Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Une comparaison d’estimateurs non paramétriques pour les fonctions de répartition de populations finies

par Leo Pasquazzi et Lucio de Capitani

Date de diffusion : le 22 juin 2016

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 95-130
Statistique Canada, nº 12-001-X au catalogue

Une comparaison d'estimateurs non paramétriques pour les fonctions de répartition de populations finies

Leo Pasquazzi et Lucio de Capitani[1]

Résumé

Le présent travail a pour objet de comparer des estimateurs non paramétriques pour des fonctions de répartition de populations finies fondés sur deux types de valeurs prédites, à savoir celles données par l'estimateur bien connu de Kuo et une version modifiée de ces dernières, qui intègre une estimation non paramétrique de la fonction de régression à la moyenne. Pour chaque type de valeurs prédites, nous considérons l'estimateur fondé sur un modèle correspondant et, après incorporation des poids de sondage, l'estimateur par la différence généralisée. Nous montrons sous des conditions assez générales que le terme principal de l'erreur quadratique moyenne sous le modèle n'est pas affecté par la modification des valeurs prédites, même si cette modification réduit la vitesse de convergence pour le biais sous le modèle. Les termes d'ordre deux des erreurs quadratiques moyennes sous le modèle sont difficiles à obtenir et ne seront pas calculés dans le présent article. La question est de savoir si les valeurs prédites modifiées offrent un certain avantage du point de vue de l'approche fondée sur un modèle. Nous examinons aussi les propriétés des estimateurs sous le plan de sondage et proposons pour l'estimateur par la différence généralisée un estimateur de variance fondé sur les valeurs prédites modifiées. Enfin, nous effectuons une étude en simulation. Les résultats des simulations laissent entendre que les valeurs prédites modifiées entraînent une réduction importante de l'erreur quadratique moyenne si l'échantillon est de petite taille.

Mots-clés : Échantillonnage en population finie; estimateur de fonction de répartition; valeur prédite; estimateur de Kuo.

1 Introduction

Depuis la publication de l'article fondamental de Chambers et Dunstan (1986), plusieurs estimateurs ont été proposés pour les fonctions de répartition de populations finies. La plupart sont fondés sur différents types de valeurs prédites ou sur différents moyens de combiner ces valeurs en un estimateur. Ainsi, l'estimateur proposé par Chambers et Dunstan (1986) s'appuie sur des valeurs prédites tirées d'un modèle de superpopulation dans lequel le lien entre la variable étudiée et une variable auxiliaire est donné par un modèle de régression linéaire à composantes d'erreur indépendantes dont les variances sont supposées connues. En remplaçant les fonctions indicatrices non observées par les valeurs prédites dans la définition de la fonction de répartition de la population de la variable étudiée, on obtient l'estimateur de Chambers et Dunstan. Rao, Kovar et Mantel (1990) intègrent les poids de sondage dans les valeurs prédites de Chambers et Dunstan, puis utilisent celles-ci dans un estimateur par la différence généralisée. Kuo (1988) recourt à la régression non paramétrique pour estimer directement la relation de régression entre les fonctions indicatrices et la variable auxiliaire, et obtient des valeurs prédites qui admettent pratiquement n'importe quel modèle de superpopulation. Comme Chambers et Dunstan, elle remplace les fonctions indicatrices non observées par les valeurs prédites correspondantes et obtient un estimateur fondé sur un modèle. Chambers, Dorfman et Wehrly (1993) combinent les valeurs prédites de Chambers et Dunstan (1986) et de Kuo (1988) et proposent un autre estimateur fondé sur un modèle qui vise à être plus efficace que l'estimateur de Kuo si le modèle de superpopulation linéaire supposé par Chambers et Dunstan est vérifié, et qui ne souffre pas d'un biais de spécification incorrecte du modèle autrement. À la suite de ces premiers travaux, un assez grand nombre de propositions ont été faites en vue de réaliser un gain d'efficacité par rapport à l'estimateur

1. Leo Pasquazzi et Lucio de Capitani, Università degli Studi di Milano-Bicocca, Milan, Italie. Courriel : leo.pasquazzi@unimib.it, lucio.decapitani1@unimib.it.

de Horvitz-Thompson, tout en préservant la robustesse de ce dernier et parfois aussi l'une de ses propriétés souhaitables suivantes ou les deux, à savoir i) le fait qu'il s'agit d'une combinaison linéaire des fonctions indicatrices dans l'échantillon dont les coefficients ne dépendent pas de la variable étudiée et ii) le fait qu'il produit toujours des estimations non décroissantes pour la fonction de répartition.

