n}^{\theta_n} s^{2(\delta-\gamma)} \left(\frac{s}{s + \xi_{s,n}}\right)^{2(\gamma+1)} \left\{\frac{1 - \mathbb{V}_n\left(1-s\right) - s}{s}\right\}^2 ds.$$

Puisque $\theta_n \to 0$ et $n\theta_n \to \infty$ quand $n \to \infty$, d'après *Wellner (1978)* [43] nous avons

$$\sup_{\theta_n \leq s \leq 1} \left|\frac{\mathbb{V}_n\left(s\right) - s}{s}\right| = o_p(1), \qquad \text{quand } n \to \infty.$$

Cela implique que

$$\sup_{\theta_n \leq s \leq 1} \left|\frac{s}{s + \xi_{s,n}}\right| = 1 + o_p(1), \qquad \text{quand } n \to \infty.$$

Soit $\beta_n\left(t\right) := \sqrt{n}\left(\mathbb{V}_n\left(t\right) - t\right) = \sqrt{n}\left(Q_n\left(t\right) - t\right), 0 < t < 1$ le processus de quantile uniforme, par la substitutions dans l'égalité suivante, on trouve :

$$T_{n9}^{*} = \gamma^2 (k/n)^{2(\gamma-\delta)} \int_{\theta_n}^{k/n} s^{2\delta} \left(s + \xi_{s,n}\right)^{-2(\gamma+1)} \left\{\beta_n (1-s)\right\}^2 ds$$

$$\stackrel{D}{=} (1 + o_p(1)) \gamma^2 (k/n)^{2(\gamma-\delta)} \int_{\theta_n}^{k/n} s^{2(\delta-\gamma-1)} \left\{\beta_n (s)\right\}^2 ds.$$

On note que d'après Théorème.(2.1) de *Csörgő, M., Gsörgö, S., Horvàth et Mason, D.* (1986) [7], ici il existe une suite de ponts Brownien [3](pour $0 \leq \upsilon \leq 1/2$)

$$\sup_{1/n \leq s \leq 1-1/n} \left| \frac{n^{1/2} \left(\mathbb{V}_n (s) - s\right) - B_n (s)}{s^{1/2-\upsilon}} \right| = O_p(n^{\upsilon}), \quad \text{quand } n \to \infty. \tag{4.6}$$

Fixons $0 < \upsilon < \delta - \gamma$, d'après (4.6) T_{n9}^{*} peut être écrit sous la forme

$$T_{n9}^{*} \stackrel{D}{=} \mathbb{Z}_{n1} + \mathbb{Z}_{n2} + \mathbb{Z}_{n3},$$

où

$$\mathbb{Z}_{n1} := (1 + o_p(1)) \gamma^2 (k/n)^{2(\gamma-\delta)} \int_{\theta_n}^{k/n} s^{2(\delta-\gamma-1)} \left\{\beta_n (s) - B_n (s)\right\}^2 ds,$$

$$\mathbb{Z}_{n2} := 2 (1 + o_p(1)) \gamma^2 (k/n)^{2(\gamma-\delta)} \int_{\theta_n}^{k/n} s^{2(\delta-\gamma-1)} B_n (s) \left\{\beta_n (s) - B_n (s)\right\} ds,$$

et

$$\mathbb{Z}_{n3} := (1 + o_p(1)) \gamma^2 (k/n)^{2(\gamma-\delta)} \int_{\theta_n}^{k/n} s^{2(\delta-\gamma-1)} \left\{B_n (s)\right\}^2 ds.$$

D'après (4.6) , il est facile de vérifier que

$$\begin{aligned} \mathbb{Z}_{n1} &= (1 + o(1)) \gamma^2 (k/n)^{2(\gamma-\delta)} \int_{\theta_n}^{k/n} s^{2(\delta-\gamma-1)} s^{1-2\upsilon} \left\{ \frac{\beta_n (s) - B_n (s)}{s^{1/2-\upsilon}} \right\}^2 ds, \\ &= O_p(k^{-2\upsilon}), \quad \text{quand} \quad n \to \infty. \end{aligned}$$

[3]Un processus stochastique $\{\mathcal{B}(t); 0 \leq t \leq 1\}$ est appelé Pont Brownien si
i- La distribution jointe de $(\mathcal{B}(t_1), \mathcal{B}(t_2), ..., \mathcal{B}(t_k))$ est Gaussienne, avec $\mathbb{E}\mathcal{B}(t) \equiv 0$,
ii- La fonction de covariance de $\mathcal{B}(t)$ est $R(r,t) = \mathbb{E}\mathcal{B}(s)\mathcal{B}(t) = s \wedge t - st$.

Donc $\mathbb{Z}_{n1} = o_p(1)$. D'autre part en utilisant l'inégalité

$$E\left|B_n\left(s\right)\right| \leq \sqrt{E\left[B_n\left(s\right)\right]^2} = \sqrt{s\left(1-s\right)} \leq s^{1/2}, \quad 0 \leq s \leq 1,$$

nous obtenons facilement $\mathbb{Z}_{n2} = O_p(k^{-\upsilon}) = o_p(1)$, quand $n \to \infty$.

Pour tout $0 < u < 1$, nous avons

$$B_n\left(u\right) \overset{\mathcal{D}}{=} W_n\left(u\right) + \xi_n u,$$

où W_n est une suite de processus de Wiener standard et ξ_n est une variable de la loi normale standard indépendante de W_n. Il suit que

$$\mathbb{Z}_{n3} \overset{D}{=} \left(1 + o_p(1)\right)\gamma^2 \left(k/n\right)^{2\gamma - 2\delta} \int_{\theta_n}^{k/n} s^{2(\delta-\gamma-1)} \left\{W_n\left(s\right) + \xi_n s\right\}^2 ds.$$

Rappelons que

$$E\left|W_n\left(u\right)\right| \leq \sqrt{E\left[W_n\left(u\right)\right]^2} = \sqrt{u} \leq s^{1/2}, \quad \text{pour tout } 0 \leq u \leq 1. \tag{4.7}$$

