Université de Montréal

Estimateur bootstrap de la variance d'un estimateur de quantile en contexte de population finie

par

Vanessa McNealis

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

20 décembre 2019

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Estimateur bootstrap de la variance d'un estimateur de quantile en contexte de population finie

présenté par

Vanessa McNealis

a été évalué par un jury composé des personnes suivantes :

David Haziza
(président-rapporteur)

Christian Léger
(directeur de recherche)

Alejandro Murua
(membre du jury)

Mémoire accepté le :
10 février 2020

Sommaire

Ce mémoire propose une adaptation lisse de méthodes bootstrap par pseudo-population aux fins d'estimation de la variance et de formation d'intervalles de confiance pour des quantiles de population finie. Dans le cas de données i.i.d., Hall et al. (1989) ont montré que l'ordre de convergence de l'erreur relative de l'estimateur bootstrap de la variance d'un quantile échantillonnal connaît un gain lorsque l'on rééchantillonne à partir d'une estimation lisse de la fonction de répartition plutôt que de la fonction de répartition expérimentale. Dans cet ouvrage, nous étendons le principe du bootstrap lisse au contexte de population finie en le mettant en œuvre au sein des méthodes bootstrap par pseudo-population. Étant donné un noyau et un paramètre de lissage, cela consiste à lisser la pseudo-population dont sont issus les échantillons bootstrap selon le plan de sondage initial. Deux plans sont abordés, soit l'échantillonnage aléatoire simple sans remise et l'échantillonnage de Poisson. Comme l'utilisation des algorithmes proposés nécessite la spécification du paramètre de lissage, nous décrivons une méthode de sélection par injection et des méthodes de sélection par la minimisation d'estimés bootstrap de critères d'ajustement sur une grille de valeurs du paramètre de lissage. Nous présentons des résultats d'une étude par simulation permettant de montrer empiriquement l'efficacité de l'approche lisse par rapport à l'approche standard pour ce qui est de l'estimation de la variance d'un estimateur de quantile et des résultats plus mitigés en ce qui concerne les intervalles de confiance.

Mots-clés: Estimation de quantiles; Estimation de la variance; Intervalles de confiance; Échantillonnage; Bootstrap par pseudo-population; Bootstrap lisse; Paramètre de lissage.

Summary

This thesis introduces smoothed pseudo-population bootstrap methods for the purposes of variance estimation and the construction of confidence intervals for finite population quantiles. In an i.i.d. context, Hall et al. (1989) have shown that resampling from a smoothed estimate of the distribution function instead of the usual empirical distribution function can improve the convergence rate of the bootstrap variance estimator of a sample quantile. We extend the smoothed bootstrap to the survey sampling framework by implementing it in pseudo-population bootstrap methods. Given a kernel function and a bandwidth, it consists of smoothing the pseudo-population from which bootstrap samples are drawn using the original sampling design. Two designs are discussed, namely simple random sampling and Poisson sampling. The implementation of the proposed algorithms requires the specification of the bandwidth. To do so, we develop a plug-in selection method along with grid search selection methods based on bootstrap estimates of two performance metrics. We present the results of a simulation study which provide empirical evidence that the smoothed approach is more efficient than the standard approach for estimating the variance of a quantile estimator together with mixed results regarding confidence intervals.

Keywords: Quantile estimation; Variance estimation; Confidence intervals; Survey sampling; Pseudo-population bootstrap methods; Smoothed bootstrap; Smoothing parameter.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Ce travail n'aurait pu être mené à bien sans l'aide financière précieuse de mon directeur de recherche, Christian Léger, ainsi que celle du Département de mathématiques et de statistique, de la Faculté des études supérieures et postdoctorales de l'Université de Montréal et de l'Institut des sciences mathématiques. Je remercie toutes ces entités de m'avoir permis de me consacrer entièrement à mes activités de recherche en plus de contribuer au financement d'activités de rayonnement.

Je dirige d'abord les autres dimensions de ma gratitude à l'endroit du professeur Christian Léger pour sa disponibilité sans égale. J'ai en esprit les paroles du professeur David Haziza reprenant à l'occasion les mots de Jean Jaurès : «On n'enseigne pas ce que l'on sait ou ce que l'on croit savoir : on n'enseigne et on ne peut enseigner que ce que l'on est.» Pour une première expérience de recherche, je n'aurais pu espérer côtoyer un meilleur modèle, qui est à la fois un statisticien et un communicateur hors pair. Sa lecture et son écoute attentives, desquelles résultent immanquablement des commentaires édifiants, m'ont aidée à rendre les idées véhiculées dans ce mémoire plus vraies, c'est-à-dire en accord avec leur objet. Lui qui aime la forêt, il m'a aussi amenée à prendre du recul alors que je me perdais à travers les feuilles et les branches. Je le remercie enfin pour sa générosité et son amitié.

J'ai énormément bénéficié des différentes amitiés formées en cours de route, qui m'ont tantôt apporté du réconfort, tantôt éveillé ma curiosité. Merci à toutes les personnes contribuant à la vie du département, qui m'ont marquée de par leur intégrité, leur passion pour les mathématiques et leur gentillesse. Je remercie tout particulièrement Victoire, Marc-Antoine, Alexis, Gabriel et Isabelle, dont la proximité durant des moments charnières a rempli ma vie d'images vocationnelles. Je désire également remercier Victoire pour sa générosité et

sa fermeté d'âme. Je remercie Marc-Antoine de toujours agir sous le commandement de la raison ainsi que pour l'odeur du café.

Enfin, je remercie mon amoureux, Bruce, qui a volontiers partagé avec moi des moments de chagrin et des moments de joie. Sa résilience et sa force enveloppées dans un tempérament doux ont contribué à propulser ma propre croissance personnelle. Il m'a donné un jour un conseil qui a été un point tournant dans ma démarche, soit que le meilleur moyen de vaincre l'adversité est d'en assumer la responsabilité, ce que l'on pourrait traduire par

«Pull yourself up by your bootstraps.»

Introduction

Dans les enquêtes menées par les agences statistiques officielles, l'estimation de la distribution de la population finie ou des quantiles de celle-ci à partir de données de sondage est une opération routinière. La médiane peut être un paramètre attrayant, puisqu'elle s'avère être dans certains cas une meilleure mesure du centre de la distribution que la moyenne. À cet effet, on peut notamment penser à la distribution du revenu d'une population. En 2019, le rapport annuel du Bureau de recensement américain concluait que le revenu médian d'un ménage aux États-Unis en 2018 n'était pas statistiquement différent de celui de 2017 (United States Census Bureau, 2019). Pour conduire une telle inférence, il fallait de toute évidence avoir préalablement obtenu une estimation de l'erreur standard de l'estimateur de médiane. Or, dans le cas d'une statistique non lisse comme un quantile, il n'est pas possible de formuler exactement l'expression de la variance de l'estimateur comme il est possible de le faire pour une combinaison linéaire d'observations telle qu'une moyenne.

L'estimation des quantiles à partir de données d'enquête occasionne donc des défis intéressants. Certaines méthodes existantes dans la littérature pour l'estimation de la variance, comme la linéarisation ou le jackknife, sont difficiles à mettre en œuvre ou même échouent dans le cas des quantiles (Chatterjee, 2011). Ces difficultés sont attribuées au fait que les quantiles ne sont pas des fonctions différentiables de totaux ou encore qu'ils ne sont pas suffisamment *lisses* (Efron et Tibshirani, 1993). C'est pourquoi les méthodes de rééchantillonnage revêtent un attrait particulier au sein des agences statistiques officielles, où il est pratique courante de se replier sur des méthodes bootstrap afin de dériver une estimation de l'erreur standard d'un estimateur de quantile.

Le bootstrap a été introduit dans un contexte classique par Efron (1979), qui a par

ailleurs montré dans le même ouvrage la convergence de l'estimateur bootstrap de la variance de la médiane. Cela reposait évidemment sur l'hypothèse d'observations indépendantes et identiquement distribuées. Afin de tenir compte de la dépendance des données d'enquête, des adaptations du bootstrap non paramétrique classique ont été proposées pour le contexte des sondages, qui ont été classifiées en trois grandes catégories dans l'ouvrage de Mashreghi et al. (2016). Ce mémoire se consacre à une classe parmi les trois, soit celle des méthodes bootstrap par pseudo-population. En reconstituant une pseudo-population à partir de l'échantillon d'enquête et en obtenant un échantillon bootstrap suivant le plan de sondage initial, ces méthodes pourvoient des estimateurs de la distribution échantillonnale d'un estimateur réussissant à capturer les particularités induites par le plan utilisé.

Certaines lacunes peuvent néanmoins être soulevées quant à l'application des méthodes par pseudo-population pour l'estimation de la variance d'un quantile échantillonnal pour des raisons théoriques. Tout comme dans un cadre classique, la variance asymptotique d'un quantile échantillonnal dépend de propriétés locales de la distribution génératrice des données, plus précisément de la fonction de densité évaluée au quantile lui-même. Cela étant, les méthodes de rééchantillonnage par pseudo-population, au même titre que le bootstrap non paramétrique classique, font intervenir une fonction de répartition discrète, à laquelle est associée une fonction de masse et non une fonction de densité. Il n'est donc pas surprenant que dans un cadre classique, la mise en œuvre du bootstrap à partir d'une estimation *lisse* de la distribution, plutôt que la fonction de répartition expérimentale, puisse s'être avérée bénéfique pour ce qui est de l'estimation de la variance d'un quantile échantillonnal. Cela est vrai à condition que le paramètre de lissage à partir duquel est définie cette estimation lisse soit choisi adéquatement (Hall et al., 1989).

Ce mémoire vise à étudier les avantages potentiels du lissage au sein des méthodes par pseudo-population pour l'estimation de la variance et la construction d'intervalles de confiance pour des quantiles de population finie. Pour ce faire, il est nécessaire de cerner les propriétés de l'objet d'application, soit l'estimateur d'un quantile. En théorie de l'échantillonnage, l'étude des propriétés des estimateurs peut être accomplie sous différentes approches. L'inférence statistique est parfois faite uniquement sur la base du plan de

sondage (*design-based* pour reprendre la terminologie de Särndal et al. (1992)). Mais puisque les quantiles sont à l'étude, nous devrons adopter un angle jumelant à la fois le modèle et sur le plan (*design-model-based* selon Särndal et al. (1992)). Dans une telle approche, on fait l'hypothèse que la population finie est elle-même la réalisation d'une superpopulation infinie (ou distribution de probabilité). Il est notamment nécessaire de faire cette supposition pour établir que la variance asymptotique d'un quantile calculé à partir de données d'enquête dépend localement de superpopulation à travers la fonction de densité évaluée en un point dans le cas de l'échantillonnage aléatoire simple sans remise (Chatterjee, 2011).

Ce mémoire suivra le cheminement suivant. Le chapitre 1 est une entrée en matière sur le sujet dans son ensemble en présentant des éléments de la théorie de l'échantillonnage axés sur le problème de l'estimation des quantiles. Le chapitre 2 se concentre sur les méthodes de rééchantillonnage par pseudo-population en détaillant des algorithmes adaptés à deux plans de sondage. Le chapitre 3 se rapporte à un cadre classique en présentant la méthode du bootstrap lisse pour des données indépendantes et identiquement distribuées, une alternative attrayante au bootstrap non paramétrique lorsqu'il s'agit d'estimer la variance d'un quantile échantillonnal. Le chapitre 4 fait l'élaboration d'adaptations lisses des algorithmes abordés au chapitre 2 tout en prenant soin de suggérer des façons systématiques de sélectionner le paramètre de lissage, qui n'est pas une tâche aisée. Ce mémoire se soldera enfin par une étude par simulation au chapitre 5 visant à attester de la performance des adaptations proposées sous une variété de conditions.

Chapitre 1

Éléments de la théorie de l'échantillonnage

Ce chapitre introduit la notation et quelques outils conventionnellement utilisés en théorie de l'échantillonnage. Deux plans de sondage seront étudiés, soit l'échantillonnage stratifié aléatoire simple sans remise et l'échantillonnage de Poisson. L'estimation de paramètres de population finie et de la variance des estimateurs sera discutée avant de cheminer vers le cas particulier des quantiles. À cet effet, les quantiles échantillonnaux sont définis à partir de la fonction de répartition échantillonnale. Pour cette raison, nous introduirons la méthode de linéarisation par série de Taylor pour l'estimation de la variance d'estimateurs qui sont des fonctions non linéaires de totaux, celle-ci étant mise à profit dans le calcul de la variance de la fonction de répartition échantillonnale.

Il sera par la suite possible de traiter de méthodes basées sur la linéarisation pour l'estimation de la variance et la formation d'intervalles de confiance pour les quantiles. Ce segment se soldera par des résultats concernant la distribution asymptotique des quantiles échantillonnaux dans le cadre d'une population finie.

1.1. Préliminaires

1.1.1. Notation

On commence par introduire la notation qui sera utilisée tout au long du parcours de cet ouvrage, qui correspond à celle adoptée par Särndal et al. (1992). De manière générale, on

considère une population $\mathcal{P}$ comprenant N unités étiquetées par $i = 1, \ldots, N$, soit explicitement $\{u_1, \ldots, u_i, \ldots, u_N\}$. Afin d'alléger l'écriture, on désignera plutôt la population par l'ensemble d'étiquettes la constituant, soit par $U = \{1, \ldots, i, \ldots, N\}$. De plus, soit y, une variable d'intérêt dans la population pour laquelle on souhaite inférer. Les valeurs de cette variable dans la population sont contenues au sein du vecteur $\boldsymbol{y} = (y_1, \ldots, y_i, \ldots, y_N)'$. On note θ le paramètre à estimer pour cette variable, qui est par nature une fonctionnelle des N observations de la population, ce qui permet d'écrire $\theta = \theta(U)$.

De cette population est tiré un échantillon aléatoire s de taille n_s représenté par l'ensemble d'indices $\{j_1, \ldots j_i, \ldots, j_{n_s}\}$. De manière analogue à la notation pour la population U, s réfère à un sous-ensemble d'étiquettes, où chaque étiquette $j_i \in U$, $i = 1, \ldots, n_s$.

Cet échantillon est généré selon un plan de sondage $p(\cdot)$, qui est une fonction prenant un échantillon s dans $\mathcal{S}$, l'ensemble de tous les échantillons qu'il est possible de former, et renvoie la probabilité associée à cet échantillon, $p(s)$. En particulier, le plan de sondage détermine la probabilité d'inclusion de l'unité $i \in U$ dans l'échantillon s, qui est notée $\pi_i = P(i \in s) = \sum_{s \ni i} p(s)$. Afin d'étudier la variance d'un estimateur, il est également requis de connaître la probabilité d'inclusion conjointe de deux unités i et j, notée $\pi_{ij} = P(i \in s,\ j \in s) = \sum_{s \ni (i \& j)} p(s)$. Cela donne lieu à une particularité du contexte des populations finies, où la covariance entre les unités d'un échantillon sélectionné sans remise peut ne pas être nulle, puisque la probabilité d'inclusion d'une unité dans l'échantillon peut différer conditionnellement à la présence d'une autre unité.

1.1.2. Plan de sondage

1.1.2.1. *Échantillonnage stratifié aléatoire simple sans remise*

Dans ce mémoire, nous considérerons en premier lieu des plans de sondage pour lesquels la taille de l'échantillon n est fixe. Un plan élémentaire est l'échantillonnage aléatoire simple sans remise (EASSR), dans lequel tous les échantillons de taille n ont la même probabilité $1/\binom{N}{n}$ d'être sélectionnés. Incidemment, chaque unité de la population a la même probabilité de se retrouver dans l'échantillon, ce qui en fait un plan à *probabilités égales*. L'EASSR est un cas particulier d'un plan stratifié dans lequel il n'y aurait qu'une seule strate. En revanche, il

est rarement utilisé en pratique, contrairement à l'échantillonnage stratifié. Dans le plan stratifié aléatoire simple sans remise, la population $\mathcal{P}$ est divisée en R strates $\mathcal{P}_1, \ldots, \mathcal{P}_r, \ldots, \mathcal{P}_R$ de tailles $N_1, \ldots, N_r, \ldots, N_R$ respectivement. Ensuite, un EASSR de taille n_r est sélectionné à partir de la strate r, $r = 1, \ldots, R$. Ce processus est fait de manière indépendante à travers les R strates. Sauf dans le cas d'une allocation proportionnelle à la taille de la strate, le plan stratifié constitue un exemple de plan de sondage à probabilités *inégales*, puisque les unités de la grande population U ne possèdent généralement pas toutes la même probabilité d'être incluses dans l'échantillon.

1.1.2.2. *Échantillonnage de Poisson*

Le plan stratifié EASSR sera d'abord considéré puisque c'est à celui-ci que peut s'appliquer la première méthode de rééchantillonnage par pseudo-population abordée au chapitre 2. Néanmoins, des extensions ont été développées pour des plans de sondage dans lesquels la taille d'échantillon est aléatoire. Un tel plan qui sera étudié dans ce mémoire est l'échantillonnage de Poisson. Dans celui-ci, on assigne à chaque unité de U la probabilité π_i d'être sélectionnée dans l'échantillon. Ensuite, N épreuves Bernoulli indépendantes paramétrées selon les probabilités de succès π_i définies préalablement sont conduites à travers la population. Il s'agit également d'un plan à probabilités *inégales*, sauf dans le cas de l'échantillonnage de Bernoulli où $\pi_i \equiv \pi, \forall i,\ i = 1, \ldots, N$.

Les probabilités de sélection dans le plan de Poisson peuvent être définies de façon à refléter un modèle générateur linéaire donné. Il y a alors d'emblée la supposition de l'existence d'une superpopulation ayant généré les observations de la population. Par exemple, on peut définir des probabilités d'inclusion proportionnelles aux valeurs d'un régresseur x qu'on soupçonne en pratique être corrélé à la variable d'intérêt y. Puisque le plan de Bernoulli consiste en des probabilités d'inclusion constantes, on peut faire correspondre à ce plan le modèle générateur suivant pour les observations $y_1, y_2, \ldots, y_N$:

$$y_i = \mu + \varepsilon_i, \tag{1.1.1}$$

où $\mu \in \mathbb{R}$ et $\varepsilon_i,\ i = 1, \ldots, N$, sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) selon une loi de probabilité $\mathcal{L}$ de moyenne 0 et de variance σ^2, aussi dénotée $\mathcal{L}(0,\sigma^2)$. En présence d'une variable auxiliaire x, on peut envisager le modèle de

régression linéaire simple sans ordonnée à l'origine

$$y_i = \beta x_i + \varepsilon_i, \tag{1.1.2}$$

où $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{L}(0,\sigma^2),\ i = 1, \ldots, N$ et $\beta \in \mathbb{R}$. Sachant que les mesures de la variable y peuvent être ainsi représentées et sous la condition que les valeurs de x soient connues pour l'ensemble de la population, l'estimateur Horvitz-Thompson d'un total (défini un peu plus loin) connaîtra un gain d'efficacité en posant des probabilités d'inclusion proportionnelles à la taille (Särndal et al., 1992), c'est-à-dire comme

$$\pi_i = n x_i / \sum_{i=1}^{N} x_i. \tag{1.1.3}$$

En rappelant que la taille d'échantillon, dénotée n_s, est aléatoire dans un tel plan, ces probabilités d'inclusion satisfont la condition que

$$\sum_{i=1}^{N} \pi_i = n,$$

où n est la taille d'échantillon espérée.

1.2. Estimation de paramètres d'une population finie

L'estimation des paramètres d'une population finie est faite sur la base de l'information contenue dans l'échantillon. À des fins pratiques, nous nous attarderons au paramètre du total de la population, défini par $\theta \equiv t = \sum_{i=1}^{N} y_i$. En réalité, les quantités d'intérêt se trouvent souvent à être des fonctions de totaux. Lors de la dérivation de propriétés des estimateurs, nous aurons recours aux opérateurs d'espérance et de variance sous le plan, dénotés $\mathbb{E}_p$ et Var_p respectivement.

1.2.1. Estimateur Horvitz-Thompson d'un total

Un estimateur linéaire sans biais du total, basé sur les probabilités d'inclusion des unités, est l'estimateur Horvitz-Thompson s'écrivant

$$\hat{t}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}. \tag{1.2.1}$$

Soit Z_i la variable indicatrice associée à l'événement que l'unité i soit incluse dans l'échantillon et de manière analogue, soit $Z_{ij} = Z_i Z_j$ la variable indicatrice associée à l'inclusion conjointe des unités i et j. On a que $\mathbb{E}[Z_i] = P(i \in s) = \pi_i$ et $\mathbb{E}[Z_{ij}] = P(i \in s, j \in s) = \pi_{ij}$.

On peut réécrire l'estimateur Horvitz-Thompson en fonction des unités au sein de la population au moyen des indicatrices comme ceci:

$$\hat{t}_{HT} = \sum_{i=1}^{N} Z_i \frac{y_i}{\pi_i}.$$

Avec cette écriture, il devient presqu'immédiat de déduire que $\mathbb{E}_p\left[\hat{t}_{HT}\right] = t$.

1.2.2. Estimation de la variance de l'estimateur Horvitz-Thompson

On formule à présent l'expression de la variance de l'estimateur Horvitz-Thompson du total. Pour ce faire, on pose $\Delta_{ij} := \mathrm{Cov}(Z_i, Z_j)$. On note que $\Delta_{ij} = \mathbb{E}[Z_{ij}] - \mathbb{E}[Z_i]\mathbb{E}[Z_j] = \pi_{ij} - \pi_i\pi_j$. Ainsi, il s'en suit que la variance de l'estimateur sur la base d'un plan de sondage $p(\cdot)$ peut s'écrire

$$\mathrm{Var}_p(\hat{t}_{HT}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{1.2.2}$$

Par exemple, dans le cas du plan de sondage EASSR, la covariance entre les observations y_i et y_j, où $i, j \in s$, est donnée par

$$\Delta_{ij} = \frac{n}{N}\frac{n-1}{N-1} - \frac{n^2}{N^2} = -\frac{1}{N-1}\left(1 - \frac{n}{N}\right)\frac{n}{N},$$

ce qui donne lieu à la variance suivante:

$$\mathrm{Var}_p(\hat{t}_{HT}) = N^2(1-f)\frac{S^2}{n}, \tag{1.2.3}$$

où $f = n/N$, $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$ et $\bar{y}_U = \sum_{i=1}^{N} y_i/N$. Sous l'échantillonnage de Poisson, puisque $\pi_{ij} = \pi_i\pi_j$, on a naturellement $\Delta_{ij} = 0$ pour $i \neq j$. Ainsi, pour ce plan de sondage, la variance de l'estimateur du total se réduit à

$$\mathrm{Var}_p(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2. \tag{1.2.4}$$

En rappelant que $\mathbb{E}[Z_iZ_j] = \pi_{ij}$, il est facile de vérifier que l'estimateur suivant pour un plan de sondage arbitraire $p(\cdot)$

$$\widehat{\mathrm{Var}}_p(\hat{t}_{HT}) = \sum_{i\in s}\sum_{j\in s} \check{\Delta}_{ij} \frac{y_i}{\pi_j}\frac{y_i}{\pi_j},$$

où $\check{\Delta}_{ij} = \Delta_{ij}/\pi_{ij} = 1 - \pi_i\pi_j/\pi_{ij}$, est sans biais pour la variance de l'estimateur Horvitz-Thompson. Dans le cas du plan EASSR, la variance estimée peut être écrite plus simplement

en tant que

$$\widehat{\mathrm{Var}}_p(\hat{t}_{HT}) = N^2 (1-f) \frac{s^2}{n}, \tag{1.2.5}$$

où $s^2 = \sum_{i \in s} (y_i - \bar{y})^2 / (n-1)$ et $\bar{y} = \sum_{i \in s} y_i / n$. Si l'échantillon est généré selon le plan de Poisson, alors l'estimateur sans biais de la variance du total se réduit à

$$\widehat{\mathrm{Var}}_p(\hat{t}_{HT}) = \sum_{i \in s} \frac{1-\pi_i}{\pi_i^2} y_i^2. \tag{1.2.6}$$

Si N est connu, l'estimateur sans biais de la moyenne de la population $\bar{y}_U = t/N$ est dérivé à partir de celui du total pour obtenir $\bar{y}_s = \hat{t}_{HT}/N$. L'expression pour sa variance peut être dérivée directement. Alternativement, que N soit connu ou non, il est possible d'avoir recours à la moyenne échantillonnale pondérée (*weighted sample mean*), qui correspond au ratio de l'estimateur du total de y, noté, $\hat{t}_{HT}$, et de l'estimateur Horvitz-Thompson du total N:

$$\begin{aligned} \tilde{y}_s &= \frac{\hat{t}_{HT}}{\hat{N}} \\ &= \frac{\sum_{i \in s} y_i/\pi_i}{\sum_{i \in s} 1/\pi_i} \end{aligned} \tag{1.2.7}$$

Dans le cas de l'EASSR et le plan stratifié aléatoire simple sans remise, les probabilités d'inclusion sont telles que les quantités $\tilde{y}_s$ et $\bar{y}_s$ coïncident. Or, ce ne sera pas nécessairement le cas dans l'échantillonnage de Poisson, puisque dans la définition du plan, les probabilités de sélection n'ont pas à satisfaire $\sum_{i \in s} 1/\pi_i = N$. Tel que discuté dans l'ouvrage de Särndal et al. (1992), il y a alors un avantage à privilégier $\tilde{y}_s$ par rapport à $\bar{y}_s$ d'un point de vue d'efficacité dans le plan de Poisson. Cela dit, l'estimateur $\tilde{y}_s$ est une fonction non linéaire des observations d'un échantillon issu d'un plan de Poisson. Par conséquent, il est approximativement sans biais pour $\bar{y}_U$ et il n'est pas possible de formuler une expression exacte pour sa variance. La prochaine discussion s'articulera autour de la linéarisation par série de Taylor, qui constitue un moyen d'étudier les propriétés des estimateurs qui sont des fonctions non linéaires de totaux.

1.2.3. Technique de linéarisation pour l'estimation de la variance d'un ratio de deux estimateurs de totaux

Nous nous appliquerons à développer une expression pour la variance de l'estimateur d'un paramètre de population finie $\theta = f(t_y, t_w)$, où y et w sont deux variables à l'étude et

$f : \mathbb{R} \times \mathbb{R} \backslash \{0\} \to \mathbb{R}$ est définie par $f(x,y) = x/y$. L'estimateur approximativement sans biais correspondant est $\hat{\theta} = f\left(\hat{t}_y, \hat{t}_w\right)$, où $\hat{t}_y$ et $\hat{t}_w$ sont les estimateurs Horvitz-Thompson des totaux de y et w dans la population respectivement. Un cas particulier qui sera d'intérêt est celui où $w_i = 1$, $i = 1, \ldots, N$ lorsqu'il sera question de la moyenne échantillonnale pondérée, qui peut être vue comme un ratio de deux totaux.

Le développement en série de Taylor de premier ordre de f autour de θ permet de dériver un pseudo-estimateur, noté $\tilde{\theta}$:

$$\begin{aligned}
\tilde{\theta} &= \theta + \frac{\partial f}{\partial \hat{t}_y}\Big|_{(\hat{t}_y,\hat{t}_w)=(t_y,t_w)} (\hat{t}_y - t_y) + \frac{\partial f}{\partial \hat{t}_w}\Big|_{(\hat{t}_y,\hat{t}_w)=(t_y,t_w)} (\hat{t}_w - t_w) \\
&= \theta + \frac{1}{t_w}(\hat{t}_y - t_y) - \frac{t_y}{t_w^2}(\hat{t}_w - t_w) \\
&= \theta + \frac{1}{t_w}(\hat{t}_y - t_y) - \frac{\theta}{t_w}(\hat{t}_w - t_w) \\
&= \theta + \frac{1}{t_w}\left[\hat{t}_y - \theta\hat{t}_w - (t_y - \theta t_w)\right].
\end{aligned}$$

Une formule approximative de la variance de $\hat{\theta}$ peut ainsi être obtenue:

$$\begin{aligned}
\mathrm{Var}_p(\hat{\theta}) \approx \mathrm{Var}_p(\tilde{\theta}) &= \frac{1}{t_w^2}\mathrm{Cov}\left(\hat{t}_y - \theta\hat{t}_w, \hat{t}_y - \theta\hat{t}_w\right) \\
&= \frac{1}{t_w^2}\sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{y_i - \theta w_i}{\pi_i}\frac{y_j - \theta w_j}{\pi_j}.
\end{aligned}$$

Il en découle l'estimateur de la variance pour un ratio de deux totaux pour un plan de sondage quelconque, qui est obtenu en *injectant* les estimateurs des quantités inconnues dans l'expression précédente.

$$\widehat{\mathrm{Var}}_p(\hat{\theta}) = \frac{1}{\hat{t}_z^2}\sum_{i\in s}\sum_{j\in s}\check{\Delta}_{ij}\frac{y_i - \hat{\theta} w_i}{\pi_i}\frac{y_j - \hat{\theta} w_j}{\pi_j}. \tag{1.2.8}$$

Si on revient à la moyenne échantillonnale pondérée, alors $\hat{\theta} = \tilde{y}_s$ en prenant $w_i \equiv 1\ \forall i \in U$ et le résultat précédent permet de formuler une expression pour sa variance estimée.

$$\widehat{\mathrm{Var}}_p(\tilde{y}_s) = \frac{1}{\hat{N}^2}\sum_{i\in s}\sum_{j\in s}\check{\Delta}_{ij}\frac{y_i - \tilde{y}_s}{\pi_i}\frac{y_j - \tilde{y}_s}{\pi_j}. \tag{1.2.9}$$

1.3. Estimation de quantiles

L'estimation de quantiles est un problème important en théorie des sondages. Dans le cas notamment d'une population très asymétrique, la médiane pourrait être préférée à la moyenne comme paramètre de localisation de la distribution. L'estimation de quantiles en contexte de population finie a fait l'objet de plusieurs ouvrages visant entre autres à étudier l'estimation de leur variance, la construction d'intervalles de confiance et leur comportement asymptotique.

1.3.1. Définition des estimateurs

Les quantiles expérimentaux sont définis à partir de la fonction de répartition expérimentale, pour laquelle on fournit une définition formelle. Rappelons que l'on s'intéresse à une variable y prenant ses valeurs sur la droite réelle.

Définition 1.3.1 (Fonction de répartition de la population)**.** *La fonction de répartition de la population évaluée en t d'un vecteur de mesures $\boldsymbol{y} = (y_1, \ldots, y_N)'$ associe à un élément $t \in \mathbb{R}$ un élément de l'ensemble* [0,1] *et est définie par*

$$F_N(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i \leq t),$$

où $\mathbb{1}(A)$ dénote la fonction indicatrice pour l'ensemble A.

Définition 1.3.2 (Quantile de la population)**.** *Le quantile de niveau p de la population est défini par la relation*

$$\xi_p = F_N^{-1}(p),$$

où F_N^{-1} est la fonction inverse de F_N, définie par

$$F_N^{-1}(p) = \inf\{t : F_N(t) \geq p\}.$$

Un quantile souvent d'intérêt est la médiane, dénotée $M := \xi_{0,5}$. La fonction de répartition de la population F_N peut être vue comme la moyenne d'une variable indicatrice. Incidemment, étant donné un échantillon aléatoire tiré de la population, l'estimateur de la moyenne

pondérée (équation 1.2.7) de F_N calculé sur la base des observations s'écrit

$$\begin{aligned}\widehat{F}(t) &= \frac{1}{\widehat{N}} \sum_{i \in s} \frac{\mathbb{1}(y_i \leq t)}{\pi_i} \\ &= \frac{\sum_{i \in s} \mathbb{1}(y_i \leq t)/\pi_i}{\sum_{i \in s} 1/\pi_i}. \end{aligned} \tag{1.3.1}$$

Dans un plan stratifié EASSR doté de R strates, chacune de taille N_r, $r = 1, \ldots, R$, l'estimateur de fonction de répartition s'écrit

$$\widehat{F}(t) = \sum_{r=1}^{R} W_h \widehat{F}_r(t), \tag{1.3.2}$$

où $W_r = N_r/N$. Remarquons que pour ce plan, chacune des variables aléatoires $\widehat{F}_r(t)$, étant la proportion des éléments de s satisfaisant la condition $y_i \leq t$, est distribuée selon une loi hypergéométrique de paramètres n_r, $N_r F_r(t)$, N_r, $r = 1, \ldots, R$. Par conséquent, $\mathbb{E}_p[\widehat{F}_r(t)] = F_r(t)$ et par suite $\mathbb{E}_p[\widehat{F}(t)] = F_N(t)$.

En se fondant sur la définition 1.3.2, il découle de l'estimateur de fonction de répartition la formule de l'estimateur d'un quantile, s'écrivant

$$\hat{\xi}_p = \widehat{F}^{-1}(p). \tag{1.3.3}$$

Remarque 1.3.3. *Nous détaillons à présent l'algorithme de calcul d'un quantile échantillonnal basé sur les poids de sondage présenté dans le chapitre 5 de Särndal et al. (1992). La première étape consiste à ordonner les valeurs de l'échantillon* $(y_{j_1}, y_{j_2}, \ldots, y_{j_n})'$ *de manière à obtenir les statistiques d'ordre*

$$y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)},$$

et les poids de sondage correspondants $\pi_{j_i}, i = 1, \ldots, n$, *de façon à obtenir*

$$\pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(n)},$$

où $\pi_{(i)}$ *est le poids de sondage associé à la* i*-ème statistique d'ordre* $y_{(i)}$, $i = 1, \ldots, n$. *On définit le poids cumulatif* $\widehat{N}_k$ *pour* $k \in \{1, \ldots, n\}$ *comme étant*

$$\widehat{N}_k = \sum_{i=1}^{k} \frac{1}{\pi_{(i)}}.$$

Celui-ci correspond à l'estimation du rang de l'unité (i) dans la population. Ainsi, le quantile échantillonnal de niveau p est donné alternativement par

$$\hat{\xi}_p = \begin{cases} y_{(k)} & si\ \widehat{N}_{k-1} < p\widehat{N} < \widehat{N}_k \\ \frac{1}{2}(y_{(k)} + y_{(k+1)}) & si\ \widehat{N}_k = p\widehat{N} \end{cases}. \tag{1.3.4}$$

1.3.2. Estimation de la variance de la fonction de répartition échantillonnale

Dans l'optique de construire un intervalle de confiance approximatif pour un quantile selon la méthode de Woodruff (1952), il sera en premier lieu nécessaire d'estimer la variance de $\widehat{F}(\xi_p)$. Nous avons vu précédemment que cette quantité est un cas particulier de la moyenne échantillonnale pondérée pour la variable d'intérêt $d = \mathbb{1}(y \leq \xi_p)$. En substituant cette dernière dans (1.2.9), on trouve la forme préliminaire de l'estimateur de la variance de l'estimateur de la fonction de répartition échantillonnale, soit

$$\widetilde{\mathrm{Var}}_p(\widehat{F}(\xi_p)) = \frac{1}{\widehat{N}^2} \sum_{i \in s} \sum_{j \in s} \breve{\Delta}_{ij} \frac{d_i - \tilde{d}_s}{\pi_i} \frac{d_j - \tilde{d}_s}{\pi_j}, \tag{1.3.5}$$

où $\tilde{d}_s = \widehat{F}(\xi_p)$. Remarquons que les quantités $d_i = \mathbb{1}(y_i \leq \xi_p),\ i = 1, \ldots, n$ et $\tilde{d}_s$ dans (1.3.5) sont inconnues. Un dernier ajustement pour estimer la variance consiste à substituer ξ_p par le quantile échantillonnal dans l'expression.

Ainsi, dans le cas du plan EASSR, l'estimateur de la variance de la fonction de répartition expérimentale est donné par (Särndal et al., 1992)

$$\widehat{\mathrm{Var}}_p\left[\widehat{F}(\xi_p)\right] = \frac{1-f}{n-1} \widehat{F}(\hat{\xi}_p)\left[1 - \widehat{F}(\hat{\xi}_p)\right]. \tag{1.3.6}$$

Une extension de ce résultat au plan stratifié EASSR comportant R strates peut être faite facilement par indépendance des tirages d'une strate à l'autre, auquel cas l'estimateur est donné par

$$\widehat{\mathrm{Var}}_p\left[\widehat{F}(\xi_p)\right] = \sum_{r=1}^{R} \frac{1-f_r}{n_r - 1} W_r^2 \widehat{F}_r(\hat{\xi}_p)\left[1 - \widehat{F}_r\left(\hat{\xi}_p\right)\right]. \tag{1.3.7}$$

Dans le cas du plan de Poisson, l'expression de l'estimateur de variance pour une fonction de répartition peut être réduite plus simplement à

$$\widehat{\mathrm{Var}}_p\left[\widehat{F}(\xi_p)\right] = \frac{1}{\widehat{N}^2} \sum_{i \in s} \frac{1-\pi_i}{\pi_i^2} \left[\mathbb{1}(y_i \leq \hat{\xi}_p) - \widehat{F}\left(\hat{\xi}_p\right)\right]^2, \tag{1.3.8}$$

où $\widehat{N} = \sum_{i \in s} \pi_i^{-1}$.