Le présent travail part de l'idée d'améliorer les valeurs prédites proposées par Kuo (1988) en y incorporant une estimation de la fonction de régression à la moyenne (voir la section 2). Cette idée, avancée dans un ouvrage récent de Chambers et Clark (2012), repose sur l'hypothèse d'un modèle de superpopulation sous-jacent caractérisé par une relation de régression lisse entre la variable étudiée et une variable auxiliaire, ainsi qu'une variation lisse des distributions des composantes de l'erreur. Selon cette idée, les valeurs prédites sont le résultat d'une procédure en deux étapes : à la première étape, la fonction de régression à la moyenne est estimée par régression paramétrique ou non paramétrique, et à la deuxième étape, en utilisant les résidus de cette régression, les fonctions de répartition des composantes de l'erreur sont estimées par régression non paramétrique afin de tenir compte de la possibilité d'une variation lisse des distributions des composantes de l'erreur. En combinant les deux estimations, on peut calculer les valeurs prédites pour les fonctions indicatrices qui figurent dans la fonction de répartition de la population finie de la variable étudiée. Chambers et Clark (2012) analysent l'estimateur fondé sur un modèle obtenu en remplaçant les fonctions indicatrices non observées par les valeurs prédites correspondantes, et ils esquissent une preuve qui mène à une expression pour la variance sous le modèle de l'estimateur résultant. Dans cette preuve, ils supposent que la fonction de régression à la moyenne est estimée au moyen d'un estimateur convergent et que la contribution de son erreur d'estimation à la variance sous le modèle de l'estimateur de la fonction de répartition finale peut être négligée. Dans le présent travail, nous considérons la régression linéaire locale pour estimer à la fois la fonction de régression à la moyenne sous le modèle et les distributions des composantes de l'erreur. Nous donnons des développements asymptotiques pour le biais et pour la variance sous le modèle de l'estimateur résultant et les comparons à ceux correspondant à l'estimateur de Kuo fondé sur la régression linéaire locale. Il s'avère que les termes principaux dans les variances sous le modèle sont les mêmes et que, pour des suites de fenêtres de lissage choisies comme il convient, le carré du biais sous le modèle des deux estimateurs tend vers zéro plus rapidement que la variance sous le modèle. Pour établir quel estimateur est asymptotiquement plus efficace du point de vue de la modélisation, il est donc nécessaire de connaître les termes d'ordre deux des variances sous le modèle. Cependant, ces derniers dépendent d'hypothèses plus précises que celles considérées dans le présent travail et, du moins pour l'estimateur fondé sur les valeurs prédites modifiées, il semble que la détermination des termes d'ordre deux des variances sous le modèle ne soit pas une tâche facile. La question de savoir quel estimateur est le plus efficace du point de vue de la modélisation reste donc à résoudre.

En plus des estimateurs fondés sur un modèle susmentionnés, nous analysons les estimateurs par la différence généralisée fondés sur les deux types de valeurs prédites dans leurs versions pondérées selon le plan de sondage. Les résultats présentés à la section 3 montrent que les vitesses de convergence de leurs biais et de leurs variances sous le modèle sont les mêmes que pour leurs équivalents fondés sur un modèle. Les propriétés sous le plan de sondage sont discutées dans une certaine mesure à la section 4, de même que la question de l'estimation de la variance. Il serait évidemment intéressant d'établir et de comparer les développements asymptotiques pour les biais et les variances sous le plan de sondage. Breidt et Opsomer (2000) obtiennent sous des conditions faibles une expression générale pour le terme d'ordre un dans l'erreur quadratique moyenne sous le plan des estimateurs de régression par polynômes locaux, dont l'estimateur par la différence généralisée fondé sur les valeurs prédites de Kuo est un cas particulier. L'estimateur par la

différence généralisée fondé sur les valeurs prédites modifiées ne rentre toutefois pas dans cette classe. À l'instar de Särndal, Swensson et Wretman (1992), nous conjecturons que, sous des conditions générales, le terme d'ordre un de son erreur quadratique moyenne sous le plan est le même que celui de l'estimateur par la différence généralisée fondé sur les valeurs prédites de Kuo. Des preuves formelles pourraient peut-être être obtenues en adaptant et en étendant certains résultats présentés dans Wang et Opsomer (2011). Pour vérifier cette conjecture et comparer la performance de l'estimateur par la différence généralisée et de l'estimateur fondé sur un modèle dans diverses conditions, nous effectuons une étude en simulation dont les résultats sont présentés à la section 5.