Alors :

$$\mathbb{Z}_{n3} = \left(1 + o_p(1)\right)\gamma^2 \int_{n\theta_n/k}^{1} s^{2(\delta-\gamma-1)} W_n^2\left(s\right) ds + o_p(1), \text{ quand } n \to \infty.$$

Observons que, lorsque $n \to \infty$

$$\mathbb{Z}_{n3} = \left(1 + o_p(1)\right)\gamma^2 \left\{\int_0^1 s^{2(\delta-\gamma-1)} W_n^2\left(s\right) ds + \int_0^{n\theta_n/k} s^{2(\delta-\gamma-1)} W_n^2\left(s\right) ds\right\} + o_p(1).$$

De plus, en utilisant (4.7), nous obtenons

$$\mathbb{Z}_{n3} = \left(1 + o_p(1)\right)\gamma^2 \left\{\int_0^1 s^{2(\delta-\gamma-1)} W_n^2\left(s\right) ds + O_p\left(\left(\frac{n\theta_n}{k}\right)^{2(\delta-\gamma)}\right)\right\} + o_p(1).$$

Depuis $n\theta_n/k = k^{-1/2} \to 0$, quand $n \to \infty$. Il suit que

$$\mathbb{Z}_{n3} \overset{\mathcal{D}}{=} \gamma^2 \int_0^1 s^{2(\delta-\gamma-1)} W^2(s)\, ds + o_p(1) \quad \text{quand } n \to \infty.$$

Cela achève la preuve du Théorème 4.2.1 ■

4.3 Simulations

Cette partie sera consacrée à la présentation des résultats de simulation de *Théorème (4.2.1)* de *Necir et Boukhetala (2005)* [40]. Pour la simulation, on utilise *le logiciel* **R** (version 2.0.1).

Nous avons décrit deux versions de ce test, qui basées sur la méthode du bootstrap paramétrique pour l'estimation des fluctuations d'échantillonnage, on a deux cas :

1erCas

La loi de simulation utilisée dans ce cas est une loi de *Burr* de paramètre (β, τ, λ) et de fonction de répartition.

$$F(x) = 1 - \left(\frac{\beta}{\beta + x^\tau}\right)^\lambda, \quad \text{pour} \quad x > 0,$$

où $\gamma = 1/(\lambda\tau)$, $\rho = -1/\lambda$, $(\beta > 0, \lambda > 0, \tau > 0)$.

Nous avons généré un échantillon de taille $n = 10000$, à partir d'une variable s de loi $\mathcal{U}([0,1])$, le modèle ajusté sera :

$$F^{-1}(s) = \left(\frac{\beta\left(1 - s^{-1/\lambda}\right)}{s^{-1/\lambda}}\right)^{1/\tau}, \qquad 0 < s < 1.$$

On applique le théorème (4.2.1) pour $\delta > \gamma$, on obtient

$$\begin{aligned} d^2_{\delta,k}(F_n, F) &= (2\delta+1)\int_0^1 s^{2\delta}\left\{F_n^{-1}(1-ks/n) - F^{-1}(1-ks/n)\right\}^2 ds. \\ &= A - 2AB + B, \end{aligned}$$

où

$$F_n^{-1}(1-ks/n) = X_{n-j+1,n}, \qquad \tfrac{j-1}{k} < s < \tfrac{j}{k},$$

et

$$\begin{aligned} A &= (2\delta+1)\int_0^1 s^{2\delta}\left(F_n^{-1}(1-ks/n)\right)^2 ds. \\ &= (2\delta+1)\sum_{j=1}^{k}\int_{\frac{j-1}{k}}^{\frac{j}{k}} s^{2\delta}\left(F_n^{-1}(1-ks/n)\right)^2 ds \\ &= \sum_{j=1}^{k}\left(\left(\frac{j}{k}\right)^{2\delta+1} - \left(\frac{j-1}{k}\right)^{2\delta+1}\right) X_{n-j+1,n}^2, \end{aligned}$$

et

$$\begin{aligned} AB &= (2\delta+1)\int_0^1 s^{2\delta} F_n^{-1}(1-ks/n)\, F^{-1}(1-ks/n)\, ds. \\ &= (2\delta+1)\sum_{j=1}^{k}\int_{(j-1)/k}^{j/k} s^{2\delta} F_n^{-1}(1-ks/n)\, F^{-1}(1-ks/n)\, ds \\ &= (2\delta+1)\sum_{j=1}^{k} X_{n-j+1,n}\int_{(j-1)/k}^{j/k} s^{2\delta} F^{-1}(1-ks/n)\, ds. \end{aligned}$$

On applique la méthode d'approximation numérique de l'intégrale

$$C = \int_{(j-1)/k}^{j/k} s^{2\delta} F^{-1}(1-ks/n)\, ds;$$

on a :

$$C = \int_{(j-1)/k}^{j/k} s^{2\delta}\left(\frac{\beta\left(1-s^{-1/\lambda}\right)}{s^{-1/\lambda}}\right)^{1/\tau} ds = \frac{1}{2}\left(\frac{j}{k} - \frac{j-1}{k}\right)(C1 + C2),$$

où

$$\begin{aligned} C1 &= \left(\frac{j}{k}\right)^{2\delta}\left(\beta\left(\frac{j}{k}\right)^{1/\lambda}\left(1-\left(\frac{j}{k}\right)^{-\frac{1}{\lambda}}\right)\right)^{1/\tau}, \\ C2 &= \left(\frac{j-1}{k}\right)^{2\delta}\left(\beta\left(\frac{j-1}{k}\right)^{1/\lambda}\left(1-\left(\frac{j-1}{k}\right)^{-1/\lambda}\right)\right)^{1/\tau}, \end{aligned}$$

et

$$B = \int_0^1 s^{2\delta} \left(F^{-1}\left(1 - ks/n\right)\right)^2 ds = \int_0^1 s^{2(\delta+\gamma)} \beta^{2/\tau} \left(1 - s^{-1/\lambda}\right)^{2/\tau} ds.$$