1.3.3. Intervalle de confiance de Woodruff pour un quantile

Woodruff (1952) a développé une méthode pour construire un intervalle de confiance pour un quantile, qui est basée sur l'inversion de l'intervalle de confiance pour une fonction de répartition, i.e.

$$P\left(c_1 \leq \hat{F}(\xi_p) \leq c_2\right) \approx P\left(\hat{F}^{-1}(c_1) \leq \xi_p \leq \hat{F}^{-1}(c_2)\right), \tag{1.3.9}$$

où $c_1, c_2 \in (0,1)$. Il suffit donc de trouver c_1 et c_2 tels que

$$P\left(c_1 \leq \hat{F}(\xi_p) \leq c_2\right) = 1 - \alpha,$$

où $1 - \alpha$ est le niveau nominal de l'intervalle de confiance. Afin de poursuivre, on fait l'hypothèse que $\widehat{F}(\xi_p) \sim \mathcal{N}(\mathbb{E}_p[\widehat{F}(\xi_p)], \mathrm{Var}_p(\widehat{F}(\xi_p))$. De fait, la normalité asymptotique de l'estimateur de fonction de répartition a été montrée par Francisco et Fuller (1991). Dans le cas des plans équilibrés par rapport à la taille de la population, c'est-à-dire satisfaisant $\sum_{i \in s} 1/\pi_i = N$, on a exactement que $\mathbb{E}_p[\widehat{F}(\xi_p)] = F(\xi_p) = p$. Dans les plans pour lesquels $\sum_{i \in s} 1/\pi_i \neq N$, l'estimateur $\widehat{F}(\xi_p)$ est seulement approximativement sans biais pour $F(\xi_p)$. Les hypothèses précédentes permettent de déduire la relation approximative suivante basée sur la distribution normale:

$$P\left(p - z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}_p(\widehat{F}(\xi_p))} \leq \widehat{F}(\xi_p) \leq p + z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}_p(\widehat{F}(\xi_p))}\right) \approx 1 - \alpha \tag{1.3.10}$$

Par la suite, la variance de l'estimateur de la fonction de répartition $\mathrm{Var}_p(\widehat{F}(\xi_p))$ constituant l'expression ci-haut doit être estimée par $\widehat{\mathrm{Var}}_p(\widehat{F}(\xi_p))$. En prenant la réciproque $\widehat{F}^{-1}$ de chaque membre à l'intérieur de la probabilité (1.3.10), on déduit enfin l'intervalle de confiance approximatif de niveau $1 - \alpha$ pour ξ_p suivant:

$$\left[\widehat{F}^{-1}\left(p - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}_p(\widehat{F}(\xi_p))}\right), \widehat{F}^{-1}\left(p + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}_p(\widehat{F}(\xi_p)}\right)\right]. \tag{1.3.11}$$

Tel que noté par Sitter et Wu (2001), pour des tailles échantillonnales modérées, la distribution échantillonnale de $\widehat{F}(\xi_p)$ sera d'autant plus asymétrique que le niveau du quantile p sera éloigné de 1/2. Les études de simulations qu'ils ont réalisées confirment que dans ces conditions, le taux d'erreur de couverture réel des intervalles pour $F(\xi_p)$ ne correspond pas au taux nominal lorsque p se rapproche de 0 ou de 1. Ainsi, pour cette

raison, il serait attendu que les intervalles de confiance de Woodruff (1.3.11) pour ξ_p aient un comportement indésirable pour de petites ou de grandes valeurs de p. Or, les mêmes simulations conduites par Sitter et Wu portent à croire que le taux d'erreur de couverture nominal de ces intervalles est atteint en réalité, en dépit de l'asymétrie de la distribution échantillonnale de $\widehat{F}(\xi_p)$. Cela suggère que l'approximation (1.3.9) et le fait de substituer ξ_p par $\hat{\xi}_p$ dans l'estimation de la variance de $\widehat{F}\,(\xi_p)$ tendent à rééquilibrer les choses de façon avantageuse. La convergence des intervalles de Woodruff a également été montrée par Francisco et Fuller (1991).

L'approche de Woodruff pour la construction d'intervalles de confiance fait l'objet de notre attention puisqu'elle constitue une alternative valide peu coûteuse aux méthodes de rééchantillonnage qui seront abordées au chapitre 2.

1.3.4. Estimation de la variance de l'estimateur d'un quantile via l'intervalle de confiance de Woodruff

Si l'on fait la supposition qu'un quantile échantillonnal $\hat{\xi}_p$ est normalement distribué, on peut lui faire correspondre un intervalle de confiance basé sur la loi normale de niveau $1-\beta$, où $\beta \in (0,1)$. Nous pouvons par la suite fournir une estimation de la variance de l'estimateur du quantile à partir de la demi-longueur de cet intervalle. En effet, de ces hypothèses découle la relation

$$\frac{L(\beta)}{2} = z_{1-\beta/2}\sqrt{\mathrm{Var}(\hat{\xi}_p)}, \tag{1.3.12}$$

où $z_{1-\beta/2}$ est le quantile $1-\beta/2$ de la distribution normale centrée réduite et $L(\beta)$ est la longueur d'un intervalle de niveau $(1-\beta)$ %. Un estimateur de la variance de ξ_p découlant de (1.3.12) est

$$\widehat{\mathrm{Var}}_\beta(\hat{\xi}_p) = \left(\frac{L(\beta)}{2z_{\beta/2}}\right)^2. \tag{1.3.13}$$

Une approche envisageable est de mettre à profit l'intervalle de Woodruff construit à la sous-section précédente pour dériver l'estimateur de variance du quantile échantillonnal. Comme cet estimateur dépend en principe du taux d'erreur nominal β de l'intervalle, à chaque valeur de β correspondra une estimation différente de la variance. Par étude de simulation, il est possible de sélectionner la valeur du paramètre de lissage β qui minimisera

le biais ou la précision de l'estimateur de variance pour une population, un plan de sondage $p(\cdot)$ et une fraction de sondage f donnés.

Les démarches vues jusqu'à présent font appel à des outils élémentaires traités en échantillonnage pouvant être appliqués à des fonctionnelles linéaires comme les fonctions de distribution (dans le cas où la taille d'échantillon est fixe). Nous terminerons ce chapitre en traitant brièvement des propriétés asymptotiques des quantiles échantillonnaux en contexte de population finie.

1.3.5. Distribution asymptotique des quantiles

Dans cette section, nous rapporterons quelques résultats démontrés par Chatterjee (2011) concernant la distribution asymptotique des quantiles échantillonnaux dans le cas du plan EASSR. Ces propriétés étant naturellement analogues à celles dérivées pour des données indépendantes et identiquement distribuées, nous rappelons d'abord des résultats dans le contexte d'une population infinie.

Soit $Y_1, Y_2, \ldots, Y_n$ un échantillon d'observations indépendantes et identiquement distribuées selon une distribution F_0 avec fonction de densité de probabilité f_0. Soit $\hat{F}_n(t)$ la fonction de répartition évaluée en $t \in \mathbb{R}$ pour cet échantillon définie par $\hat{F}_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(Y_i \leq t)$. Alors, en vertu du théorème limite central, il est facile de vérifier que

$$\sqrt{n}(\hat{F}_n(t) - F_0(t)) \xrightarrow{d} \mathcal{N}(0, F_0(t)\,[1 - F_0(t)]).$$

Soit $\tilde{\xi}_p = F_0^{-1}(p)$ le quantile de niveau p de la distribution F_0. En définissant $\check{\xi}_p = \hat{F}_n^{-1}(p)$ le quantile de niveau p associé à l'échantillon d'observations i.i.d., en supposant que $f_0(\tilde{\xi}_p) > 0$, il est possible de montrer que

$$\sqrt{n}(\check{\xi}_p - \tilde{\xi}_p)) \xrightarrow{d} \mathcal{N}(0, \rho^2),$$

où $\rho^2 = p(1-p)/f_0^2(\tilde{\xi}_p)$. Ce résultat peut être montré au moyen de la méthode delta en imposant des conditions de régularité sur f_0. Une façon alternative de montrer ce résultat consiste à étudier la distribution de la représentation de Bahadur du quantile échantillonnal

(Francisco et Fuller, 1991; Sitter et Wu, 2001), donnée par

$$\check{\xi}_p = \tilde{\xi}_p - \left[f_0(\tilde{\xi}_p)\right]^{-1} [\hat{F}_n(\tilde{\xi}_p) - F_0(\tilde{\xi}_p)] + o_p(n^{-1/2}). \tag{1.3.14}$$

En se remémorant que la variable aléatoire $\hat{F}_n(\tilde{\xi}_p)$ est distribuée selon une loi binomiale de paramètres n et p, on voit intuitivement la correspondance entre la variance asymptotique de $\check{\xi}_p$ et celle du membre de droite de (1.3.14).

De façon générale, si l'on s'intéresse aux propriétés asymptotiques d'estimateurs de paramètres de population finie, il nous faut approfondir le cadre théorique déjà établi. On suppose à présent que la population dont l'échantillon s est tiré est elle-même un échantillon aléatoire d'une superpopulation distribuée selon une fonction de répartition F_0 avec fonction de densité f_0. L'opérateur d'espérance dans la superpopulation est noté $\mathbb{E}_0$. De plus, on fait l'hypothèse que cette superpopulation a pour espérance $\mu < \infty$ et variance $\sigma^2 < \infty$.

Maintenant, soit un indice $r \in \{1, 2, \ldots\}$ et soit $U_{N_r} = \{1, \ldots, i, \ldots, N_r\}$ une population constituée d'observations i.i.d. issues de la distribution F_0 et s_{n_r} un échantillon aléatoire simple sans remise tiré de la population U_{N_r}. Suivant ce nouveau cadre théorique, pour $t \in \mathbb{R}$, les fonctions de répartition de la population et échantillonnale pour le plan EASSR sont redéfinies comme étant respectivement

$$F_{N_r}(t) = \frac{1}{N_r} \sum_{i=1}^{N_r} \mathbb{1}(y_i \leq t) \quad \text{et} \quad \widehat{F}_{n_r}(t) = \frac{1}{n_r} \sum_{i \in s_{n_r}} \mathbb{1}(y_i \leq t).$$

Les quantiles de niveau p correspondants sont $\xi_{p,N_r} = F_{N_r}^{-1}(p)$ et $\hat{\xi}_{p,n_r} = \widehat{F}_{n_r}^{-1}(p)$. Pour tout $r \in \mathbb{N}$, on définit le quantile standardisé comme

$$\zeta_r = a_r^{-1} \left(\hat{\xi}_{p,n_r} - \xi_{p,N_r}\right). \tag{1.3.15}$$

où $f_r = n_r/N_r$ correspond à la fraction de sondage et $a_r = n_r/(1 - f_r)$ est le facteur de normalisation. Pour poursuivre, on doit introduire la distribution échantillonnale de ζ_r étant donné U_{N_r}

$$G_{n_r}(t) = P_{\cdot|U_{N_r}}(\zeta_r \leq t),$$

qui n'est rien d'autre que la distribution sous le plan pour cette population fixe. De façon analogue, les quantités $\mathbb{E}_{\cdot|U_{N_r}}$ et $\mathrm{Var}_{\cdot|U_{N_r}}$ dénotent respectivement l'espérance et la variance conditionnelles à cette population. On dénote le quantile de niveau p de la superpopulation

$\tilde{\xi}_p = F_0^{-1}(p)$ afin de le distinguer de celui de la population finie. Les résultats suivants, montrés par Chatterjee (2011), caractérisent la distribution asymptotique de ζ_r sous le plan EASSR.

Théorème 1.3.4. *En supposant que* $f_0^2(\tilde{\xi}_p) > 0$ *et que*

$$\lim_{r\to\infty} f_r = f, \quad \textit{pour } f \in (0,1), \tag{1.3.16}$$

alors ζ_r *converge en loi vers une distribution* $\mathcal{N}(0,\rho^2)$ *presque sûrement, où* $\rho^2 = p(1-p)/f_0^2(\tilde{\xi}_p)$.

Théorème 1.3.5. *En supposant qu'il existe* $\alpha > 0$ *tel que* $\mathbb{E}_0\left[|Y|^\alpha\right] < \infty$ *et que la condition (1.3.16) du théorème 1.3.4 est satisfaite, alors pour tout* $\delta \in (0,\infty)$

$$\sup_{r\geq 1} \mathbb{E}_{\cdot|U_{N_r}}\left[|\zeta_r|^{2+\delta}\right] < \infty, \tag{1.3.17}$$

presque sûrement.

En particulier, on peut déduire du théorème précédent que $\mathbb{E}_{\cdot|U_{N_r}}\left[|\zeta_r|^2\right] < \infty$, ce qui mène au corollaire suivant.

Corollaire 1.3.6. *Si les conditions des théorèmes 1.3.4 et 1.3.5 sont satisfaites, alors*

$$\mathrm{Var}_{\cdot|U_{N_r}}(\zeta_r) \to \rho^2 \quad \textit{lorsque} \quad r \to \infty, \tag{1.3.18}$$

presque sûrement.

Pour future référence, il convient d'alléger la notation utilisée pour désigner la variance asymptotique de $\hat{\xi}_p$ sous le plan de sondage EASSR. Nous laisserons de côté la suite de populations indicées par r, tout en gardant en mémoire que la taille échantillonnale n et la taille de la population N tendent toutes deux vers l'infini avec une fraction de sondage qui converge vers $f \in (0,1)$. Ainsi, les résultats précédents permettent d'établir

$$\mathrm{Var}_p(\hat{\xi}_p \mid U) \to \frac{1-f}{n}\rho^2 \quad \text{lorsque} \quad n \to \infty, \tag{1.3.19}$$

presque sûrement, où $\mathrm{Var}(\hat{\xi}_p \mid U)$ est la variance sous le plan EASSR conditionnellement à ce que la population finie soit U.

Chapitre 2

Bootstrap en contexte de population finie

Nous avons vu précédemment qu'il était possible de construire un intervalle de confiance et d'estimer la précision des estimateurs pour le problème particulier de l'estimation de quantiles de population finie. Comme pour toutes les fonctionnelles qui ne sont pas linéaires, les méthodes analytiques reposent sur une série d'approximations. Cette section vise à survoler quelques méthodes de rééchantillonnage utilisées pour estimer la distribution échantillonnale d'un estimateur lorsque la population est finie. Nous commençons par introduire le bootstrap non paramétrique d'Efron (1979) pour des données i.i.d. pour ensuite poursuivre avec des méthodes de rééchantillonnage tenant compte du plan de sondage.

2.1. Méthode bootstrap pour des données i.i.d.

La méthode du bootstrap introduite par Efron (1979) vise à fournir une estimation de la distribution échantillonnale d'une statistique donnée. Supposons que l'on dispose d'un échantillon d'observations i.i.d. $Y_1, Y_2, \ldots, Y_n$ issues d'une distribution inconnue F_0. Une certaine fonctionnelle d'intérêt $\theta = \theta(F_0)$, telle que l'espérance ou la variance associée à cette loi de probabilité, est alors estimée par $\hat{\theta} = \theta(\hat{F})$, où $\hat{F}$ est une estimation de F_0 basée sur l'échantillon. La distribution de $\hat{\theta}$ est à toutes fins pratiques inconnue au même titre que F_0. Le bootstrap consiste à simuler des échantillons i.i.d. notés $Y_1^*, Y_2^*, \ldots, Y_n^*$ à partir de $\hat{F}$ et à estimer la distribution de $\hat{\theta}$ par celle de $\hat{\theta}^*$, qui est la même statistique calculée à partir de $Y_1^*, Y_2^*, \ldots, Y_n^*$. Le principe d'injection sous-jacent à cette méthode étant applicable à grande échelle, il est possible de le mettre en œuvre pour une grande variété de fonctionnelles, incluant des fonctionnelles non linéaires comme les quantiles.

Des arguments théoriques garantissent la convergence de la méthode pour les quantiles et d'autres statistiques usuelles. À cet égard, la médiane a reçu l'attention d'Efron (1979) dans son ouvrage fondateur.

Comme en atteste la notation utilisée ci-haut, le principe d'estimation derrière le bootstrap est de nature injective (ou *plug-in*), alors que l'estimation est effectuée au niveau de la distribution elle-même. En particulier, le bootstrap non paramétrique d'Efron prescrit d'estimer la distribution F_0 par la fonction de répartition expérimentale de l'échantillon $\hat{F}_n$, qui attribue le poids $1/n$ à chaque observation de l'échantillon $Y_1, Y_2, \ldots, Y_n$ (Efron et Tibshirani, 1993). Ceci étant, conditionnellement à l'observation de $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, la distribution de la variable aléatoire $\hat{\theta}$ est estimée par celle de $\hat{\theta}^* = \theta(\hat{F}_n^*)$, où $\hat{F}_n^*$ est la fonction de répartition expérimentale d'un échantillon bootstrap. Des attributs spécifiques de la distribution de $\hat{\theta}$ peuvent alors être estimés par le biais de la distribution bootstrap. Soit $J_n(t, F_0) = \text{Prob}_{F_0}(\hat{\theta} - \theta \leq t)$ la fonction de répartition associée à la distribution échantillonnale de $\hat{\theta}$. Dans ce mémoire, un intérêt sera porté à deux caractéristiques de cette distribution, soit la variance associée à J_n, dénotée $\alpha(F_0)$, ainsi que la fonction quantile, $J_n^{-1}(t, F_0)$.

L'estimation bootstrap de la variance de l'estimateur est obtenue en injectant la fonction de répartition expérimentale dans la fonctionnelle $\alpha(\cdot)$, donnant $\alpha(\hat{F}_n)$. L'estimateur $\alpha(\hat{F}_n)$ équivaut à la variance de $\hat{\theta}^*$ conditionnellement à l'échantillon initial $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, que l'on note $V^* = \text{Var}(\hat{\theta}^* \mid y_1, \ldots, y_n)$. Comme il n'est pas toujours possible de formuler analytiquement les mesures de précision de la distribution bootstrap $J_n(t, \hat{F}_n)$, une approximation Monte Carlo consiste à générer un grand nombre B d'échantillons de taille n à partir de $\widehat{F}_n$. L'algorithme se décline comme suit.

Algorithme du bootstrap non paramétrique avec remise

(1) Tirer un échantillon de taille n à partir de la distribution $\hat{F}_n$, ce qui équivaut à tirer un échantillon de taille n avec remise à partir de $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. De l'échantillon bootstrap $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$ peut être calculé l'estimateur bootstrap de θ, i.e. $\hat{\theta}^*$.

(2) L'étape 1 est répétée un grand nombre de fois B de façon à obtenir la collection $(\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*)'$, qui résulte en une estimation bootstrap de la distribution de $\hat{\theta}$.

À présent, rappelons que notre problématique se situe en contexte de population finie. Il est parfaitement possible de mettre en œuvre l'algorithme du bootstrap non paramétrique dans le cas où la population dont est tiré l'échantillon est finie. Cela dit, il faudrait alors faire comme si les observations étaient indépendantes et identiquement distribuées, hypothèse qui n'est pas satisfaite si le plan de sondage est sans remise ou si les probabilités d'inclusion sont inégales. Le cas échéant, le fait que les échantillons bootstrap soient tirés avec remise dans l'algorithme à partir de s ne permet pas de refléter les particularités de l'enquête. Cette assertion est illustrée au moyen de l'exemple de l'estimateur Horvitz-Thompson sous le plan de sondage EASSR, pour lequel une formule pour $\mathrm{Var}\left(\hat{t}_{HT}^* \mid y_1, \ldots, y_n\right)$ peut être dérivée explicitement.

2.1.1. Cas de l'estimation de la variance de l'estimateur Horvitz-Thompson du total

Proposition 2.1.1 (Estimateur bootstrap de la variance pour un total)**.** *Dans le cas du plan de sondage EASSR , l'estimateur bootstrap de la variance de l'estimateur Horvitz-Thompson du total est*

$$\mathrm{Var}^*\left(\hat{t}_{HT}^* \mid y_1, \ldots, y_n\right) = N^2 \left(\frac{n-1}{n}\right) \frac{s^2}{n}.$$

Démonstration. En premier lieu, pour $i \in s$,

$$\begin{aligned}\mathbb{E}^*[y_i^*] := \mathbb{E}^*\left[y_i^* \mid y_1, \ldots, y_n\right] &= \sum_{j \in s} y_j P(y_i^* = y_j) \\ &= \bar{y}.\end{aligned}$$

De là suit que

$$\begin{aligned}
\mathrm{Var}^*\left(\hat{t}^*_{HT} \mid y_1, \ldots, y_n\right) &= \frac{N^2}{n^2} \sum_{i=1}^{n} \mathrm{Var}^*\left(y_i^* \mid y_1, \ldots, y_n\right) \\
&= \frac{N^2}{n^2} \sum_{i=1}^{n} \mathbb{E}^*\left[\left(y_i^* - \mathbb{E}^*[y_i^*]\right)^2 \mid y_1, \ldots, y_n\right] \\
&= \frac{N^2}{n^2} \sum_{i=1}^{n} \left\{ \sum_{j \in s} \left(y_j - \bar{y}\right)^2 P(y_i^* = y_j) \right\} \\
&= \frac{N^2}{n^2} \sum_{j \in s} \left(y_j - \bar{y}\right)^2 \\
&= N^2 \left(\frac{n-1}{n}\right) \frac{s^2}{n}.
\end{aligned}$$

□

Cet estimateur de la variance pour $\hat{t}_{HT}$ est biaisé, son expression ne rendant pas compte du facteur de correction pour population finie $(1 - f)$ normalement présent (équation 1.2.5).

2.1.2. Cas de l'estimation de la variance d'un estimateur de quantile

L'intuition porte à croire que l'estimateur de variance donné par le bootstrap non paramétrique standard sera également gonflé dans le cas d'un quantile de population finie de niveau p, ξ_p, puisque le mécanisme de tirage avec remise ne permet généralement pas de tenir compte de la dépendance entre les observations. De plus, rappelons que la distribution échantillonnale du quantile de niveau p est intimement liée à celle de la fonction de répartition expérimentale évaluée en ξ_p, comme en témoigne la représentation de Bahadur (équation 1.3.14). Or, la fonction de répartition expérimentale évalulée en ξ_p n'est rien d'autre que la moyenne échantillonnale de la variable indicatrice $y \leq \xi_p$. Par la proposition 2.1.1, il suit de ce raisonnement qu'une inflation de l'estimateur de variance pour $\widehat{F}_n(\xi_p)$ entraînera une inflation de l'estimateur de variance pour $\widehat{\xi}_p$ lorsque le moyen d'estimation utilisé est celui du bootstrap non paramétrique avec remise.

Chatterjee (2011) montre formellement qu'une application naïve du bootstrap classique échoue à estimer la variance de l'estimateur d'un quantile en contexte de population finie dans le cas du plan EASSR. On réinvoque le cadre asymptotique décrit à la sous-section

1.3.5, où l'on étudie des suites d'entiers indicés par r $\{n_r\}$, $\{N_r\}$ que l'on fait tendre à l'infini sous condition que $\lim_{r\to\infty} n_r/N_r = f$ avec $f \in (0,1)$. Soit $s^*_{n_r}$ un échantillon bootstrap, donné explicitement par $(y^*_{j_1}, \ldots, y^*_{j_{n_r}})'$, tiré avec remise de l'échantillon s_{n_r} et soit la fonction de répartition expérimentale pour cet échantillon

$$\widehat{F}^*_{n_r}(t) = \frac{1}{n_r} \sum_{k \in s^*_{n_r}} \mathbb{1}(y^*_k \leq t), \quad t \in \mathbb{R}.$$

Le quantile bootstrap de niveau p est naturellement dénoté comme $\hat{\xi}^*_{p,n_r} = \widehat{F}^{*-1}_{n_r}(p)$. En utilisant (1.3.15), on définit à présent une version bootstrap pour ζ_r conditionnellement à l'échantillon s_{n_r} comme

$$\zeta^*_r = a_r^{-1} \left(\hat{\xi}^*_{p,n_r} - \hat{\xi}_{p,n_r} \right), \tag{2.1.1}$$

en rappelant que $a_r = \sqrt{n_r/(1-f_r)}$. Le résultat ci-dessous de Chatterjee (2011) atteste de l'échec du bootstrap i.i.d. pour une fraction de sondage $f \in (0,1)$.

Théorème 2.1.2. *Si la condition (1.3.16) du théorème 1.3.4 est satisfaite, alors ζ^*_r converge en loi vers la distribution $\mathcal{N}(0, \rho^2/(1-f))$.*

Ainsi, la variance de la distribution limite bootstrap de ζ^*_r correspond à celle mentionnée dans le premier chapitre pour la distribution du quantile échantillonnal standardisé ζ_r, donnée au théorème 1.3.4, multipliée par un facteur $(1-f)^{-1} > 1$. Cet estimateur de variance est donc gonflé par rapport à la variance asymptotique du quantile standardisé sous le plan EASSR.

2.2. Méthodes bootstrap par pseudo-population

Les méthodes bootstrap par pseudo-population forment une classe de méthodes de rééchantillonnage en contexte de population finie qui permettent de tenir compte du plan de sondage. En reconstituant une pseudo-population et en rééchantillonnant à partir de celle-ci selon le plan de sondage de l'enquête, on peut s'attendre à ce que l'estimateur de variance bootstrap reflète naturellement la variance véritable de l'estimateur échantillonnal, qui comporte le facteur de correction pour population finie. Formellement, le rôle joué précédemment par F_0 est à présent tenu par la population U et le paramètre que l'on souhaite estimer est de la forme $\theta = \theta(U)$ (Booth et al., 1994). Nous disposons d'un échantillon de n observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ issu de U selon le plan de sondage $p(\cdot)$. Conformément au principe d'injection, une fonctionnelle d'intérêt reposant sur une

population finie U, dénotée $\alpha(U)$, est estimée par $\alpha(\widehat{U})$, où $\widehat{U}$ est une estimation de la population basée sur l'information apportée par l'échantillon $\boldsymbol{y}$. La fonctionnelle $\alpha(U)$ pourrait être à titre d'exemple la variance de l'estimateur $\hat{\theta}$.

De manière analogue à la méthode bootstrap dans le cadre d'une population infinie, en pratique, des échantillons bootstrap sont générés à partir d'une pseudo-population U^*, qui est une possibilité pour $\widehat{U}$. Cela revient à substituer U par U^* dans la fonctionnelle $\alpha(\cdot)$. Afin de refléter la réalité de l'enquête, les échantillons bootstrap sont générés selon le plan de sondage $p(\cdot)$ utilisé initialement (Booth et al., 1994). Puis, la distribution échantillonnale de l'estimateur $\hat{\theta} = \hat{\theta}(s)$ est estimée par celle de l'estimateur bootstrap $\hat{\theta}^* = \hat{\theta}(s^*)$, qui est la même statistique obtenue sur la base des observations contenues dans un échantillon bootstrap s^*. La littérature existante sur les méthodes bootstrap par pseudo-population adresse entre autres la manière de former la pseudo-population U^*. La prochaine sous-section sera dédiée aux méthodes par pseudo-population adaptées à l'échantillonnage aléatoire simple sans remise et de façon plus générale, au plan stratifié aléatoire simple sans remise. Pour une revue des méthodes bootstrap appliquées à l'échantillonnage, incluant les méthodes par pseudo-population, voir Mashreghi et al. (2016).

2.2.1. Cas de l'échantillonnage aléatoire simple sans remise

Dans le cas où $1/f \in \mathbb{N}$, Gross (1980) propose de former une pseudo-population de taille N en répliquant chaque unité de l'échantillon un même nombre de fois m, où $m = 1/f$. Cette idée semble naturelle puisque chaque unité au sein de la population possède la même probabilité de se retrouver dans l'échantillon. Cela dit, on conviendra que cette méthode est plutôt contraignante, puisque peu de situations en pratique présentent une valeur entière pour m. Booth et al. (1994) ont développé une extension de cette méthode à l'échantillonnage stratifié aléatoire simple sans remise. Nous nous restreindrons à cette méthode, bien qu'il en existe plusieurs autres (Mashreghi et al., 2016). Cet algorithme est conduit d'une strate à l'autre de manière indépendante. Celui-ci est donc décrit pour le cas où il n'y a qu'une seule strate, donc le cas de l'échantillonnage aléatoire simple sans remise.

Leur approche consiste à former U^f en répliquant les observations de l'échantillon

EASSR $\boldsymbol{y} = (y_1, \ldots, y_n)'$ respectivement un nombre $m = \lfloor N/n \rfloor$ fois, où $\lfloor \cdot \rfloor$ correspond à la partie entière. La pseudo-population U^* est obtenue en ajoutant $n' = N - nm$ unités à l'ensemble U^f de façon à ce qu'elle comporte N unités au même titre que la population originale U. Plusieurs auteurs se sont penchés sur la manière de sélectionner ces n' unités. Dans ce mémoire, nous nous restreindrons à celle proposée par Booth et al. (1994), qui consiste à tirer un échantillon aléatoire simple sans remise de n' unités parmi les n unités de s. On fournit une description séquentielle de l'algorithme BPP-EASSR proposé par Booth et al. (1994).

Algorithme 2.2.1 (BPP-EASSR)**.**

(1) Former la partie fixe de la pseudo-population, dénotée U^f en répliquant chaque unité de l'échantillon s un nombre $m = \lfloor N/n \rfloor$ fois.

(2) Compléter la pseudo-population en tirant un échantillon aléatoire simple sans remise, dénoté U^{c*} de taille $n' = N - nm$ à partir de l'échantillon s. La pseudo-population correspond ainsi à $U^* = U^f \cup U^{c*}$. On pose θ^* comme étant le paramètre bootstrap de la pseudo-population.

(3) Tirer un échantillon aléatoire simple sans remise s^* de taille n à partir de U^*.

(4) Calculer l'estimateur bootstrap $\hat{\theta}^*$ à partir des observations de s^*.

(5) Répéter les étapes 2 à 4 un grand nombre B de fois de manière à obtenir les collections

$$(\theta_1^*, \ldots, \theta_B^*)' \quad et \quad (\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*)'.$$

On dénote $\boldsymbol{y}^* = (y_{j_1}^*, \ldots, y_{j_n}^*)', j_1, \ldots, j_n \in \{1, \ldots, n\}$, l'échantillon bootstrap obtenu à la troisième étape de l'algorithme et l'ensemble d'indices associé s^*. On remarque que si $\lfloor N/n \rfloor$ n'est pas un entier, la pseudo-population U^* variera au fil des B itérations de l'algorithme bootstrap.

2.2.1.1. *Estimation de la variance*

Pour un plan de sondage arbitraire p, en supposant que l'estimateur $\hat{\theta}$ est sans biais pour le paramètre de population finie θ, la variance de $\hat{\theta}$ correspond tout simplement à l'erreur quadratique de l'estimateur et peut s'écrire

$$\mathrm{Var}_p\left(\hat{\theta}\right) = \mathbb{E}_p\left[(\hat{\theta} - \theta)^2\right]. \tag{2.2.1}$$

Un premier candidat pour un estimateur de variance bootstrap peut être

$$\tilde{V}_B^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*\right)^2, \tag{2.2.2}$$

où $\hat{\theta}_{(\cdot)}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$. Celui-ci correspond à l'approximation Monte Carlo de la variance totale bootstrap, qui dépend de deux mécanismes aléatoires. D'une part, celui qui intervient en complétant la partie fixe de la pseudo-population et d'autre part, celui derrière la génération des échantillons bootstrap tirés sans remise de la pseudo-population. On désigne ces deux mécanismes par u^* et p^* respectivement. La variance totale peut être décomposée comme

$$\begin{aligned}
\tilde{V}^*\left(\hat{\theta}^*\right) &= \mathrm{Var}^*\left(\hat{\theta}^* |\ y_1, \ldots, y_n\right) \\
&= \mathbb{E}_{u^*p^*}\left[\hat{\theta}^* - \mathbb{E}_{u^*p^*}\left(\hat{\theta}^*\right)\right]^2 \\
&= \mathbb{E}_{u^*}\left[V_{p^*}\left(\hat{\theta}^* \mid U^*\right)\right] + V_{u^*}\left(\mathbb{E}_{p^*}\left[\hat{\theta}^* \mid U^*\right]\right).
\end{aligned}$$

Si $m = 1/f$ est un entier, le deuxième terme de la variance vaudra 0. Autrement, si $\hat{\theta}^*$ est sans biais pour θ^*, $\tilde{V}_B^*$ capturera non seulement la variance de $\hat{\theta}^*$ autour du paramètre de pseudo-population θ^* qui lui correspond, mais aussi la variance de θ^* autour de sa moyenne. Cette dernière pourrait être perçue comme un terme de variabilité parasitaire (Mashreghi et al., 2016). Ainsi, dans les situations où la pseudo-population est aléatoire, un estimateur bootstrap de la variance alternatif peut être considéré, soit

$$V^*\left(\hat{\theta}^*\right) = \mathbb{E}_{u^*}\left[V_{p^*}\left(\hat{\theta}^* \mid U^*\right)\right], \tag{2.2.3}$$

ce qui représente la variabilité de l'estimateur bootstrap $\hat{\theta}^*$ autour du paramètre de pseudo-population θ^* qui lui est associé. Cet estimateur est approximé au moyen de l'algorithme BPP-EASSR en calculant

$$\hat{V}^* = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \theta_b^*)^2. \tag{2.2.4}$$

Si $\hat{\theta}^*$ est sans biais pour θ^* par rapport au plan de sondage, Mashreghi et al. (2016) argumentent que cet estimateur induit un biais d'ordre moins important que (2.2.2) lorsque la fraction de sondage f est non négligeable. Ainsi, bien que nous nous intéressions aux quantiles, nous tendrons à privilégier l'estimateur (2.2.4) puisqu'il est asymptotiquement sans biais pour $\mathrm{Var}_p\left(\hat{t}_{HT}\right)$ quelle que soit f.

Si toutefois $\hat{\theta}^*$ comporte un biais par rapport au paramètre de pseudo-population θ^*,

alors (2.2.4) ne correspond plus à l'approximation de (2.2.3) mais plutôt à l'approximation d'un estimateur bootstrap de l'erreur quadratique moyenne de l'estimateur, soit

$$\mathrm{EQM}^*\left(\hat{\theta}^*\right) = \mathbb{E}_{u^*}\left[\mathrm{EQM}_{p^*}\left(\hat{\theta}^* \mid U^*\right)\right]. \tag{2.2.5}$$

Tel que souligné par Mashreghi et al. (2016), Booth et al. (1994) ne se sont pas intéressés explicitement à l'estimation de la variance d'un estimateur lorsqu'ils ont introduit l'algorithme BPP-EASSR. En revanche, dans l'optique de construire des intervalles de confiance par bootstrap, ils ont centré l'estimateur bootstrap $\hat{\theta}^*$ par rapport au paramètre de pseudo-population θ^*. Cela suggère que s'ils avaient eu à formuler un estimateur de variance, ce dernier aurait probablement été (2.2.3). La prochaine section sera dédiée à la construction d'intervalles de confiance.

2.2.1.2. *Construction d'intervalles de confiance*

Il existe plusieurs approches pour construire un intervalle de confiance à partir de la distribution bootstrap d'un estimateur. Au sein de ce mémoire, trois d'entre elles seront étudiées par simulations numériques.

Si l'on souhaite construire un intervalle de niveau $1-\alpha$, la première méthode consiste à obtenir une estimation bootstrap de la variance $\widehat{V}$ de $\hat{\theta}$ et à dériver les quantiles α et $1-\alpha/2$ de la distribution centrée $(\hat{\theta}-\theta)/\sqrt{\widehat{V}}$. Une approximation de la distribution de $(\hat{\theta}-\theta)/\sqrt{\widehat{V}}$ par la loi normale produit par inversion l'intervalle de confiance asymptotique de niveau $1-\alpha$ suivant

$$\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{V}}, \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{V}}\right], \tag{2.2.6}$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile de niveau $1-\alpha/2$ de la loi normale centrée réduite.

Une autre approche générale, soit l'intervalle bootstrap de base, repose sur l'inversion du pivot $\hat{\theta}-\theta$. On définit la fonction de répartition associée à la distribution échantillonnale de $\hat{\theta}$ dans un contexte de population finie comme $J_n(t,U) = \mathrm{Prob}_U\left(\hat{\theta}-\theta \leq t\right)$. La forme souhaitée pour l'intervalle de confiance de niveau α est justifiée par la relation suivante

$$\begin{aligned} & P\left(J_n^{-1}(\alpha/2,U) \leq \hat{\theta}-\theta \leq J_n^{-1}(1-\alpha/2,U)\right) = \alpha \\ \iff\ & P\left(\hat{\theta} - J_n^{-1}(1-\alpha/2,U) \leq \theta \leq \hat{\theta} - J_n^{-1}(\alpha/2,U)\right) = \alpha. \end{aligned}$$

Les quantités $J_n^{-1}(\alpha/2, U)$ et $J_n^{-1}(1-\alpha/2, U)$ sont à toutes fins pratiques inconnues. Conformément au principe du bootstrap, l'intervalle bootstrap de base résulte de la substitution de U par la pseudo-population U^* dans $J_n(t,\cdot)$. Cela donne $J_n(t, U^*) = \text{Prob}_{U^*}\left(\hat{\theta}^* - \theta^* \leq t\right)$. Ainsi, en remplaçant les quantiles inconnus par ceux de la distribution bootstrap $J_n(t, U^*)$, l'intervalle de niveau $1-\alpha$ suivant est obtenu

$$\left[\hat{\theta} - J_n^{-1}(1-\alpha/2, U^*),\, \hat{\theta} - J_n^{-1}(\alpha/2, U^*)\right]. \tag{2.2.7}$$

Remarquons que lorsque $n' = N - nm$ est différent de 0, entraînant la génération de pseudo-populations aléatoires, le paramètre θ^* varie à chaque itération de l'algorithme. Maintenant, une variante de cet intervalle consiste à inverser les quantiles et à changer le signe, de façon à avoir

$$\left[\hat{\theta} + J_n^{-1}(\alpha/2, U^*),\ \hat{\theta} + J_n^{-1}(1-\alpha/2, U^*)\right]. \tag{2.2.8}$$

L'intervalle (2.2.8) est dit de type *percentile*. Dans le cas où $n' = N - nm = 0$, le paramètre de pseudo-population θ^* est invariant et équivaut toujours à $\hat{\theta}$. En définissant les statistiques d'ordre $\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \ldots \leq \hat{\theta}^*_{(B)}$ correspondant à la distribution bootstrap de $\hat{\theta}$, l'intervalle (2.2.8) peut alors être réécrit plus simplement comme

$$\left[\hat{\theta}^*_{B\frac{\alpha}{2}},\ \ \hat{\theta}^*_{B\left(1-\frac{\alpha}{2}\right)}\right], \tag{2.2.9}$$

où $\hat{\theta}^*_{B\alpha/2}$ et $\hat{\theta}^*_{B(1-\alpha/2)}$ sont les quantiles expérimentaux de la distribution de $\hat{\theta}^*$ de niveaux $\alpha/2$ et $1-\alpha/2$ respectivement. Si la pseudo-population U^* varie au fil des itérations de l'algorithme BPP-EASSR, en raison du terme de variance parasitaire invoqué plus tôt (voir l'équation 2.2.1), l'intervalle en (2.2.9) surcouvrira le paramètre θ (Mashreghi et al., 2016).