2 Définition des estimateurs

Soit (y_i, x_i) les valeurs prises par une variable étudiée Y et une variable auxiliaire X sur l'unité i d'une population finie $U := \{1, 2, \ldots, N\}$. Supposons que

$$y_i = m(x_i) + \varepsilon_i, \qquad i \in U, \tag{2.1}$$

où $m(x)$ est une fonction lisse et où les ε_i sont des variables aléatoires indépendantes de moyenne nulle dont les fonctions de répartition $P(\varepsilon_i \le \varepsilon) = G(\varepsilon | x_i)$ varient continûment en fonction de x_i. Soit $s \subset U$ un échantillon tiré de la population U selon un certain plan de sondage. Comme d'habitude dans le contexte de l'information auxiliaire complète, nous supposons que les valeurs x_i sont connues pour toutes les unités de la population, tandis que les valeurs y_i sont observées uniquement pour les unités de la population qui appartiennent à l'échantillon s.

Pour estimer la fonction de répartition inconnue de la population

$$F_N(t) := \frac{1}{N} \sum_{i \in U} I(y_i \le t),$$

Kuo (1988) propose l'estimateur donné par

$$\hat{F}(t) := \frac{1}{N} \left(\sum_{j \in s} I\left(y_j \le t\right) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(y_j \le t\right) \right), \tag{2.2}$$

où, à la place de $w_{i,j}$, elle propose d'utiliser soit les poids de régression constants locaux

$$w_{i,j} := \frac{K\left(\frac{x_i - x_j}{\lambda}\right)}{\sum_{k \in s} K\left(\frac{x_i - x_k}{\lambda}\right)}$$

avec une fonction noyau (intégrable) à la place de $K(u)$ et $\lambda > 0$, soit les poids des k plus proches voisins

$$w_{i,j} := \begin{cases} 1/k, & \text{si } x_j \text{ est l'un des } k \text{ plus proches voisins de } x_i \\ 0, & \text{sinon.} \end{cases}$$

Notons que, dans la définition $\hat{F}(t)$,

$$\hat{G}_i(t) := \sum_{j \in s} w_{i,j} I\left(y_j \le t\right) \tag{2.3}$$

est utilisé comme valeur prédite remplaçant la fonction indicatrice non observée $I(y_i \le t)$ pour $i \notin s$.

En nous inspirant d'une idée avancée dans l'ouvrage de Chambers et Clark (2012), nous allons analyser un estimateur de $F_N(t)$ basé sur des valeurs prédites de rechange qui intègrent une estimation non paramétrique de la fonction de régression à la moyenne $m(x)$. Les valeurs prédites en question sont données par

$$\hat{G}_i^*(t) := \sum_{j \in s} w_{i,j} I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right) \tag{2.4}$$

où

$$\hat{m}_i := \sum_{k \in s} w_{i,j} y_j$$

est un estimateur non paramétrique de $m(x)$ à $x = x_i$, et l'estimateur résultant de $F_N(t)$ est donné par

$$\hat{F}^*(t) := \frac{1}{N}\left(\sum_{j \in s} I\left(y_j \le t\right) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right)\right). \tag{2.5}$$

Les valeurs prédites en (2.3) et (2.4), ou leurs versions modifiées de manière appropriée comprenant l'intégration des probabilités d'inclusion dans l'échantillon dans les poids de régression $w_{i,j}$, peuvent de toute évidence être calculées également pour $i \in s$, et elles peuvent être utilisées, par exemple, dans les estimateurs par la différence généralisée (Särndal et coll. 1992, page 221) ou dans les estimateurs calés sur un modèle (voir par exemple Wu et Sitter 2001; Chen et Wu 2002; Wu 2003; Montanari et Ranalli 2005; Rueda, Martínez, Martínez et Arcos 2007; Rueda, Sànchez-Borrego, Arcos et Martínez 2010). En plus des estimateurs fondés sur un modèle donné en (2.2) et (2.5), nous examinerons les estimateurs par la différence généralisée donnés par

$$\tilde{F}(t) := \frac{1}{N}\left(\sum_{i \in U} \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)\right) + \sum_{i \in s} \pi_i^{-1}\left(I\left(y_i \le t\right) - \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)\right)$$