Alors

$$\begin{aligned} d^2_{\delta,k}\left(F_n, F\right) &:= (2\delta+1) \int_0^{k/n} s^{2\delta} \left\{F_n^{-1}\left(1 - ks/n\right) - F^{-1}\left(1 - ks/n\right)\right\}^2 ds \\ &= \sum_{j=1}^{k} \left[\left(\tfrac{j}{k}\right)^{2\delta+1} - \left(\tfrac{j-1}{k}\right)^{2\delta+1}\right] X^2_{n-j+1,n} + \\ &(2\delta+1) \left[\sum_{j=1}^{k} X_{n-j+1,n} C + \int_0^1 s^{2\delta} \left(\frac{\beta\left(1 - s^{-1/\lambda}\right)}{s^{-1/\lambda}}\right)^{2/\tau} ds\right]. \end{aligned}$$

On utilise la méthode du bootstrap pour simuler un échantillon d'une loi de $burr\,(1,2,1)$ $(\gamma = 0.5, \rho = -1)$, $burr\,(1,1/3,4)$ $(\gamma = 0.75, \rho = -1.33)$, $burr\,(1,1/2,2)$ $(\gamma = 1, \rho = -1)$ avec $B = 100$ réplications bootstrap, pour calculer la distance $d^2_{\delta,k}$ nous choisissons les valeurs de δ et γ telle que $\delta > \gamma > 0$. On trace la distance $d^2_{\delta,k}$ en fonction de $theta = \alpha$, $theta = \frac{1}{m} : \frac{m}{m}$ et $m = 1000$ avec $k_n = k = [n^\alpha]$ telle que k_n varié de 1 à n. Les graphes qui présentent les valeurs moyennes des distances bootstrapées sont obtenus d'après l'algorithme suivant :

ALGORITHME : Par conséquent, l'algorithme bootstrap utilisé pour tester la distance pondérée est le suivant :

1) Exécution des étapes (2) à (9), pour $b = 1 : B$.

2) Génération de l'échantillon $(x_1, ..., x_n)$ de la loi $burr\,(\beta, \tau, \lambda)$.

3) Calcul de A^*, pour chaque k_n, échantillonnage aléatoire d'une valeur à partir d'une loi inverse de $Burr\,(\beta, \tau, \lambda)$ tel que :

$$A^* = \sum_{j=1}^{k} \left(\left(\frac{j}{k}\right)^{2\delta+1} - \left(\frac{j-1}{k}\right)^{2\delta+1}\right) X^{*2}_{n-j+1,n}.$$

4) Application de la méthode d'approximation numérique pour le calcul de l'intégrale

$$\int_{\frac{j-1}{k}}^{\frac{j}{k}} s^{2\delta} F^{-1}\left(1 - ks/n\right) ds.$$

5) Calcul de l'intégrale

$$(AB)^* = (2\delta + 1) \sum_{j=1}^{k} X^*_{n-j+1,n} \int_{\frac{j-1}{k}}^{\frac{j}{k}} s^{2\delta} \left(F^*\right)^{-1} (1 - ks/n)\, ds.$$

6) Calcul numérique de l'intégrale

$$B^* = \int_0^1 s^{2\delta} \left(\left(F^*\right)^{-1} (1 - ks/n)\right)^2 ds.$$

7) Calcul de la distance $d^{*2}_{\delta,k}\left(F^*_n, F^*\right)$ telle que $\delta > \gamma > 0$

$$d^{*2}_{\delta,k}\left(F^*_n, F^*\right) := A^* - 2(AB)^* + B^*.$$

8) Détermination de la moyenne : $d^*_1 = \dfrac{1}{B} \sum_{k=1}^{B} d^*_k$.

9) Traçage des graphes correspondants.

Pour γ fixé

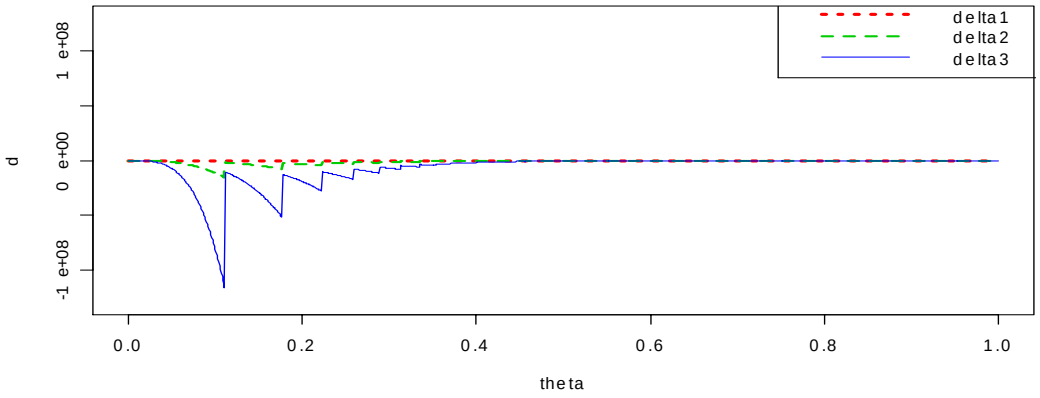


FIG. 4.5. La distance bootstrapée pour $\delta = (\delta_1, \delta_2, \delta_3) = (1, 1.3, 1.5)$, $\gamma = 0.5$.

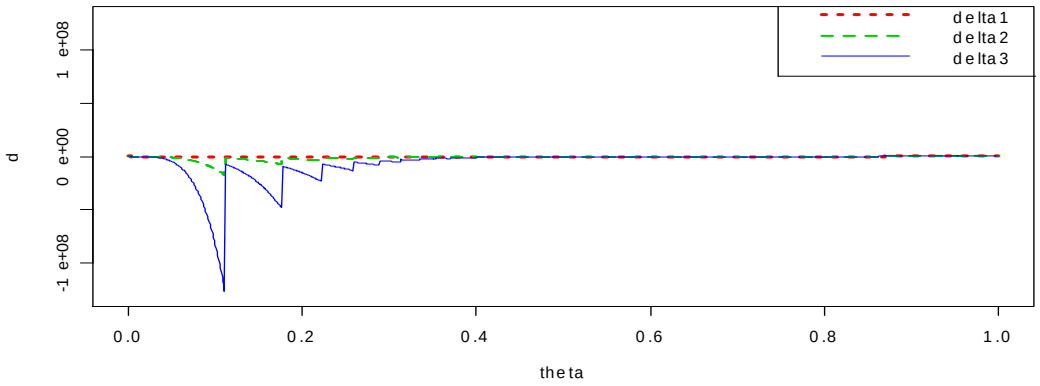


FIG. 4.6. La distance bootstrapée pour $\delta = (\delta_1, \delta_2, \delta_3) = (1, 1.3, 1.5)$, $\gamma = 0.75$.