2.2.2. Cas de l'échantillonnage de Poisson

Le principe du bootstrap par pseudo-population peut aussi être mis en œuvre dans le cas de l'échantillonnage de Poisson, un plan à probabilités inégales. De plus, rappelons que sous ce plan de sondage, chaque unité de la population est sélectionnée de manière indépendante selon une épreuve de Bernoulli avec probabilité de succès π_i. Par conséquent, la taille de l'échantillon résultant est une variable aléatoire. Comme pour un plan de sondage aussi élémentaire que l'échantillonnage aléatoire simple sans remise, le principe consiste toujours à effectuer le rééchantillonnage en reproduisant les conditions ayant permis d'obtenir

l'échantillon s initial. L'algorithme suivant rapporté par Mashreghi et al. (2016) décrit la procédure proposée par Chauvet (2007).

Algorithme 2.2.2 (BPP-Poisson)**.**

(1) Répliquer les unités (y_i, π_i), $i \in s$ *un nombre* $\lfloor \pi_i^{-1} \rfloor$ *de fois, où* $\lfloor \cdot \rfloor$ *est la partie entière, de façon à obtenir la partie fixe de la pseudo-population, notée* U^f.

(2) Afin de compléter la pseudo-population, former l'ensemble U^{c*} *en échantillonnant des unités à partir de* s *selon le plan de Poisson avec probabilité d'inclusion* $\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$ *pour la* i*-ème unité* (y_i, π_i), $i \in s$. *La pseudo-population ainsi obtenue est notée* $U^* = U^f \cup U^{c*}$ *et est constituée des paires notées* $(\breve{y}_i, \breve{\pi}_i), i \in U^*$, *où* $\breve{y}_i$ *se trouve à être l'une des valeurs présentes dans* s *et* $\breve{\pi}_i$ *est le poids de sondage correspondant. On pose* θ^* *comme étant le paramètre bootstrap calculé à partir de la pseudo-population.*

(3) Obtenir l'échantillon bootstrap s^* *en conduisant l'échantillonnage de Poisson à travers l'ensemble* U^* *avec les poids de sondage* $\breve{\pi}_i$, $i \in U^*$, *énoncés à l'étape précédente.*

(4) Calculer l'estimateur bootstrap $\hat{\theta}^*$ *à partir des observations contenues dans* s^*.

(5) Répéter les étapes 2 à 4 un grand nombre B *de fois de manière à obtenir les collections*

$$(\theta_1^*, \ldots, \theta_B^*)' \quad et \quad (\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*)'.$$

Cet algorithme mène à l'estimateur de variance (2.2.4). Si l'estimateur $\hat{\theta}$ n'est pas biaisé par rapport à θ, l'estimateur bootstrap de la variance préconisé se trouve encore à être $\mathbb{E}_{u^*}\left[V_{p^*}\left(\hat{\theta}^* \mid U^*\right)\right]$, qui est approximé par $\widehat{V}_B^*$ obtenu au terme de l'algorithme. Dans le cas contraire, ce dernier est une estimation de l'erreur quadratique moyenne de l'estimateur. Dans le cas de l'estimateur du total, Chauvet (2007) montre analytiquement que $\mathbb{E}_{u^*}\left[V_{p^*}\left(\hat{t}_{HT}^* \mid U^*\right)\right]$ se réduit à l'estimateur habituel de la variance sous l'échantillonnage de Poisson donné par (1.2.6). Il faut noter que la taille de la pseudo-population formée à l'étape 2 peut différer de N.

Tel que suggéré par Mashreghi et al. (2016), cet algorithme constitue une extension de l'algorithme de Booth et al. (1994) formulé dans le contexte du plan EASSR à l'échantillonnage de Poisson en calculant pour chaque pseudo-population formée $b = 1, \ldots, B$ le paramètre boostrap θ^* qui lui est associé. Ainsi, tout comme dans le plan précédent, en exploitant la distribution bootstrap $(\hat{\theta}^* - \theta^*)$, il sera non seulement possible de dériver

l'intervalle de confiance asymptotique mais aussi les intervalles de type percentile et de base discutés à la sous-section 2.2.1.2.

2.2.3. Le bootstrap appliqué à des statistiques non lisses

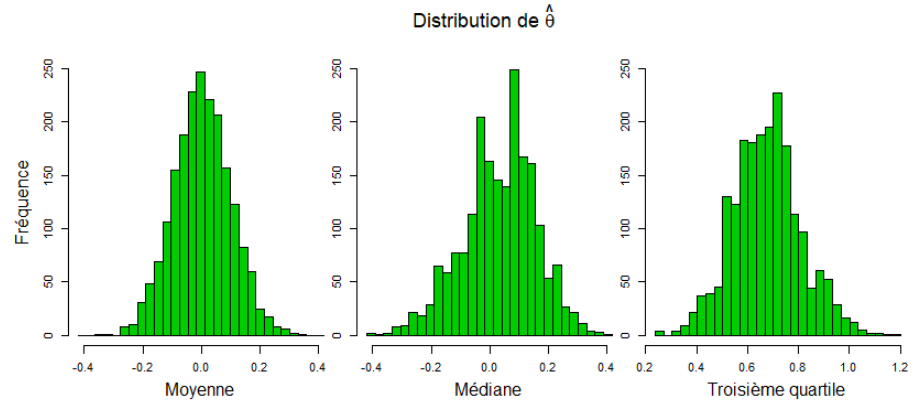
Les mécanismes de rééchantillonnage utilisés dans la méthode du bootstrap non paramétrique et dans les méthodes par pseudo-population occasionnent des répétitions de valeurs au sein des échantillons bootstrap. Dans le premier cas, le rééchantillonnage avec remise est en cause; dans le deuxième, la répétition est attribuable à la partie fixe de la pseudo-population, créée en répliquant chaque unité un nombre de fois qui dépend de sa probabilité d'inclusion de premier ordre. Les répétitions impliquent un espace échantillonnal réduit dans le cas de certaines statistiques. Ceci ne devrait pas affecter outre mesure l'histogramme bootstrap d'une moyenne, qui est une statistique dite lisse. Des changements mineurs dans les données n'entraîneront alors que de petites variations dans la statistique (Efron et Tibshirani, 1993).

En revanche, le nombre restreint de valeurs distinctes affectera davantage des statistiques non lisses telles que les quantiles, dont la distribution asymptotique dépend de caractéristiques locales de la distribution F_0. Ce phénomène est d'autant plus perceptible si l'on prend l'exemple de la médiane et une taille d'échantillon n impaire. Dans ce cas, le support échantillonnal est alors exactement l'échantillon initial $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ et comportera au plus n valeurs distinctes.

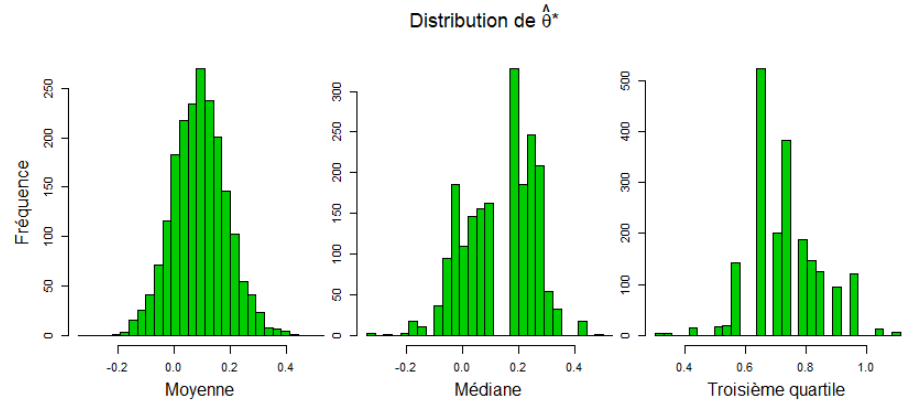
Les trois ensembles d'histogrammes exhibés à la figure 2.1 dressent la comparaison des distributions échantillonnales d'une statistique lisse, la moyenne $\bar{y}$, à deux statistiques non lisses, soit la médiane $\hat{\xi}_{0,50}$ et le troisième quartile $\hat{\xi}_{0.75}$. Le premier panneau illustre la distribution échantillonnale des trois statistiques pour 2 000 EASSR de taille $n = 97$ tirés d'une population finie de taille $N = 1\ 428$ unités. Cette population finie est elle-même un échantillon i.i.d. de la distribution $\mathcal{N}(0,1)$. Quant au deuxième panneau, celui-ci contient des approximations basées sur 2 000 échantillons bootstrap des distributions bootstrap de $\bar{y}^*$, $\hat{\xi}^*_{0,50}$ et $\hat{\xi}^*_{0.75}$, qui ont été générées à partir de l'un des 2 000 échantillons initiaux. Celle de la moyenne bootstrap compte 2 000 valeurs distinctes, tandis que l'on distingue seulement

30 et 27 valeurs uniques pour les distributions de $\hat{\xi}^*_{0,50}$ et $\hat{\xi}^*_{0.75}$ respectivement.

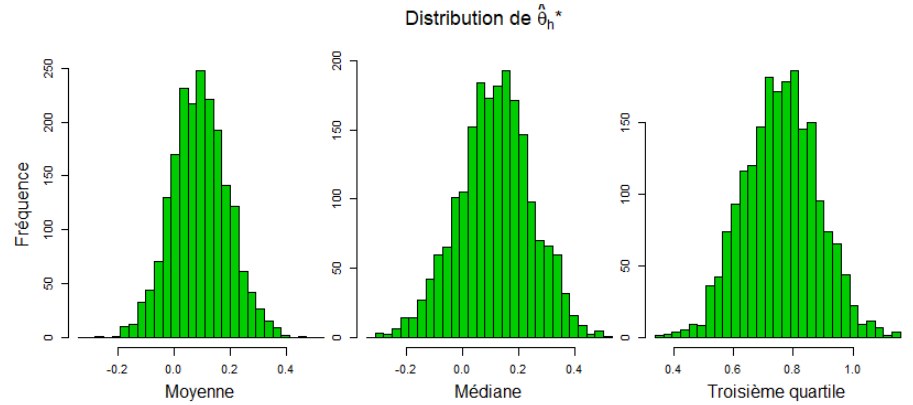
Enfin, le dernier panneau d'histogrammes permet d'illustrer les bénéfices apportés par l'introduction du lissage au sein de la méthode bootstrap par pseudo-population, méthode qui sera introduite au chapitre 4. Dans cet exemple, un paramètre de lissage $h = 0{,}30$ a été utilisé. En enrichissant le support échantillonnal de l'estimateur bootstrap d'un quantile, une adaptation lisse de l'algorithme laisse entrevoir un potentiel d'amélioration au regard de l'estimation de la variance et de la formation d'intervalles de confiance. Maintes questions restent à être élucidées quant à la mise en œuvre de la méthode, le choix du paramètre de lissage et les circonstances dans lesquelles elle peut être avantageuse par rapport à la méthode standard. La méthode du bootstrap lisse pour des données i.i.d. a fait l'objet de l'attention de plusieurs auteurs, notamment Efron et Tibshirani (1993) et Hall et al. (1989). Celle-ci sera étudiée dans le contexte d'une population infinie au prochain chapitre dans le but d'en saisir les tenants et les aboutissants avant de l'étendre au contexte de population finie au chapitre 4.



(a) Histogramme de $\hat{\theta}$ basé sur $S = 2\ 000$ EASSR de taille $n = 97$.



(b) Histogramme bootstrap de $\hat{\theta}^*$ construit à partir de $B = 2\ 000$ réplicats.



(c) Histogramme bootstrap de $\hat{\theta}_h^*$ ($h = 0{,}3$) construit à partir de $B = 2\ 000$ réplicats.

Fig. 2.1. Distributions de $\hat{\theta}$, $\hat{\theta}^*$ et $\hat{\theta}_h^*$ ($h = 0{,}3$) pour trois statistiques ($\bar{y}$, $\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$). Les histogrammes bootstrap ont été engendrés à partir d'un échantillon aléatoire simple sans remise de taille $n = 97$ d'une population finie de $N = 1\ 428$ unités provenant de la distribution $\mathcal{N}(0{,}1)$.

Chapitre 3

Méthode du bootstrap lisse pour des données indépendantes et identiquement distribuées

À la fin du chapitre précédent, nous avons soulevé des difficultés rencontrées au niveau du support de la distribution bootstrap lorsque les méthodes de rééchantillonnage classiques sont appliquées à des statistiques non lisses telles que les quantiles. Dans le cadre d'une population infinie, ce problème peut être enrayé par la méthode du bootstrap lisse, qui fera l'objet du présent chapitre. Le principe de cette alternative au bootstrap non paramétrique usuel de même que sa mise en œuvre seront discutés.

Les estimateurs bootstrap lisses étant définis à un paramètre de lissage près, il s'ensuivra une discussion sur la sélection de celui-ci pour le cas particulier de l'estimation de la variance d'un quantile échantillonnal. Par le fait même, quelques éléments de la théorie sur l'estimation par le noyau de la fonction de densité en un point seront présentés. Nous serons alors en mesure d'optimiser la performance de l'estimateur de variance d'un quantile échantillonnal en faisant des rapprochements avec le cas de l'estimateur non paramétrique de fonction de densité.

3.1. Introduction

La méthode du bootstrap non paramétrique pour des données indépendantes et identiquement distribuées selon F_0 a été décrite à la section 2.1. À défaut de pouvoir connaître la distribution échantillonnale d'une statistique $\hat{\theta}$, donnée par $J_n(t, F_0) = \text{Prob}_{F_0}(\hat{\theta} - \theta \leq t)$,

nous avons vu que la méthode standard du bootstrap consiste à simuler des échantillons à partir de la fonction de répartition expérimentale $\hat{F}_n$ et à calculer la statistique $\hat{\theta}$ à partir des échantillons simulés afin d'obtenir une approximation de cette distribution. La simulation est mise en œuvre en tirant n unités uniformément avec remise de l'échantillon initial $Y_1, Y_2, \ldots, Y_n$ un grand nombre de fois. Cela étant, les échantillons ainsi obtenus auront la particularité de présenter des valeurs identiques à celles de l'échantillon initial, qui se répéteront en nombre variable (Silverman, 1986).

Le bootstrap lisse constitue une approche permettant d'éviter que ces propriétés se manifestent dans la génération des échantillons bootstrap. Il s'agit d'estimer une certaine fonctionnelle $\alpha(F_0)$ par $\alpha(\hat{F}_h)$ plutôt que $\alpha(\hat{F}_n)$, où $\hat{F}_h : \mathbb{R} \to [0,1]$ est une estimation lisse de F_0. Une possibilité pour $\hat{F}_h$ est l'estimateur par le noyau, ce qui en fait une approche tout aussi non paramétrique que le recours à la fonction de répartition expérimentale. Avec ce choix, $\hat{F}_h$ est défini comme

$$\hat{F}_h(y) = \int_{-\infty}^{y} \hat{f}_h(t)dt, \tag{3.1.1}$$

où $\hat{f}_h(y)$ est l'estimateur de fonction de densité de Rosenblatt-Parzen (Silverman, 1986), donné par

$$\hat{f}_h(t) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{t - Y_i}{h}\right), \tag{3.1.2}$$

qui dépend d'un paramètre de lissage h, soit le voisinage effectif autour de t de l'estimateur de fonction de densité, et d'une fonction de noyau k. Dans ce mémoire, les propriétés de l'estimateur seront étudiées en se restreignant aux fonctions k satisfaisant

$$\int k(t)dt = 1, \quad k(x) \geq 0 \; \forall \; x, \quad \int tk(t)dt = 0, \quad \int t^2 k(t)dt =: \kappa_2 < \infty, \tag{3.1.3}$$

c'est-à-dire des fonctions de densité de probabilité dotées d'un deuxième moment fini. En contraignant k à être une fonction de densité de probabilité, il sera possible d'approximer les estimateurs bootstrap par simulation Monte Carlo en rééchantillonnant à partir de $\hat{F}_h$ de la manière usuelle (Hall et al., 1989). En explicitant (3.1.1), on obtient l'équivalence

$$\hat{F}_h(y) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right), \tag{3.1.4}$$

où $K(y) = \int_{-\infty}^{y} k(t)dt$ est la fonction de répartition associée à k. Ainsi, plutôt que d'être générées à partir de $\hat{F}_n$, les observations bootstrap sont issues de la distribution $\hat{F}_h$ et à présent

dénotées $Y^*_{1,h}, Y^*_{2,h}, \ldots, Y^*_{n,h}$. Incidemment, la distribution de $\hat{\theta}$ est estimée par celle de $\hat{\theta}^*_h$, qui est la même statistique calculée sur la base des observations bootstrap $Y^*_{1,h}, Y^*_{2,h}, \ldots, Y^*_{n,h}$. Remarquons que la distribution $\hat{F}_h$ au même titre que les variables aléatoires qui en découlent sont régies par le paramètre de lissage h et une fonction de noyau k, qui doivent tous deux être fixés par l'utilisateur.

3.2. Mise en œuvre

Dans l'optique d'appliquer l'algorithme du bootstrap lisse pour le problème d'estimation de la variance d'un quantile, nous éluciderons d'abord la question de générer des données à partir de $\hat{F}_h$. Les propositions suivantes permettent de justifier l'algorithme qui sera décrit en fin de section. La première établit la fonction de répartition d'une convolution de deux variables aléatoires à valeurs réelles indépendantes, disons X et Y.

Proposition 3.2.1. *Soit X et Y deux variables aléatoires indépendantes dotées d'un support à valeurs réelles dont les fonctions de répartition sont respectivement notées F_X et F_Y. On suppose également que Y est une variable aléatoire continue dont la fonction de densité est dénotée f_Y. En posant la convolution des deux variables comme étant $Z = X + Y$, alors la fonction de répartition de Z peut être écrite comme étant*

$$F_Z(z) = \int_{\mathbb{R}} F_X(z - y) f_Y(y) dy.$$

Démonstration. Directement, on a

$$\begin{aligned} F_Z(z) &= F_{X+Y}(z) \\ &= P(X + Y \leq z) \\ &= \mathbb{E}_Y \left[P(X + y \leq z) \mid Y = y \right] \\ &= \int_{\mathbb{R}} F_X(z - y) f_Y(y) dy. \end{aligned}$$

□

La fonction de répartition de la convolution $X + Y$ est dénotée alternativement $F_X \star F_Y$. La proposition suivante établit que l'estimation lisse $\hat{F}_h$ de F_0 est la fonction de répartition de la

convolution d'une variable aléatoire ayant pour fonction de répartition K_h et d'une variable aléatoire ayant pour fonction de répartition $\hat{F}_n$.

Proposition 3.2.2. *Soit K_h une fonction de répartition K avec paramètre de lissage h, définie comme $K_h : \mathbb{R} \to [0,1]$, $K_h(t) = K(t/h) = \int_{-\infty}^{t} h^{-1}k(t/h)dt$. De plus, soit X une variable aléatoire distribuée selon $\hat{F}_n$, la fonction de répartition expérimentale calculée à partir d'un échantillon i.i.d. $X_1, \ldots, X_n$. L'estimateur par le noyau défini à partir d'une fonction de noyau $k = K'$ et un paramètre de lissage h est la convolution de X avec une variable aléatoire ayant pour fonction de répartition K_h.*

Démonstration. D'abord, on note $k_h = K_h'$. Dès lors, en utilisant la fonction de répartition d'une convolution de deux variables aléatoires indépendantes donnée par la proposition 3.2.1, on a

$$\begin{aligned}
\left(\hat{F}_n \star K_h\right)(x) &= \int_{\mathbb{R}} \hat{F}_n(x-t)k_h(t)\,dt \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} \mathbb{1}(X_i \leq x-t)\frac{1}{h}k\left(\frac{t}{h}\right)dt \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\int_{-\infty}^{x-X_i} \frac{1}{h}k\left(\frac{t}{h}\right)dt \\
&= \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \\
&= \hat{F}_h(x).
\end{aligned}$$

□

Ainsi, un échantillon i.i.d. issu de $\hat{F}_h$ peut être obtenu en tirant un échantillon $Y_1^*, Y_2^*, \ldots, Y_n^*$ avec remise à partir de l'échantillon initial et en y additionnant le vecteur aléatoire $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$, où les observations ε_i^*, $i = 1, \ldots n$, sont indépendantes et identiquement distribuées selon la loi déterminée par K_h. Cela consiste en une légère modification de l'algorithme du bootstrap non paramétrique standard comme en témoigne la description de l'adaptation lisse qui suit.

Algorithme 3.2.3 (Bootstrap lisse pour des données i.i.d.)**.**

(1) Tirer un échantillon de taille n à partir de la distribution $\hat{F}_n$, c'est-à-dire tirer un échantillon de taille n avec remise à partir de l'échantillon observé $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. Cet échantillon est noté $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)'$.

(2) Générer n observations indépendantes et identiquement distribuées $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ à partir de la distribution déterminée par K.

(3) Les n observations issues de $\hat{F}_h$ sont obtenues en calculant $y_{i,h}^* = x_i^* + h\varepsilon_i^*$, $i = 1, \ldots, n$. L'estimateur bootstrap lisse $\hat{\theta}_h^*$ est calculé à partir de l'échantillon bootstrap lisse $\boldsymbol{y}_h^* = (y_{1,h}^*, y_{2,h}^*, \ldots, y_{n,h}^*)'$.

(4) Les étapes 1 à 3 sont répétées un grand nombre de fois B de façon à obtenir la collection $(\hat{\theta}_{1,h}^*, \hat{\theta}_{2,h}^*, \ldots, \hat{\theta}_{B,h}^*)'$.

Remarquons que la distribution bootstrap $(\hat{\theta}_{1,h}^*, \hat{\theta}_{2,h}^*, \ldots, \hat{\theta}_{B,h}^*)'$ est indicée par la taille de fenêtre h, laquelle doit être fixée par l'utilisateur. Une valeur $h = 0$ correspond à la méthode du bootstrap non paramétrique standard. Silverman (1986, chap. 6) rapporte que le choix du paramètre de lissage est souvent fait de manière arbitraire dans la méthode du boostrap lisse à défaut de cerner précisément les conditions sous lesquelles le lissage permet de surpasser la méthode classique. La prochaine section se consacrera à la sélection du paramètre de lissage et plus particulièrement au choix de la taille de fenêtre pour le problème de l'estimation de la variance du quantile échantillonnal.

3.3. Choix de la taille de la fenêtre

Il peut être utile de s'attarder au paramètre de lissage optimal pour l'estimation de la fonction de densité f_0 en vue de dresser un parallèle avec le paramètre de lissage optimal pour l'estimation de la variance de la médiane échantillonnale. Les démonstrations des différents résultats énoncés peuvent être retrouvées dans le chapitre 3 de Silverman (1986).

3.3.1. Choix de la taille de la fenêtre pour l'estimation de la fonction de densité en un point

Supposons que nous disposions d'un échantillon i.i.d. $Y_1, Y_2, \ldots, Y_n$ provenant d'une distribution F_0 avec fonction de densité de probabilité f_0. Celle-ci est estimée par l'estimateur de Rosenblatt-Parzen défini en (3.1.2), qui dépend d'un noyau k satisfaisant les conditions (3.1.3) et d'une taille de fenêtre h. Si l'impact du noyau sur la performance de l'estimateur

de densité est faible, celui de la taille de la fenêtre revêt une grande importance (Silverman, 1986). Afin de minimiser le critère d'ajustement qui suivra, le paramètre de lissage devra être d'un ordre de grandeur précis. En prenant comme fonction de perte la distance quadratique, ou la norme L_2, une mesure d'ajustement ponctuelle est l'erreur quadratique moyenne en un point y du support donnée par

$$\begin{aligned}\mathrm{EQM}(\hat{f}_h(y)) &= \mathbb{E}\left[\hat{f}_h(y) - f_0(y)\right]^2 \\ &= \mathrm{Var}\left(\hat{f}_h(y) - f_0(y)\right) + \left(\mathbb{E}\left[\hat{f}_h(y) - f_0(y)\right]\right)^2 \\ &= \mathrm{Var}(\hat{f}_h(y)) + \mathrm{Biais}^2(\hat{f}_h(y)). \end{aligned} \tag{3.3.1}$$

Le paramètre de lissage optimal pour l'estimation de $f_0(y)$ sera celui minimisant $\mathrm{EQM}(\hat{f}_h(y))$. Nous devrons donc expliciter l'expression de l'erreur quadratique moyenne de $\hat{f}_h(y)$. Une expression approximative pour le biais de $\hat{f}_h(y)$ basée sur un développement en série de Taylor d'ordre 2 est

$$\mathrm{Biais}(\hat{f}_h(y)) \approx \frac{1}{2}h^2 f_0''(y)\kappa_2, \tag{3.3.2}$$

où $\kappa_2 = \int t^2 k(t)dt$. De même, la variance de l'estimateur de Rosenblatt-Parzen s'écrit approximativement

$$\mathrm{Var}(\hat{f}_h(y)) \approx \frac{1}{nh} f_0(y)\kappa_1, \tag{3.3.3}$$

où $\kappa_1 = \int k^2(t)dt$. En remplaçant les expressions approximatives du biais et de la variance de $\hat{f}(x)$ dans (3.3.1), on obtient finalement

$$\mathrm{EQM}(\hat{f}_h(y)) \approx \frac{1}{nh} f_0(y)\kappa_1 + \frac{1}{4}h^4 \left(f_0''(y)\right)^2 \kappa_2^2. \tag{3.3.4}$$

Remarquons que le problème de minimisation de (3.3.4) consiste en l'atteinte d'un équilibre entre le terme du biais et le terme de la variance. Supposons que $h = C \cdot n^{-p}$, avec $C > 0, p > 0$. Selon l'ordre de grandeur n^{-p} choisi pour h, l'un ou l'autre des termes dominera. Si $p < 1/5$, le terme de variance diminuera plus rapidement que le terme du biais. Inversement, si $p > 1/5$, le terme du biais sera négligeable par rapport à la variance. L'ordre de grandeur de la taille de la fenêtre sera choisi de manière à ce que les deux termes constituant (3.3.4) diminuent à la même vitesse à mesure que la taille d'échantillon croît. Le lemme suivant (Parzen, 1962), qui peut être vérifié facilement, pourvoit une solution formelle à ce problème d'optimisation en établissant la valeur optimale de la constante C pour une taille échantillonnale n donnée.

Lemme 3.3.1. *Soit A, B, α et β des nombres positifs donnés. Alors,*

$$\min_{x>0} Ax^{\alpha} + Bx^{-\beta} = A(1+\alpha/\beta)(\beta B/\alpha A)^{\alpha/(\alpha+\beta)},$$

et

$$\arg\min_{x>0} Ax^{\alpha} + Bx^{-\beta} = \left(\frac{\beta B}{\alpha A}\right)^{\frac{1}{\alpha+\beta}}.$$

Ainsi, le problème de minimisation de (3.3.4) en tant que fonction de h est résolu en appliquant le lemme ci-dessus, avec $A = \frac{1}{4}\left(f_0''(y)\right)^2 \kappa_2^2$, $B = n^{-1} f_0(y)\kappa_1$, $\alpha = 4$ et $\beta = 1$. À condition que $f_0''(y) \neq 0$, la valeur optimale de h pour l'estimation de la fonction de densité en y est donc donnée par

$$h_{\text{opt}}(y) = \kappa_2^{-2/5} {\kappa_1}^{1/5} \left[f_0(y)\right]^{1/5} \left[f_0''(y)\right]^{-2/5} n^{-1/5}. \tag{3.3.5}$$

3.3.2. Choix de la taille de la fenêtre pour l'estimation de la variance d'un quantile échantillonnal

Comme la figure 2.1 permet d'illustrer, l'utilisation de $\hat{F}_h$ plutôt que $\hat{F}_n$ agit de façon à enrichir le support de la distribution échantillonnale bootstrap d'une statistique. Il y a néanmoins lieu de se questionner à savoir si cette technique peut réellement améliorer la performance des estimateurs bootstrap et dans quelle mesure.

Hall et al. (1989) argumentent que l'introduction du lissage au sein du bootstrap peut s'avérer particulièrement bénéfique lorsque les quantités à l'étude dépendent d'une certaine façon de propriétés locales de la distribution sous-jacente F_0. Il s'agit précisément du cas de la variance du quantile de niveau p d'un échantillon $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_0$, dénoté ici $\check{\xi}_p$, dont la variance asymptotique est donnée par $n^{-1}\rho^2 = n^{-1}p(1-p)/f_0^2(\tilde{\xi}_p)$. Cette dernière dépend localement de la distribution initiale F_0 à travers la fonction de densité f_0 évaluée en $\tilde{\xi}_p$. Incidemment, conditionnellement à un ordre de grandeur approprié pour le paramètre de lissage, nous verrons que le bootstrap lisse accélère la convergence de l'estimateur de la variance du quantile échantillonnal par rapport au bootstrap non paramétrique standard.

Soit $\text{Var}_h^*(\check{\xi}_p)$ l'estimateur de variance du quantile échantillonnal $\check{\xi}_p$ obtenu au terme de l'algorithme du bootstrap lisse décrit à la section 3.2. La preuve du théorème suivant

est contenue dans l'ouvrage de Hall et al. (1989). Ce résultat fait office de point de départ quant à la formulation d'un critère de performance de l'estimateur $\mathrm{Var}_h^*(\check{\xi}_p)$.

Théorème 3.3.2. *Soit* $Y_1, Y_2, \ldots Y_n \overset{i.i.d.}{\sim} F_0$. *On fait l'hypothèse que la distribution* F_0 *a une fonction de densité* f_0 *bornée satisfaisant également les conditions que* f_0' *soit bornée dans un voisinage de* $\tilde{\xi}_p$ *et que* f_0'' *soit uniformément continue. On suppose qu'il existe* $\alpha > 0$ *tel que* $\mathbb{E}_0\left[|Y|^\alpha\right] < \infty$. *Enfin, supposons que* k *est une fonction de densité de probabilité et que le paramètre de lissage* $h = h(n)$ *satisfait* $h \to 0$ *et* $nh^3 \log n \to \infty$ *tandis que* $n \to \infty$. *Alors,*

$$n\left(\mathrm{Var}_h^*(\check{\xi}_p) - n^{-1}\rho^2\right) = -2f_0(\tilde{\xi}_p)^{-3}\left[(nh)^{-1/2}Z + \frac{h^2}{2}\kappa_2\left\{f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right\}\right] + o_p\{(nh)^{-1/2} + h^2\} \quad (3.3.6)$$

presque sûrement, où $Z = (nh)^{1/2}\left[\hat{f}_h(\tilde{\xi}_p) - \mathbb{E}_0[\hat{f}_h(\tilde{\xi}_p)]\right]$, $\hat{f}_h$ *est l'estimateur de Rosenblatt-Parzen et* $\kappa_2 = \int t^2 k(t)dt$.

Ainsi, tel que souligné par Hall et al. (1989), le problème de minimisation de EQM $\left(\mathrm{Var}_h^*(\check{\xi}_p)\right)$ par rapport à h pour une taille échantillonnale n fixée peut être appréhendé en minimisant le carré de l'espérance du terme encadré par des crochets dans (3.3.6). Par le théorème limite central, la variable aléatoire Z est asymptotiquement normalement distribuée avec espérance $\mathbb{E}[Z] = 0$ et variance $\mathrm{Var}(Z) = nh\mathrm{Var}(\hat{f}_h(\tilde{\xi}_p)) = f_0(\tilde{\xi}_p)\kappa_1$, où $\kappa_1 = \int k^2(t)dt$. Il s'ensuit que

$$\mathbb{E}_0\left[\left((nh)^{-1/2}Z + \frac{h^2}{2}\kappa_2\left\{f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right\}\right)^2\right] = (nh)^{-1}f_0(\tilde{\xi}_p)\kappa_1 + \frac{h^4}{4}\kappa_2^2\left\{f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right\}^2. \quad (3.3.7)$$

Le lemme 3.3.1 introduit précédemment peut être utilisé afin de déterminer le paramètre de lissage minimisant l'erreur quadratique moyenne de $\mathrm{Var}_h^*(\check{\xi}_p)$ ou minimisant (3.3.7) de façon équivalente. Avec $A = \frac{1}{4}\kappa_2^2\left\{f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right\}^2$, $B = n^{-1}f_0(\tilde{\xi}_p)\kappa_1$, $\alpha = 4$ et $\beta = 1$, si $f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1} \neq 0$ la solution s'écrit

$$h_{\mathrm{opt}}(\tilde{\xi}_p) = \kappa_2^{-2/5}{\kappa_1}^{1/5}\left[f_0(\tilde{\xi}_p)\right]^{1/5}\left[f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right]^{-2/5} n^{-1/5}. \quad (3.3.8)$$

Dans le cas de la médiane et d'une fonction de densité symétrique, il est intéressant de remarquer que le paramètre de lissage optimisant la performance de $\mathrm{Var}_h^*(\check{\xi}_{0,50})$ est égal à celui pour l'estimateur de fonction de densité $\hat{f}_h(\tilde{\xi}_{0,50})$. En effet, il suffit d'observer que si f_0 est symétrique, alors $f_0'(\tilde{\xi}_{0,50}) = 0$, ce qui permet de retrouver comme paramètre de lissage (3.3.5), c'est-à-dire celui minimisant $\mathrm{EQM}(\hat{f}_h(\tilde{\xi}_{0,50}))$.

En réécrivant la taille de fenêtre optimale comme $h_{\mathrm{opt}}(\tilde{\xi}_p) = C_{\mathrm{opt}} \cdot n^{-1/5}$, où $C_{\mathrm{opt}} \equiv C_{\mathrm{opt}}(\tilde{\xi}_p) = \kappa_2^{-2/5}\kappa_1{}^{1/5}\left[f_0(\tilde{\xi}_p)\right]^{1/5}\left[f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right]^{-2/5}$ et en l'injectant dans (3.3.6), l'erreur relative suivante pour $\mathrm{Var}_h^*(\check{\xi}_p)$ est obtenue:

$$\begin{aligned}&\frac{\mathrm{Var}_h^*(\check{\xi}_p) - n^{-1}\rho^2}{n^{-1}\rho^2}\\ &= -\frac{2}{\rho^2} f_0(\tilde{\xi}_p)^{-3}\left[n^{-2/5} C_{\mathrm{opt}}^{-1/2} Z + \frac{1}{2} n^{-2/5} C_{\mathrm{opt}}^2 \kappa_2 \left\{f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1}\right\}\right] + o_p\{n^{-2/5}\}.\end{aligned} \tag{3.3.9}$$

L'équation 3.3.9 établit que l'ordre de l'erreur relative de l'estimateur lisse de variance d'un quantile est de $n^{-2/5}$ avec un choix de noyau k satisfaisant les conditions (3.1.3). À toutes fins pratiques, toute taille de fenêtre de la forme $h = C \cdot n^{-1/5},\ C \in \mathbb{R}^+$ satisfera cette vitesse de convergence. Le résultat suivant, tiré de la remarque 2.1 de Hall et Martin (1988), détermine la vitesse de convergence lorsque l'on choisit $h = 0$, ou autrement dit, lorsque les échantillons bootstrap sont simulés à partir de $\hat{F}_n$. Sous certaines conditions de régularité sur la fonction de densité f_0, Hall et Martin (1988) montrent que

$$\frac{\mathrm{Var}^*(\check{\xi}_p) - n^{-1}\rho^2}{n^{-1}\rho^2} = O_p(n^{-\frac{1}{4}}), \tag{3.3.10}$$

où $\mathrm{Var}^*(\check{\xi}_p)$ est l'estimateur de variance de $\check{\xi}_p$ résultant de l'algorithme du bootstrap non paramétrique standard ($h = 0$). Il s'ensuit que pour le quantile échantillonnal de niveau p, l'introduction du lissage peut avoir un impact important sur la performance de l'estimateur de variance bootstrap, faisant passer l'ordre de l'erreur relative de $n^{-\frac{1}{4}}$ à $n^{-\frac{2}{5}}$. Cela étant dit, dans le cas de statistiques linéaires, Hall et al. (1989) font remarquer que le paramètre de lissage optimal sera celui qui fera en sorte que l'estimateur lisse de variance converge à une vitesse $n^{-1/2}$, ce qui correspond à l'ordre de l'estimateur non lisse ($h = 0$). Le lissage ne peut donc avoir qu'un effet de second ordre sur la performance des estimateurs si ceux-ci

peuvent s'exprimer comme des fonctions différentiables de moyennes.