et par

$$\tilde{F}^*(t) := \frac{1}{N}\left(\sum_{i \in U} \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right)\right) + \sum_{i \in s} \pi_i^{-1}\left(I\left(y_i \le t\right) - \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right)\right)$$

où π_i désigne les probabilités d'inclusion d'ordre un dans l'échantillon, $\tilde{w}_{i,j}$ désigne les poids de régression pondérés selon le plan de sondage dont la définition est donnée plus bas, et $\tilde{m}_i := \sum_{k \in s} \tilde{w}_{i,k} y_k$. Notons que, $\tilde{F}(t)$ et $\tilde{F}^*(t)$ sont fondés sur les équivalents des valeurs prédites pondérées selon le plan de sondage $\hat{G}_i(t)$ et $\hat{G}_i^*(t)$ qui sont donnés par

$$\tilde{G}_i(t) := \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j \le t\right)$$

et

$$\tilde{G}_i^*(t) := \sum_{j \in s} \tilde{w}_{i,j} I\left(y_j - \tilde{m}_j \le t - \tilde{m}_i\right),$$

respectivement.

Quant aux poids de régression $w_{i,j}$ et $\tilde{w}_{i,j}$, nous les remplaçons dans le présent travail par des poids de régression linéaires locaux. Dans la suite de l'exposé, $w_{i,j}$ et $\tilde{w}_{i,j}$, sont donc définis par

$$w_{i,j} := \frac{1}{n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{M_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) M_{1,s}(x_i)}{M_{2,s}(x_i) M_{0,s}(x_i) - M_{1,s}^2(x_i)}$$

et

$$\tilde{w}_{i,j} := \frac{1}{\pi_j n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{\tilde{M}_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) \tilde{M}_{1,s}(x_i)}{\tilde{M}_{2,s}(x_i) \tilde{M}_{0,s}(x_i) - \tilde{M}_{1,s}^2(x_i)},$$

où n est le nombre d'unités dans l'échantillon s,

$$M_{r,s}(x) := \sum_{k \in s} \frac{1}{n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2,$$

et

$$\tilde{M}_{r,s}(x) := \sum_{k \in s} \frac{1}{\pi_k n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2.$$

Il convient de souligner que les estimateurs non paramétriques de la présente section ne sont pas bien définis si les poids de régression $w_{i,j}$ et $\tilde{w}_{i,j}$, inclus dans leurs définitions ne sont pas bien définis. Ce problème se pose, par exemple, quand le support de la fonction noyau $K(u)$ est donné par l'intervalle [–1,1] (par exemple, noyau uniforme, noyau d'Epanechnikov), et quand il n'existe pas au moins deux $j \in s$ tels que $\left|x_i - x_j\right| < \lambda$. Pour contourner ce problème, on peut utiliser une fonction noyau dont le support correspond à la courbe réelle entière (par exemple, noyau gaussien) ou choisir la fenêtre de lissage de manière adaptative. La dernière solution peut aussi aboutir à des estimateurs plus efficaces (voir par exemple, Fan et Gijbels 1992). Pour ce qui est des estimateurs $\hat{F}^*(t)$ et $\tilde{F}^*(t)$ fondés sur les valeurs prédites modifiées, il convient en outre de noter que l'on pourrait en principe appliquer différentes fenêtres de lissage et (ou) différents poids de régression aux valeurs y_i et aux fonctions indicatrices. Par souci de simplicité, nous ne considérerons ici ni la sélection adaptative de la fenêtre de lissage ni la possibilité de différents poids de régression pour estimer la fonction de régression à la moyenne et les distributions des composantes de l'erreur.

Si l'on compare les définitions des estimateurs fondés sur les deux types de valeurs prédites, il saute aux yeux que $\hat{F}(t)$ et $\tilde{F}(t)$ sont plus faciles à calculer puisqu'il s'agit de combinaisons linéaires des fonctions indicatrices observées $I(y_j \le t)$. Les coefficients de ces combinaisons linéaires ne dépendent pas de la variable étudiée Y et peuvent par conséquent être utilisés pour estimer les moyennes d'autres fonctions que les fonctions indicatrices, ou de fonctions de plusieurs variables étudiées, en particulier quand il y a tout lieu de croire que ces dernières sont reliées à la variable auxiliaire X. Ce fait est particulièrement précieux pour les praticiens qui veulent que les estimations reliées à plusieurs variables étudiées soient cohérentes. Néanmoins, il existe aussi un argument puissant en faveur des estimateurs $\hat{F}^*(t)$ et $\tilde{F}^*(t)$ fondés sur les valeurs prédites modifiées : si $y_i = a + bx_i$ pour tout $i \in U$, il s'ensuit que $\hat{F}^*(t) = \tilde{F}^*(t) = F_N(t)$ pour chaque échantillon s tel que les estimateurs sont bien définis. On s'attendrait donc à ce que $\hat{F}^*(t)$ et $\tilde{F}^*(t)$ soient plus efficaces que $\hat{F}(t)$ et $\tilde{F}(t)$ quand il existe une forte relation de régression entre Y et X.