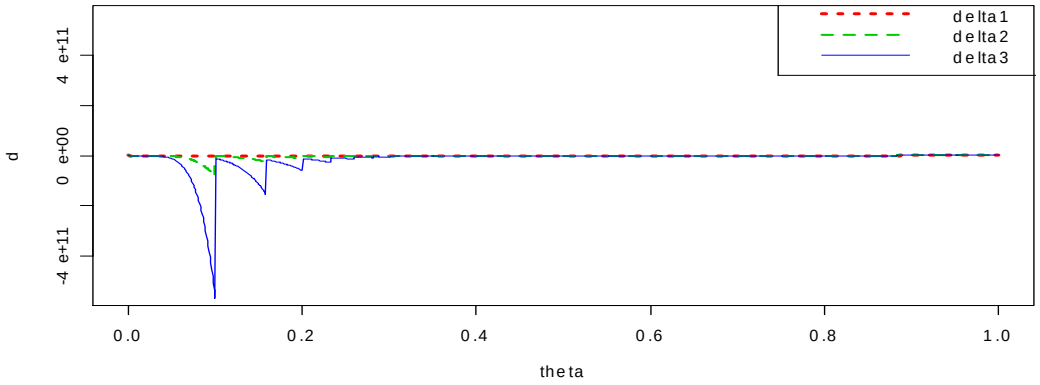


FIG. 4.7. La distance bootstrapée pour $\delta = (\delta_1, \delta_2, \delta_3) = (1.5, 2, 2.5)$, $\gamma = 1$.

Pour δ fixé

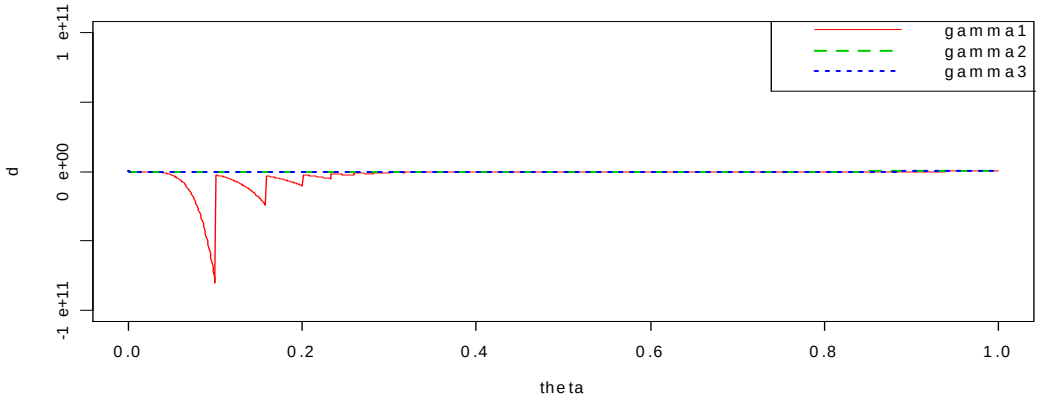


FIG. 4.8. La distance bootstrapée pour $\delta = 1.5$, $\gamma = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 0.75, 1)$.

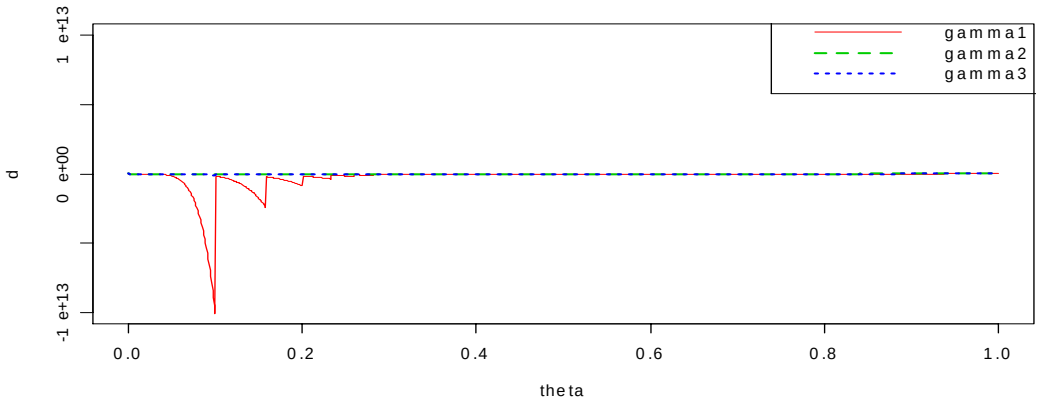


FIG. 4.9. La distance bootstrapée pour $\delta = 2$, $\gamma = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 0.75, 1)$.

2èmeCas

Inversement„ pour obtenir un bon estimateur d'une distribution à queue lourde, on utilise la procédure de bootstrap paramétrique.