Remarque 3.3.3. *Pour certaines distributions, la condition* $f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1} \neq 0$ *ne sera pas vérifiée en certains points du support. Par exemple, supposons que* $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim}$ Lognormale(μ,1) *et que nous nous intéressions à la médiane de la distribution,* $\tilde{\xi}_{0,50} = \exp(\mu)$. *Alors, on peut facilement vérifier à l'aide des propriétés de la loi lognormale exhibées dans l'annexe A.2 que* $f_0'(e^\mu) = -e^{-\mu} f_0(e^\mu)$ *et que* $f_0''(e^\mu) = e^{-2\mu} f_0(e^\mu)$, *donnant ainsi* $f_0''(\tilde{\xi}_p) - f_0'(\tilde{\xi}_p)^2 f_0(\tilde{\xi}_p)^{-1} = 0$. *Dans cette instance, il n'existe donc pas de paramètre de lissage optimal en fondant l'expression de* $\mathrm{EQM}(\mathrm{Var}_h^*(\check{\xi}_{0,50}))$ *sur un développement en série de Taylor d'ordre 2 tel que l'ont fait Hall et al. (1989). Une valeur pourrait en revanche être dérivée si des termes d'ordre supérieur étaient retenus dans le développement, ce qui aurait une incidence sur la puissance de n qui serait optimale pour la taille de fenêtre.*

Chapitre 4

Bootstrap lisse en contexte de population finie

Les chapitres précédents ont jeté les bases nécessaires à l'étude d'une méthode bootstrap lisse adaptée au contexte de population finie. Au chapitre 2, nous avons vu le principe injectif de la méthode du bootstrap non paramétrique être transposé à l'instance où le paramètre d'intérêt est une fonctionnelle des éléments d'un ensemble fini. La pseudo-population que font intervenir les méthodes de rééchantillonnage en population finie est précisément le dispositif qui rend ce parallèle possible. Au chapitre 3, l'apport du lissage au sein du bootstrap a été étudié pour des données indépendantes et identiquement distribuées. Le présent chapitre sera consacré à une idée naturelle, constituant par ailleurs l'innovation proposée dans ce mémoire, qui est celle d'étendre le principe du bootstrap lisse aux algorithmes décrits au chapitre 2.

Les algorithmes de rééchantillonnage proposés seront décrits pour les deux plans de sondage abordés précédemment. La mise en œuvre de ces algorithmes requiert une valeur pour le paramètre de lissage. De façon à guider le choix de celui-ci en pratique, le reste du chapitre se concentre sur deux méthodes de sélection, l'une faisant recours à des hypothèses sur la distribution de la superpopulation et l'autre étant intégralement non paramétrique. Par ailleurs, la première méthode, une sélection par injection, s'articule autour du résultat théorique énoncé au chapitre précédent concernant le paramètre de lissage optimal pour l'estimation de la variance d'un quantile échantillonnal. La seconde méthode, une sélection par bootstrap, vise à explorer une grille de valeurs pour h afin d'identifier celles minimisant

deux critères d'optimisation. Le premier critère consiste en une estimation bootstrap de l'erreur quadratique moyenne de l'estimateur de variance. Le deuxième est basé sur une estimation bootstrap d'une distance entre l'erreur de couverture expérimentale et l'erreur de couverture nominale d'intervalles de confiance.

4.1. Mise en œuvre

Nous avons vu que le fait de tirer des observations avec remise à partir d'un échantillon $Y_1, Y_2, \ldots, Y_n$ peut être décrit par la fonction de répartition expérimentale $\hat{F}_n$. De la même manière, l'échantillonnage à partir d'une pseudo-population U^* définit une fonction de répartition. L'extension du bootstrap lisse dans le contexte des sondages tient de ce principe. Ainsi, si l'on rééchantillonnait à partir d'une estimation lisse de F_0 au chapitre 3, dénotée $\hat{F}_h$, la modification proposée des algorithmes de rééchantillonnage abordés au chapitre 2 consiste à tirer les échantillons bootstrap d'une version lisse de la pseudo-population, dénotée U_h^*.

En se remémorant la mise en œuvre du bootstrap lisse pour des données i.i.d., il est facile de se représenter la façon de construire U_h^*. Puisque la pseudo-population U^* définit une fonction de répartition, la convolution de celle-ci avec une variable aléatoire ayant pour fonction de densité k_h, où k est une fonction de noyau et h est le paramètre de lissage, constituera la fonction de répartition lisse souhaitée (voir la proposition 3.2.2). Une façon directe de reproduire cette convolution est d'additionner à chacune des unités incluses dans U^* la quantité $h\varepsilon$, où ε est une variable aléatoire réelle ayant pour fonction de densité k. La modification proposée consiste donc à introduire le lissage au moment de la création de la pseudo-population U^* dont il est question dans les algorithmes 2.2.1 et 2.2.2. Par la suite, les échantillons bootstrap sont tirés de la pseudo-population lisse U_h^* selon le plan de sondage initial.

En raison de ses propriétés mathématiques intéressantes, les algorithmes proposés seront décrits et étudiés en se restreignant au cas $k = \phi$, où ϕ est la fonction de densité gaussienne standardisée. Il est à noter que k aurait pu être toute autre fonction de densité de probabilité satisfaisant les conditions (3.1.3), de façon à ce que l'ordre de $n^{-2/5}$ pour

l'erreur relative de l'estimateur bootstrap lisse de variance d'un estimateur de quantile (voir l'équation 3.3.9) soit respecté dans le cadre classique. Dans le cas du noyau gaussien, il suit que $\kappa_1 = \int k^2(t)dt = [2\sqrt{\pi}]^{-1}$ et $\kappa_2 = \int t^2 k(t)dt = 1$, deux quantités jouant un rôle dans la formulation du paramètre de lissage optimal pour l'estimation de la variance d'un quantile échantillonnal. Nous utiliserons ces valeurs à l'avenir.

Nous introduisons à présent l'algorithme de bootstrap par pseudo-population lisse pour le plan de sondage EASSR.

Algorithme 4.1.1 (BPP-EASSR lisse)**.**

(1) Former la partie fixe de la pseudo-population, dénotée U^f en répliquant chaque unité de l'échantillon s un nombre $m = \lfloor N/n \rfloor$ fois.

(2) Compléter la pseudo-population en tirant un échantillon aléatoire simple sans remise, dénoté U^{c*} de taille $n' = N - nm$ à partir de l'échantillon s. La pseudo-population correspond ainsi à $U^* = U^f \cup U^{c*}$. Les valeurs d'une variable y dans U^* sont contenues au sein du vecteur $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_N^*)'$.

(3) Générer N observations indépendantes et identiquement distribuées $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_N^*)'$, où $\varepsilon_i^* \sim \mathcal{N}(0,1)$, $i = 1, \ldots, N$. Obtenir le vecteur $\boldsymbol{y}_h^* = \boldsymbol{y}^* + h\boldsymbol{\varepsilon}^*$, où $h > 0$ est le paramètre de lissage, donnant la pseudo-population lisse U_h^*. Le paramètre bootstrap associé est donné par $\theta_h^* = \theta(U_h^*)$.

(4) Tirer un échantillon aléatoire simple sans remise s_h^* de taille n à partir de U_h^*.

(5) Calculer l'estimateur bootstrap lisse $\hat{\theta}_h^*$ à partir des observations de s_h^*.

(6) Répéter les étapes 2 à 6 un grand nombre B de fois de manière à obtenir les collections

$$(\theta_{1,h}^*, \ldots, \theta_{B,h}^*)' \quad et \quad (\hat{\theta}_{1,h}^*, \ldots, \hat{\theta}_{B,h}^*)'.$$

Remarquons que les étapes (1) et (2) de l'algorithme 4.1.1 sont identiques à celles de l'algorithme 2.2.1. Seule l'étape (3) est réellement nouvelle, tandis qu'on lisse la pseudo-population U^* afin qu'elle devienne U_h^*. Celle-ci est indicée par h pour mettre en évidence que la pseudo-population obtenue dépend de la valeur du paramètre de lissage h utilisée. Ensuite, les étapes (4) à (6) sont identiques aux étapes (3) à (5) de l'algorithme initial si ce n'est que les différentes quantités sont maintenant indicées par h. Si plusieurs valeurs de h doivent être comparées, comme c'est le cas dans la sélection par bootstrap

qui sera décrite ultérieurement, il n'est pas nécessaire de simuler de nouvelles variables aléatoires $\boldsymbol{\varepsilon}^*$ à chaque fois. Pour une même itération de l'algorithme, il suffit de simuler $\boldsymbol{\varepsilon}^*$ une seule fois et de conserver la réalisation du vecteur aléatoire pour le calcul des différentes pseudo-populations U_h^* induites par chaque valeur de h.

L'algorithme de bootstrap par pseudo-population lisse pour le plan de Poisson se décline comme suit.

Algorithme 4.1.2 (BPP-Poisson lisse)**.**

(1) Répliquer les unités (y_i, π_i), $i \in s$ *un nombre* $\lfloor \pi_i^{-1} \rfloor$ *de fois, où* $\lfloor \cdot \rfloor$ *est la partie entière, de façon à obtenir la partie fixe de la pseudo-population, notée* U^f.

(2) Afin de compléter la pseudo-population, former l'ensemble U^{c*} *en échantillonnant des unités à partir de* s *selon le plan de Poisson avec probabilité d'inclusion* $\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$ *pour la* i*-ème unité* (y_i, π_i), $i \in s$. *La pseudo-population ainsi obtenue est notée* $U^* = U^f \cup U^{c*}$ *et est constituée des paires notées* $(\breve{y}_i, \breve{\pi}_i), i \in U^*$, *où* $\breve{y}_i$ *se trouve à être l'une des valeurs présentes dans* s *et* $\breve{\pi}_i$ *est le poids de sondage initial correspondant. Les valeurs d'une variable* y *dans* U^* *sont contenues au sein du vecteur* $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_{N^*}^*)'$, *où* N^* *est la taille de la pseudo-population* U^*, *qui est aléatoire.*

(3) Générer N^* *observations indépendantes et identiquement distribuées* $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_{N^*}^*)'$, *où* $\varepsilon_i^* \sim \mathcal{N}(0,1)$, $i = 1,2,\ldots,N^*$. *Obtenir le vecteur* $\boldsymbol{y}_h^* = \boldsymbol{y}^* + h\boldsymbol{\varepsilon}^*$, *où* $h > 0$ *est le paramètre de lissage, donnant la pseudo-population lisse* U_h^*. *Le paramètre bootstrap associé est donné par* $\theta_h^* = \theta(U_h^*)$.

(4) Obtenir l'échantillon bootstrap s_h^* *en conduisant l'échantillonnage de Poisson à travers l'ensemble* U_h^* *avec les poids de sondage* $\breve{\pi}_i$, $i \in U_h^*$.

(5) Calculer l'estimateur bootstrap lisse $\hat{\theta}_h^*$ *à partir des observations contenues dans* s_h^*.

(6) Répéter les étapes 2 à 6 un grand nombre B *de fois de manière à obtenir les collections*

$$(\theta_{1,h}^*, \ldots, \theta_{B,h}^*)' \quad et \quad (\hat{\theta}_{1,h}^*, \ldots, \hat{\theta}_{B,h}^*)'.$$

De façon analogue au plan précédent, la nouveauté réside dans le lissage de la pseudo-population U^*, qui engendre subséquemment la pseudo-population lisse U_h^* ainsi que les autres quantités indicées par h.

À partir des distributions engendrées par les algorithmes 4.1.1 et 4.1.2, un estimateur de variance bootstrap lisse analogue à celui présenté à la sous-section 2.2.1.1 peut être dérivé. Une formulation explicite de l'estimateur résultant est

$$\hat{V}_h^* = \hat{V}_h^*(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{b,h}^* - \theta_{b,h}^*\right)^2. \tag{4.1.1}$$

Une fois de plus, dans l'éventualité où $\hat{\theta}$ n'est pas sans biais pour le paramètre dans la population θ, (4.1.1) consistera en un estimateur de l'erreur quadratique moyenne de l'estimateur plutôt que de sa variance. Les distributions bootstrap $(\theta_{1,h}^*, \ldots, \theta_{B,h}^*)'$ et $(\hat{\theta}_{1,h}^*, \ldots, \hat{\theta}_{B,h}^*)'$ peuvent par le fait même être exploitées afin d'obtenir des intervalles de confiance bootstrap asymptotiques, de base et percentiles pour θ tel que décrit à la sous-section 2.2.1.2.

Observons qu'en posant $h = 0$, cela revient à n'appliquer aucun lissage et donc à recourir aux algorithmes 2.2.1 ou 2.2.2 selon le plan de sondage. Une valeur $h > 0$ quelconque définira une toute nouvelle distribution bootstrap. Cela étant, une taille de fenêtre inappropriée peut mener à de pires performances que les méthodes sans lissage, soit au regard de l'estimation de la variance ou du taux de couverture des intervalles de confiance. Il conviendra donc de se doter de méthodes raisonnées pour la sélection du paramètre de lissage plutôt que d'avoir recours à une sélection arbitraire. Nous définirons ainsi des critères à optimiser, soit l'erreur quadratique moyenne de l'estimateur de variance d'un estimateur $\hat{\theta}$ et la distance entre les taux de couverture nominaux et expérimentaux d'intervalles de confiance pour θ. La prochaine section concerne une méthode de sélection se focalisant précisément sur l'optimisation de l'erreur quadratique moyenne de $\hat{V}_h^*(\hat{\xi}_p)$.

4.2. Sélection du paramètre de lissage par principe d'injection

Dans cette section, nous devrons faire la supposition que les mesures comprises dans la population U, soit $y_1, y_2, \ldots, y_N$, sont des observations i.i.d. issues d'une superpopulation F_0 avec fonction de densité f_0. Tout comme dans la section 1.3.5, on distingue les opérateurs d'espérance et de variance dans la superpopulation, dénotés $\mathbb{E}_0$ et Var_0 respectivement, de ceux sous le plan, dénotés $\mathbb{E}_p$ et Var_p respectivement.

Une sélection par injection ou *plug-in* peut être réalisée lorsqu'une expression analytique du paramètre de lissage minimisant un certain critère (l'erreur quadratique moyenne, par exemple) est à notre disposition. Les résultats énoncés dans la sous-section 3.3.2 mènent à une expression pour le paramètre de lissage optimal pour minimiser l'erreur quadratique de l'estimateur de variance bootstrap lisse de $\check{\xi}_p$ dans le cas de données indépendantes et identiquement distribuées (voir l'équation 3.3.8). En utilisant la fonction de noyau $k = \phi$, la taille de fenêtre optimale pour ce critère s'écrit désormais

$$h_{\text{opt,i.i.d.}}(\tilde{\xi}_p) = \left[f_0(\tilde{\xi}_p)\right]^{1/5} \left[2\sqrt{\pi}\right]^{-1/5} \left[f_0''(\tilde{\xi}_p) - \left(f_0'(\tilde{\xi}_p)\right)^2 f_0(\tilde{\xi}_p)^{-1}\right]^{-2/5} n^{-1/5}, \tag{4.2.1}$$

où $\tilde{\xi}_p$ est le quantile de niveau p de la distribution F_0. Nous mettrons à profit ce résultat en introduisant une méthode de sélection *plug-in* dans le cadre d'une population finie. Afin d'appliquer la méthode de sélection par injection à des données d'un plan de sondage EASSR, on doit donc postuler que (4.2.1), qui est la taille optimale pour minimiser l'erreur relative asymptotique de l'estimateur de variance bootstrap lisse dans le cas i.i.d., est également la taille de fenêtre optimale pour minimiser l'erreur relative de $\hat{V}_h^* = \hat{V}_h^*(\hat{\xi}_p)$ par rapport à la variance sous le plan conditionnellement à U.

Cette approche peut être justifiée par le fait que les propriétés asymptotiques de l'estimateur d'un quantile dans le cadre i.i.d. et sous le plan EASSR sont comparables. Pour des données i.i.d., rappelons que la taille de fenêtre (4.2.1) est dérivée en considérant l'estimateur de variance bootstrap comme un estimateur de la variance asymptotique de $\check{\xi}_p$. Celle-ci est égale à

$$\text{Var}_{\text{asy}}(\check{\xi}_p) = \frac{\rho^2}{n}, \tag{4.2.2}$$

où $\rho^2 = p(1-p)/f_0^2(\tilde{\xi}_p)$. Dans le cas du plan EASSR, en réinvoquant l'existence d'une superpopulation, il a été établi au chapitre 1 que

$$\text{Var}_p(\hat{\xi}_p \mid U) \to \frac{1-f}{n}\rho^2 \quad \text{lorsque} \quad n \to \infty \text{ et } N \to \infty, \tag{4.2.3}$$

presque sûrement, où la variance asymptotique contient le facteur de correction pour population finie, $(1-f)$. La variance asymptotique dans (4.2.3) est par conséquent égale à (4.2.2) à un facteur multiplicatif près. Si l'on désirait mieux refléter (4.2.3), une avenue serait substituer la taille échantillonnale n par la *taille échantillonnale effective*, définie

comme $n/(1-f)$, dans l'expression de (4.2.1). De cette façon, une expression alternative du paramètre de lissage optimal serait

$$h_{\mathrm{opt},f}(\tilde{\xi}_p) = \left[f_0(\tilde{\xi}_p)\right]^{1/5} \left[2\sqrt{\pi}\right]^{-1/5} \left[f_0^{(2)}(\tilde{\xi}_p) - \left(f_0'(\tilde{\xi}_p)\right)^2 f_0(\tilde{\xi}_p)^{-1}\right]^{-2/5} \left[\frac{n}{1-f}\right]^{-1/5}, \quad (4.2.4)$$

où f dans $h_{\mathrm{opt},f}$ fait référence à la fraction de sondage du plan. Le mode conditionnel est employé puisque le recours à (4.2.4) plutôt que (4.2.1) a une incidence marginale sur la performance de l'estimateur $\hat{V}_h^*$, tel que des simulations pour le plan EASSR analogues à celles présentées au chapitre 5 ont pu permettre de constater. Cela peut être appréhendé en faisant remarquer que $h_{\mathrm{opt},f}(\tilde{\xi}_p) = (1-f)^{1/5} h_{\mathrm{opt,\ i.i.d.}}(\tilde{\xi}_p)$. Pour des fractions de sondage de 7% et de 30%, le facteur $(1-f)^{1/5}$ vaut approximativement 0,985 et 0,931 respectivement. Qui plus est, pour que (4.2.4) vaille la moitié de (4.2.1) en maintenant n fixe, il faudrait que $f = 31/32$, ce qui est près d'être un recensement. Pour ces raisons, nous ne retiendrons que l'expression (4.2.1) pour l'évaluation de la méthode de sélection par injection dans le cas du plan EASSR.

La quantité (4.2.1) dépend de quantités inconnues, à commencer par la fonction de densité f_0. Une approche envisageable est de faire de cette valeur une cible à atteindre en remplaçant les quantités inconnues dans l'expression par leurs analogues échantillonnaux, d'où le nom de *plug-in*. De manière similaire à ce que propose Silverman (1986) pour le problème d'estimation de la fonction de densité, on fait d'abord la supposition que les données proviennent de la distribution (superpopulation) $\mathcal{N}(\mu, \sigma^2)$ en posant $f_0(x) = \sigma^{-1}\phi(z)$, où ϕ est la fonction de densité gaussienne standardisée et $z \equiv z(x, \mu, \sigma^2) = (x-\mu)/\sigma$. Alors, l'expression (4.2.1) ne dépendra que du quantile théorique $\tilde{\xi}_p$ et des paramètres de localisation et d'échelle de la distribution. À partir des propriétés de la fonction de densité de la distribution normale exhibées à l'annexe A.1, on peut écrire la taille de fenêtre optimale pour la distribution normale comme suit

$$h_{\mathrm{opt,norm}}(z) = \left[\frac{1}{\sigma}\phi(z)\right]^{1/5} \left[2\sqrt{\pi}\right]^{-1/5} \left[\frac{1}{\sigma}\phi''(z) - \left(\frac{1}{\sigma}\phi'(z)\right)^2 \left(\frac{1}{\sigma}\phi(z)\right)^{-1}\right]^{-2/5} n^{-1/5}, \quad (4.2.5)$$

où $z \equiv z(\tilde{\xi}_p, \mu, \sigma^2) = (\tilde{\xi}_p - \mu)/\sigma$. Conditionnellement à l'observation d'un échantillon $y_1, y_2, \ldots, y_n$, une estimation de la taille de fenêtre optimale pourra alors être obtenue en substituant $\tilde{\xi}_p$ par $\hat{\xi}_p$, μ par l'estimateur à vraisemblance maximale $\hat{\mu} = \bar{y} = n^{-1}\sum_{i=1}^n y_i$ et σ^2 par l'estimateur sans biais $\hat{\sigma}^2 = s_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$ dans (4.2.5). L'estimateur

plug-in de la taille de fenêtre optimale basée sur la normalité est donné explicitement comme

$$\hat{h}_{\text{plug-in}}(\hat{z}) = \left[\frac{1}{\hat{\sigma}}\hat{\phi}(\hat{z})\right]^{1/5}\left[2\sqrt{\pi}\right]^{-1/5}\left[\frac{1}{\hat{\sigma}}\phi''(\hat{z}) - \left(\frac{1}{\hat{\sigma}}\phi'(\hat{z})\right)^2\left(\frac{1}{\hat{\sigma}}\phi(\hat{z})\right)^{-1}\right]^{-2/5} n^{-1/5}, \quad (4.2.6)$$

où $\hat{z} \equiv \hat{z}(\hat{\xi}_p, \hat{\mu}, \hat{\sigma}^2) = (\hat{\xi}_p - \hat{\mu})/\hat{\sigma}$.

En l'absence d'un résultat théorique pour la variance asymptotique de $\hat{\xi}_p$ dans le cas du plan de Poisson, il est plus difficile de spéculer quant au paramètre de lissage optimal. La variance sous le plan dépend alors des probabilités d'inclusion de premier ordre $\pi_i,\ i = 1, \ldots, N$, qui peuvent elles-mêmes dépendre des observations d'une variable auxiliaire x pour un plan proportionnel à la taille. Les cas de figure sont donc multiples sous ce plan et la variance asymptotique a donc le potentiel de s'éloigner grandement de (4.2.2). Par conséquent, la méthode de sélection par injection sera écartée lors de l'étude par simulation pour le plan de Poisson.

La section suivante porte sur une méthode de sélection par bootstrap pouvant être utilisée pour optimiser deux critères: l'erreur quadratique moyenne de l'estimateur de variance lisse d'une statistique $\hat{\theta}$ et une mesure de distance entre les taux d'erreur de couverture expérimentale et nominale d'intervalles de confiance pour θ. Bien qu'elle soit numériquement intensive, elle comporte l'avantage de pouvoir être mise en œuvre quels que soient le plan de sondage ou la statistique examinés.

4.3. Sélection du paramètre de lissage par bootstrap

Nous cheminons à présent vers une méthode de sélection qui ne nécessite pas de formuler d'hypothèses quant au modèle générateur des observations dans la population. Elle peut être mise en œuvre quelle que soit la statistique examinée, par opposition à la méthode précédente qui reposait sur un résultat théorique concernant uniquement les quantiles. Le principe général consiste en premier lieu à définir une fonction mesurant la qualité de l'estimateur que l'on souhaitera minimiser par rapport à h. Puisque celle-ci dépend le plus souvent de quantités inconnues, nous nous appliquerons à minimiser une estimation de cette mesure d'ajustement obtenue au moyen du bootstrap. Cette méthode fait appel à

une grille de valeurs pour le paramètre de lissage $h > 0$ que l'on dénote $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$.

4.3.1. Optimisation de l'erreur quadratique moyenne de l'estimateur de variance bootstrap

En prenant comme fonction de perte la distance quadratique, ou la norme L_2, une mesure d'ajustement de l'estimateur $\hat{V}_h^*(\hat{\xi}_p)$ par rapport à la variance véritable sous le plan est donnée par

$$\mathrm{EQM}\left(\hat{V}_h^*\right) = \mathbb{E}_p\left[\left(\hat{V}_h^* - \mathrm{Var}_p(\hat{\xi}_p)\right)^2\right]. \tag{4.3.1}$$

Supposons que l'on souhaite sélectionner h de manière à minimiser (4.3.1). Cette dernière quantité étant inconnue, il s'agit d'une tâche impossible à réaliser. Nous pourrions en contrepartie minimiser une estimation bootstrap de (4.3.1), s'écrivant

$$\widehat{\mathrm{EQM}}\left(\hat{V}_h^*\right) = \mathbb{E}^*\left[\left(\hat{V}_h^{**} - \hat{V}_h^*\right)^2\right], \tag{4.3.2}$$

où $\hat{V}_h^{**} = \hat{V}_h^{**}(\hat{\xi}_p)$ est l'estimateur (double) bootstrap de $\hat{V}_h^*$.

Si l'on prend l'exemple de l'EASSR, en se référant à l'algorithme 4.1.1, il suffirait de générer une distribution bootstrap de deuxième niveau indicée par un paramètre de lissage h à partir de l'échantillon s_h^* généré à l'étape (4). Cela consiste en d'autres mots à réappliquer l'algorithme 4.1.1 un nombre D de fois à s_h^* de manière imbriquée, de façon à obtenir les collections $(\theta_{1,h}^{**}, \ldots, \theta_{D,h}^{**})'$ et $(\hat{\theta}_{1,h}^{**}, \ldots, \hat{\theta}_{D,h}^{**})'$. Les collections des paramètres bootstrap et des estimateurs bootstrap de deuxième niveau permettront alors d'évaluer

$$\hat{V}_h^{**} = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_{d,h}^{**} - \theta_{d,h}^{**}\right)^2, \tag{4.3.3}$$

qui est l'estimateur bootstrap de $\hat{V}_h^*$. Si ceci est accompli pour chacun des B échantillons bootstrap générés au premier niveau de l'algorithme 4.1.1, cela nous conduira à la distribution $(\hat{V}_{1,h}^{**}, \ldots, \hat{V}_{B,h}^{**})'$ associée à l'estimateur $\hat{V}_h^*(\hat{\xi}_p)$ résultant du premier niveau. Ici, $\hat{V}_{i,h}^{**}$ est l'estimateur (4.3.3) associé au $i^{\text{ème}}$ échantillon bootstrap (de premier niveau), $i = 1, \ldots, B$. La même approche peut être réalisée dans le cas de l'algorithme 4.1.2 pour l'échantillonnage de Poisson.

Dans les deux cas, on estimera la valeur optimale du paramètre de lissage par celle minimisant l'estimation bootstrap de l'erreur quadratique de $\hat{V}_h^*$, c'est-à-dire

$$\hat{h}_{\text{boot,var}} = \arg\min_{h\in\mathcal{H}} \widehat{\text{EQM}}\left(\hat{V}_h^*\right) \tag{4.3.4}$$

$$= \arg\min_{h\in\mathcal{H}} \frac{1}{B}\sum_{b=1}^{B}\left[\hat{V}_{h,b}^{**} - \hat{V}_h^*\right]^2, \tag{4.3.5}$$

où $\hat{V}_h^*$ est donné par (4.1.1).

4.3.2. Optimisation du taux de couverture d'intervalles de confiance

Comme nous le verrons dans le chapitre 5, en sélectionnant la valeur de $h \in \mathcal{H}$ minimisant l'estimation bootstrap du critère (4.3.1), les intervalles de confiance lisses résultants ne possèdent pas nécessairement de bonnes probabilités de couverture. Rien n'oblige en effet que la valeur de h optimisant un critère relatif à un estimateur de variance soit également la valeur optimale pour un critère relatif à la probabilité de couverture d'un intervalle de confiance. Les intervalles de confiance étant également des objets d'intérêt dans ce mémoire, cela nous conduit à considérer un critère de qualité adapté à cette visée particulière. On souhaitera à présent choisir h de façon à ce que la probabilité de couverture d'un intervalle de confiance lisse pour θ se rapproche de la probabilité de couverture nominale.

Soit I_h^* un intervalle de confiance bootstrap lisse de niveau $1-\alpha$ pour un paramètre $\theta = \theta(U)$ construit à partir de données de sondage $\boldsymbol{y} = (y_1, \ldots, y_n)'$ et défini à partir d'un paramètre de lissage h. Supposons que l'on fixe $1-\alpha = 0{,}95$. S'il y a sur-couverture pour I_h^*, celle-ci ne peut dépasser 5% alors qu'une sous-couverture pourrait aller jusqu'à 95%. Ainsi, une sur-couverture de $\beta = 100 \cdot |P(\theta \in I_h^*) - (1-\alpha)|\%$ ne devrait pas être pénalisée de la même façon qu'une sous-couverture du même niveau. Pour cette raison, nous avons recours à une autre fonction de distance que la valeur absolue de la différence, qui a été introduite par Calonico et al. (2018). Cette fonction de distance est $\mathcal{L} : [0,1] \to \mathbb{R}^+$, définie par $\mathcal{L}(e) = \mathcal{L}_\tau(e) = e\,(\tau - \mathbb{1}(e<0))$ et $\tau \in (0,1)$. Celle-ci joue un rôle au sein du critère suivant:

$$L(I_h^*, \theta) = \mathcal{L}\left[P(\theta \in I_h^*) - (1-\alpha)\right], \tag{4.3.6}$$

La fonction (4.3.6) est tracée pour des valeurs de $P(\theta \in I_h^*)$ comprises entre 0,8 et 1,0 à la figure 4.1 pour trois valeurs du paramètre τ.

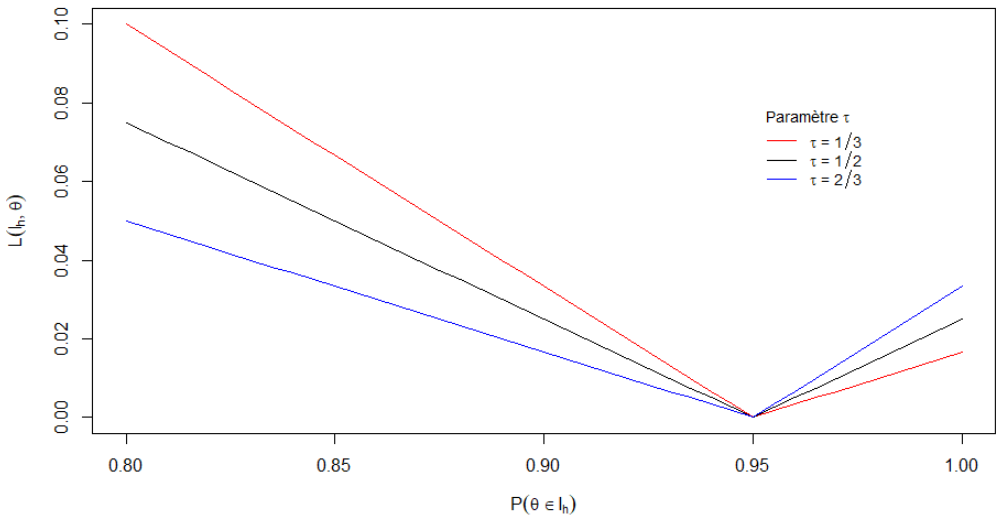


Fig. 4.1. Graphe de la fonction de distance L en fonction de la probabilité de couverture $P(\theta \in I_h^*)$ pour des valeurs de paramètre τ de $1/3,\ 1/2,\ 2/3$ et un niveau de confiance de $1 - \alpha = 0{,}95$.

Nous pouvons observer que le paramètre τ régit la tangente de la fonction de part et d'autre du point d'inflexion $1 - \alpha$, une valeur de $1/2$ produisant une symétrie. Nous nous attardons sur ce point puisque nous pourrions opter pour des valeurs de τ différentes de $1/2$ selon le coût que nous attribuons à diverses décisions. Par exemple, si l'on veut faire correspondre une plus grande perte à une sur-couverture de $\beta\%$ qu'à une sous-couverture de même niveau, nous privilégierons des valeurs de τ supérieures à $1/2$. Dans la mise en œuvre de la méthode, nous choisirons $\tau = 2/3$, qui correspond à la courbe en bleu dans la figure 4.1. Avec ce choix, pour une probabilité de couverture nominale de $1 - \alpha = 0{,}95$, nous obtenons des valeurs de L de 0,02 et 0,01 pour des probabilités de couverture de 0,98 et 0,92 respectivement.

Si $P(\theta \in I_h^*)$ était connue, le paramètre de lissage serait choisi de telle sorte que

$$h_0 = \arg\min_{h \in \mathcal{H}} L(I_h^*, \theta). \tag{4.3.7}$$

Or, la probabilité de couverture de I_h^* est bien sûre inconnue et ce faisant, $L(I_h^*, \theta)$ l'est tout autant. En pratique, un utilisateur pourra estimer $L(I_h, \theta)$ par bootstrap pour chaque valeur h comprise dans une grille et ainsi sélectionner la valeur minimisant l'estimation du critère.

De façon plus explicite, l'évaluation de l'estimation bootstrap de $L(I_h^*, \theta)$ se fera au moyen des estimateurs bootstrap lisses de deuxième niveau $\hat{V}_h^{**}$ discutés à la section précédente (voir l'équation 4.3.3). Pour chaque échantillon bootstrap $b = 1, \ldots, B$ du premier niveau de l'algorithme lisse, on peut obtenir l'intervalle de confiance asymptotique de deuxième niveau dénoté $I_{b,h}^{**}$ et donné par

$$\left[\hat{\theta}_b^* - z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}_{b,h}^{**}}, \hat{\theta}_b^* + z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}_{b,h}^{**}}\right], \tag{4.3.8}$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile de niveau $1-\alpha/2$ de la distribution gaussienne standardisée. Le calcul de l'intervalle (4.3.8) plutôt qu'un intervalle percentile ou de base procure l'avantage de réduire l'intensité de calcul de l'algorithme du double bootstrap. Un intervalle percentile ou de base de deuxième niveau serait basé sur les quantiles expérimentaux de niveaux $\alpha/2$ et $1-\alpha/2$ de la distribution $(\hat{\theta}_h^{**} - \theta_h^{**})$. Tel qu'établi par Efron et Tibshirani (1993), le nombre de réplicats D devrait être alors de l'ordre de 1 000 et encore davantage s'il s'agit d'estimer les percentiles situés dans les ailes de la distribution (notamment les percentiles 2,5% et 97,5%), où il y a moins d'occurrences. Dans le cas présent, l'intervalle (4.3.8) se fonde sur une estimation bootstrap de deuxième niveau de la variance $\hat{V}_h^*$, laquelle nécessite aussi peu de réplicats que $D = 50$ pour obtenir une précision suffisante (Efron et Tibshirani, 1993).

Avec la collection $(I_{1,h}^*, \ldots, I_{B,h}^*)'$ ainsi formée, une estimation bootstrap de $P(\theta \in I_h^*)$ est donnée par

$$\hat{P}(\theta \in I_h^*) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\theta_b^* \in I_{b,h}^{**}), \tag{4.3.9}$$

où θ_b^* est le paramètre de la pseudo-population avant lissage obtenue à la b-ième itération de l'algorithme 4.1.1. Celui-ci est mis en place du vrai paramètre θ dans le monde bootstrap. Notons que nous avons également considéré d'évaluer $\hat{P}(\theta \in I_h^*)$ en cherchant plutôt à couvrir $\hat{\theta}_b$, qui constitue également une approche raisonnable. Or, contrairement à θ_b^*, celui-ci est invariant à travers les itérations de l'algorithme. Après avoir obtenu des résultats

numériques similaires pour les deux façons, seule l'approche donnée par (4.3.9) a été retenue.

Enfin, on estimera la valeur optimale du paramètre de lissage pour ce critère, donnée par (4.3.7), par celle minimisant l'estimation bootstrap $\hat{L}(I_h^*, \theta)$, soit

$$\hat{h}_{\text{boot,ic}} = \arg\min_{h\in\mathcal{H}} \hat{L}(I_h^*, \theta) \tag{4.3.10}$$

$$= \arg\min_{h\in\mathcal{H}} \mathcal{L}\left[\hat{P}(\theta \in I_h^*) - (1-\alpha)\right] \tag{4.3.11}$$

$$= \arg\min_{h\in\mathcal{H}} \mathcal{L}\left[\frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(\theta_b^* \in I_h^{**}) - (1-\alpha)\right]. \tag{4.3.12}$$

La prochaine sous-section vise à synthétiser l'approche de sélection du paramètre de lissage venant d'être discutée en l'intégrant au sein de la description de l'algorithme par pseudo-population lisse.

4.3.3. Algorithme du double bootstrap

La sélection du paramètre de lissage optimal par bootstrap pour les critères discutés dans les sous-sections 4.3.1 et 4.3.2 peut être faite au sein d'une même exécution de l'algorithme lisse. La mise en œuvre de l'approche du double bootstrap est décrite explicitement ci-dessous en prenant l'exemple de l'algorithme lisse pour le plan EASSR (algorithme 4.1.1).

Algorithme 4.3.1 (Double BPP-EASSR lisse)**.**

(1) Constituer une grille de valeurs pour le paramètre de lissage $h > 0$, dénotée $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$.

(2) Former la partie fixe de la pseudo-population, dénotée U^f, en répliquant chaque unité de l'échantillon s un nombre $m = \lfloor N/n \rfloor$ fois.