3 Propriétés sous le modèle

À la présente section, nous donnons des développements asymptotiques pour le biais et la variance sous le modèle des estimateurs présentés à la section précédente. Ces développements s'appuient sur les hypothèses suivantes :

(C1) $N \to \infty$ et les suites de valeurs x_i et de plans de sondage sont telles que

$$H_{N,s}(x) := \frac{1}{n}\sum_{i \in s} I(x_i \le x)$$

et

$$H_{N,\bar{s}}(x) := \frac{1}{N-n}\sum_{i \notin s} I(x_i \le x)$$

convergent vers des fonctions de répartition absolument continues $H_s(x) := \int_a^x h_s(z)\,dz$ et $H_{\bar{s}}(x) := \int_a^x h_{\bar{s}}(z)\,dz$ respectivement. Le support de $H_s(x)$ et $H_{\bar{s}}(x)$ est donné par un intervalle borné $[a,b]$ et les dérivées premières des fonctions de densité $h_s(x)$ et $h_{\bar{s}}(x)$ sont bornées pour $x \in (a,b)$. $h_s(x)$ possède une borne inférieure strictement positive.

(C2) La fonction noyau $K(u)$ est symétrique, a pour support $[-1,1]$ et possède une dérivée bornée pour $u \in (-1,1)$. La suite de fenêtres de lissage λ tend vers zéro suffisamment lentement pour que

$$\alpha := \max\left\{\sup_{x \in [a,b]} \left|H_{N,s}(x) - H_s(x)\right|, \sup_{x \in [a,b]} \left|H_{N,\bar{s}}(x) - H_{\bar{s}}(x)\right|\right\}$$

soit d'ordre $o(\lambda)$.

(C3) Les valeurs y_i de la population sont générées à partir du modèle (2.1). La fonction $m(x)$ est telle que

$$\left| m(x) - m(x_0) - m'(x_0)(x - x_0) - \frac{1}{2} m''(x_0)(x - x_0)^2 \right| \leq C \,|\, x - x_0 \,|^{2+\delta}$$

pour un certain $\delta > 0,$ et la famille des fonctions de répartition des composantes de l'erreur $G(\varepsilon \,|\, x)$ est telle que

$$\begin{aligned} &\left| G(\varepsilon \,|\, x) - G(\varepsilon_0 | x_0) - G^{(1,0)}(\varepsilon_0 | x_0)(\varepsilon - \varepsilon_0) - G^{(0,1)}(\varepsilon_0 | x_0)(x - x_0) \right. \\ &\left. - \frac{1}{2}\left(G^{(2,0)}(\varepsilon_0 | x_0)(\varepsilon - \varepsilon_0)^2 + 2G^{(1,1)}(\varepsilon_0 | x_0)(\varepsilon - \varepsilon_0)(x - x_0) + G^{(0,2)}(\varepsilon_0 | x_0)(x - x_0)^2 \right) \right| \\ &\quad \leq C\left(|\, \varepsilon - \varepsilon_0 \,|^{2+\delta} + |\, x - x_0 \,|^{2+\delta} \right) \end{aligned}$$

pour $C > 0$ et $\delta > 0,$ où

$$G^{(r,s)}(\varepsilon \,|\, x) := \partial^{r+s} G(\varepsilon \,|\, x) \big/ \left(\partial \varepsilon^r \partial x^s \right) \quad \text{pour} \quad r, s = 0, 1, 2.$$