Soit $F \in D(G_\gamma)$ pour $\gamma > 0$, d'après *Drees (2002)* [14] nous avons :

$$\frac{F^{-1}(1-\lambda s)}{F^{-1}(1-\lambda)} - s^{-\gamma} \to 0, \quad s > 0, \tag{4.8}$$

où $\lambda = k_n/n$, $1 \leq k_n < n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$.

$$\begin{aligned} x_{p_n} := F^{-1}(1-p_n) &\approx F^{-1}\left(1-\frac{k_n}{n}\right)\left(\frac{np_n}{k_n}\right)^{-\gamma} \\ &\approx X_{n-k_n,n}\left(\frac{np_n}{k_n}\right)^{-\widehat{\gamma}_n} \\ &= X_{n-k_n,n}s^{-\widehat{\gamma}_n} =: \hat{x}_{p_n}^{k_n} = \hat{x}_{p_n}. \end{aligned}$$

On déduit pour $k = k_n$ et $\delta > \gamma > 0$

$$d^2_{\delta,k}\left(F_n, F\right) := (2\delta+1)\sum_{j=1}^{k}\int_{(j-1)/k}^{j/k} s^{2\delta}\left\{F_n^{-1}\left(1-ks/n\right) - F^{-1}\left(1-ks/n\right)\right\}^2 ds$$

$$= (2\delta+1)\sum_{j=1}^{k}\int_{(j-1)/k}^{j/k} s^{2\delta}\left\{X_{n-j+1,n} - X_{n-k,n}s^{-\widehat{\gamma}}\right\}^2 ds$$

$$= A - 2AB + B,$$

telle que :

$$A := (2\delta+1)\sum_{j=1}^{k} X^2_{n-j+1,n}\int_{(j-1)/k}^{j/k} s^{2\delta} ds$$

$$= \sum_{j=1}^{k}\left[\left(\tfrac{j}{k}\right)^{2\delta+1} - \left(\tfrac{j-1}{k}\right)^{2\delta+1}\right] X^2_{n-j+1,n},$$

et

$$AB := (2\delta+1)\sum_{j=1}^{k} X_{n-j+1,n}\left[\int_{(j-1)/k}^{j/k} s^{2\delta}\left(X_{n-k,n}s^{-\widehat{\gamma}}\right) ds\right]$$

$$= (2\delta+1)\, X_{n-k,n}\sum_{j=1}^{k} X_{n-j+1,n}\int_{(j-1)/k}^{j/k} s^{2\delta-\hat{\gamma}} ds,$$

$$= X_{n-k,n}\sum_{j=1}^{k}\frac{(2\delta+1)}{(2\delta-\hat{\gamma}+1)}\left[\left(\tfrac{j}{k}\right)^{(2\delta-\hat{\gamma}+1)} - \left(\tfrac{j-1}{k}\right)^{(2\delta-\hat{\gamma}+1)}\right] X_{n-j+1,n},$$

et enfin

$$B := (2\delta+1)\sum_{j=1}^{k}\int_{(j-1)/k}^{j/k} s^{2\delta}\left(X_{n-k,n}s^{-\hat{\gamma}}\right)^2 ds$$

$$= (2\delta+1)\, X^2_{n-k,n}\sum_{j=1}^{k}\int_{(j-1)/k}^{j/k} s^{2(\delta-\hat{\gamma})} ds$$

$$= X^2_{n-k,n}\sum_{j=1}^{k}\frac{(2\delta+1)}{2\left(\delta-\hat{\gamma}\right)+1}\left[\left(\tfrac{j}{k}\right)^{2(\delta-\hat{\gamma})+1} - \left(\tfrac{j-1}{k}\right)^{2(\delta-\hat{\gamma})+1}\right].$$

Dans ce cas, nous appliquons la méthode du bootstrap pour les même données simulées de taille $n = 10000$ pour des valeurs différentes de γ et δ avec $\delta > \gamma > 0$ pour $k_n = k = [n^\alpha]$ telle que k_n varié de 1 à n. On trace la distance $d^2_{\delta,k}$ de l'échantillon de la loi de $Burr\,(1,2,1)$ $(\gamma = 0.5, \rho = -1)$, $Burr\,(1,1/3,4)$ $(\gamma = 0.75, \rho = -1.33)$, $Burr\,(1,1/2,2)$ $(\gamma = 1, \rho = -1)$ en fonction de $theta = \frac{1}{m} : \frac{m}{m}$ et $m = 1000$ avec $B = 100$ réplications bootstrap.

Pour γ fixé

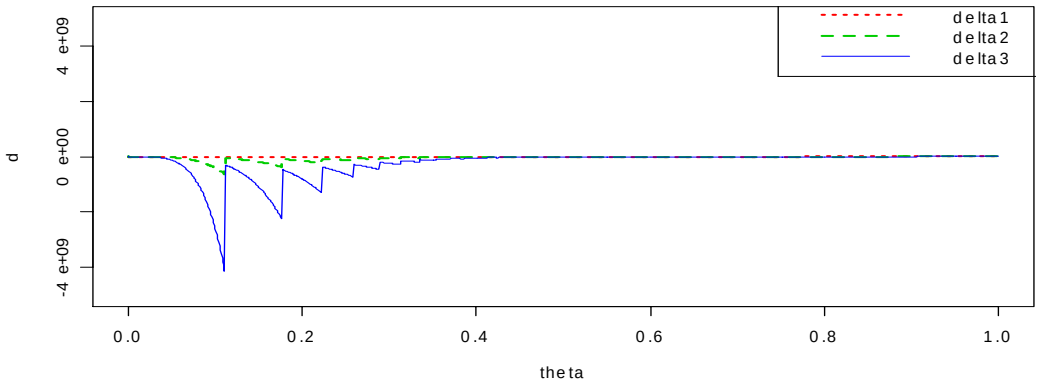


FIG. 4.10. La distance bootstrapée pour $\delta = (\delta_1, \delta_2, \delta_3) = (1.5; 2; 2.5)$, $\gamma = 0.5$.

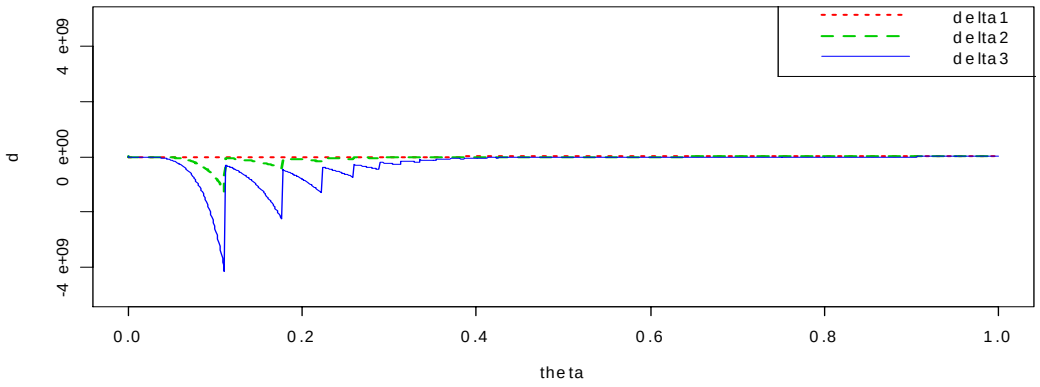


FIG. 4.11. La distance bootstrapée pour $\delta = (\delta_1, \delta_2, \delta_3) = (1.2; 1.5; 2)$, $\gamma = 0.75$.