(3) Compléter la pseudo-population en tirant un échantillon aléatoire simple sans remise, dénoté U^{c*}, de taille $n' = N - nm$ à partir de l'échantillon s. La pseudo-population correspond ainsi à $U^* = U^f \cup U^{c*}$. Les valeurs d'une variable y dans U^* sont contenues au sein du vecteur $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_N^*)'$. Le paramètre bootstrap non lisse est donné par $\theta^* = \theta(U^*)$

(4) *Générer N observations indépendantes et identiquement distribuées $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_N^*)'$, où $\varepsilon_i^* \sim \mathcal{N}(0,1)$, $i = 1, \ldots, N$. Pour chaque $h \in \mathcal{H}$, obtenir le vecteur $\boldsymbol{y}_h^* = \boldsymbol{y}^* + h\boldsymbol{\varepsilon}^*$, donnant la pseudo-population lisse U_h^* et le paramètre bootstrap associé $\theta_h^* = \theta(U_h^*)$.*

(5) *Tirer un échantillon aléatoire simple sans remise s^* de taille n à partir de U^*. Pour chaque $h \in \mathcal{H}$, tirer un échantillon aléatoire simple sans remise s_h^* de taille n à partir de U_h^*.*

(6) *Calculer l'estimateur bootstrap non lisse $\hat{\theta}^*$ à partir des observations de s^* de même que les estimateurs bootstrap lisses $\hat{\theta}_h^*$ à partir des observations des échantillons lisses $s_h^*, h \in \mathcal{H}$.*

(7) *Appliquer les étapes (2) à (6) à l'échantillon s^* avec la grille $\mathcal{H}$ et répéter un nombre D fois de manière à obtenir, pour chaque $h \in \mathcal{H}$, les collections d'estimateurs bootstrap $(\theta_{1,h}^{**}, \ldots, \theta_{D,h}^{**})'$ et de paramètres bootstrap $(\hat{\theta}_{1,h}^{**}, \ldots, \hat{\theta}_{D,h}^{**})'$ de deuxième niveau. Pour chaque $h \in \mathcal{H}$, il en découle*

$$\hat{V}_h^{**} = \frac{1}{D} \sum_{d=1}^{D} \left(\hat{\theta}_{d,h}^{**} - \theta_{d,h}^{**} \right)^2,$$

soit l'estimateur bootstrap de variance de deuxième niveau ainsi que l'intervalle de confiance asymptotique de deuxième niveau dénoté I_h^{**} et donné par

$$\left[\hat{\theta}^* - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_h^{**}}, \hat{\theta}^* + z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_h^{**}} \right].$$

(8) *Répéter les étapes 2 à 7 un grand nombre B de fois de manière à obtenir les collections*

$$(\theta_1^*, \ldots, \theta_B^*)', (\theta_{1,h}^*, \ldots, \theta_{B,h}^*)', (\hat{\theta}_{1,h}^*, \ldots, \hat{\theta}_{B,h}^*)', (\hat{V}_{1,h}^{**}, \ldots, \hat{V}_{B,h}^{**})' \text{ et } (I_{1,h}^{**}, \ldots, I_{B,h})'.$$

(9) *Pour $h \in \mathcal{H}$, à partir des collections $(\theta_{1,h}^*, \ldots, \theta_{B,h}^*)'$ et $(\hat{\theta}_{1,h}^*, \ldots, \hat{\theta}_{B,h}^*)'$, calculer l'estimateur de variance bootstrap de premier niveau $\hat{V}_h^*$, donné par (4.1.1).*

(10) *À partir de $\hat{V}_h^*$ et la collection $(\hat{V}_{1,h}^{**}, \ldots, \hat{V}_{B,h}^{**})', h \in \mathcal{H}$, calculer $\hat{h}_{boot,var}$ donné par (4.3.5), puis obtenir $\hat{h}_{boot,ic}$ donné par (4.3.12) à partir des collections $(\theta_1^*, \ldots, \theta_B^*)'$ et $(I_{1,h}^{**}, \ldots, I_{B,h})', h \in \mathcal{H}$.*

(11) *Parmi les estimateurs de variance de premier niveau $\hat{V}_h^*, h \in \mathcal{H}$, sélectionner $\hat{V}_{\hat{h}_{boot,var}}^*$ et construire les intervalles de confiance asymptotique, de base et percentile pour θ comme décrit à la sous-section 2.2.1.2 à partir des histogrammes bootstrap*

$$(\theta_{1;\hat{h}_{boot,ic}}^*, \ldots, \theta_{B;\hat{h}_{boot,ic}}^*)' \quad et \quad (\hat{\theta}_{1;\hat{h}_{boot,ic}}^*, \ldots, \hat{\theta}_{B;\hat{h}_{boot,ic}}^*)'.$$

Les étapes (2) à (6) de l'algorithme 4.3.1 sont identiques à celles de l'algorithme 4.1.1 (sans deuxième niveau de bootstrap) à l'exception faite que certaines font intervenir des quantités dérivant de la pseudo-population non lisse (U^*, s^*) et qu'il y a autant de pseudo-populations lisses qu'il n'y a de paramètres de lissage dans la grille $\mathcal{H}$. L'étape (7) amène le deuxième niveau d'imbrication du bootstrap. C'est là que des estimateurs bootstrap de variance de deuxième niveau indicés par $h \in \mathcal{H}$ seront obtenus et que des intervalles de confiance de deuxième niveau indicés par $h \in \mathcal{H}$ seront construits à partir de l'estimateur bootstrap $\hat{\theta}^*$ du premier niveau de bootstrap. Pour chaque taille de fenêtre h, les étapes (8) et (9) introduisent les quantités intermédiaires pour le calcul de $\hat{h}_{\text{boot,var}}$ et de $\hat{h}_{\text{boot,ic}}$ se faisant à l'étape (10). Enfin, l'étape (11) de l'algorithme renvoie l'estimateur de variance et les intervalles de confiance sélectionnés, qui s'expriment en fonction de $\hat{h}_{\text{boot,var}}$ et de $\hat{h}_{\text{boot,ic}}$. Nous laissons le soin au lecteur de transposer les étapes relatives au double bootstrap à l'algorithme 4.1.2 afin d'obtenir un analogue de l'algorithme 4.3.1 pour l'échantillonnage de Poisson.

Remarque 4.3.2 (Choix de la grille à explorer)**.** *La performance de cette méthode repose en partie sur la grille d'exploration $\mathcal{H}$ pour le paramètre de lissage $h > 0$. L'étendue de valeurs couverte par $\mathcal{H}$ de même que l'espacement entre les points (en nombre fini) peuvent avoir un impact. En pratique, une première exécution de l'algorithme 4.3.1 donnera accès à des versions bruitées des graphes de ces fonctions. Il sera possible de vérifier graphiquement que la grille $\mathcal{H}$ contient bel et bien les minimums globaux de (4.3.2) et de (4.3.9). Dans le cas contraire, il faudra étendre la grille à gauche ou à droite jusqu'à ce que l'on devine une tangente négative à gauche et une tangente positive à droite d'un point $h \in \mathcal{H}$.*

Chapitre 5

Étude par simulation

Une étude par simulation est présentée pour la médiane et le troisième quartile afin d'attester de la performance relative de l'innovation proposée au regard de l'estimation de l'erreur quadratique moyenne et la formation d'intervalles de confiance. Différents scénarios sont explorés en faisant varier le plan de sondage, la superpopulation génératrice, la taille échantillonnale et la fraction de sondage. À travers les différents cas de figure, l'adaptation lisse des méthodes de Booth et al. (1994) ou de Chauvet (2007) selon le plan étudié sera comparée à l'algorithme original (sans lissage) ainsi qu'à la méthode de Woodruff (1952). L'algorithme lisse est décliné selon les différentes façons de sélectionner la taille de la fenêtre.

Le cheminement de ce chapitre va comme suit. Les protocoles de génération de populations finies utilisés sont discutés avant tout. Pour les deux plans de sondage, les populations fixes générées lors des simulations basées sur le plan sont décrites à l'aide de représentations graphiques. La section suivante rassemble les définitions des mesures de performance qui permettront la comparaison entre les différentes méthodes d'estimation de variance et de formation d'intervalles de confiance. Ensuite, dans le but de clarifier la présentation des résultats, une section est consacrée à la description et à la dénomination des méthodes comparées en faisant référence aux chapitres précédents. Les résultats sont divisés en deux sections consacrées aux deux plans de sondage discutés. Dans chacune, les mesures de performance relatives à l'estimation de l'erreur quadratique moyenne et aux taux de couverture d'intervalles de confiance constituent les deux sous-sections principales. En ce qui a trait à l'estimation de l'erreur quadratique moyenne, les tableaux de résultats sont

complémentés par des représentations graphiques des courbes d'instabilité de l'estimateur. Afin d'évaluer l'optimalité du paramètre de lissage pour données i.i.d. dans un contexte de population finie, d'autres courbes d'instabilité seront produites à partir de simulations alliant le modèle (superpopulation) et le plan conjointement.

5.1. Génération des populations finies

Nous considérons quatre scénarios d'enquête issus du croisement de deux fractions de sondage, $f_1 = 0{,}07$, $f_2 = 0{,}30$, et de deux tailles échantillonnales, $n_1 = 100$, $n_2 = 500$. Pour chaque superpopulation F_0 étudiée, un seul échantillon i.i.d. de taille $\widetilde{N} = \lfloor \max(n_1, n_2)/\min(f_1, f_2) \rfloor = \lfloor 500/0{,}07 \rfloor$, où $\lfloor \cdot \rfloor$ est la fonction partie entière, est généré à partir de la distribution. Puis, cet échantillon est utilisé pour former des populations finies de différentes tailles $N = \lfloor n/f \rfloor$ en sélectionnant les N premiers indices. À partir de chaque population finie, S échantillons sont générés à partir du plan de sondage examiné.

L'étude par simulation ainsi construite se rapproche de la réalité de l'échantillonnage pour populations finies tel que pratiqué par des agences statistiques officielles comme Statistique Canada. On fixe alors une seule population U particulière, par opposition à une infinité d'autres populations qui auraient pu exister. On dénomme ce type de simulation comme étant une simulation *basée sur le plan*. Ce protocole de simulation a également été adopté par Chatterjee (2011), bien que le cadre théorique de l'article repose sur l'hypothèse que la population U est une réalisation aléatoire d'une superpopulation F_0. Les résultats principaux contenus dans ce chapitre auront été obtenus selon une simulation orientée sur le plan, c'est-à-dire en fixant préalablement une population finie.

Cela dit, en amont de ce chapitre, il a aussi été question d'une population U aléatoire issue d'une superpopulation F_0. Des résultats ont été établis en calculant l'espérance par rapport à ce processus aléatoire par le biais de l'opérateur $\mathbb{E}_0$. Un exemple est notamment l'expression du paramètre de lissage optimal, qui est valide si l'on considère une infinité de populations contrairement à une seule. En vue d'éclaircir la question de l'optimalité du paramètre de lissage trouvé dans le contexte des sondages, pour chaque

scénario d'enquête, des simulations secondaires seront conduites conformément au processus $\mathbb{E}_0\mathbb{E}_p$. Cela sera réalisé en générant S populations finies de taille N à partir de F_0, puis en tirant à partir de chacune un échantillon selon le plan de sondage examiné.

La sous-section suivante s'applique à décrire les populations finies fixes utilisées lors des simulations basées sur le plan pour les deux plans d'enquête abordés dans ce mémoire.

5.1.1. Échantillonnage aléatoire simple sans remise

Afin de tenir compte de l'influence de l'asymétrie de la distribution sur les mesures de performance, la distribution $\mathcal{N}(0,1)$ et la distribution Lognormale(0,1), font office de superpopulations pour ce plan de sondage. Pour chacune, les quatre populations fixées engendrées par les scénarios $\{n_1, n_2\} \times \{f_1, f_2\}$ sont décrites au moyen d'histogrammes de fréquences aux figures 5.1 et 5.2.

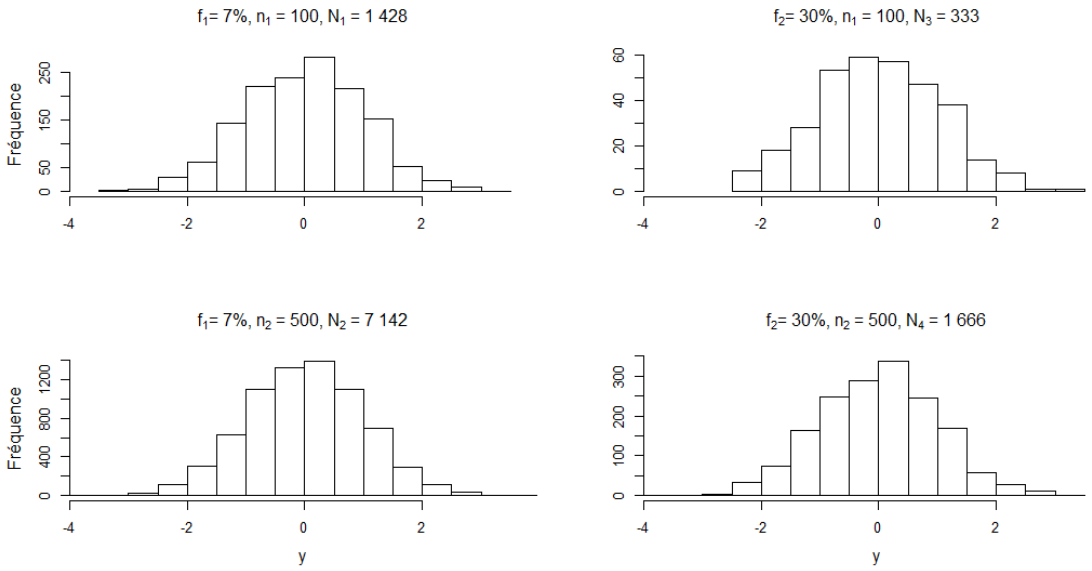


Fig. 5.1. Histogrammes de fréquence des quatre populations correspondant aux quatre scénarios d'enquête et formées à partir d'un échantillon i.i.d. de taille $\widetilde{N} = 7\ 142$ de la distribution $\mathcal{N}(0,1)$.

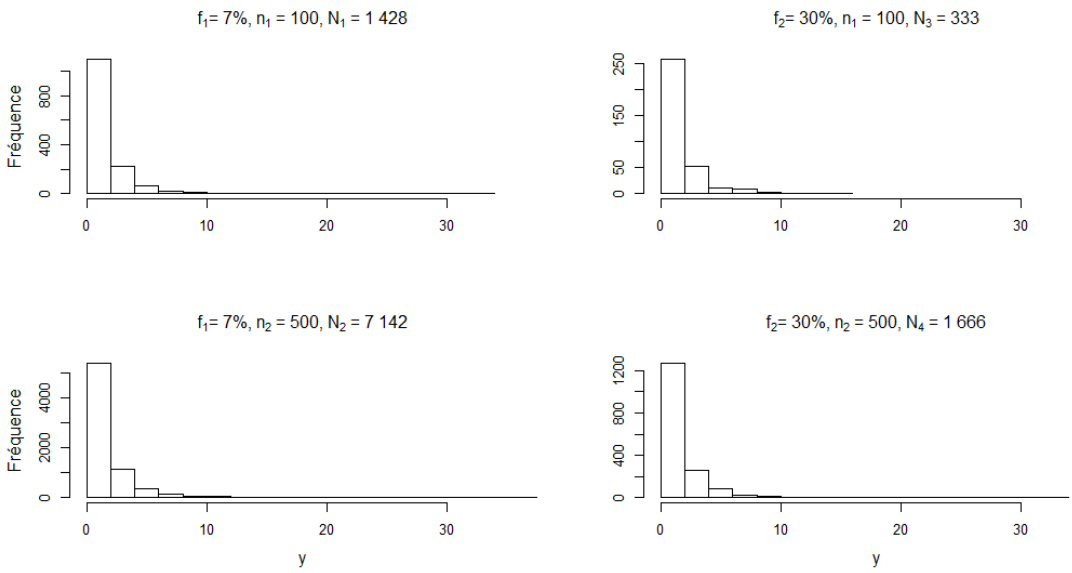


Fig. 5.2. Histogrammes de fréquence des quatre populations correspondant aux quatre scénarios d'enquête et formées à partir d'un échantillon i.i.d. de taille $\widetilde{N} = 7\ 142$ de la distribution Lognormale(0,1).

5.1.2. Échantillonnage de Poisson

Dans ce plan, n correspond à la taille échantillonnale espérée. La grande population de taille $\widetilde{N}$ est générée selon le modèle suivant

$$Y_i = \gamma X_i + \sigma\varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad X_i \overset{\text{i.i.d.}}{\sim} \chi^2_\nu, \quad i = 1,\ldots,\widetilde{N}, \tag{5.1.1}$$

avec $\gamma = 0{,}6$, $\nu = 100$ et $\sigma = 12$. Avec ce choix de paramètres, le coefficient de corrélation entre les réalisations $\boldsymbol{y} = (y_1, y_2, \ldots, y_{\widetilde{N}})'$ et $\boldsymbol{x} = (x_1, x_2, \ldots, x_{\widetilde{N}})'$ est de 0,5795, cette valeur étant basée sur la population de taille $\widetilde{N} = 7\ 142$ observations. Ainsi, pour chacun des quatre scénarios $\{n_1, n_2\} \times \{f_1, f_2\}$, les probabilités d'inclusion sont calculées en utilisant $\pi_i = nx_i/(\sum_{j=1} x_j),\ i = 1,\ldots,N$ donnant un plan proportionnel à la taille. Comme $x_i > 0\ \forall i$, les probabilités d'inclusion sont strictement positives. Les distributions des quatre populations engendrées sont décrites à la figure 5.3.

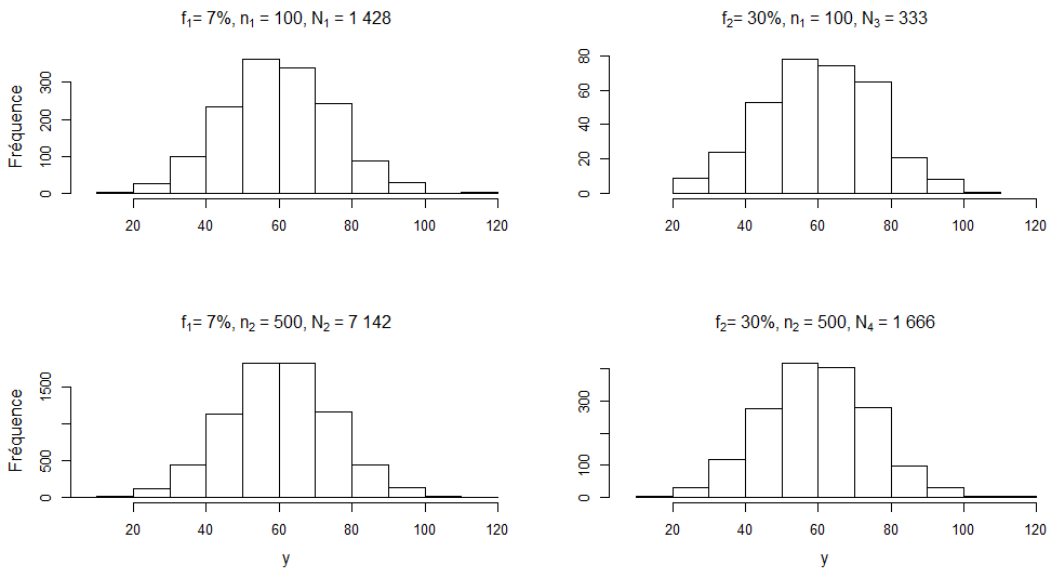


Fig. 5.3. Histogrammes de fréquence des quatre populations correspondant aux quatre scénarios d'enquête et formées à partir d'un échantillon i.i.d. de taille $\widetilde{N} = 7\ 142$ issu du modèle de régression (5.1.1).

5.2. Mesures de performance

5.2.1. Simulation basée sur le plan

La plupart des mesures de performance qui suivront sont définies à partir de l'erreur quadratique moyenne sous le plan d'un estimateur $\hat{\theta}$, qui est soit l'estimateur de la médiane, $\hat{\xi}_{0,50}$, ou l'estimateur du troisième quartile, $\hat{\xi}_{0,75}$. Une approximation Monte Carlo de celle-ci est calculée en simulant 3 000 échantillons à partir d'une population U fixée selon le plan de sondage examiné puis en évaluant

$$\mathrm{EQM}_{\mathrm{MC}} \equiv \mathrm{EQM}_{\mathrm{MC}}(\hat{\theta}) = \frac{1}{3000}\sum_{i=1}^{3000}(\hat{\theta}_i - \theta)^2, \tag{5.2.1}$$

où $\hat{\theta}_i = \hat{\theta}(s_i)$ est l'estimation associée au i-ème échantillon et θ est le paramètre dans la population finie. La simulation consiste par la suite à générer $S = 2\ 000$ échantillons selon le plan EASSR ou le plan de Poisson et à appliquer à chacun d'eux les méthodes étudiées. Il est à noter que l'approximation Monte Carlo (5.2.1) a été calculée avec plus de précision puisque le coût de ce calcul est bien moindre que celui représenté par les méthodes de

rééchantillonnage bootstrap.

Les différents estimateurs d'erreur quadratique moyenne de $\hat{\theta}$, dénotés $\hat{V} = \hat{V}(\hat{\theta})$, peuvent être comparés entre eux au moyen du biais relatif et de la racine carrée de l'erreur quadratique moyenne relative, définis comme

$$\text{biais rel.} = \frac{1}{S}\sum_{s=1}^{S}\frac{\hat{V}_s - \text{EQM}_{\text{MC}}}{\text{EQM}_{\text{MC}}} \tag{5.2.2}$$

et

$$\text{RREQM} = \frac{1}{\text{EQM}_{\text{MC}}}\sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{V}_s - \text{EQM}_{\text{MC}}\right)^2} \tag{5.2.3}$$

respectivement.

En dénotant un intervalle de confiance pour θ comme $[\theta_{\text{Inf}}, \theta_{\text{Sup}}]$, nous introduisons les taux d'erreur de couverture inférieure, supérieure et bilatérale, s'écrivant

$$L = \frac{1}{S}\sum_{i=1}^{S}\mathbb{1}(\theta < \theta_{\text{Inf},s}), \quad U = \frac{1}{S}\sum_{i=1}^{S}\mathbb{1}(\theta > \theta_{\text{Sup},s}), \quad T = L + U. \tag{5.2.4}$$

Enfin, la longueur relative d'un intervalle de confiance est définie comme

$$\text{longueur} = \frac{1}{S}\sum_{s=1}^{S}\frac{\theta_{\text{Sup},s} - \theta_{\text{Inf},s}}{\text{longueur}_{\text{Woodruff}}}, \tag{5.2.5}$$

où $\text{longueur}_{\text{Woodruff}}$ est la longueur moyenne d'un intervalle de confiance de Woodruff calculée à partir des S échantillons. Les méthodes de rééchantilonnage ont été mises en œuvre avec $B = 1\ 000$ échantillons bootstrap. Lors de la sélection du paramètre de lissage par double bootstrap, $D = 50$ réplicats ont été utilisés au deuxième niveau de rééchantillonnage. Le taux d'erreur de couverture total nominal pour les intervalles de confiance a été fixé à $\alpha = 5\%$. Ce faisant, avec $S = 2\ 000$ échantillons, les taux d'erreur de couverture inférieure et supérieure expérimentaux se situeront entre 1,82% et 3,18% 95 fois sur 100 si la vraie probabilité de couverture est de 2,5%. Une région d'acceptation de niveau 95% pour un taux nominal d'erreur de couverture bilatérale de 5% est donnée par $[4{,}04; 5{,}96]\%$.

5.2.2. Simulation basée sur le modèle et sur le plan

Afin de répondre à des questions relatives au paramètre de lissage optimal, une partie de l'analyse fera appel au cadre où l'on considère une infinité de populations par opposition à

une seule fixée. À cette fin, on définit une mesure de stabilité de l'estimateur de variabilité d'un estimateur $\hat{\theta}$ adaptée au cadre de simulation où les S échantillons d'enquête proviennent de S populations finies distinctes. En premier lieu, on doit approximer l'erreur quadratique moyenne sous le modèle et sous le plan de $\hat{\theta}$, dénotée $\mathbb{E}_0\mathbb{E}_p\left[(\hat{\theta}-\theta)^2\right]$. L'erreur quadratique sous le modèle et sous le plan de $\hat{\theta}$ entre dans la composition de la *variance anticipée* de $\hat{\theta}$ (Särndal et al., 1992). Cette appellation due à Isaki et Fuller (1982) désigne

$$AV(\hat{\theta}) = \mathbb{E}_0\mathbb{E}_p\left[(\hat{\theta}-\theta)^2\right] - \left(\mathbb{E}_0\mathbb{E}_p[\hat{\theta}-\theta]\right)^2. \tag{5.2.6}$$

Ainsi, si $\mathbb{E}_0\mathbb{E}_p[\hat{\theta}-\theta]=0$, la variance anticipée et l'erreur quadratique moyenne sous le plan et sous le modèle coïncident. Afin d'approximer $\mathbb{E}_0\mathbb{E}_p\left[(\hat{\theta}-\theta)^2\right]$, 3 000 populations finies sont générées à partir de la distribution F_0, puis un échantillon est tiré à partir de chacune selon un plan de sondage donné. L'approximation est obtenue en évaluant

$$\mathrm{EQM}_{0,\mathrm{MC}} \equiv \mathrm{EQM}_{0,\mathrm{MC}}(\hat{\theta}) = \frac{1}{3000}\sum_{i=1}^{3000}(\hat{\theta}_i-\theta_i)^2, \tag{5.2.7}$$

où θ_i est le paramètre dans la $i^{\text{ème}}$ population finie et $\hat{\theta}_i = \hat{\theta}(s_i)$ est l'estimation basée sur le plan de θ_i associée au $i^{\text{ème}}$ échantillon, $i = 1,\ldots,3\ 000$. La simulation se poursuit en générant $S = 2\ 000$ populations finies à partir de F_0, puis en générant un échantillon à partir de chacune conformément au plan. La mesure d'instabilité d'un estimateur $\hat{V}$ sous le modèle et sous le plan, qui présente la même forme qu'en (5.2.3), est donnée par

$$\mathrm{RREQM}_0 = \frac{1}{\mathrm{EQM}_{0,\mathrm{MC}}}\sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{V}_s - \mathrm{EQM}_{0,\mathrm{MC}}\right)^2}, \tag{5.2.8}$$

où $\hat{V}_s$ est l'estimateur d'erreur quadratique moyenne calculé à partir de l'échantillon s, $s = 1,\ldots,S$.

5.3. Méthodes évaluées

Les particularités des méthodes évaluées par simulation de même que leur dénomination sont énumérées ci-dessous. La plupart d'entre elles sont communes aux deux plans de sondage, à l'exception de la méthode de sélection du paramètre de lissage par injection, qui n'est évaluée que pour le plan EASSR dans cette étude. De plus, la méthode de rééchantillonnage par pseudo-population standard (sans lissage) est dénommée selon l'auteur de l'algorithme,

qui diffère selon le plan de sondage. Par ailleurs, notons que l'on peut faire correspondre à chaque méthode étudiée une valeur d'un paramètre de lissage ou d'ajustement.

- **Woodruff**: Cette dénomination fait référence à la méthode de Woodruff (1952) pour la formation d'intervalles de confiance et l'estimation de la variance pour des quantiles. L'intervalle de confiance de Woodruff pour ξ_p est unique à un échantillon donné, ce qui n'est pas le cas de l'estimateur de variance de Woodruff (voir la sous-section 1.3.4). En effet, celui-ci est défini à partir de la demi-longueur d'un intervalle de confiance de Woodruff de niveau $(1-\beta)$, ce qui fait de β un paramètre d'ajustement. À chaque valeur de $\beta \in (0,1)$ correspond donc un estimateur de variance de Woodruff différent, dénoté $\hat{V}_\beta$. Les mesures de performance pour cette méthode seront présentées pour $\beta \in \{0{,}01; 0{,}025; 0{,}05; 0{,}1; 0{,}2\}$, qui correspond à l'ensemble de valeurs rapportées par Sitter (1992).
- **Booth et al.** (EASSR) ou **Chauvet** (Poisson) : Ces dénominations désignent l'estimateur non lisse $\hat{V}^*$ donné par (2.2.4), obtenu via l'algorithme 2.2.1 proposé par Booth et al. (1994) pour le plan EASSR ou à partir de l'algorithme 2.2.2 introduit par Chauvet (2007) pour l'échantillonnage de Poisson.
- **Lisse plug-in norm** (EASSR): La sélection du paramètre de lissage par injection (*plug-in*) est décrite à la section 4.2. L'algorithme 4.1.1 est exécuté avec un paramètre de lissage égal à $\hat{h}_{\text{plug-in}}$, dont l'expression est explicitée en (4.2.6). L'estimateur de l'erreur quadratique moyenne de $\hat{\xi}_p$ correspondant est dénoté $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$.
- **Lisse boot var**: Tel que décrit à la sous-section 4.3.3, une grille $\mathcal{H}(n) = \{h_1, h_2, \ldots, h_K\}$ est parcourue afin de sélectionner la valeur de h minimisant une estimation bootstrap du critère de l'erreur quadratique moyenne, donné par (4.3.1). La valeur sélectionnée pour le paramètre de lissage est $\hat{h}_{\text{boot,var}}$, donnée par (4.3.5), et l'estimateur de l'erreur quadratique moyenne résultant est dénoté $\hat{V}^*_{\hat{h}_{\text{boot,var}}}$.
- **Lisse boot IC**: De manière à optimiser le taux de couverture des intervalles de confiance, le double bootstrap est utilisé pour identifier la valeur de $h \in \mathcal{H}(n)$ minimisant une estimation bootstrap du critère L, donné par (4.3.6), avec $\tau = 2/3$. La valeur sélectionnée pour le paramètre de lissage est $\hat{h}_{\text{boot,ic}}$, donnée par (4.3.12), et l'estimateur de l'erreur quadratique moyenne résultant est dénoté $\hat{V}^*_{\hat{h}_{\text{boot,ic}}}$.

Les tableaux 5.1 et 5.2 renferment les grilles d'exploration $\mathcal{H}(n)$ intervenant dans les méthodes *Lisse boot var* et *Lisse boot IC* pour l'échantillonnage aléatoire simple sans remise et l'échantillonnage de Poisson respectivement. Chaque entrée correspond à un scénario, soit le croisement entre l'un des deux estimateurs de quantile et une superpopulation F_0. La colonne $C_{\text{opt}}(\tilde{\xi}_p)$ désigne la valeur de la constante optimale théorique pour l'estimation de la variance d'un quantile échantillonnal pour la distribution F_0, à savoir la constante multipliant $n^{-1/5}$ dans l'expression du paramètre de lissage optimal donnée par (4.2.1). L'expression de cette constante est explicitée à nouveau comme suit

$$C_{\text{opt}}(\tilde{\xi}_p) = \left[f_0(\tilde{\xi}_p)\right]^{1/5} \left[2\sqrt{\pi}\right]^{-1/5} \left[f_0''(\tilde{\xi}_p) - \left(f_0'(\tilde{\xi}_p)\right)^2 f_0(\tilde{\xi}_p)^{-1}\right]^{-2/5}. \tag{5.3.1}$$

On distingue cette constante de $C_{\text{norm}}(\tilde{\xi}_p)$ figurant en deuxième colonne, que l'on définit comme le facteur multipliant $n^{-1/5}$ dans l'expression de la fenêtre optimale basée sur la distribution $\mathcal{N}(\mu, \sigma^2)$ donnée par (4.2.5), à savoir

$$C_{\text{norm}}(\tilde{\xi}_p) = \left[\frac{1}{\sigma}\phi\left(z\right)\right]^{1/5} \left[2\sqrt{\pi}\right]^{-1/5} \left[\frac{1}{\sigma}\phi''\left(z\right) - \left(\frac{1}{\sigma}\phi'\left(z\right)\right)^2 \left(\frac{1}{\sigma}\phi\left(z\right)\right)^{-1}\right]^{-2/5}, \tag{5.3.2}$$

où $z \equiv z(\tilde{\xi}_p, \mu, \sigma^2) = (\tilde{\xi}_p - \mu)/\sigma$ et $\tilde{\xi}_p$, μ et σ sont respectivement le quantile de niveau p, le premier moment et l'écart type de la superpopulation F_0. Rappelons que l'expression de $\hat{h}_{\text{plug-in}}$ jouant un rôle dans la méthode *lisse plug-in* est basée sur une estimation de (5.3.2). En dernière colonne des tableaux 5.1 et 5.2, on retrouve les grilles $\mathcal{H}(n)$, qui constituent des vecteurs de constantes fixes multipliant la quantité $n^{-1/5}$. Elles permettent donc d'accommoder diverses tailles échantillonnales. Notons que l'étendue de valeurs d'une grille $\mathcal{H}(n)$ est déterminée de façon à couvrir (5.3.1) ou encore (5.3.2) dans les cas où (5.3.1) n'existe pas.

Scénario	$C_{\text{opt}}(\tilde{\xi}_p)$	$C_{\text{norm}}(\tilde{\xi}_p)$	$\mathcal{H}(n)$
$\hat{\xi}_{0,50}$, $\mathcal{N}(0,1)$	0,93	Même	$\{h_k \mid h_k = [0{,}01 + (k-1)\,0{,}05] \cdot n^{-1/5},\ 1 \leq k \leq 50,\ k \in \mathbb{N}\}$
$\hat{\xi}_{0,75}$, $\mathcal{N}(0,1)$	0,98	Même	$\{h_k \mid h_k = [0{,}01 + (k-1)\,0{,}05] \cdot n^{-1/5},\ 1 \leq k \leq 50,\ k \in \mathbb{N}\}$
$\hat{\xi}_{0,50}$, Lognormale(0,1)	∞	2,03	$\{h_k \mid h_k = [0{,}01 + (k-1)\,0{,}05] \cdot n^{-1/5},\ 1 \leq k \leq 50,\ k \in \mathbb{N}\}$
$\hat{\xi}_{0,75}$, Lognormale(0,1)	2,24	2,02	$\{h_k \mid h_k = [1{,}00 + (k-1)\,0{,}05] \cdot n^{-1/5},\ 1 \leq k \leq 51, k \in \mathbb{N}\}$

Tableau 5.1. Grilles d'exploration pour les différents scénarios étudiés dans le plan EASSR.

Scénario	$C_{\text{opt}}(\tilde{\xi}_p)$	$C_{\text{norm}}(\tilde{\xi}_p)$	$\mathcal{H}(n)$
$\hat{\xi}_{0.50}$	13,71	Même	$\{h_k \mid h_k = [0{,}100000 + (k-1)\,0{,}557551] \cdot n^{-1/5},\ 1 \leq k \leq 50, k \in \mathbb{N}\}$
$\hat{\xi}_{0.75}$	14,35	Même	$\{h_k \mid h_k = [0{,}1000000 + (k-1)\,0{,}5836735] \cdot n^{-1/5},\ 1 \leq k \leq 50, k \in \mathbb{N}\}$

Tableau 5.2. Grilles d'exploration pour les différents scénarios étudiés dans l'échantillonnage de Poisson.

À la lecture du tableau 5.1, on note que dans le cas de la superpopulation Lognormale(0,1) et de la médiane, la valeur de $C_{\text{opt}}(\tilde{\xi}_p)$ n'existe pas tel que spécifié par la remarque 3.3.3. En ce qui concerne le tableau 5.2, il convient de préciser que les constantes optimales théoriques pour le modèle donné par (5.1.1) ont été calculées en utilisant l'approximation normale pour la loi du χ^2. Le fait que le régresseur dans le modèle (5.1.1) soit distribué selon la loi χ^2_{100} indépendamment de l'erreur $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{\tilde{N}})'$ permet de déduire que les variables aléatoires indépendantes $Y_1, Y_2, \ldots, Y_{\tilde{N}}$ de la population sont approximativement distribuées selon la loi $\mathcal{N}(\gamma\nu, 2\nu\gamma^2 + \sigma^2)$. Dans ces cas, la constante a donc été évaluée en prenant la densité de la loi normale avec moyenne $\mu = \gamma\nu$ et écart type $\tau = \sqrt{2\nu\gamma^2 + \sigma^2}$, où $\gamma = 0{,}6$, $\nu = 100$ et $\sigma = 12$.

Si l'on souhaitait appliquer les méthodes *Lisse boot var* et *Lisse boot IC* à un jeu de données particulier en dehors de l'environnement contrôlé d'une simulation, la remarque 4.3.2 décrit une approche ne nécessitant pas de connaître F_0 pour former la grille de valeurs à explorer pour la sélection du paramètre de lissage.

5.4. Résultats pour l'échantillonnage aléatoire simple sans remise

Dans cette section, la performance des méthodes énumérées dans la section 5.3 est étudiée en tirant les S échantillons selon le plan EASSR. Les résultats pour l'estimation de l'erreur quadratique moyenne et les intervalles de confiance sont déclinés selon les deux statistiques d'intérêt, $\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$, les différents scénarios d'enquête et les deux superpopulations génératrices utilisées dans ce plan de sondage.

5.4.1. Estimation de l'erreur quadratique moyenne

Les tableaux 5.3 et 5.4 exhibent les mesures d'adéquation des estimateurs d'erreur quadratique moyenne pour une réalisation de la superpopulation $F_0 = \mathcal{N}(0,1)$ et une réalisation de la superpopulation $F_0 = \text{Lognormale}(0,1)$ respectivement. Dans chacun des tableaux, les deux branches verticales principales du tableau correspondent à la médiane et au troisième quartile. À l'intérieur de chacune, on retrouve les quatre cellules formées par les scénarios d'enquête $\{n_1, n_2\} \times \{f_1, f_2\}$.