L'hypothèse (C1) impose une contrainte sur la façon dont les valeurs x_i dans l'échantillon et hors de celui-ci sont générées. Conjuguée à l'hypothèse (C2), elle fait en sorte que les erreurs d'estimation des estimateurs à noyau de la densité pour $h_s(x)$ et $h_{\bar{s}}(x)$ tendent vers zéro uniformément pour $x \in [a + \lambda, b - \lambda]$ et qu'elles sont bornées uniformément pour $x \in [a, b].$ Le remplacement de (C1) par des hypothèses plus précises pourrait permettre de relâcher (C2) et d'accroître la vitesse de convergence uniforme pour l'erreur d'estimation des estimateurs à noyau de la densité (voir par exemple les résultats dans Hansen 2008). Enfin, l'hypothèse (C3) est nécessaire pour que les erreurs quadratiques moyennes des deux estimateurs sous le modèle convergent vers zéro. Elle peut être relâchée au prix d'une réduction des vitesses de convergence. En plus des hypothèses (C1) à (C3), nous aurons besoin de l'hypothèse (C4) qui suit pour nous assurer que les erreurs quadratiques moyennes des estimateurs par la différence généralisée sous le modèle tendent vers zéro :

(C4) Les probabilités d'inclusion d'ordre un dans l'échantillon sont données par

$$\pi_i := n^* \frac{\pi(x_i)}{\sum_{j \in U} \pi(x_j)}, \qquad i \in U,$$

où n^* est la taille d'échantillon espérée et $\pi(x)$ est une fonction dont la borne inférieure est strictement positive et qui possède une dérivée première bornée pour $x \in (a, b).$

Proposition 1. *Sous les hypothèses (C1) à (C3), il s'ensuit que :*

$$E\left(\hat{F}(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}(t-m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t-m(x)|x)m''(x) - 2G^{(1,1)}(t-m(x)|x)m'(x) + G^{(0,2)}(t-m(x)|x)\Big]h_{\bar{s}}(x)\,dx + o\left(\lambda^2\right)$$

et

$$\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) = \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t-m(x)|x)-G^2(t-m(x)|x)\Big]\Big[h_{\bar{s}}(x)/h_s(x)\Big]h_{\bar{s}}(x)\,dx + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t-m(x)|x)-G^2(t-m(x)|x)\Big]h_{\bar{s}}(x)\,dx + o\left(n^{-1}\right),$$

où $\mu_r := \int_{-1}^{-1} K(u)u^r du$ *pour* $r = 0,1,2.$

En ajoutant l'hypothèse (C4), on peut montrer que

$$E\left(\tilde{F}(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}(t-m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t-m(x)|x)m''(x) - 2G^{(1,1)}(t-m(x)|x)m'(x) + G^{(0,2)}(t-m(x)|x)\Big]h(x)\,dx + o\left(\lambda^2\right),$$

où

$$h(x) := h_{\bar{s}}(x) + \left(1-\pi^{-1}(x)\right)h_s(x),$$

et l'on peut montrer que

$$\operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) = \operatorname{var}\left(\hat{F}(t)-F_N(t)\right) + o\left(n^{-1}\right).$$

Proposition 2. *Sous les hypothèses (C1) à (C3) et en supposant que*

i) la fonction

$$\sigma^2(x) := \int_{-\infty}^{\infty} \varepsilon^2 dG(\varepsilon|x)$$

possède une dérivée première bornée pour $x \in (a,b)$,

ii)

$$\sup_{x\in[a,b]} \int_{-\infty}^{\infty} \varepsilon^4 dG(\varepsilon|x) < \infty,$$

on peut montrer que

$$E\left(\hat{F}^*(t)-F_N(t)\right) = \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h_{\bar{s}}(x)dx$$

$$+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h_{\bar{s}}(x)dx\right.$$

$$\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h_{\bar{s}}(x)dx\right]+o\left(\lambda^2+(n\lambda)^{-1}\right),$$

où $\kappa := \int_{-1}^{1}K^2(u)du$ *et* $\theta := \int_{-1}^{1}K(v)\int_{-1}^{1}K(u+v)K(u)dudv$, *et on peut montrer que*

$$\mathrm{var}\left(\hat{F}^*(t)-F_N(t)\right)=\mathrm{var}\left(\hat{F}(t)-F_N(t)\right)+o\left(n^{-1}+\lambda^5\right).$$

En ajoutant l'hypothèse (C4), on peut également montrer que

$$E\left(\tilde{F}^*(t)-F_N(t)\right) = \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h(x)dx$$

$$+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h(x)dx\right.$$

$$\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h(x)dx\right]$$

$$+o\left(\lambda^2+(n\lambda)^{-1}\right)$$

et que

$$\mathrm{var}\left(\tilde{F}^*(t)-F_N(t)\right)=\mathrm{var}\left(\hat{F}(t)-F_N(t)\right)+o\left(n^{-1}+\lambda^5\right).$$

Les preuves des propositions sont données en annexe. Dorfman et Hall (1993) ont dérivé des développements similaires pour l'estimateur de Kuo en utilisant des poids de régression locaux constants au lieu de linéaires.