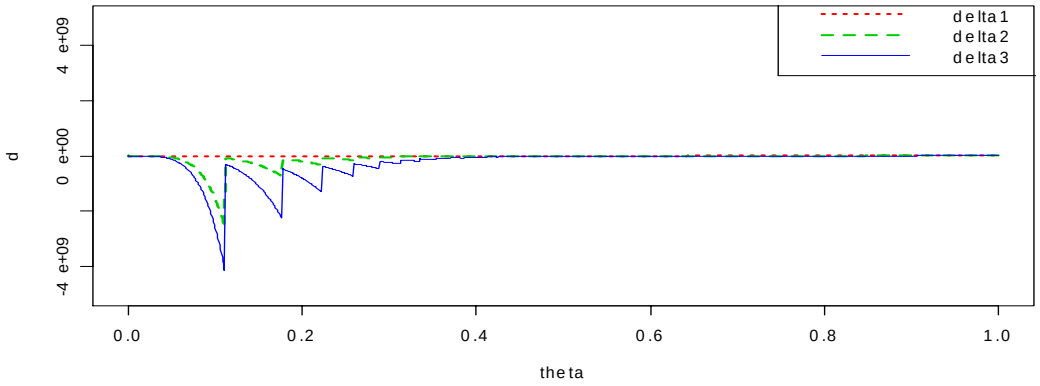


FIG. 4.12. La distance bootstrapée pour $\delta = (\delta_1, \delta_2, \delta_3) = (1.5; 2; 2.5)$, $\gamma = 1$.

Pour δ fixé

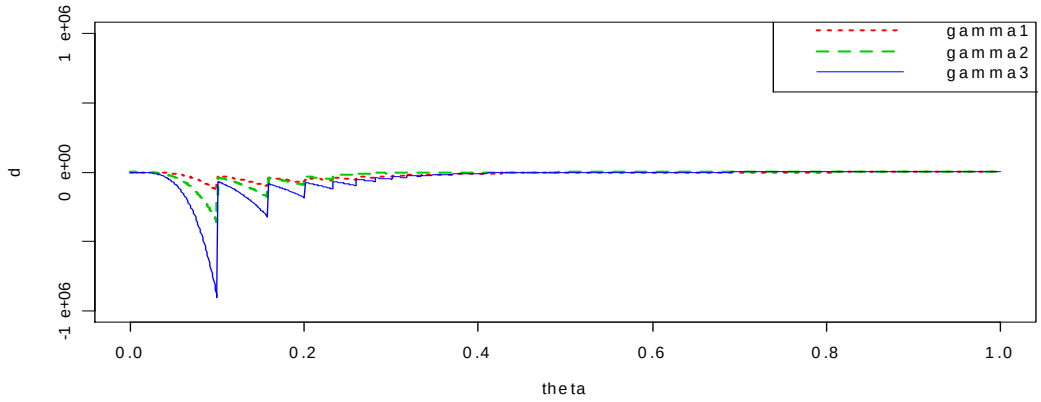


FIG. 4.13. La distance bootstrapée pour $\delta = 1.2$ et $\gamma = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 0.75, 1)$.

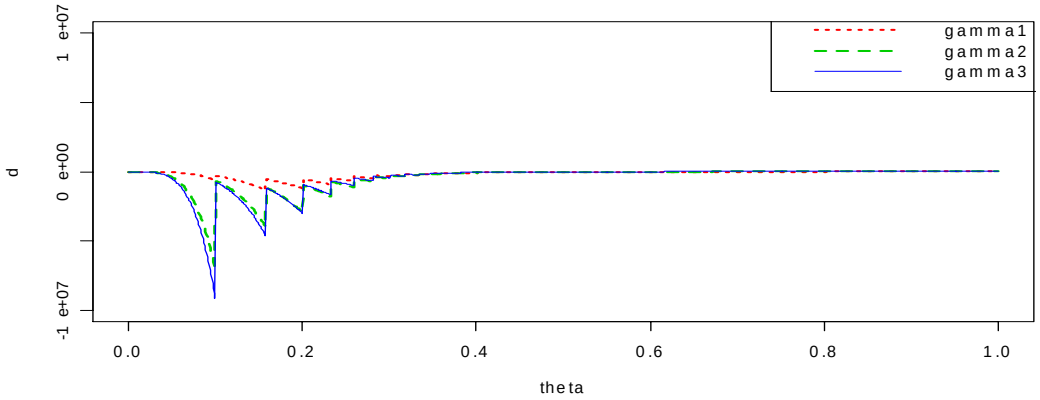


FIG. 4.14. La distance bootstrapée pour $\delta = 1.5$, $\gamma = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 0.75, 1)$.

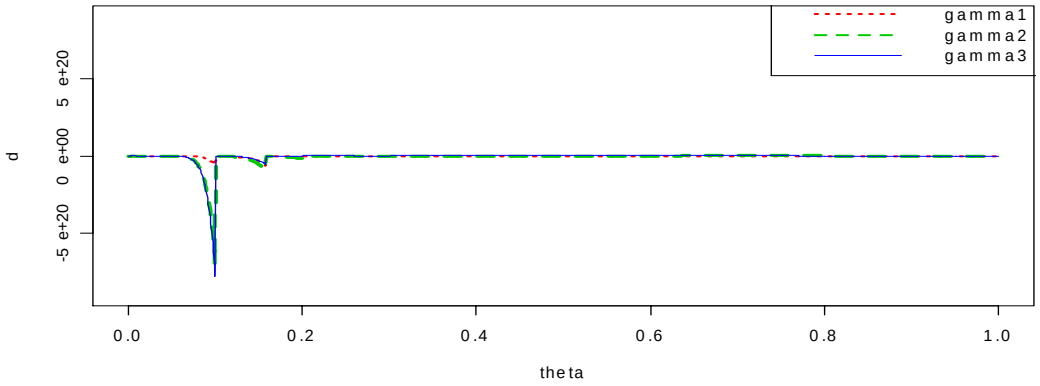


FIG. 4.15. La distance bootstrapée pour $\delta = 5$, $\gamma = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 0.75, 1)$.