En se focalisant d'abord sur la superpopulation $F_0 \equiv \mathcal{N}(0,1)$, on observe que les méthodes basées sur le lissage de la pseudo-population introduisent un biais d'estimation de $\text{EQM}_p[\hat{\xi}_{0,50}]$ par rapport à l'algorithme de Booth et al. (1994). La magnitude du biais est par ailleurs plus grande pour la plus petite taille échantillonnale ($n = 100$). Cela est compensé par une variance généralement inférieure des estimateurs bootstrap $\hat{V}_h^*(\hat{\xi}_{0,50})$ en comparaison à $\hat{V}^*(\hat{\xi}_{0,50})$, conférant une plus grande stabilité aux estimateurs lisses. En effet, dans tous les scénarios d'enquête, la valeur de RREQM est inférieure à celle de l'estimateur non lisse pour les méthodes de sélection de h basées sur le critère de l'erreur quadratique moyenne de l'estimateur de variance (méthodes *lisse plug-in norm* et *lisse boot var*). Cela dit, une plus grande stabilité est attribuable à *lisse plug-in norm*. Il en va de même pour la statistique $\hat{\xi}_{0,75}$, si ce n'est que le biais des méthodes *lisse plug-in norm* et *lisse boot var* est également moins grand que celui de l'estimateur non lisse pour $f_1 = 7\%$. Pour les deux quantiles, les conclusions au regard du critère RREQM sont mitigées en ce qui concerne la méthode de sélection *lisse boot IC*, qui est basée sur un critère de distance entre les taux de couverture nominaux et expérimentaux. L'instabilité de l'estimateur dérivé de cette approche est près de celle de Booth et al. (1994). Enfin, dans tous les scénarios d'enquête, il est remarquable que la méthode de Woodruff (1952) surpasse la méthode proposée par Booth et al. (1994) au regard de la stabilité de l'estimateur pour une grande étendue de valeurs de β. L'étude par simulation conduite par Sitter (1992) menait à un constat similaire lorsque la méthode de Woodruff (1952) était comparée aux méthodes de rééchantillonnage standards. En revanche, l'estimateur $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$ introduit dans ce mémoire constitue souvent une amélioration très importante par rapport à la meilleure valeur β possible de l'estimateur de variance de Woodruff parmi les valeurs présentées. Même pour

la médiane sous le scénario $n_1 = 100, f_2 = 30\%$, Woodruff a soit une instabilité à peine plus petite, soit identique ou encore supérieure (pour les valeurs plus élevées de β). La méthode *lisse boot var*, ayant une stabilité comparable à la méthode *lisse plug-in* surpasse également la méthode de Woodruff. Comme pour *lisse plug-in*, Woodruff fait légèrement mieux pour quelques valeurs de β dans l'instance de la médiane avec les modalités $n_1 = 100, f_2 = 30$, mais aussitôt que $n = 500$, l'instabilité de *lisse boot var* devient bien moindre que le meilleur estimateur $\hat{V}_\beta$ présenté.

Tableau 5.3. Mesures de performance des estimateurs d'erreur quadratique moyenne pour l'échantillonnage aléatoire simple sans remise et une superpopulation $\mathcal{N}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$\hat{\xi}_{0,50}$				$\hat{\xi}_{0,75}$			
		$f_1 = 7\ \%$		$f_2 = 30\ \%$		$f_1 = 7\ \%$		$f_2 = 30\ \%$	
n	Méthode	biais rel.	RREQM	biais rel.	RREQM	biais rel.	RREQM	biais rel.	RREQM
	Woodruff								
	$\beta = 0{,}01$	0,10	0,39	0,10	0,40	0,14	0,41	0,22	0,51
100	$\beta = 0{,}025$	0,09	0,42	0,08	0,42	0,14	0,46	0,20	0,53
	$\beta = 0{,}05$	0,08	0,45	0,06	0,43	0,15	0,51	0,18	0,55
	$\beta = 0{,}1$	0,08	0,49	0,06	0,47	0,16	0,58	0,15	0,59
	$\beta = 0{,}2$	0,09	0,56	0,06	0,55	0,17	0,68	0,12	0,65
	Booth et al.	0,17	0,56	0,15	0,52	0,25	0,68	0,24	0,65
	Lisse plug-in norm	0,17	0,31	0,34	0,43	0,19	0,33	0,24	0,34
	Lisse boot var	0,26	0,39	0,37	0,51	0,25	0,41	0,29	0,39
	Lisse boot IC	0,24	0,55	0,24	0,56	0,32	0,67	0,34	0,67
	Woodruff								
	$\beta = 0{,}01$	-0,00	0,25	-0,05	0,25	-0,05	0,26	0,08	0,30
500	$\beta = 0{,}025$	-0,00	0,27	-0,05	0,28	-0,05	0,28	0,06	0,30
	$\beta = 0{,}05$	-0,00	0,29	-0,04	0,31	-0,04	0,30	0,05	0,32
	$\beta = 0{,}1$	0,01	0,32	-0,04	0,32	-0,04	0,33	0,04	0,35
	$\beta = 0{,}2$	0,02	0,37	-0,03	0,35	-0,03	0,38	0,03	0,38
	Booth et al.	0,04	0,32	0,01	0,32	0,01	0,34	0,09	0,35
	Lisse plug-in norm	0,07	0,15	0,04	0,12	-0,00	0,13	0,13	0,19
	Lisse boot var	0,07	0,18	0,02	0,15	-0,00	0,16	0,13	0,20
	Lisse boot IC	0,09	0,31	0,06	0,30	0,05	0,31	0,16	0,36

D'autres conclusions peuvent être tirées pour la superpopulation Lognormale(0,1) à la lumière des résultats du tableau 5.4. En examinant ceux pour $\hat{\xi}_{0,50}$, la valeur minimale du biais en valeur absolue et du critère RREQM est généralement atteinte par l'un des estimateurs $\hat{V}_\beta$. Ceci dit, sauf pour $f_1 = 30\%$, $n_2 = 100$ la valeur du RREQM associée à *lisse boot var* est dans l'étendue couverte par les estimateurs de Woodruff et non loin de la meilleure valeur. De plus, avec les modalités $f_1 = 7\%$, $n_2 = 500$, *Lisse boot var* fait beaucoup mieux au regard de la stabilité. Il convient également de remarquer que sauf dans ce dernier scénario, les méthodes de sélection *lisse plug-in norm* et *lisse boot var* mènent à des estimateurs de variance aussi sinon plus instables que l'algorithme de Booth et al. (1994). Les mesures d'instabilité les plus élevées sont observées pour l'approche *lisse plug-in norm*, qui est, rappelons-le, basée sur la constante optimale pour la loi $\mathcal{N}(0,1)$ (équation 5.3.2). Sur ce point, les résultats associés à la méthode *lisse plug-in norm* sont de meilleur augure pour le troisième quartile. Ceci s'explique par le fait que la constante optimale pour la distribution Lognormale(0,1) évaluée en $\tilde{\xi}_{0,75}$ et la constante optimale basée sur la loi $\mathcal{N}(0,1)$ ont des valeurs numériques assez proches. Tel que vu dans le tableau 5.1, la première vaut 2,24, tandis que la deuxième, obtenue en évaluant (5.3.2) à partir du troisième quartile, vaut 2,02. La méthode *lisse boot var* mène à des résultats comparables et bien que plus instable, la méthode *lisse boot ic* fait quand même mieux que l'homologue non lisse dans toutes les instances. Au regard du RREQM, la méthode de Woodruff est surclassée en ce qui concerne le troisième quartile.

Tableau 5.4. Mesures de performance des estimateurs d'erreur quadratique moyenne pour l'échantillonnage aléatoire simple sans remise et une superpopulation Lognormale(0,1) ($S =$ 2 000, $B = 1$ 000 et $D = 50$).

		$\hat{\xi}_{0,50}$				$\hat{\xi}_{0,75}$			
		$f_1 = 7$ %		$f_2 = 30$ %		$f_1 = 7$ %		$f_2 = 30$ %	
n	Méthode	biais rel.	RREQM	biais rel.	RREQM	biais rel.	RREQM	biais rel.	RREQM
	Woodruff								
	$\beta = 0{,}01$	0,22	0,62	0,38	0,78	0,45	0,91	0,48	0,87
100	$\beta = 0{,}025$	0,18	0,60	0,29	0,72	0,36	0,84	0,40	0,85
	$\beta = 0{,}05$	0,16	0,61	0,22	0,68	0,30	0,81	0,33	0,84
	$\beta = 0{,}1$	0,15	0,63	0,17	0,68	0,25	0,80	0,27	0,87
	$\beta = 0{,}2$	0,14	0,69	0,12	0,68	0,21	0,85	0,21	0,90
	Booth et al.	0,24	0,70	0,28	0,75	0,32	0,86	0,34	0,89
	Lisse plug-in norm	1,14	1,46	1,61	1,74	0,18	0,47	0,18	0,44
	Lisse boot var	0,43	0,67	0,80	1,05	0,32	0,54	0,29	0,47
	Lisse boot IC	0,33	0,73	0,44	0,89	0,30	0,60	0,32	0,61
	Woodruff								
	$\beta = 0{,}01$	-0,05	0,24	0,21	0,38	0,06	0,34	0,13	0,37
500	$\beta = 0{,}025$	-0,04	0,26	0,18	0,38	0,05	0,36	0,10	0,36
	$\beta = 0{,}05$	-0,03	0,28	0,16	0,39	0,04	0,37	0,08	0,36
	$\beta = 0{,}1$	-0,02	0,31	0,15	0,41	0,04	0,40	0,07	0,38
	$\beta = 0{,}2$	0,00	0,37	0,12	0,44	0,03	0,43	0,06	0,41
	Booth et al.	0,03	0,34	0,19	0,42	0,08	0,41	0,12	0,40
	Lisse plug-in norm	0,21	0,27	0,96	1,00	-0,01	0,19	0,25	0,31
	Lisse boot var	-0,04	0,17	0,33	0,43	0,01	0,18	0,26	0,33
	Lisse boot IC	0,05	0,30	0,31	0,50	0,04	0,24	0,28	0,36

L'échec de la méthode *lisse plug-in norm* a été constaté dans le cas de la médiane pour cette réalisation de la loi Lognormale(0,1). Tel que suggéré précédemment, la plus grande instabilité associée à la sélection par injection peut être attribuée au fait que l'on estime une constante basée sur la loi $\mathcal{N}(\mu, \sigma^2)$, donnée par (5.3.2), au lieu d'une constante basée sur la loi Lognormale(μ,σ^2). En revanche, un second facteur pouvant influencer les résultats est l'étendue de valeurs de h pour lesquelles une amélioration est possible au regard de l'instabilité de l'estimateur $\hat{V}_h^*$ donné par (4.1.1). Le succès des méthodes de sélection automatiques

proposées repose donc entre autres sur le potentiel d'amélioration pré-existant que recèle une population finie. La sous-section suivante servira d'appui à la présente discussion en illustrant graphiquement ce potentiel d'amélioration pour les différents cas de figure.

5.4.1.1. *Représentations graphiques*

Le gain pourvu par l'algorithme par pseudo-population lisse par rapport à l'algorithme standard peut être apprécié au moyen de représentations graphiques de l'instabilité. En traçant la courbe du RREQM de l'estimateur $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$, en fonction de $C > 0$, il est également possible d'évaluer la magnitude de l'amélioration potentielle ainsi que la valeur de h maximisant le gain.

Les figures 5.4 et 5.5 représentent ces graphiques pour $\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$ sous le scénario $\mathcal{N}(0,1)$. Le trait plein en noir, la courbe pleine en bleu et le trait pointillé en rouge représentent l'approximation numérique de l'instabilité de $\hat{V}^*$, de $\hat{V}_h^*$ et de $\hat{V}_{\hat{h}_{\text{plug-in}}}^*$ respectivement. Dans chaque figure, les graphiques sont déclinés en fonction des quatre scénarios d'enquête. On définit le potentiel d'amélioration comme la distance maximale entre la courbe en bleu et le trait en noir. Par exemple, dans le cas de la médiane et d'une fraction de sondage de 7%, la réduction maximale associée au lissage est 44,8% pour $n_1 = 100$ et d'au plus 54,4% pour $n_2 = 500$. De façon générale, le gain potentiel augmente tandis que la taille échantillonnale croît. À la vue de ces graphiques, on déduit que la méthode de sélection par injection du paramètre de lissage se comporte tel que souhaité lorsque les données sont normalement distribuées. Dans tous les cas, le critère RREQM mesuré pour $\hat{V}_{\hat{h}_{\text{plug-in}}}^*$ est essentiellement identique à la valeur mesurée pour $h = C_{\text{opt}}(\tilde{\xi}_p)n^{-1/5},\ p = 0{,}50; 0{,}75$.

En revanche, le minimum de la courbe en bleu n'est pas toujours atteint en $C_{\text{opt}}(\tilde{\xi}_p)$. La différence est manifeste dans le cas de la médiane et de la population finie de taille $N = 333$ (scénario $n_1 = 100$, $f_1 = 30\%$), qui correspond à la plus petite d'entre les quatre. Afin d'éliminer l'effet de considérer une seule population finie U par scénario d'enquête, nous procédons dans un deuxième temps à l'évaluation de RREQM_0 donnée par (5.2.8), à savoir la mesure d'instabilité de $\hat{V}_h^*$ par rapport au modèle et au plan conjointement. Rappelons que cela consiste en une modification du processus de génération des

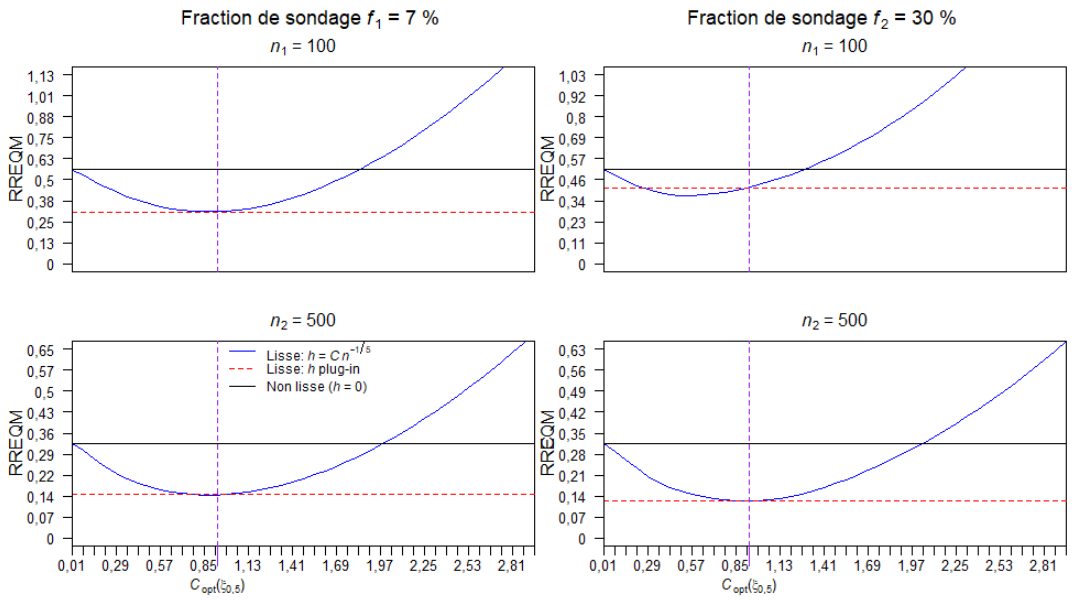


Fig. 5.4. RREQM d'estimateurs bootstrap de $\mathrm{EQM}_p(\hat{\xi}_{0,50})$ en fonction de $C > 0$ sous le scénario $F_0 = \mathcal{N}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$). Trois estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$), $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$, et $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$.

$S = 2\ 000$ échantillons, qui sont alors issus de S réalisations i.i.d. de taille N de $F_0 \equiv \mathcal{N}(0,1)$.

Les figures 5.6 et 5.7 illustrent l'approximation numérique de RREQM_0 de l'estimateur $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$, en fonction de plusieurs valeurs de $C > 0$ (courbe en bleu), ainsi que la même mesure pour l'estimateur $\hat{V}^*$ (trait en noir) pour les quantiles $\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$ respectivement. On constate que les différences entre les fractions de sondage $f_1 = 7\%$ et $f_2 = 30\%$ observées dans les figures 5.4 et 5.5 sont devenues presqu'imperceptibles. Qui plus est, en dépit du bruit aléatoire dû au fait qu'un nombre fini d'échantillons sont considérés, le minimum de la courbe en bleu est essentiellement atteint en $C_{\text{opt}}(\tilde{\xi}_p)$ pour les deux tailles échantillonnales, $p = 0{,}50; 0{,}75$. Ainsi, en considérant une infinité de populations finies, ces graphiques suggèrent que le comportement de l'estimateur $\hat{V}_h^*$ reproduit celui de l'estimateur bootstrap lisse de la variance d'un quantile échantillonnal considéré par Hall et al. (1989) dans un cadre classique.

Les figures 5.8 et 5.9 représentent l'approximation numérique de la mesure d'instabilité *sous le plan seulement* de $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$, en fonction de différentes valeurs de $C > 0$ pour

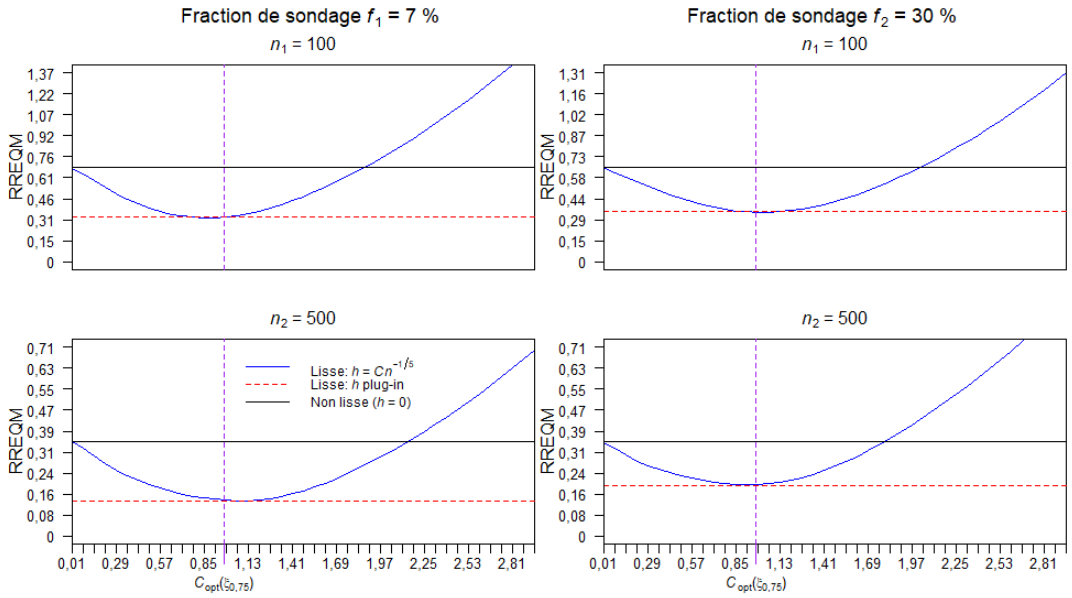


Fig. 5.5. RREQM d'estimateurs bootstrap de $\mathrm{EQM}_p(\hat{\xi}_{0,75})$ en fonction de $C > 0$ sous le scénario $F_0 = \mathcal{N}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$). Trois estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$), $\hat{V}^*_h$, où $h = C \cdot n^{-1/5}$, et $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$.

$\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$ avec la superpopulation $F_0 \equiv$ Lognormale(0,1). La courbe en bleu correspond à $\hat{V}^*_h$, tandis que les traits en noir et en rouge renvoient à $\hat{V}^*$ et $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$ respectivement. Il est d'abord remarquable que les graphiques pour la médiane se démarquent particulièrement de ce qui est observé pour la réalisation de $F_0 \equiv \mathcal{N}(0,1)$ au regard du potentiel d'amélioration par le lissage et ce particulièrement pour la fraction de sondage $f_2 = 30\%$. Pour cette réalisation de la distribution Lognormale(0,1), on constate que l'étendue de valeurs de C résultant en une amélioration est assez restreinte avec la modalité $f_2 = 30\%$. Quant au troisième quartile, une diminution importante du potentiel d'amélioration survient pour la fraction de sondage $f_2 = 30\%$ lorsque l'on passe de la population correspondant à $n_1 = 100$ à la population correspondant à $n_2 = 500$.

Tel que rapporté dans le tableau 5.1, la constante $C_{\text{opt}}(\tilde{\xi}_{0,50})$ pour la superpopulation Lognormale(0,1) n'existe pas et ne figure donc pas dans les graphiques de la figure 5.8. La quantité $C_{\text{norm}}(\tilde{\xi}_{0,50})$ (équation 5.3.2), figurant en abscisse a une valeur numérique de 2,03. Pour $n_1 = 100$, nous voyons que le critère RREQM associé à $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$ est assez éloigné de la valeur attendue pour $C_{\text{norm}}(\tilde{\xi}_{0,50})$. Pour une plus grande taille échantillonnale, la méthode

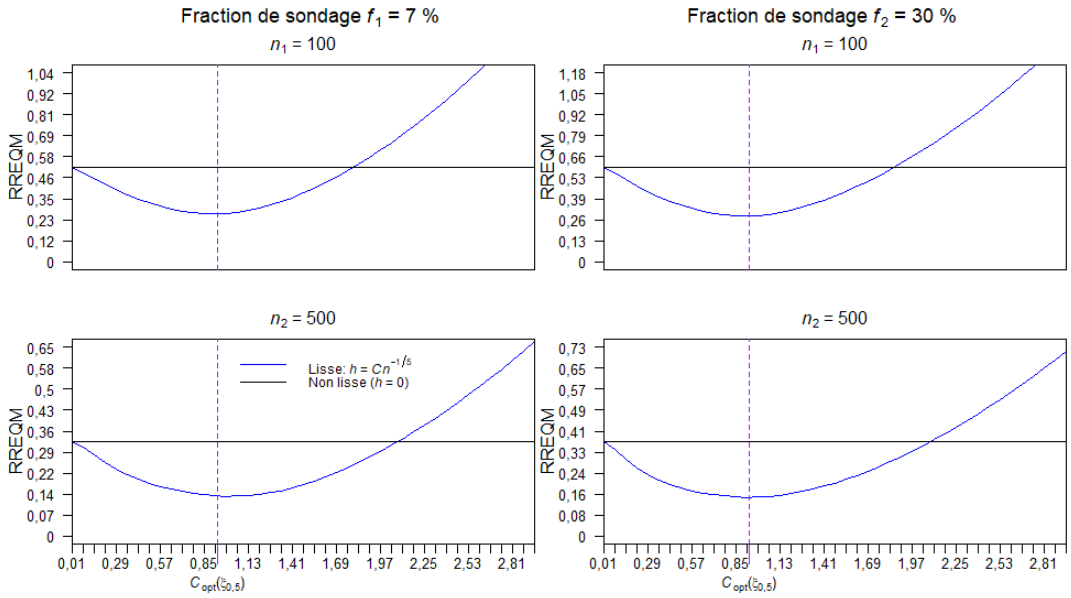


Fig. 5.6. RREQM_0 d'estimateurs bootstrap de $\mathbb{E}_0[\mathrm{EQM}_p(\hat{\xi}_{0,50})]$ en fonction de $C > 0$ sous le scénario $F_0 = \mathcal{N}(0,1)$ ($S = 2\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

atteint la cible, mais ce n'est que pour $f_1 = 7\%$ que le critère RREQM associé à la sélection par injection surpasse marginalement l'algorithme non lisse. Quant au troisième quartile, l'évaluation de $C_{\mathrm{opt}}(\tilde{\xi}_{0,75})$ (présente dans le tableau 5.1) se fait sans problème et celle-ci figure donc sur les graphiques correspondants en plus de $C_{\mathrm{norm}}(\tilde{\xi}_{0,75})$. On constate que les valeurs de $C_{\mathrm{opt}}(\tilde{\xi}_{0,75})$ et de $C_{\mathrm{norm}}(\tilde{\xi}_{0,75})$ sont assez proches, ce qui fait que l'instabilité associée à la méthode *plug-in* se trouve aussi dans la région d'amélioration.

Il y a lieu de voir si ces motifs persistent lorsque la mesure d'instabilité est calculée sur la base du plan et du modèle $F_0 \equiv$ Lognormale(0,1). Les figures 5.10 et 5.11 représentent l'approximation numérique de RREQM_0 de l'estimateur $\hat{V}_h^*$ en fonction de plusieurs valeurs de C (courbe en bleu), où $h = C \cdot n^{-1/5}$, ainsi que la même mesure pour l'estimateur $\hat{V}^*$ (courbe en noir) pour les quantiles $\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$ respectivement. Encore une fois, l'effet présumément dû à la fraction de sondage lorsque l'on considérait une seule réalisation de $F_0 \equiv$ Lognormale(0,1) s'est effacé. Cela suggère qu'il s'agissait de différences dues à l'échantillonnage d'une population finie à partir d'une superpopulation F_0, dont la taille pouvait être aussi petite que $N = 333$.

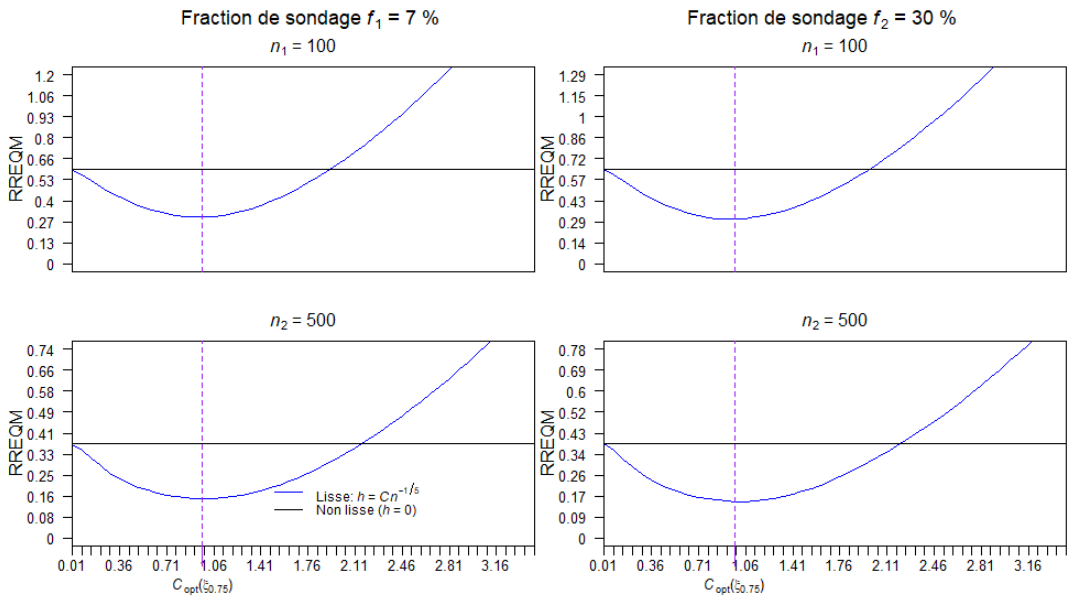


Fig. 5.7. RREQM_0 d'estimateurs bootstrap de $\mathbb{E}_0[\mathrm{EQM}_p(\hat{\xi}_{0,75})]$ en fonction de $C > 0$ sous le scénario $F_0 = \mathcal{N}(0,1)$ ($S = 2\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

À d'autres égards, les courbes de RREQM_0 basées sur le plan et le modèle pour la superpopulation Lognormale(0,1) se distinguent de celles de la superpopulation $\mathcal{N}(0,1)$. Pour cette dernière, tel que vu précédemment, le critère RREQM_0 est à toutes fins pratiques minimisé par la constante théorique. À la lumière de la remarque 3.3.3, il était attendu que la constante C minimisant RREQM_0 varie à mesure que n croît dans le cas de la médiane et de $F_0 \equiv$ Lognormale(0,1), puisque l'ordre du développement pour l'erreur relative de l'estimateur de variance lisse était insuffisant. La puissance de n dans $h = C \cdot n^{-1/5}$ n'était pas appropriée dans ce cas précis, ce que les graphiques corroborent. En revanche, le même phénomène se présente pour le troisième quartile dans la figure 5.11, ce qui n'était pas prévu par la théorie. Bien que la distribution $F_0 \equiv$ Lognormale(0,1) satisfasse les conditions de régularité du théorème 3.3.2 et que la constante théorique $C_{\mathrm{opt}}(\tilde{\xi}_{0,75})$ existe, la constante minimisant la courbe en bleu s'éloigne de $C_{\mathrm{opt}}(\tilde{\xi}_{0,75})$ alors que n croît.

Afin d'étayer l'analyse pour le troisième quartile, nous quittons un instant le cadre de l'échantillonnage pour nous transporter dans le cadre classique. Soit

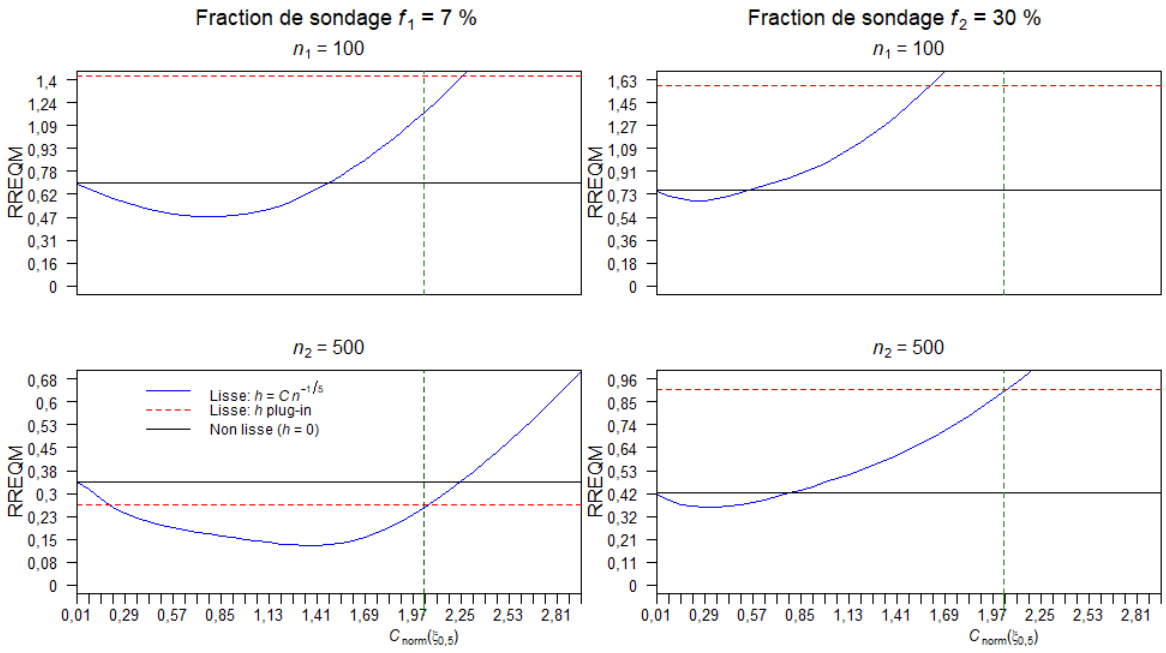


Fig. 5.8. RREQM d'estimateurs bootstrap de $\text{EQM}_p(\hat{\xi}_{0,50})$ en fonction de $C > 0$ sous le scénario $F_0 = \text{Lognormale}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$). Trois estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$), $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$, et $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$.

$Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_0 \equiv \text{Lognormale}(0,1)$. Le troisième quartile d'un tel échantillon est dénoté $\check{\xi}_{0,75}$. Considérons $S = 1\ 000$ échantillons de la forme donnée. Pour chacun des échantillons simulés, l'estimateur $\text{Var}_h^*(\check{\xi}_{0,75})$ sur lequel se concentre le théorème 3.3.2 est obtenu pour chaque valeur de $h = C \cdot n^{-1/5}$, $C > 0$ en suivant l'algorithme 3.2.3, tandis que l'estimateur non lisse, $\text{Var}^*(\check{\xi}_{0,75})$, est obtenu selon l'algorithme du bootstrap non paramétrique classique. De plus, une approximation Monte Carlo de la variance véritable, $\text{Var}_{\text{i.i.d.}}(\check{\xi}_{0,75})$, est calculée à partir de S échantillons, ce qui permet d'évaluer l'instabilité de $\text{Var}_h^*(\check{\xi}_{0,75})$ et de $\text{Var}^*(\check{\xi}_{0,75})$. La figure 5.12 illustre le critère $\text{RREQM}_{\text{i.i.d.}}$ de l'estimateur $\text{Var}_h^*(\check{\xi}_{0,75})$, où $h = C \cdot n^{-1/5}$, pour plusieurs valeurs de $C > 0$ que l'on compare à celui de l'estimateur non lisse $\text{Var}^*(\check{\xi}_{0,75})$. Quatre tailles échantillonnales y sont représentées, soit $n = 50, 100, 500, 1\ 000$.

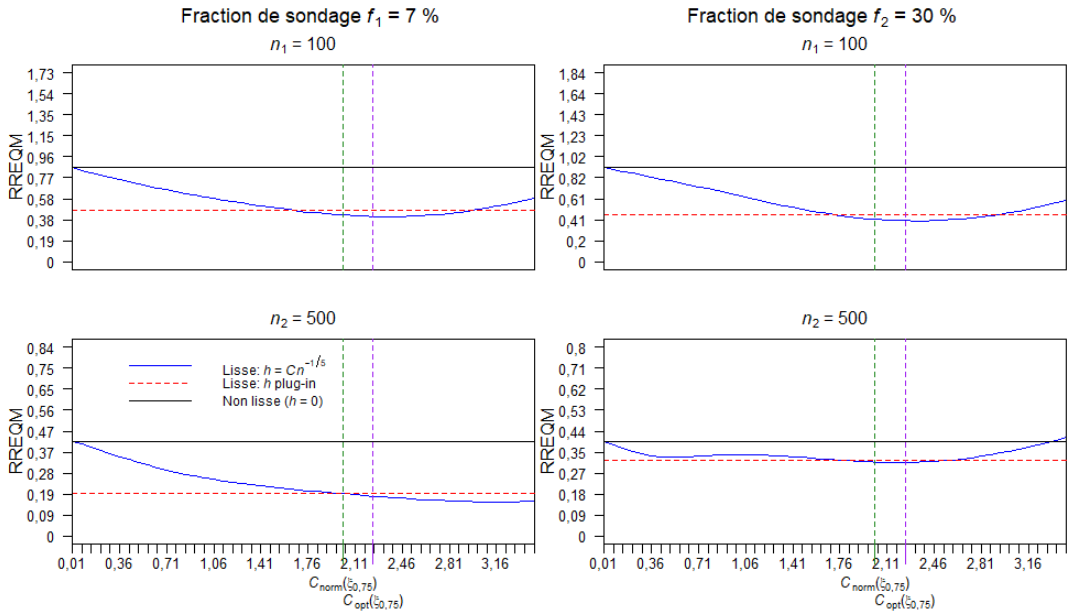


Fig. 5.9. RREQM d'estimateurs bootstrap de $\mathrm{EQM}_p(\hat{\xi}_{0,75})$ en fonction de $C > 0$ sous le scénario $F_0 = \text{Lognormale}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$). Trois estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$), $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$, et $\hat{V}^*_{\hat{h}_{\text{plug-in}}}$.