Notons qu'étant donné les développements asymptotiques, il est possible de choisir des suites de fenêtres de lissage λ de manière à être certain que les carrés des biais de modèle soient d'un ordre de grandeur inférieur aux variances sous le modèle correspondantes. Pour les estimateurs fondés sur les valeurs prédites de Kuo, cette condition est réalisée quand $\lambda = o\left(n^{-1/4}\right)$, tandis que pour les estimateurs utilisant les valeurs prédites modifiées, cela exige que λ tende vers zéro plus rapidement que $O\left(n^{-1/4}\right)$ et plus lentement que $O\left(n^{-1/2}\right)$. Les vitesses de convergence pour les biais des derniers estimateurs sous le modèle sont optimisées quand $\lambda = O\left(n^{-1/3}\right)$ et, dans ce cas, les biais sous le modèle résultants sont tous deux d'ordre $O\left(n^{-2/3}\right)$. Pour les estimateurs fondés sur les valeurs prédites de Kuo, la convergence des biais sous le modèle peut être rendue plus rapide, en fonction des suites $H_{N,s}(x)$ et $H_{N,\bar{s}}(x)$ et de la suite de fenêtres de lissage λ.

Étant donné les considérations susmentionnées concernant les biais sous le modèle et vu que les termes principaux des variances sous le modèle sont les mêmes pour les deux types de valeurs prédites, il serait intéressant de connaître les termes d'ordre deux de ces variances afin d'établir quel estimateur est le plus efficace sous l'angle de l'approche fondée sur un modèle. Les preuves présentées en annexe font toutefois penser que les termes d'ordre deux dépendent d'hypothèses plus spécifiques que (C1) à (C3) et que, en particulier pour les estimateurs fondés sur les valeurs prédites modifiées, ils sont difficiles à déterminer.

4 Propriétés sous le plan de sondage

À la section précédente, nous avons montré que les estimateurs fondés sur le modèle $\hat{F}(t)$ et $\hat{F}^*(t)$ sont asymptotiquement sans biais sous le modèle et convergents en termes d'erreur quadratique moyenne sous le modèle. Cependant, ils ne sont pas sans biais sous le plan de sondage en général et ne devraient donc pas être utilisés quand les probabilités d'inclusion dans l'échantillon ne sont pas constantes. Dans ces cas, il convient de se servir des estimateurs par la différence généralisée $\tilde{F}(t)$ et $\tilde{F}^*(t)$. En fait, il découle des résultats présentés dans Breidt et Opsomer (2000) que, sous des conditions assez générales, $\tilde{F}(t)$ est asymptotiquement sans biais sous le plan de sondage et que son erreur quadratique moyenne sous le plan est donnée par

$$E_d\left(\left|\tilde{F}(t)-F_N(t)\right|^2\right)=\frac{1}{N^2}\sum_{i,j\in U}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\left[I(y_i\le t)-\bar{G}_i(t)\right]\left[I\left(y_j\le t\right)-\bar{G}_j(t)\right]+o\left(n^{-1}\right),$$

où $E_d(\cdot)$ désigne l'espérance par rapport au plan de sondage, $\pi_{i,j}$ désigne la probabilité d'inclusion conjointe des unités i et j dans l'échantillon (il est entendu que $\pi_{i,i}=\pi_i\big)$, et où

$$\bar{G}_i(t):=\sum_{j\in U}\bar{w}_{i,j}I\left(y_j\le t\right).$$

Les poids de régression $\bar{w}_{i,j}$ qui figurent dans la définition de $\bar{G}_i(t)$ s'appliquent à la population finie entière U et sont donnés par

$$\bar{w}_{i,j}:=\frac{1}{N\lambda}K\left(\frac{x_i-x_j}{\lambda}\right)\frac{\bar{M}_{2,s}(x_i)-\left(\frac{x_i-x_j}{\lambda}\right)\bar{M}_{1,s}(x_i)}{\bar{M}_{2,s}(x_i)\bar{M}_{0,s}(x_i)-\bar{M}_{1,s}^2(x_i)},$$

où

$$\bar{M}_{r,s}(x):=\sum_{k\in U}\frac{1}{N\lambda}K\left(\frac{x-x_k}{\lambda}\right)\left(\frac{x-x_k}{\lambda}\right)^r,\qquad r=0,1,2.$$

En outre, selon Breidt et Opsomer (2000),

$$\tilde{V}\left(\tilde{F}(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}\left[I\left(y_i \le t\right)-\tilde{G}_i(t)\right]\left[I\left(y_j \le t\right)-\tilde{G}_j(t)\right]$$

est un estimateur convergent pour l'erreur quadratique moyenne sous le plan de $\tilde{F}(t)$.