Résultats de simulation

Afin de comparer les résultats graphiques obtenus dans les deux cas de simulation précédents, on a utilisé des méthodes du bootstrap sur la loi de $Burr\,(\beta, \tau, \lambda)$, on remarque que la distance entre les quantiles et l'estimateur du paramètre γ pour la classe des distributions a queues lourdes (la distance entre deux fonctions inverse) tend vers zéro dans tous les cas ; puisque la valeur δ est près de la valeur γ, la distance est plus petite δ écrase γ dans ce cas.

CONCLUSION

Le bootstrap est une méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formules à beaucoup de questions statistiques réelles.

L'utilisation des techniques de rééchantillonnage a été rendue possible grâce à la généralisation des moyens de calculs performants, ces techniques reposent, au départ, sur des idées simples. Toutefois, il faut bien admettre que les développements apportés aux méthodes de base leur ont fait perdre une partie de cette simplicité.

Dans ce mémoire, nous nous sommes limité au problème de l'estimation du biais, de l'erreur-standard d'un paramètre, et à la détermination des limites de confiance d'un paramètre. Il ne s'agit cependant pas des seules applications des méthodes de rééchantillonnage. Celles-ci peuvent, en effet, aussi être utilisées pour la réalisation de différents tests d'hypothèses, pour le choix des variables et l'estimation de l'erreur de prédiction en régression. L'application des méthodes de bootstrap sur les modèles de régression fournit une approximation de la distribution des erreurs de prédiction par leur distribution empirique lorsque celle-ci est inconnue. Le bootstrap est ainsi particulièrement utile lorsque les échantillons de données sont de petite taille et qu'il n'est pas possible de formuler l'hypothèse d'une distribution gaussienne du terme d'erreur. Le nombre de réplications peut être déterminé à partir des coefficients de variation de l'étendue de l'intervalle de confiance des coefficients, ou de l'intervalle de prédiction lorsque ceux-ci deviennent peu variants. L'utilisation des techniques de bootstrap permet donc une meilleure copie de l'information contenue dans l'échantillon initial pour les intervalles de prédiction.

Bien qu'elles puissent être utilisées dans des situations très variées, leur mise en oeuvre ne présente guère d'intérêt lorsque l'inférence statistique peut être réalisée par des méthodes analytiques classiques, pour lesquelles les conditions d'application sont remplies. Elles ne sont donc pas destinées à remplacer les méthodes d'inférence statistique classiques lorsque celles-ci sont applicables mais plutôt à fournir des réponses à des questions pour les quelles les méthodes classiques sont inapplicables ou non disponibles.

À la fin on a appliquer la méthode du bootstrap aux queues de distributions, aux valeurs extrêmes dans l'estimation a queues lourdes et on a trouver des bons résultats pour améliorer la précision de cette estimation.

Bibliographie

[1] Andrews, D.W.K. (2000), "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space", Econometrica 68 :399-405.

[2] Athreya, K. (1987), "Bootstrap of the mean in the infinite variance case", Annals of Statistics 15 :724-731.

[3] Beran, R. & G.R. Ducharme (1991), Asymptotic Theory for Bootstrap Methods in Statistics, (Les Publications CRM, Centre de recherches mathematiques, Universite de Montreal, Montreal, Canada).

[4] Bickel, P.J. & D.A. Freedman (1981), "Some asymptotic theory for the bootstrap", Annals of Statistics 9, 1196-1217.

[5] Brown, B.W. (1999), "Simulation variance reduction for bootstrapping", in R. Mariano, T.

[6] Chernick MR. (1999). "Bootstrap methods : a practitioner's guide". New York : Wiley, 264 p.

[7] Csörgő, M., Gsörgö, S., Horvàth, L. & Mason, D. (1986). "Weighted empirical and quantile processec. Ann. Probab., 14, 31-85.

[8] Danielsson, J., de Hann, L., Peng, L. & de Vries, C. G. (2001). "Using a bootstrap Method to choose the sample fraction in tail estimation". J. Multivariare Annal., 76, 226-248.

[9] Davison, A.C. & D.V. Hinkley (1997), Bootstrap methods and their application (Cambridge University Press, Cambridge, U.K).

[10] Dagnelie P. (1998)." Statistique théorique et appliquée". Tome1 : statistique descriptive et bases de l'inférence statistique. Bruxelles : De Boeck et Larcier, 508 p.

[11] De Angelis, D., P. Hall, & G.A. Young (1993), "Analytical and bootstrap approximations to estimator distributions in L^1 regression", Journal of the American Statistical Association 88 :1310-1316.

[12] de Wet, T. (2002). Goodness-of-fit tests for location and scale families based on a weighted $\mathbb{L}_2$- Wasserstein distance measure. Test, 11, 89-107.

[13] Donald, S.G. & H.J. Paarsch (1996), "Identification, estimation, and testing in empirical models of auctions within the independent private values paradigm", Econometric Theory 12 :517-567.

[14] Drees, H. (2002)." Extreme quantile estimation for dependent data with application to finance". université de Saarland, (soumis).

[15] Efron, B. (1979), "Bootstrap methods : another look at the jackknife", Annals of Statistics 7 :1-26.

[16] Efron, B. (1987), "Better bootstrap confidence intervals", Journal of the American Statistical Association 82 :171-185.

[17] Efron, B. & R.J. Tibshirani (1993), "An Introduction to the Bootstrap". (Chapman & Hall, New York).