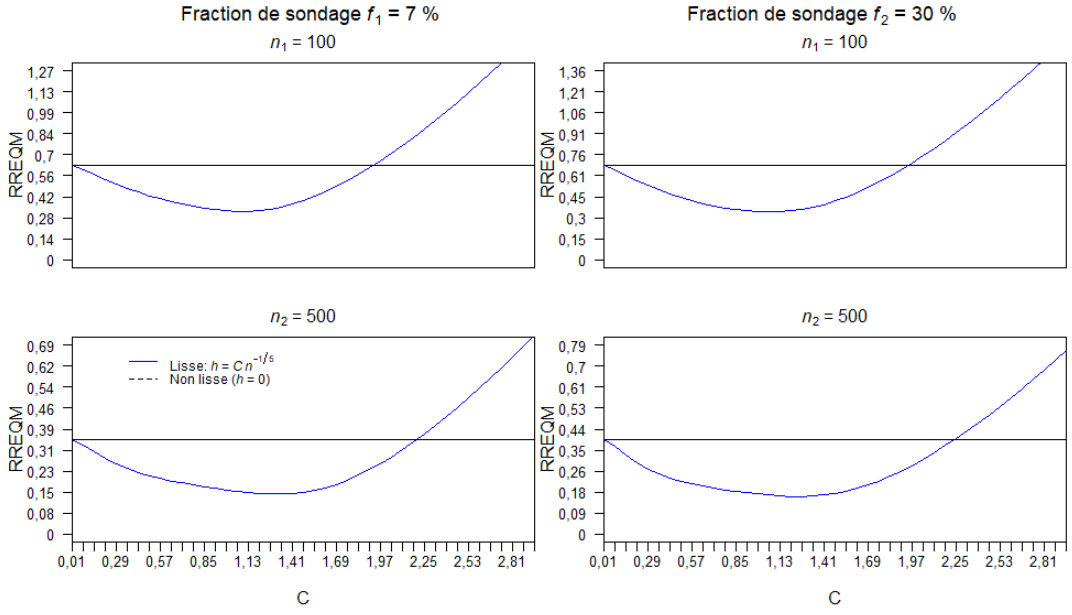


Fig. 5.10. RREQM_0 d'estimateurs bootstrap de $\mathbb{E}_0[\mathrm{EQM}_p(\hat{\xi}_{0,50})]$ en fonction de $C > 0$ sous le scénario $F_0 = \text{Lognormale}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

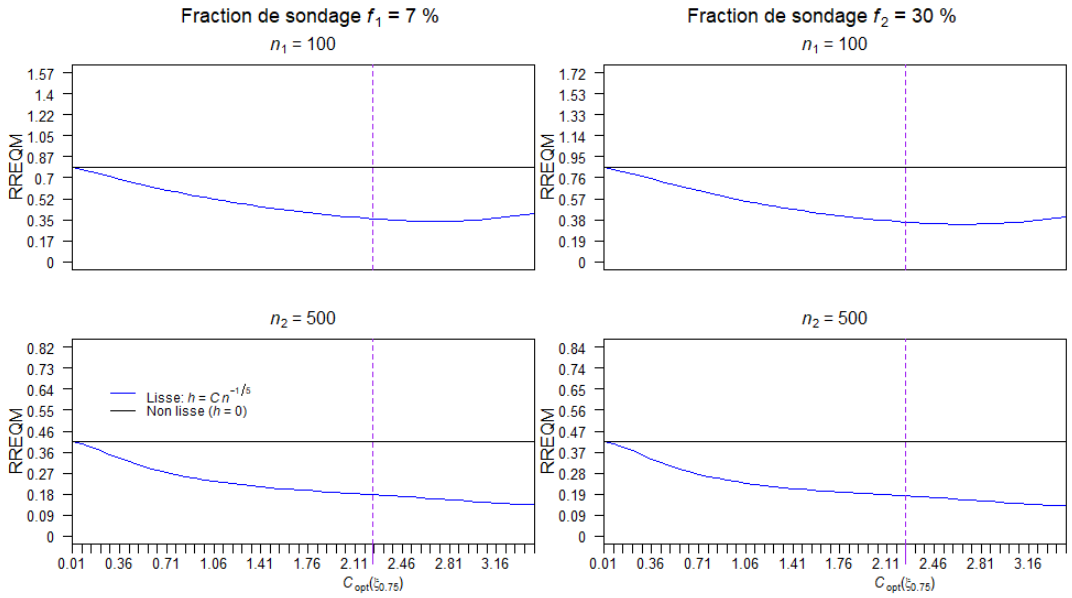


Fig. 5.11. RREQM_0 d'estimateurs bootstrap de $\mathbb{E}_0[\mathrm{EQM}_p(\hat{\xi}_{0,75})]$ en fonction de $C > 0$ sous le scénario $F_0 = \text{Lognormale}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

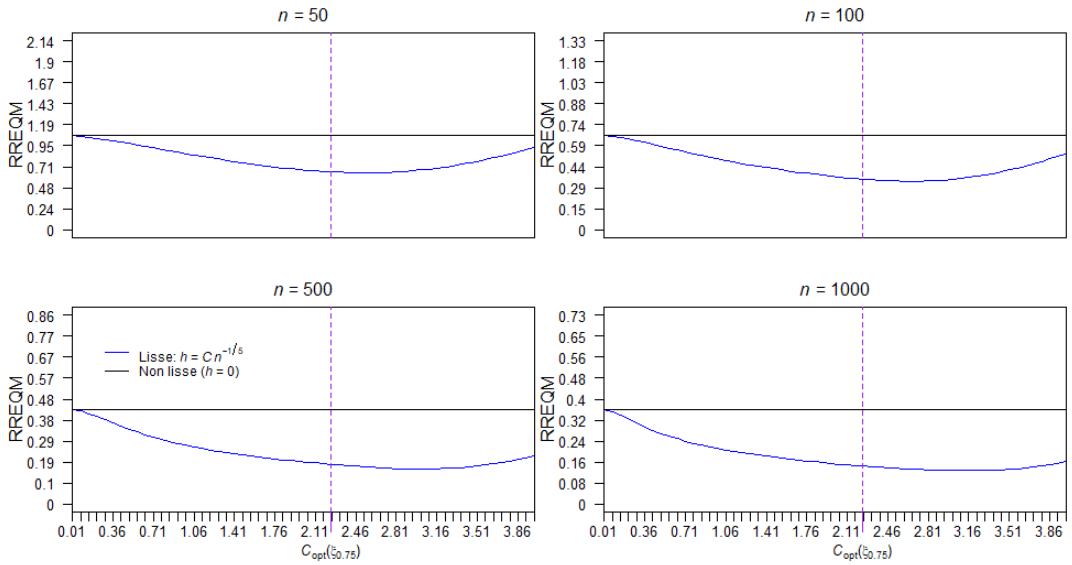


Fig. 5.12. $\mathrm{RREQM}_{\text{i.i.d.}}$ d'estimateurs bootstrap de $\mathrm{Var}_{\text{i.i.d.}}(\check{\xi}_{0,75})$ en fonction de $C > 0$ sous le scénario $F_0 = \text{Lognormale}(0,1)$ ($S = 1\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\mathrm{Var}^*(\check{\xi}_{0,75})$ ($h = 0$) et $\mathrm{Var}_h^*(\check{\xi}_{0,75})$, où $h = C \cdot n^{-1/5}$.

Le motif que l'on devinait à la figure 5.11 se représente à la figure 5.12, qui est pourtant entièrement conforme au cadre soutenant le théorème 3.3.2, faisant appel à des conditions de régularité sur f_0 et à des données indépendantes et identiquement distribuées. Cela remet-il en cause la validité du théorème de Hall et al. (1989) établissant la forme du paramètre de lissage optimal? Deux causes potentielles peuvent être suggérées afin d'expliquer la déviation inattendue du $\text{RREQM}_{\text{i.i.d.}}$ pour le troisième quartile. La première est que les tailles échantillonnales considérées sont possiblement insuffisantes pour cette distribution et que la convergence du minimum vers $C_{\text{opt}}(\tilde{\xi}_{0,75})$ pourrait avoir lieu pour des tailles supérieures. La deuxième explication nous amène à revoir les étapes inférentielles suivant le théorème 3.3.2 pour en arriver à l'expression de $h_{\text{opt}}(\tilde{\xi}_p)$ donnée par (3.3.8). Le théorème 3.3.2 établit une expression pour l'erreur relative de $\text{Var}_h^*(\check{\xi}_p)$ donnée par (3.3.6) et faisant intervenir une variable aléatoire Z. De fait, la variable aléatoire Z converge en loi vers la distribution $\mathcal{N}(0, \text{Var}_0(Z))$, où $\text{Var}_0(Z) = f_0(\tilde{\xi}_p)\kappa_1$ avec $\kappa_1 = \int k^2(t)dt$. On utilise cela pour inférer la convergence de $\mathbb{E}_0[Z^2]$ vers $f_0(\tilde{\xi}_p)\kappa_1$, ce qui aboutit à la quantité fixe étant minimisée par (3.3.8). Or, la convergence en loi n'implique pas la convergence des moments, ce qui pourrait expliquer l'instabilité du minimum observée.

5.4.2. Taux de couverture d'intervalles de confiance

Les intervalles de confiance de niveau 95% issus des différentes méthodes sont comparés au moyen des taux d'erreur de couverture inférieure, supérieure et bilatérale définis en (5.2.4) et de la mesure de longueur donnée par (5.2.5). Les tableaux 5.5 et 5.6 présentent ces mesures d'adéquation pour la médiane et le troisième quartile dans le cas d'une superpopulation $F_0 \equiv \mathcal{N}(0,1)$, alors que les tableaux 5.7 et 5.8 exhibent les résultats pour ces deux quantiles dans le cas d'une superpopulation $F_0 \equiv$ Lognormale(0,1). Les deux branches horizontales principales de ces tableaux correspondent aux deux tailles échantillonnales $n_1 = 100$, $n_2 = 500$, tandis que les deux branches principales verticales représentent les deux fractions de sondage $f_1 = 0{,}07$, $f_2 = 0{,}30$. Dans chacune des quatre cellules, les quatre mesures de performance sont rapportées pour les divers intervalles de confiance. La méthode de Woodruff présente invariablement une longueur de 1,00, puisque la longueur d'un intervalle donné est rapportée à la longueur d'un intervalle

de confiance de Woodruff. Pour les méthodes de rééchantillonnage, les trois types d'intervalle de confiance vus précédemment (asymptotique, de base et percentile) sont présentés.

L'étude du cas du centre d'une population issue d'une distribution symétrique donne lieu à des comparaisons intéressantes. Dans cette instance, on remarque d'abord à partir du tableau 5.5 que les taux d'erreur de couverture expérimentaux des intervalles de confiance de Woodruff correspondent immanquablement aux taux nominaux. Pour tous les scénarios d'enquête, les taux d'erreur de couverture unilatérale et bilatérale associés à cette méthode se trouvent à l'intérieur des régions d'acceptation, c'est-à-dire qu'ils ne sont pas statistiquement différents des taux nominaux de 2,5% et 5%. Il en va de même pour les intervalles asymptotiques calculés via l'algorithme de Booth et al. (1994), sauf pour le cas $n_2 = 500$, $f_2 = 30\%$, où une asymétrie significative entre les erreurs de couverture à gauche et à droite est observée. L'erreur de couverture bilatérale n'étant jamais significativement différente de 5% pour les intervalles asymptotiques de Booth et al. (1994), le lissage de la pseudo-population occasionne des intervalles plus longs que ceux de la méthode standard ainsi que des taux d'erreur de couverture bilatérale uniformément plus petits. De plus, les méthodes lisses entraînent un déséquilibre significatif entre les taux d'erreur de couverture unilatérale des intervalles asymptotiques pour la plus petite taille échantillonnale ($n_1 = 100$). Cela étant dit, le taux d'erreur de couverture bilatérale associé à la méthode *lisse boot ic* n'est pas significativement différent de 5% dans trois scénarios d'enquête sur quatre. C'est en examinant les intervalles de base qu'on constate un avantage important associé au lissage de la pseudo-population. En effet, avec la méthode standard ($h = 0$), les taux d'erreur de couverture bilatérale de ces intervalles sont non seulement statistiquement différents de 5%, mais sont aussi considérablement plus élevés à travers tous les cas de figure. L'introduction du lissage fait alors diminuer le taux d'erreur de couverture dans le cas des intervalles de base. Par ailleurs, les taux d'erreur de couverture bilatérale associés à la méthode *lisse boot ic* se trouvent dans la région d'acceptation d'un taux nominal de 5%, à l'exception du scénario $n_1 = 100$, $f_1 = 7\%$, pour lequel le taux expérimental vaut 6,35% comparativement à 10,85% pour la méthode de Booth et al. (1994). Enfin, bien que de même longueur que les intervalles de base par construction, les intervalles percentiles calculés via Booth et al. (1994) présentent une performance comparable à celle des intervalles asymptotiques. De

plus, on observe encore un certain équilibre entre l'erreur de couverture à gauche et celle à droite lorsque $h = 0$. Les taux d'erreur de couverture bilatérale des méthodes lisses sont généralement en dehors des régions d'acceptation pour ces intervalles, à l'exception faite de la méthode *lisse boot ic* dans le scénario $n_2 = 500$, $f_2 = 30\%$. Concernant la longueur des intervalles, une règle générale est que les intervalles asymptotiques de Booth et al. (1994) sont un peu plus longs que ceux de Woodruff (1952), alors que les intervalles de base et percentiles sont à toutes fins pratiques aussi longs. Tel que mentionné plus haut, le lissage de la pseudo-population engendre de manière générale des intervalles plus longs que ceux de (Booth et al., 1994), indifféremment du type d'intervalle considéré.

D'autres conclusions peuvent être portées lorsque l'on s'attarde au troisième quartile de la population issue de la distribution $\mathcal{N}(0,1)$ faisant l'objet du tableau 5.6. Relevons d'abord la performance légèrement moins reluisante de la méthode de Woodruff, pour laquelle on observe notamment un taux d'erreur de couverture bilatérale en dehors de la région d'acceptation pour un taux nominal de 5% pour le scénario $n_1 = 100$, $f_1 = 7\%$. Dans deux instances sur quatre, les intervalles asymptotiques de Booth et al. (1994) montrent des taux d'erreur de couverture bilatérale significativement différents de 5%. Ainsi, pour le troisième quartile, il peut y avoir un avantage à recourir aux intervalles asymptotiques lisses. En particulier, les taux d'erreur de couverture bilatérale des intervalles asymptotiques construits à partir des méthodes lisses se trouvent tous à l'intérieur de la région d'acceptation sans exception dans le scénario $n_2 = 500$, $f_1 = 7\%$, là où l'intervalle asymptotique de Booth et al. (1994) échoue. La piètre performance des intervalles de base selon Booth et al. (1994) se reproduit dans cette instance, tandis que les intervalles de base selon la méthode *lisse boot IC* font bien au regard de l'erreur de couverture bilatérale dans tous les scénarios à l'exception du scénario d'enquête $n_2 = 500$, $f_1 = 7\%$. Les intervalles percentiles de Booth et al. (1994) se conforment un peu mieux aux taux nominaux que les intervalles asymptotiques correspondants, bien que le taux d'erreur de couverture bilatérale soit également plus élevé que 5% dans le scénario $n_2 = 500$, $f_1 = 7\%$. Encore une fois, pour ce dernier scénario, les intervalles percentiles lisses montrent alors des taux uniformément près du taux nominal d'erreur de couverture bilatérale. En dernier lieu, il est intéressant d'observer la forte asymétrie entre les taux d'erreur de couverture à gauche et à droite

lorsque l'on transitionne de $n_1 = 100$ à $n_2 = 500$ dans le cas des intervalles asymptotiques et de base et ce quel que soit l'algorithme de rééchantillonnage utilisé. Pour les intervalles percentiles, cette asymétrie fait son apparition même dans le cas de $n_1 = 100$. La remarque faite plus tôt à l'égard de la longueur des intervalles dans le cas de la médiane d'une population symétrique peut également s'appliquer au troisième quartile, à l'exception du scénario $n_1 = 100$, $f_1 = 7\%$, où l'intervalle asymptotique associé à la méthode *lisse plug-in norm* est un peu plus court que l'homologue non lisse.

Tableau 5.5. Mesures de performance relatives aux intervalles de confiance de niveau 95% pour $\xi_{0,50}$ et une superpopulation $\mathcal{N}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$f_1 = 7\ \%$				$f_1 = 30\ \%$			
n	Méthode	% L	% U	% L+U	longueur	% L	% U	% L+U	longueur
100	Woodruff	2,00	2,65	4,65	1,00	2,70	2,75	5,45	1,00
	Intervalles asymptotiques								
	Booth et al.	2,40	2,70	5,10	1,03	1,95	2,90	4,85	1,04
	Lisse plug-in norm	1,00	3,00	4,00	1,06	0,85	1,95	2,80	1,14
	Lisse boot var	0,90	2,65	3,55	1,09	0,90	2,20	3,10	1,15
	Lisse boot IC	2,05	2,60	4,65	1,07	1,65	2,50	4,15	1,08
	Intervalles de base								
	Booth et al.	4,55	6,30	10,85	1,00	3,20	5,10	8,30	1,01
	Lisse plug-in norm	1,40	3,20	4,60	1,06	0,90	2,15	3,05	1,16
	Lisse boot var	1,10	2,80	3,90	1,09	1,00	2,20	3,20	1,17
	Lisse boot IC	2,90	3,45	6,35	1,07	1,90	2,95	4,85	1,09
	Intervalles percentiles								
	Booth et al.	2,15	2,30	4,45	1,00	2,65	2,50	5,15	1,01
	Lisse plug-in norm	0,95	2,80	3,75	1,06	0,70	1,75	2,45	1,16
	Lisse boot var	0,90	2,55	3,45	1,09	0,75	1,50	2,25	1,17
	Lisse boot IC	1,55	2,25	3,80	1,07	1,70	2,10	3,80	1,09
500	Woodruff	1,85	2,60	4,45	1,00	2,65	2,65	5,30	1,00
	Intervalles asymptotiques								
	Booth et al.	2,75	2,15	4,90	1,02	3,90	1,75	5,65	1,03
	Lisse plug-in norm	2,00	1,55	3,55	1,04	1,50	1,50	3,00	1,05
	Lisse boot var	1,95	1,55	3,50	1,04	1,85	1,50	3,35	1,04
	Lisse boot IC	1,95	1,85	3,80	1,04	3,05	1,55	4,60	1,05
	Intervalles de base								
	Booth et al.	5,35	3,20	8,55	1,00	6,55	1,70	8,25	1,00
	Lisse plug-in norm	2,10	1,70	3,80	1,04	1,50	1,50	3,00	1,06
	Lisse boot var	2,15	1,70	3,85	1,04	1,95	1,65	3,60	1,05
	Lisse boot IC	2,50	2,25	4,75	1,04	3,65	1,45	5,10	1,05
	Intervalles percentiles								
	Booth et al.	2,05	2,90	4,95	1,00	2,60	2,65	5,25	1,00
	Lisse plug-in norm	1,75	1,55	3,30	1,04	1,40	1,50	2,90	1,06
	Lisse boot var	1,90	1,55	3,45	1,04	1,80	1,70	3,50	1,05
	Lisse boot IC	1,70	2,15	3,85	1,04	2,35	1,75	4,10	1,05

Régions d'acceptation pour les taux d'erreur de couverture unilatérale (2,5%) et les taux d'erreur de couverture bilatérale (5%): [1,82; 3,18]% et [4,04; 5,96]%.

Tableau 5.6. Mesures de performance relatives aux intervalles de confiance de niveau 95% pour $\xi_{0,75}$ et une superpopulation $\mathcal{N}(0,1)$ ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$f_1 = 7$ %				$f_1 = 30$ %			
n	Méthode	% L	% U	% L+U	longueur	% L	% U	% L+U	longueur
100	Woodruff	1,75	2,10	3,85	1,00	2,45	2,60	5,05	1,00
	Intervalles asymptotiques								
	Booth et al.	1,85	3,10	4,95	1,04	1,85	1,80	3,65	1,03
	Lisse plug-in norm	1,90	1,95	3,85	1,03	1,55	2,55	4,10	1,05
	Lisse boot var	1,80	1,55	3,35	1,05	1,30	2,30	3,60	1,07
	Lisse boot IC	1,40	2,55	3,95	1,07	1,50	2,15	3,65	1,07
	Intervalles de base								
	Booth et al.	6,05	6,20	12,25	0,99	4,35	2,80	7,15	1,01
	Lisse plug-in norm	2,60	1,95	4,55	1,03	2,45	2,25	4,70	1,06
	Lisse boot var	2,40	1,55	3,95	1,05	2,20	2,20	4,40	1,08
	Lisse boot IC	2,45	3,25	5,70	1,06	2,85	2,00	4,85	1,08
	Intervalles percentiles								
	Booth et al.	1,15	3,70	4,85	0,99	1,05	3,60	4,65	1,01
	Lisse plug-in norm	1,35	2,10	3,45	1,03	0,90	2,45	3,35	1,06
	Lisse boot var	1,45	1,75	3,20	1,05	0,95	2,40	3,35	1,08
	Lisse boot IC	1,10	2,50	3,60	1,06	0,95	2,75	3,70	1,08
500	Woodruff	2,50	3,20	5,70	1,00	2,00	2,65	4,65	1,00
	Intervalles asymptotiques								
	Booth et al.	1,50	5,25	6,75	1,02	1,60	3,15	4,75	1,02
	Lisse plug-in norm	0,70	4,30	5,00	1,03	1,05	2,70	3,75	1,04
	Lisse boot var	0,70	4,50	5,20	1,03	1,00	2,80	3,80	1,05
	Lisse boot IC	1,00	4,70	5,70	1,05	1,25	2,85	4,10	1,05
	Intervalles de base								
	Booth et al.	2,45	8,00	10,45	0,99	2,50	4,05	6,55	1,00
	Lisse plug-in norm	1,00	4,30	5,30	1,03	1,25	2,75	4,00	1,05
	Lisse boot var	0,70	4,20	4,90	1,03	1,15	2,75	3,90	1,05
	Lisse boot IC	1,30	4,80	6,10	1,05	1,40	2,75	4,15	1,06
	Intervalles percentiles								
	Booth et al.	2,00	4,15	6,15	0,99	1,75	3,60	5,35	1,00
	Lisse plug-in norm	0,65	4,80	5,45	1,03	0,65	2,65	3,30	1,05
	Lisse boot var	0,70	4,55	5,25	1,03	0,80	2,60	3,40	1,05
	Lisse boot IC	1,15	4,40	5,55	1,05	1,20	2,60	3,80	1,06

Régions d'acceptation pour les taux d'erreur de couverture unilatérale (2,5%) et les taux d'erreur de couverture bilatérale (5%): [1,82; 3,18]% et [4,04; 5,96]%.

Les résultats du tableau 5.7 portant sur la médiane d'une réalisation de la distribution Lognormale(0,1) permettent de relever des différences par rapport à ce qui a été observé pour une population finie symétrique. Dans ces conditions, la méthode de Woodruff supplante encore toutes les autres, tant du point de vue de la couverture unilatérale que celui de la couverture bilatérale. Des différences sont manifestes en ce qui concerne les intervalles de confiance construits au moyen du rééchantillonnage. Par exemple, les intervalles asymptotiques de Booth et al. (1994) montrent des taux d'erreur de couverture bilatérale en deçà du taux nominal dans les deux scénarios pour lesquels $f_2 = 30\%$, ce qui n'était pas le cas pour la réalisation de la superpopulation $\mathcal{N}(0,1)$. Alors que ces mêmes taux se trouvent dans la région d'acceptation lorsque $f_2 = 7\%$, il en va de même pour les taux d'erreur de couverture bilatérale des intervalles percentiles associés à la méthode *lisse boot IC*. Par ailleurs, lorsque $f_2 = 30\%$, les intervalles de base construits sans lissage affichent des résultats opposés à ceux pour la population finie symétrique. On note en effet des taux d'erreur de couverture bilatérale bien moindres (3,75% et 5,00%), dont l'un se situe dans la région d'acceptation. Ce faisant, dans ces cas, l'introduction du lissage n'améliore pas la donne en engendrant des intervalles de base encore plus conservateurs. Les taux élevés pour les intervalles de base observés précédemment dans le cas de la population finie normale se représentent néanmoins dans les scénarios pour lesquels $f_2 = 7\%$. De surcroît, dans ces cas, certains intervalles de base lisses se conforment au taux nominal (à la fois *lisse boot var* et *lisse boot IC* pour une taille échantillonnale $n_2 = 500$). Dans l'ensemble des scénarios d'enquête, les intervalles percentiles de Booth et al. (1994) affichent des taux d'erreur de couverture bilatérale compris dans la région d'acceptation. En optant pour un intervalle percentile lisse, cela est vrai seulement lorsque $n_2 = 500$, $f_1 = 7\%$ pour les méthodes *lisse boot var* et *lisse boot IC*. Notons dernièrement les taux d'erreur de couverture bilatérale particulièrement petits et les longueurs particulièrement élevées associés à la méthode *lisse plug-in norm*, qui fait intervenir une taille de fenêtre *plug-in* basée sur la normalité.

À la lecture du tableau 5.8, il semble que le clivage entre les résultats de couverture des deux fractions de sondage est moins prononcé dans le cas du troisième quartile et de la superpopulation Lognormale(0,1). Autrement dit, les valeurs ponctuelles associées aux deux fractions de sondage sont plus rapprochées que pour le cas de la médiane d'une

population asymétrique à droite. La méthode de Woodruff est égale à elle-même. Du côté des intervalles asymptotiques construits par rééchantillonnage, au moins l'une des méthodes lisses se conforme au taux nominal pour l'erreur de couverture bilatérale dans trois scénarios sur quatre. Pour le scénario dérogeant à la règle, soit $f_2 = 30\%$, $n_2 = 500$, la méthode de Booth et al. (1994) affiche un taux d'erreur de couverture bilatérale égal à 4,35%. De plus, dans trois scénarios sur quatre, les intervalles asymptotiques associés à la méthode *lisse plug-in norm* sont plus longs que ceux de Booth et al. (1994). Le taux d'erreur de couverture bilatérale de l'intervalle de base de Booth et al. (1994) se trouve dans la région d'acceptation uniquement pour le scénario d'enquête $f_2 = 30\%$, $n_2 = 500$. Dans les autres cas de figure, comme il est devenu habituel d'observer, le taux est supérieur au taux nominal. Dans deux de ces cas, soit les scénarios associés à $f_1 = 7\%$, au moins l'une des méthodes lisses présentent un taux non différent de 5% (*lisse boot var* dans les deux cas). Les intervalles percentiles montrent une performance comparable à celle des intervalles asymptotiques. Une différence survient toutefois pour la méthode standard de Booth et al. (1994), pour laquelle le taux d'erreur de couverture bilatérale se situe immanquablement dans la région d'acceptation, contrairement à ce qui en était pour les intervalles asymptotiques. Pour finir, les intervalles de confiance construits à l'aide des méthodes par pseudo-population montrent une asymétrie généralisée entre les taux d'erreur de couverture à gauche et à droite. À ce titre, le taux d'erreur de couverture à droite est généralement plus élevé que celui à gauche.

Tableau 5.7. Mesures de performance relatives aux intervalles de confiance de niveau 95% pour $\xi_{0,50}$ et une superpopulation Lognormale(0,1) ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$f_1 = 7$ %				$f_1 = 30$ %			
n	Méthode	% L	% U	% L+U	longueur	% L	% U	% L+U	longueur
100	Woodruff	1,95	2,10	4,05	1,00	2,95	2,35	5,30	1,00
	Intervalles asymptotiques								
	Booth et al.	1,45	3,85	5,30	1,03	0,80	1,65	2,45	1,02
	Lisse plug-in norm	0,25	0,45	0,70	1,37	0,05	0,40	0,45	1,49
	Lisse boot var	0,65	2,40	3,05	1,12	0,20	1,50	1,70	1,23
	Lisse boot IC	0,95	3,80	4,75	1,07	0,50	1,60	2,10	1,08
	Intervalles de base								
	Booth et al.	2,50	9,65	12,15	0,99	1,05	2,70	3,75	1,00
	Lisse plug-in norm	0,25	1,00	1,25	1,37	0,05	0,45	0,50	1,52
	Lisse boot var	0,65	3,65	4,30	1,12	0,05	1,85	1,90	1,25
	Lisse boot IC	1,00	6,30	7,30	1,07	0,40	2,15	2,55	1,09
	Intervalles percentiles								
	Booth et al.	2,15	2,35	4,50	0,99	2,60	2,05	4,65	1,00
	Lisse plug-in norm	0,45	0,30	0,75	1,37	0,20	0,20	0,40	1,52
	Lisse boot var	0,75	1,55	2,30	1,12	0,25	1,20	1,45	1,25
	Lisse boot IC	1,40	2,50	3,90	1,07	1,50	1,35	2,85	1,09
500	Woodruff	2,35	2,20	4,55	1,00	2,30	2,05	4,35	1,00
	Intervalles asymptotiques								
	Booth et al.	3,20	2,20	5,40	1,03	1,25	2,25	3,50	1,01
	Lisse plug-in norm	1,30	0,90	2,20	1,12	0,25	0,90	1,15	1,31
	Lisse boot var	2,45	1,75	4,20	1,00	0,95	2,80	3,75	1,08
	Lisse boot IC	2,55	2,00	4,55	1,04	1,10	2,30	3,40	1,06
	Intervalles de base								
	Booth et al.	5,20	3,90	9,10	0,99	1,50	3,50	5,00	1,00
	Lisse plug-in norm	1,10	1,30	2,40	1,12	0,20	1,10	1,30	1,32
	Lisse boot var	2,25	2,60	4,85	1,00	0,75	3,20	3,95	1,08
	Lisse boot IC	2,65	3,00	5,65	1,04	0,80	2,95	3,75	1,06
	Intervalles percentiles								
	Booth et al.	2,60	2,05	4,65	0,99	1,95	2,20	4,15	1,00
	Lisse plug-in norm	1,75	0,80	2,55	1,12	0,30	0,70	1,00	1,32
	Lisse boot var	3,35	1,50	4,85	1,00	1,05	2,35	3,40	1,08
	Lisse boot IC	2,60	1,70	4,30	1,04	1,55	1,90	3,45	1,06

Régions d'acceptation pour les taux d'erreur de couverture unilatérale (2,5%) et les taux d'erreur de couverture bilatérale (5%): [1,82; 3,18]% et [4,04; 5,96]%.

Tableau 5.8. Mesures de performance relatives aux intervalles de confiance de niveau 95% pour $\xi_{0,75}$ et une superpopulation Lognormale(0,1) ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$f_1 = 7\ \%$				$f_1 = 30\ \%$			
n	Méthode	% L	% U	% L+U	longueur	% L	% U	% L+U	longueur
100	Woodruff	2,05	2,00	4,05	1,00	1,95	3,45	5,40	1,00
	Intervalles asymptotiques								
	Booth et al.	0,75	4,60	5,35	1,01	0,60	5,90	6,50	1,00
	Lisse plug-in norm	0,65	5,30	5,95	0,97	0,95	6,30	7,25	0,97
	Lisse boot var	0,50	3,75	4,25	1,03	1,05	4,85	5,90	1,01
	Lisse boot IC	0,50	4,25	4,75	1,02	0,80	4,90	5,70	1,02
	Intervalles de base								
	Booth et al.	2,25	10,85	13,10	0,97	1,35	8,50	9,85	0,98
	Lisse plug-in norm	0,35	7,35	7,70	0,97	0,85	6,90	7,75	0,98
	Lisse boot var	0,50	4,65	5,15	1,03	0,95	5,20	6,15	1,03
	Lisse boot IC	0,65	5,60	6,25	1,02	0,75	5,90	6,65	1,03
	Intervalles percentiles								
	Booth et al.	1,30	3,25	4,55	0,97	0,80	5,50	6,30	0,98
	Lisse plug-in norm	1,05	3,55	4,60	0,97	1,15	4,70	5,85	0,98
	Lisse boot var	0,65	3,55	4,20	1,03	1,10	4,20	5,30	1,03
	Lisse boot IC	0,60	3,35	3,95	1,02	0,80	4,05	4,85	1,03
500	Woodruff	2,25	2,20	4,45	1,00	2,30	2,60	4,90	1,00
	Intervalles asymptotiques								
	Booth et al.	1,60	2,85	4,45	1,01	1,75	2,60	4,35	1,02
	Lisse plug-in norm	1,65	3,10	4,75	0,98	0,55	2,50	3,05	1,08
	Lisse boot var	1,75	2,75	4,50	0,99	0,70	2,35	3,05	1,09
	Lisse boot IC	1,40	2,50	3,90	1,01	0,65	2,20	2,85	1,10
	Intervalles de base								
	Booth et al.	3,15	5,45	8,60	0,99	2,15	2,85	5,00	1,00
	Lisse plug-in norm	1,25	3,55	4,80	0,98	0,50	2,90	3,40	1,09
	Lisse boot var	1,55	3,20	4,75	0,99	0,55	2,50	3,05	1,10
	Lisse boot IC	1,25	3,15	4,40	1,00	0,55	2,30	2,85	1,10
	Intervalles percentiles								
	Booth et al.	1,95	2,60	4,55	0,99	1,80	3,30	5,10	1,00
	Lisse plug-in norm	2,10	2,40	4,50	0,98	0,60	2,40	3,00	1,09
	Lisse boot var	2,05	2,30	4,35	0,99	0,70	2,30	3,00	1,10
	Lisse boot IC	1,70	2,05	3,75	1,00	0,85	2,05	2,90	1,10

Régions d'acceptation pour les taux d'erreur de couverture unilatérale (2,5%) et les taux d'erreur de couverture bilatérale (5%): [1,82; 3,18]% et [4,04; 5,96]%.

5.5. Résultats pour l'échantillonnage de Poisson

Dans cette section, la performance des méthodes énumérées dans la section 5.3 au regard de l'estimation de l'erreur quadratique moyenne et de la formation d'intervalles de confiance est étudiée en tirant les S échantillons selon le plan de Poisson. Rappelons que pour l'étude de ce plan, la population finie a été générée selon le modèle de régression décrit en (5.1.1). Les probabilités de sélection $\pi_i,\ i = 1, \ldots, n$, du plan sont incidemment définies à partir de la réalisation considérée du régresseur X. Encore une fois, les résultats pour l'estimation de l'erreur quadratique moyenne et les intervalles de confiance sont déclinés selon les deux statistiques d'intérêt, $\hat{\xi}_{0,50}$ et $\hat{\xi}_{0,75}$, et les différents scénarios d'enquête.

5.5.1. Estimation de l'erreur quadratique moyenne

Le tableau 5.9 renferme les mesures de biais et de RREQM des différents estimateurs d'erreur quadratique moyenne pour la population finie considérée dans l'étude du plan de Poisson. La structure du tableau suit celle des tableaux 5.3 et 5.4 vus lors de l'étude de l'échantillonnage aléatoire simple sans remise. La remarque faite à l'endroit du biais associé aux méthodes lisses lors de l'étude du plan EASSR tient toujours pour ce plan. Dans tous les cas de figure, on observe un biais aussi sinon plus élevé en magnitude pour les estimateurs bootstrap lisses que pour l'estimateur bootstrap de Chauvet (2007). Cela s'accompagne en revanche invariablement d'une réduction de la variance, qui se traduit par des valeurs numériques de RREQM invariablement plus faibles pour les méthodes lisses que pour l'algorithme standard. Tout comme pour l'étude de la population symétrique sous le plan EASSR, l'instabilité de l'estimateur non lisse peut valoir jusqu'à deux fois celle de l'estimateur *lisse boot var* dans le cas notamment de la médiane sous le scénario $n_2 = 500$, $f_2 = 30\ \%$. Quelle que soit la valeur de β parmi les valeurs présentées, l'estimateur d'erreur quadratique moyenne $\hat{V}_\beta$ (méthode de Woodruff) comporte toujours un biais plus faible que les méthodes de rééchantillonnage par pseudo-population. De plus, lorsque l'on porte son regard sur le critère RREQM, cet estimateur indicé par β est aussi plus stable que la méthode par pseudo-population standard (Chauvet, 2007). Ce constat fait écho à ce qui a été observé pour une population finie symétrique lors de l'étude du plan EASSR. L'estimateur $\hat{V}_{\hat{h}_{\text{boot,var}}}$ surpasse généralement les estimateurs de Woodruff $\hat{V}_\beta$ du point de vue de l'instabilité, à l'exception de quelques cas où la performance est assez semblable. Du côté de la médiane, cela survient

pour les deux scénarios d'enquête correspondant à $n_1 = 100$. En ce qui a trait au troisième quartile, la méthode *lisse boot var* fait toujours mieux que la méthode de Woodruff.

Tableau 5.9. Mesures de performance des estimateurs d'erreur quadratique moyenne pour l'échantillonnage de Poisson et la superpopulation donnée par (5.1.1) ($S = 2\,000$, $B = 1\,000$ et $D = 50$).

		$\hat{\xi}_{0.50}$				$\hat{\xi}_{0.75}$			
		$f_1 = 7\ \%$		$f_2 = 30\ \%$		$f_1 = 7\ \%$		$f_2 = 30\ \%$	
n	Méthode	biais rel.	RREQM	biais rel.	RREQM	biais rel.	RREQM	biais rel.	RREQM
	Woodruff								
	$\beta = 0{,}01$	0,06	0,42	0,11	0,43	0,17	0,48	-0,06	0,33
100	$\beta = 0{,}025$	0,04	0,45	0,09	0,45	0,15	0,50	-0,05	0,36
	$\beta = 0{,}05$	0,03	0,46	0,07	0,47	0,14	0,53	-0,03	0,41
	$\beta = 0{,}1$	0,04	0,51	0,06	0,51	0,15	0,60	-0,00	0,47
	$\beta = 0{,}2$	0,06	0,60	0,08	0,61	0,14	0,67	0,03	0,57
	Chauvet	0,15	0,57	0,20	0,58	0,22	0,60	0,13	0,49
	Lisse boot var	0,25	0,43	0,33	0,47	0,25	0,42	0,01	0,24
	Lisse boot IC	0,17	0,49	0,25	0,51	0,22	0,49	0,08	0,37
	Woodruff								
	$\beta = 0{,}01$	0,07	0,29	-0,01	0,23	0,12	0,33	0,09	0,29
500	$\beta = 0{,}025$	0,07	0,30	0,00	0,25	0,11	0,34	0,08	0,31
	$\beta = 0{,}05$	0,07	0,32	0,01	0,28	0,10	0,36	0,08	0,33
	$\beta = 0{,}1$	0,07	0,36	0,02	0,32	0,09	0,39	0,07	0,36
	$\beta = 0{,}2$	0,07	0,40	0,02	0,38	0,08	0,44	0,07	0,42
	Chauvet	0,11	0,38	0,04	0,33	0,13	0,40	0,10	0,39
	Lisse boot var	0,13	0,22	-0,02	0,16	0,21	0,31	0,19	0,27
	Lisse boot IC	0,13	0,28	0,05	0,21	0,19	0,34	0,17	0,30

5.5.1.1. *Représentations graphiques*

Les figures 5.13 et 5.14 permettent de mettre en relation quelques résultats numériques vus dans le tableau 5.9 dans le cas de $\hat{\xi}_{0,50}$ et de $\hat{\xi}_{0,75}$ respectivement. Ces graphiques sont analogues à ceux présentés dans la section des résultats pour l'échantillonnage simple sans remise. Encore une fois, la courbe en bleu fait référence à la mesure RREQM de l'estimateur bootstrap indicé par $h = C \cdot n^{-1/5}$, soit $\hat{V}_h^*$, tandis que le trait en noir se rapporte à la

mesure RREQM de l'estimateur bootstrap non lisse, dénoté $\hat{V}^*$. Les quatres scénarios d'enquête sont déclinés en quatre panneaux.

À la figure 5.13, on aperçoit entre autres le grand potentiel d'amélioration apporté par le lissage de la pseudo-population dans le scénario $n_2 = 500$, $f_2 = 30$ % pour la médiane, ce qui rappelle les résultats observés dans la cellule correspondante dans le tableau 5.9. Par ailleurs, c'est seulement pour ce dernier scénario d'enquête que le minimum de la courbe coïncide avec la constante théorique pour $\tilde{\xi}_{0,50}$ sous le modèle (5.1.1) pouvant être retrouvée dans le tableau 5.2. Pour la médiane, on remarque que le potentiel d'amélioration croît alors que la taille échantillonnale augmente, ce qui a été observé pour la population finie symétrique considérée lors de l'étude du plan EASSR. Du côté du troisième quartile, le potentiel le plus élevé est observé pour le scénario d'enquête $n_1 = 100$, $f_2 = 30$ %, ce qui se reflète également dans les résultats du tableau 5.9 à travers la différence marquée entre la valeur du RREQM de la méthode non lisse et celle correspondant à la méthode *lisse boot var*. En revanche, la relation entre la taille échantillonnale et le potentiel d'amélioration est moins claire que pour le cas de la médiane. Qui plus est, une proximité entre la constante $C_{\text{opt}}(\tilde{\xi}_{0,75})$ et le minimum de la courbe en bleu est seulement aperçue pour le scénario $n_1 = 100$, $f_2 = 7$ %. Afin de pouvoir porter des conclusions quant à l'optimalité des contantes théoriques, il convient encore une fois de se pencher sur résultats de simulation alliant à la fois le modèle et le plan de sondage.

Considérons à cette fin l'approximation numérique de RREQM_0 de l'estimateur $\hat{V}_h^*$ pour le plan de Poisson, qui est représentée aux figures 5.15 (quantile $\hat{\xi}_{0,50}$) et 5.16 (quantile $\hat{\xi}_{0,75}$) pour une grille de valeurs de C, où $h = C \cdot n^{-1/5}$. La courbe en bleu correspond au critère RREQM_0 de l'estimateur $\hat{V}_h^*$, tandis que la droite en noire se rattache à l'estimateur non lisse $\hat{V}^*$. Maintenant que nous considérons un grand nombre de réalisations du modèle (5.1.1) ($S = 2\ 000$) par opposition à une seule, les différences relevées entre les fractions de sondage 7 % et 30 % se sont estompées pour les deux quantiles étudiés. De plus, tant pour la médiane et le troisième quartile, les minimums des courbes en bleu se sont rapprochés des constantes théoriques. Ces constats ne sont pas sans rappeler le cas de la superpopulation $\mathcal{N}(0,1)$ lors de l'étude de l'échantillonnage aléatoire simple sans remise.

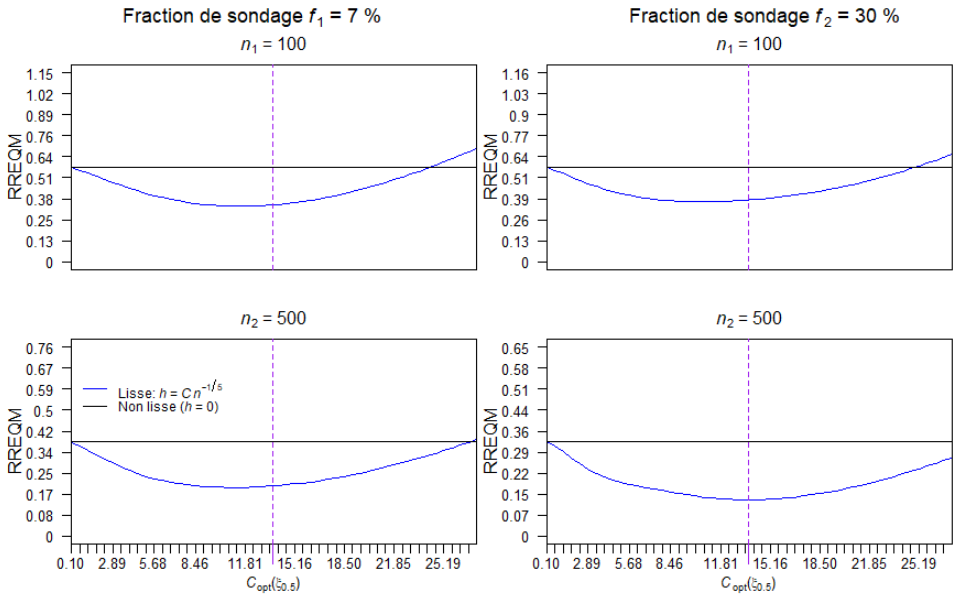


Fig. 5.13. RREQM d'estimateurs bootstrap de $\mathrm{EQM}_p(\hat{\xi}_{0,50})$ en fonction de $C > 0$ pour l'échantillonnage de Poisson et la superpopulation donnée par (5.1.1) ($S = 2\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

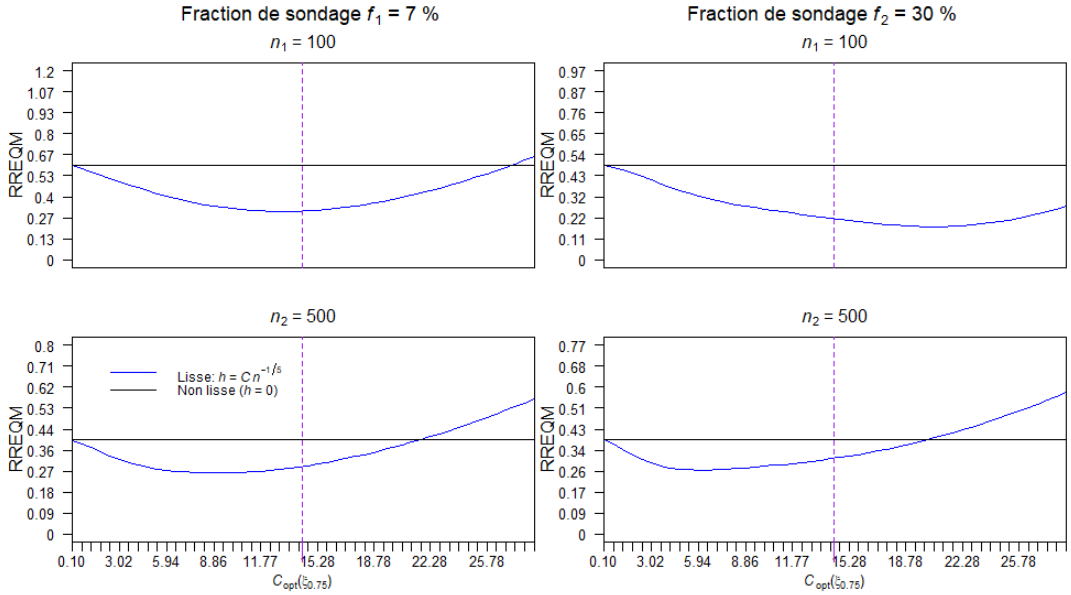


Fig. 5.14. RREQM d'estimateurs bootstrap de $\mathrm{EQM}_p(\hat{\xi}_{0,75})$ en fonction de $C > 0$ pour l'échantillonnage de Poisson et la superpopulation donnée par (5.1.1) ($S = 2\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

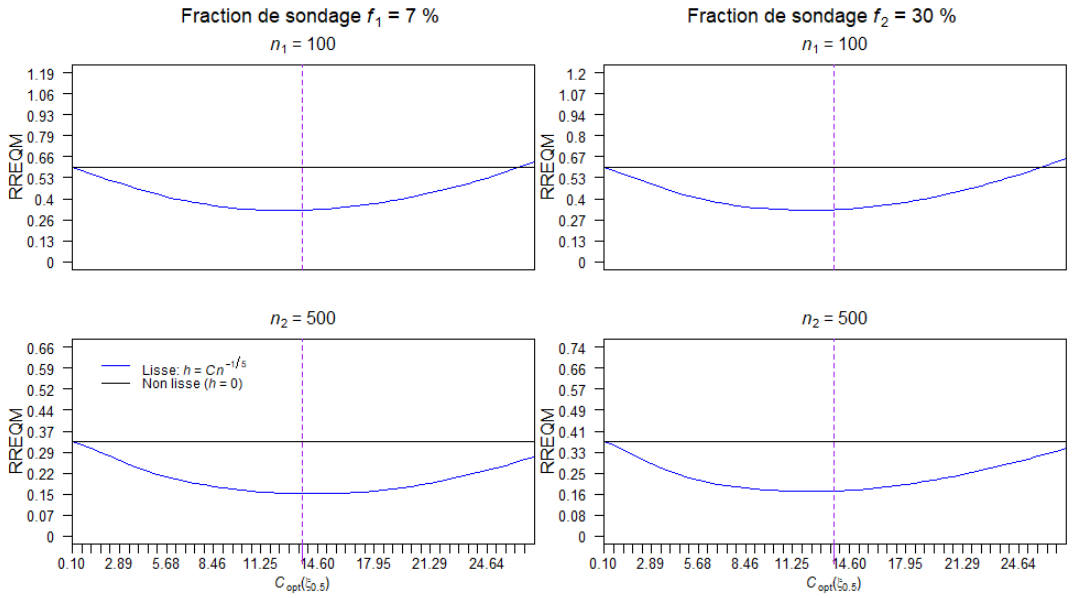


Fig. 5.15. RREQM_0 d'estimateurs bootstrap de $\mathbb{E}_0[\mathrm{EQM}_p(\hat{\xi}_{0,50})]$ en fonction de $C > 0$ pour l'échantillonnage de Poisson et la superpopulation donnée par (5.1.1) ($S = 2\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

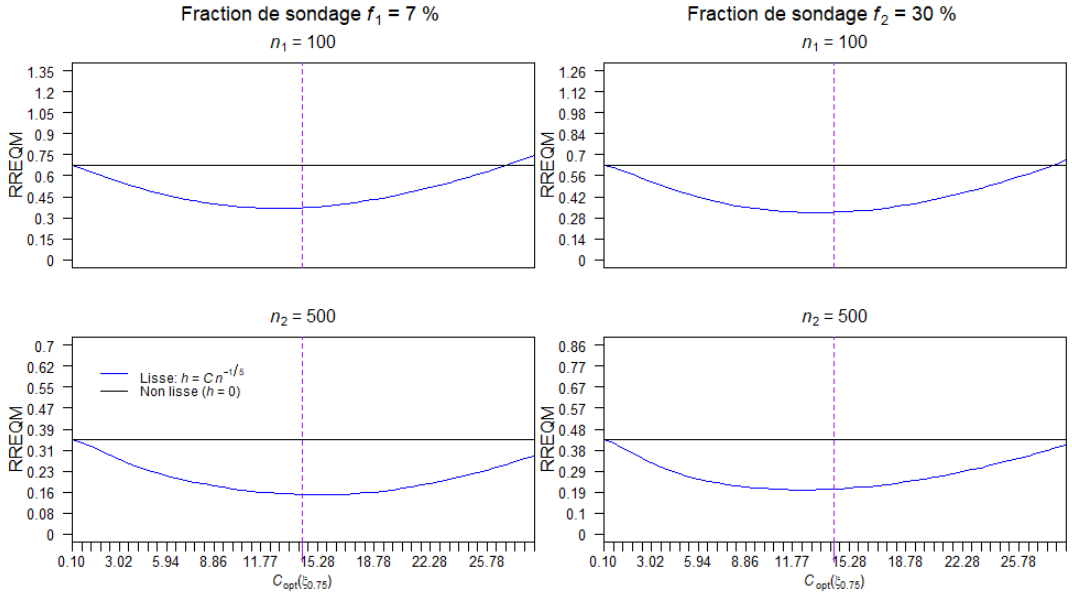


Fig. 5.16. RREQM_0 d'estimateurs bootstrap de $\mathbb{E}_0[\mathrm{EQM}_p(\hat{\xi}_{0,75})]$ en fonction de $C > 0$ pour l'échantillonnage de Poisson et la superpopulation donnée par (5.1.1) ($S = 2\,000$, $B = 1\,000$). Deux estimateurs sont représentés, soit $\hat{V}^*$ ($h = 0$) et $\hat{V}_h^*$, où $h = C \cdot n^{-1/5}$.

5.5.2. Taux de couverture d'intervalles de confiance

Les mesures d'adéquation des intervalles de confiance des différentes méthodes étudiées pour l'échantillonnage de Poisson sont rapportées dans les tableaux 5.10 et 5.11, qui se rattachent respectivement à la médiane et au troisième quartile d'une population finie issue du modèle (5.1.1).

En portant d'abord notre attention sur le cas de la médiane, nous voyons à partir du tableau 5.10 que les intervalles asymptotiques construits via les méthodes par pseudo-population sont toujours plus longs que ceux associés à la méthode de Woodruff. Les intervalles asymptotiques de Chauvet (2007) comportent des taux d'erreur de couverture bilatérale qui sont toujours compris dans la région d'acceptation de $[4{,}04; 5{,}96]\%$ pour un taux nominal de 5%. Pour ces intervalles, on décèle une légère asymétrie entre les taux d'erreur de couverture à gauche et à droite dans le cas du scénario $n_1 = 100$, $f_2 = 30$ %. En somme, la performance de cette méthode non lisse est donc analogue à celle observée plus tôt dans des conditions similaires lors de l'étude du plan EASSR. Lorsque $n_1 = 100$, les méthodes lisses ($h > 0$) mènent à des intervalles asymptotiques plus longs et aux taux d'erreur de couverture bilatérale significativement plus petits que 5%. Pour la taille échantillonnale supérieure, les intervalles asymptotiques associés à la méthode *lisse boot ic*, qui se fonde sur un critère d'erreur de couverture estimée par bootstrap, se conforment au taux nominal de 5%. Le tableau 5.10 montre aussi une résurgence de la piètre performance des intervalles de base non lisses, avec des taux d'erreur de couverture bilatérale parfois aussi élevés que 13,60%. Dans trois scénarios sur quatre, la méthode *lisse boot ic* parvient encore une fois à ramener ce taux dans les eaux de 5%. Les taux d'erreur de couverture à gauche et à droite des intervalles de base connaissent un déséquilibre généralisé. On ne peut en dire autant des intervalles percentiles de Chauvet (2007), dont l'erreur de couverture à gauche est sensiblement égale à celle à droite. De plus, les intervalles percentiles non lisses montrent une performance analogue à celle des intervalles asymptotiques non lisses du point de vue du taux d'erreur de couverture bilatérale. Le même constat peut être fait vis-à-vis les intervalles percentiles construits via *lisse boot ic*. Toujours est-il que les intervalles de Woodruff offrent un excellent rendement sur tous les plans.

Tableau 5.10. Mesures de performance relatives aux intervalles de confiance de niveau 95% pour $\xi_{0,50}$ dans le cas de l'échantillonnage de Poisson et de la superpopulation donnée par (5.1.1) ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$f_1 = 7$ %				$f_1 = 30$ %			
n	Méthode	% L	% U	% L+U	longueur	% L	% U	% L+U	longueur
100	Woodruff	2,75	2,20	4,95	1,00	2,70	2,45	5,15	1,00
	Intervalles asymptotiques								
	Chauvet	2,05	3,10	5,15	1,05	1,65	2,80	4,45	1,05
	Lisse boot var	1,00	1,50	2,50	1,12	0,50	2,80	3,30	1,13
	Lisse boot IC	1,55	2,20	3,75	1,07	1,00	2,75	3,75	1,09
	Intervalles de base								
	Chauvet	4,15	9,45	13,60	0,99	3,45	6,10	9,55	1,01
	Lisse boot var	1,15	1,55	2,70	1,12	0,55	2,90	3,45	1,15
	Lisse boot IC	1,85	3,00	4,85	1,07	1,05	3,15	4,20	1,10
	Intervalles percentiles								
	Chauvet	2,45	2,45	4,90	0,99	3,05	2,30	5,35	1,01
	Lisse boot var	1,00	1,30	2,30	1,12	0,40	2,35	2,75	1,15
	Lisse boot IC	1,30	1,90	3,20	1,07	1,20	2,30	3,50	1,10
500	Woodruff	1,80	3,10	4,90	1,00	3,10	2,75	5,85	1,00
	Intervalles asymptotiques								
	Chauvet	2,70	2,50	5,20	1,02	1,95	3,90	5,85	1,02
	Lisse boot var	1,70	2,30	4,00	1,04	1,50	4,95	6,45	1,00
	Lisse boot IC	2,30	2,35	4,65	1,03	1,75	4,15	5,90	1,03
	Intervalles de base								
	Chauvet	5,00	3,65	8,65	1,00	2,65	7,10	9,75	1,00
	Lisse boot var	1,80	2,35	4,15	1,03	1,60	5,00	6,60	1,01
	Lisse boot IC	2,55	2,75	5,30	1,03	1,80	4,35	6,15	1,04
	Intervalles percentiles								
	Chauvet	2,05	3,15	5,20	1,00	2,65	3,00	5,65	1,00
	Lisse boot var	1,70	2,35	4,05	1,03	1,60	4,85	6,45	1,01
	Lisse boot IC	2,00	2,45	4,45	1,03	1,65	3,80	5,45	1,04

Régions d'acceptation pour les taux d'erreur de couverture unilatérale (2,5%) et les taux d'erreur de couverture bilatérale (5%): [1,82; 3,18]% et [4,04; 5,96]%.

Tableau 5.11. Mesures de performance relatives aux intervalles de confiance de niveau 95% pour $\xi_{0,75}$ dans le cas de l'échantillonnage de Poisson et de la superpopulation donnée par (5.1.1) ($S = 2\ 000$, $B = 1\ 000$ et $D = 50$).

		$f_1 = 7\ \%$				$f_1 = 30\ \%$			
n	Méthode	% L	% U	% L+U	longueur	% L	% U	% L+U	longueur
100	Woodruff	2,85	1,85	4,70	1,00	2,90	1,55	4,45	1,00
	Intervalles asymptotiques								
	Chauvet	1,95	3,20	5,15	1,04	3,25	3,35	6,60	1,07
	Lisse boot var	1,30	3,10	4,40	1,07	1,10	4,60	5,70	1,04
	Lisse boot IC	1,80	3,15	4,95	1,05	2,30	3,30	5,60	1,06
	Intervalles de base								
	Chauvet	4,25	6,25	10,50	0,98	7,60	8,75	16,35	1,02
	Lisse boot var	1,45	3,35	4,80	1,07	1,25	4,65	5,90	1,06
	Lisse boot IC	2,10	3,60	5,70	1,05	3,75	4,70	8,45	1,07
	Intervalles percentiles								
	Chauvet	2,20	3,35	5,55	0,98	2,40	1,85	4,25	1,02
	Lisse boot var	1,00	2,95	3,95	1,07	0,85	4,25	5,10	1,06
	Lisse boot IC	1,40	2,70	4,10	1,05	1,30	2,30	3,60	1,07
500	Woodruff	2,75	2,15	4,90	1,00	2,95	2,65	5,60	1,00
	Intervalles asymptotiques								
	Chauvet	2,45	2,85	5,30	1,01	1,60	4,10	5,70	1,02
	Lisse boot var	1,40	2,40	3,80	1,06	1,25	4,15	5,40	1,07
	Lisse boot IC	1,90	2,40	4,30	1,05	1,30	4,20	5,50	1,06
	Intervalles de base								
	Chauvet	4,25	3,70	7,95	0,99	3,30	6,10	9,40	1,00
	Lisse boot var	1,65	2,35	4,00	1,05	1,30	4,10	5,40	1,08
	Lisse boot IC	1,90	2,60	4,50	1,04	1,25	4,05	5,30	1,07
	Intervalles percentiles								
	Chauvet	2,35	2,85	5,20	0,99	2,40	3,55	5,95	1,00
	Lisse boot var	1,35	2,35	3,70	1,05	1,05	4,25	5,30	1,08
	Lisse boot IC	1,75	2,25	4,00	1,04	1,15	4,25	5,40	1,07

Régions d'acceptation pour les taux d'erreur de couverture unilatérale (2,5%) et les taux d'erreur de couverture bilatérale (5%): [1,82; 3,18]% et [4,04; 5,96]%.

La discussion des résultats se termine avec le cas du troisième quartile de la réalisation considérée du modèle de régression (5.1.1). Les méthodes de rééchantillonnage par pseudo-population lisse s'illustrent alors davantage, c'est-à-dire que l'une des deux (*lisse boot var* ou *lisse boot ic*) fait généralement aussi bien ou mieux encore que la méthode de Chauvet (2007). En effet, on obtient un taux d'erreur de couverture bilatérale non différent de 5% avec l'une ou l'autre des méthodes lisses quel que soit le type d'intervalle. Une exception à la règle survient pour l'intervalle percentile du scénario $n_2 = 500, f_2 = 7\%$. Il va sans dire que l'on tire un avantage marqué à recourir aux algorithmes mis de l'avant dans ce mémoire dans le cas des intervalles de base. Les intervalles de Woodruff demeurent cependant imbattables tout en étant numériquement moins intensifs.

Conclusion

Le présent ouvrage avait pour but d'étudier une adaptation lisse de méthodes de rééchantillonnage par pseudo-population existantes, qui peuvent être mises à profit pour estimer la distribution échantillonnale d'estimateurs complexes de paramètres de population finie. La modification proposée comporte un attrait particulier lorsque l'estimateur sous la loupe est celui d'un quantile de population finie. C'est pourquoi ce mémoire s'est focalisé sur le problème de l'estimation de la variance et la construction d'intervalles de confiance pour des quantiles.

Une discussion sur l'estimation de quantiles en population finie a permis, dans un premier temps, d'appréhender les difficultés survenant au moment de développer des mesures de précision de l'estimateur du quantile, qui est une fonction non linéaire de totaux. Basée sur la linéarisation de l'estimateur de la fonction de répartition dans la population, la méthode approximative de Woodruff (1952) mène à un intervalle de confiance unique pour un quantile de population finie et à une famille d'estimateurs de la variance de l'estimateur de quantile indicés par un paramètre. En termes d'efforts de calculs, cette méthode est peu coûteuse en comparaison avec les méthodes de rééchantillonnage constituant le cœur de cet ouvrage. Ainsi, non seulement les méthodes bootstrap ont-elles été comparées entre elles, mais aussi avec la méthode de Woodruff.

D'entre toutes les méthodes discutées s'appuyant sur la puissance de calcul informatique, nous avons relevé l'importance de recourir à des méthodes de rééchantillonnage reflétant le plan de sondage utilisé lors de l'enquête. Notamment, la méthode du bootstrap non paramétrique d'Efron (1979) fait fi de la covariance entre les unités issues d'une population finie en les tenant pour indépendantes et identiquement distribuées. Dans le cas

précis de l'estimateur d'un quantile, Chatterjee (2011) établit explicitement que l'estimateur de variance pourvu par la méthode du bootstrap non paramétrique d'Efron (1979) ne converge pas vers la variance asymptotique de l'estimateur même dans le cas d'un plan de sondage aussi élémentaire que celui de l'EASSR.

Les méthodes de rééchantillonnage par pseudo-population émulent les conditions de l'enquête en proposant de tirer les échantillons bootstrap selon le plan de sondage utilisé initialement à partir d'une pseudo-population. Nous les avons décrites en nous concentrant sur les algorithmes de Booth et al. (1994) et de Chauvet (2007) pour l'échantillonnage aléatoire simple sans remise et pour l'échantillonnage de Poisson respectivement. Mais que dire à propos de la qualité des estimateurs issus de ces méthodes dans le cas des quantiles? L'illustration de la distribution bootstrap des estimateurs de la médiane et du troisième quartile engendrée par l'algorithme de Booth et al. (1994) a permis de sceller la problématique de ce mémoire. Un support échantillonnal très pauvre est observé et vient compromettre la qualité des estimateurs bootstrap résultants.

La distribution asymptotique d'un quantile échantillonnal a la particularité de dépendre localement de la distribution dans le cas i.i.d. ou de la distribution de la superpopulation dans le cas d'une population finie. Incidemment, la variance du quantile échantillonnal est une fonctionnelle qualifiée de *non lisse*. Dans le cadre i.i.d., celle-ci a par ailleurs servi d'exemple à Hall et al. (1989) au moment de relever une instance où il y avait avantage à recourir au bootstrap lisse. À ce titre, un théorème démontré par Hall et al. (1989) a permis d'établir que

(1) Le rééchantillonnage à partir d'une estimation par le noyau $\hat{F}_h$ de F_0 peut avoir un impact important lorsque la fonctionnelle à l'étude dépend de propriétés locales de la distribution, comme c'est le cas de la variance d'un quantile échantillonnal.
(2) Un paramètre de lissage de l'ordre $h = C \cdot n^{-1/5},\ C > 0$, combiné à un noyau K gaussien fait passer l'ordre de l'erreur relative de l'estimateur bootstrap de la variance d'un quantile échantillonnal de $n^{-1/4}$ à $n^{-2/5}$.
(3) Pour une taille échantillonnale n fixée, une expression de la constante C minimisant l'erreur relative peut être obtenue et cette constante dite optimale dépend de la

fonction de densité f_0 caractérisant la distribution des données.

Le boostrap lisse a été étendu au contexte de l'échantillonnage dans le chapitre 4 de ce mémoire. L'idée du lissage de la pseudo-population dans les méthodes bootstrap par pseudo-population a alors été mise au jour. Tel qu'illustré à la fin du chapitre 2, cela a pour conséquence d'enrichir le support de la distribution bootstrap du quantile échantillonnal. À notre connaissance, la mise en œuvre du bootstrap lisse pour des données d'enquête constitue une idée nouvelle. Nous avons procédé à la description d'une adaptation lisse de l'algorithme de Booth et al. (1994) pour le plan EASSR et de l'algorithme de Chauvet (2007) pour l'échantillonnage de Poisson. Ce faisant, nous avons introduit une famille d'estimateurs de variance bootstrap ainsi qu'une famille d'intervalles de confiance bootstrap indicés par un paramètre de lissage h. Le degré de lissage étant déterminant au niveau de la qualité des estimateurs bootstrap résultants, nous avons orienté la suite du mémoire autour de cette question.

Deux méthodes de sélection du paramètre de lissage mettant à profit l'information fournie par l'échantillon ont été introduites de manière à compléter les algorithmes proposés, afin qu'ils puissent être utilisés en pratique. La première, la sélection par injection, se fonde sur l'expression de la constante optimale pour l'estimation de la variance d'un quantile développée dans le cadre classique pour des données normalement distribuées. La seconde, la sélection par bootstrap, ne fait aucune hypothèse distributionnelle et vise à minimiser une estimation bootstrap d'une fonction de perte. Deux possibilités de fonction de perte ont été présentées pour rencontrer les objectifs de ce mémoire, l'une étant l'erreur quadratique moyenne de l'estimateur bootstrap de la variance et l'autre étant une distance entre les taux d'erreur de couverture expérimental et nominal d'un intervalle de confiance.

Une étude par simulation pour la médiane et le troisième quartile a été menée à bien dans le but de comparer les divers estimateurs d'erreur quadratique moyenne des estimateurs de quantiles et les différents intervalles de confiance. Nous avons opté pour une simulation basée sur le plan, consistant à effectuer le tirage des échantillons à partir d'une population finie fixée. Deux distributions se différenciant du point de vue de l'asymétrie ont

été couvertes, de même que plusieurs scénarios d'enquête.

Plusieurs constats d'intérêt émergent de l'étude axée sur le plan EASSR. En premier lieu, les estimateurs bootstrap de l'erreur quadratique moyenne sous le plan de l'estimateur de la médiane et de l'estimateur du troisième quartile obtenus via le lissage de la pseudo-population sont généralement plus stables que leur homologue non lisse et parfois de loin. Dans bien des cas, ils surpassent aussi les estimateurs de Woodruff (1952) pour les valeurs fixes du paramètre d'ajustement considérées. Les méthodes de sélection automatiques du paramètre de lissage h réussissent généralement à capturer une valeur pour le paramètre de lissage permettant de réduire substantiellement la racine carrée de l'erreur quadratique moyenne relative de l'estimateur. Cela est particulièrement vrai pour la méthode de sélection par injection lorsqu'elle est mise à exécution avec des données normalement distribuées, puisque la fonction de densité postulée est alors la densité véritable. Quant à la population log-normale, les résultats obtenus donnent à croire qu'il est plus avantageux de se replier sur la sélection par bootstrap du paramètre de lissage, laquelle permet généralement d'accroître la stabilité de l'estimateur. En outre, il est intéressant de noter que pour la réalisation considérée de la distribution asymétrique, les bénéfices sont plus importants pour le troisième quartile.

En périphérie, les résultats de la simulation basée sur le plan montrent que le minimum observé du critère RREQM ne correspond pas minimum théorique prévu par Hall et al. (1989) dans plusieurs scénarios. La question de vérifier empiriquement l'optimalité de la constante théorique s'est donc imposée naturellement. Une simulation jumelant à la fois le plan et le modèle a servi à cette fin. Sous cet autre procédé de génération, les résultats empiriques suggèrent que le minimum observé et la constante théorique coïncident à toutes fins pratiques. Cela s'explique par le fait que ce type de simulation reproduit sensiblement le cadre i.i.d., soit le cadre dans lequel les résultats de Hall et al. (1989) ont été démontrés. Il convient toutefois de relever l'exception notable du troisième quartile d'une population log-normale, pour lequel le minimum se déplaçait vers la droite à mesure que la taille échantionnale augmentait et ce même pour des données i.i.d. et malgré l'utilisation d'un paramètre de lissage doté de l'ordre de grandeur prescrit par le théorème. Nous avons

formulé une hypothèse pouvant expliquer cette incongruité, qui renvoie aux suppositions faites pour dériver le paramètre de lissage optimal. La valeur optimale est celle qui minimise la variance asymptotique d'une variable aléatoire qui est une approximation asymptotique de l'erreur relative de l'estimateur de variance bootstrap. Or, même en ignorant l'impact des approximations sur la variance, la variance d'une variable aléatoire ne converge pas nécessairement vers sa variance asymptotique.

Un second volet de l'étude de la performance des méthodes discutées portait sur le taux de couverture des intervalles de confiance. Les résultats de simulations agrégés pour le plan EASSR suggèrent d'abord qu'il est difficile de surpasser les intervalles de confiance construits selon la méthode de Woodruff (1952), qui non seulement représentent une faible intensité de calcul, mais montrent des taux de couverture expérimentaux se conformant presque toujours aux taux nominaux. Ceci tient tant pour la population symétrique que pour celle asymétrique. Les intervalles asymptotiques et percentiles calculés selon la méthode de Booth et al. (1994) ont généralement une bonne couverture bilatérale, par opposition aux intervalles de base construits selon le même algorithme. C'est pour l'intervalle de base que l'on tire particulièrement avantage à lisser la pseudo-population. Comme les intervalles asymptotiques et percentiles non lisses sont déjà satisfaisants, le fait de lisser n'a pour effet que de rendre ces deux types d'intervalles encore plus conservateurs, ce qui n'est pas forcément souhaitable. Quel que soit le type d'intervalle, il convient généralement mieux de sélectionner le paramètre de lissage via la sélection par bootstrap avec le critère des taux de couverture. C'est d'ailleurs dans ce but que le critère pour les intervalles de confiance a été introduit, suite à la performance moins reluisante constatée pour la sélection basée sur le critère de la stabilité de l'estimateur d'erreur quadratique moyenne.

Cette étude par simulation a été reproduite pour le second plan de sondage abordé dans ce mémoire, l'échantillonnage de Poisson, dans lequel le tirage des observations de l'échantillon se fait de manière indépendante, mais où les probabilités de sélection de premier ordre peuvent être inégales. À cet effet, nous avons adopté un plan proportionnel à la taille en obtenant la population finie à partir d'un modèle de régression. Celle-ci affichait une distribution symétrique et approximativement normale. Au regard des deux aspects

relatifs à la performance des algorithmes proposés, à savoir la stabilité de l'estimateur d'erreur quadratique moyenne bootstrap et le taux de couverture des intervalles de confiance bootstrap, les conclusions sont essentiellement identiques à celles portées envers la population finie symétrique dans le cas du plan EASSR. Notamment, la stabilité de l'estimateur d'erreur quadratique moyenne lisse avec un paramètre de lissage sélectionné par bootstrap est supérieure à celle de l'estimateur résultant de l'algorithme de Chauvet (2007).

En conclusion, cet ouvrage fournit des preuves empiriques des bénéfices du lissage de la pseudo-population dans la mise en œuvre des algorithmes de rééchantillonnage utilisés dans le contexte des sondages. Les avantages se situent surtout au niveau de la stabilité des estimateurs d'erreur quadratique moyenne en tant que mesure de précision de statistiques non lisses comme des quantiles échantillonnaux. Les expériences réalisées ont par ailleurs permis de nous doter d'une intuition quant aux circonstances favorisant le succès des méthodes lisses. Nous croyons toutefois qu'il serait important d'étayer ces connaissances empiriques au moyen d'arguments théoriques, bien que cela représenterait des défis techniques de taille.

Références bibliographiques

James G. Booth, Ronald W. Butler, et Peter Hall. Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89(428):1282–1289, 1994.

Sebastian Calonico, Matias D. Cattaneo, et Max H. Farrell. Coverage error optimal confidence intervals. *arXiv preprint arXiv:1808.01398*, 2018.

Arindam Chatterjee. Asymptotic properties of sample quantiles from a finite population. *Annals of the Institute of Statistical Mathematics*, 63(1):157–179, 2011.

Guillaume Chauvet. *Méthodes de bootstrap en population finie*. Thèse de doctorat, Université de Rennes 2, 2007.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Bradley Efron et Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London, 1993.

Carol A. Francisco et Wayne A. Fuller. Quantile estimation with a complex survey design. *The Annals of Statistics*, 19(1):454–469, 1991.

Peter Hall et Michael A. Martin. Exact convergence rate of bootstrap quantile variance estimator. *Probability Theory and Related Fields*, 80(2):261–268, 1988.

Peter Hall, Thomas J. DiCiccio, et Joseph P. Romano. On smoothing and the bootstrap. *The Annals of Statistics*, 17(2):692–704, 1989.

Cary T. Isaki et Wayne A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

Zeinab Mashreghi, David Haziza, et Christian Léger. A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10:1–52, 2016.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Carl E. Särndal, Bengt Swensson, et Jan Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New-York, 1992.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

Randy R. Sitter. Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20(2):135–154, 1992.

Randy R. Sitter et Changbao Wu. A note on Woodruff confidence intervals for quantiles. *Statistics & Probability Letters*, 52(4):353–358, 2001.

United States Census Bureau. U.S. Median Household Income Was $63,179 in 2018, Not Significally Different From 2017. "`https://www.census.gov/library/stories/2019/09/us-median-household-income-not-significantly-different-from-2017.html`", Septembre 2019. [En ligne; visité le 16-octobre-2019].

Ralph S. Woodruff. Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47(260):635–646, 1952.

Annexe A

Propriétés utiles de fonctions de densité de probabilité de lois connues

A.1. Loi normale

Soit X une variable aléatoire issue de la loi $\mathcal{N}(\mu,\sigma^2)$. La fonction de densité de probabilité de X est donnée par

$$\begin{aligned} f_0(x) &= \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \end{aligned}$$

où ϕ est la fonction de densité gaussienne standardisée. On peut vérifier que les dérivées première et seconde de f_0, étant utiles au calcul du paramètre de lissage optimal, sont les suivantes:

$$\begin{aligned} f_0'(x) &= \frac{1}{\sigma}\phi'\left(\frac{x-\mu}{\sigma}\right) \\ &= -f_0(x)\frac{x-\mu}{\sigma^2}, \\ f_0^{(2)}(x) &= \frac{1}{\sigma}\phi''\left(\frac{x-\mu}{\sigma}\right) \\ &= f_0(x)\frac{(x-\mu)^2-\sigma^2}{\sigma^4}. \end{aligned}$$

A.2. Loi log-normale

Soit X une variable aléatoire issue de la loi Lognormale(μ,σ^2). La fonction de densité de probabilité de X est donnée par

$$f_0(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}$$

Nous procédons aux calculs des dérivées première et seconde de f_0, étant utiles au calcul du paramètre de lissage optimal:

$$\begin{aligned}
f_0'(x) &= \frac{1}{\sigma\sqrt{2\pi}}\left[-\frac{1}{x^2}\exp -\frac{(\log x-\mu)^2}{2\sigma^2} - \frac{1}{x}\exp\left\{-\frac{(\log x-\mu)^2}{2\sigma^2}\right\}\frac{\log x-\mu}{\sigma^2}\frac{1}{x}\right] \\
&= -\frac{1}{x}f_0(x)\left[1+\frac{\log x-\mu}{\sigma^2}\right], \\
f_0^{(2)}(x) &= \frac{1}{x^2}f_0(x)\left[1+\frac{\log x-\mu}{\sigma^2}\right] - \frac{1}{x}f_0'(x)\left[1+\frac{\log x-\mu}{\sigma^2}\right] - \frac{1}{x}f_0(x)\left[\frac{1}{x\sigma^2}\right] \\
&= \frac{1}{x^2}f_0(x)\left\{\left[1+\frac{\log x-\mu}{\sigma^2}\right]^2 + \left[1+\frac{\log x-\mu}{\sigma^2}\right] - \frac{1}{\sigma^2}\right\}.
\end{aligned}$$