[18] Freedman, D.A. (1981), "Bootstrapping regression models", Annals of Statistics 9 :1218-1228.

[19] Hall, P. (1985), "Resampling a coverage process", Stochastic Process Applications 19 :259-269.

[20] Hall, P. (1986 a), "On the number of bootstrap simulations required to construct a confidence interval", Annals of Statistics 14 :1453-1462.

[21] Hall, P. (1986 b), "On the bootstrap and confidence intervals", Annals of Statistics 14 :1431-1452.

[22] Hall, P. (1988), "Theoretical comparison of bootstrap confidence intervals", Annals of Statistics 16 :927-953.

[23] Hall, P. (1990), "Asymptotic properties of the bootstrap for heavy-tailed distributions", Annals of Probability 18 :1342-1360.

[24] Hall, P. (1990), "Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametrique problems, J. Multivariate Anal.32 (1990), 177-203.

[25] Hall, P. (1992 a), "The Bootstrap and Edgeworth Expansion ". (Springer-Verlag, New York).

[26] Hall, P. (1992b), "Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density", Annals of Statistics 20 :675-694.

[27] Hall, P. (1994), "Methodology and theory for the bootstrap", in R.F. Engle and D.F. McFadden, eds., Handbook of Econometrics, vol. 4 (Elsevier Science, B.V, Amsterdam).

[28] Hall, P. & J.L. Horowitz (1996), "Bootstrap critical values for tests based on generalized- method of-moments estimators", Econometrica 64 :891-916.

[29] Hill. B.M. (1975), "A Simple approach to inference aAbout tail of a distribution", Ann. Statist, 3,1163-1174.

[30] Horowitz, J.L. (1997), "Bootstrap methods in econometrics : theory and numerical performance", in D.M. Kreps and K.F. Wallis, eds. Advances in Economics and Econometrics : Theory and Applications, Seventh World Congress, vol. 3 (Cambridge University Press, Cambridge, U.K).

[31] Horowitz, J.L. (2000), "The bootstrap". Department of Economics University of Iowa Iowa City, IA 52242.

[32] Lecoutre. J.P. & Tassi Ph. (1987), " Statistique non paramétrique et robustesse". Ed éconmica. Paris.

[33] Léger C., Politis DHN. & Romano JP. (1992), "Bootstrap technology and applications". Technometrics 34,p.378-398.

[34] Lo, A.Y. (1991), " Bayesian Bootstrap Clones and a Biometry Function". Sankhya A, 53, pp. 320-333.

[35] Maddala, G.S. & J. Jeong (1993), "A perspective on application of bootstrap methods in econometrics", in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., Handbook of Statistics vol. 11 (North-Holland, Amsterdam).

[36] Mammen, E. (1992), "When Does Bootstrap Work? Asymptotic Results and Simulations". (Springer-Verlag, New York).

[37] Manly BFJ. (1997). "Randomization, bootstrap and Monte Carlo methods in biology ". New York : Chapman and Hall, 399 p.

[38] Mason, D. & Newton, M. A. (1992)," A Rank Statistic Approach to the Consistency of a General bootstrap", Ann. Statist., 20, pp. 1611-1624.

[39] **N**ecir, A. & Boukhetala, K. (2004). "Estimating the risk adjusted premium for the largest Claims reinsurance covers". Compstat 2004. Edition physica-Verlag Heidelberg 2004/Springer, Volume1. pp.1577-1585.

[40] Necir, A. et Boukhetala, K. (2005). "A goodness-of-fit test of tail indices via a weighted $\mathbb{L}^2$- Wasserstein distance" (soumis).

[41] Shorack, G. R. & Wellner, J. A. (1986). " Empirical processus with applications to statistics". Wiley, New York.

[42] Vinod, H. (1993) "Bootstrap Methods : Applications in Econometrics", in Handbook of Statistics, 11, North-Holland, 629-661.

[43] Wellner, J. A. (1978). "Limit theorems for the ratio of the empirical distrbution function to the true distrbution function. Z Wahrsch. verw. Gebiete, 45, 73-88.

RESUME

Ce mémoire présente une étude théorique sur les différentes méthodes du Bootstrap et l'utilisation de cette technique de ré échantillonnage dans la statistique d'inférence pour le calcul de l'erreur standard, du biais d'un estimateur et la détermination de l'intervalle de confiance d'un paramètre estimé. Nous appliquons ces méthodes dans les testes des modèles de régression et les modèle de type Pareto, ce qui donne des meilleurs approximations. Du coté numérique, notre étude est fonder sur le logiciel statistique R version 2.0.1.

Mots- clés: Bootstrap, erreur standard, biais, jackknife, moyenne, médiane, variance, quantile, intervalle de confiance, régression, valeur extrême, queue indexé, queue lourde, bon ajustement, distance de Wasserstein.

ABSTRACT

This memory presents a theatrical study on various methods of Bootstrap and the uses of this technique to resampling statistics of inference for computing the standard error, bias of an estimator and the determination of the confidence interval of an estimated parameter. We apply these methods to testing models of regression and the Pareto type distribution, which gives better approximations. The numerical study is based on the statistical software R version 2.0.1.

Key words: Bootstrap, standard error, bias, jackknife, mean, median, variance, quantile, confidence intervals, regression, extreme value, tail index, heavy tails, goodness of fit, Wasserstein distance.

ملخـــــص

قمنا في هذه المذكرة بعرض الدراسة النظرية لمختلف طرق العينة المتكررة، إذ تطرقنا لمختلف استعمالاتها في الإحصاء الاستدلالي و ذالك لإيجاد الخطأ المعياري، الانحراف، مجالات الثقة و في الانحدارات. كما طبقنا طرق العينة المتكررة على أذناب التوزيعات من أجل نموذج باريتو لإيجاد أحسن التقريبات لاختبار عينة كبيرة الحجم ذات متغيرات مستقلة.

الكلمات الجوهرية: الخطأ المعياري، الانحراف، مجالات الثقة، المعدل، المتوسط، التغاير، الانحدار، القيمة الحدية، دليل أذناب التوزيعات، أذناب كبيرة.