Malheureusement, on ne peut appliquer les résultats de Breidt et Opsomer (2000) à l'estimateur par la différence généralisée $\tilde{F}^*(t)$, puisque celui-ci ne rentre pas dans la classe des estimateurs de régression par polynômes locaux en raison de la présence des estimateurs des fonctions de régression $\tilde{m}_i$ et $\tilde{m}_j$ à l'intérieur des fonctions indicatrices dans les valeurs prédites $\tilde{G}_i^*(t)$. Cependant, les résultats pour $\tilde{F}(t)$ donnent à penser que, dans les grands échantillons, $\tilde{G}_i^*(t)$ et

$$\bar{G}_i^*(t) := \sum_{j\in U}\bar{w}_{i,j}I\left(y_j-\bar{m}_j \le t-\bar{m}_i\right),$$

où les $\bar{m}_i := \sum_{j\in U}\bar{w}_{i,j}y_j$, sont approximativement les mêmes et que

$$E_d\left(\left|\tilde{F}^*(t)-F_N(t)\right|^2\right) = \frac{1}{N^2}\sum_{i,j\in U}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\left[I\left(y_i \le t\right)-\bar{G}_i^*(t)\right]\left[I\left(y_j \le t\right)-\bar{G}_j^*(t)\right]+o\left(n^{-1}\right).$$

Partant de cette conjecture, nous avons testé

$$\tilde{V}\left(\tilde{F}^*(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}\left[I\left(y_i \le t\right)-\tilde{G}_i^*(t)\right]\left[I\left(y_j \le t\right)-\tilde{G}_j^*(t)\right]$$

comme estimateur pour l'erreur quadratique moyenne sous le plan de l'estimateur par la différence généralisée $\tilde{F}^*(t)$ dans l'étude en simulation décrite à la section suivante.

5 Étude en simulation

À la présente section, nous analysons certains résultats de simulation. Notre objectif est de comparer l'efficacité par rapport au plan de sondage des estimateurs des fonctions de répartition présentés à la section 2 et des estimateurs de la variance présentés à la section 4. Les résultats des simulations s'appliquent à l'échantillonnage aléatoire simple sans remise et à l'échantillonnage de Poisson avec probabilités d'inclusion inégales. À titre de référence, nous avons également inclus dans l'étude en simulation l'estimateur de la fonction de répartition de Horvitz-Thompson

$$\hat{F}_\pi(t) := \frac{1}{N}\sum_{j\in s}\pi_j^{-1}I\left(y_j \le t\right)$$

et l'estimateur de variance correspondant

$$\tilde{V}\left(\hat{F}_\pi(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}I\left(y_i \le t\right)I\left(y_j \le t\right).$$

Nous avons considéré des populations artificielles ainsi que réelles. Les premières ont été obtenues en générant $N = 1000$ valeurs x_i à partir de variables aléatoires i.i.d. de loi uniforme avec support sur l'intervalle $(0,1)$ et en les combinant avec trois types de fonction de régression $m(x)$ et deux types de composantes de l'erreur ε_i. Les fonctions de régression sont i) $m(x) = 0$ (uniforme), ii) $m(x) = 10x$ (linéaire) et iii) $m(x) = 10x^{1/4}$ (concave), tandis que les composantes de l'erreur ε_i sont soit des réalisations indépendantes tirées d'une loi t de Student unique à $\nu = 5$ dl, ou des réalisations indépendantes tirées de N lois t de Student non centrales décalées à $\nu = 5$ dl et avec paramètres de non-centralité donnés par $\mu = 15x_i$. Les décalages appliqués aux composantes de l'erreur dans le dernier cas font en sorte que les moyennes des lois t de Student non centrales à partir desquelles elles sont générées soient nulles. Les populations artificielles sont présentées aux figures 5.1 à 5.3. En ce qui concerne les populations réelles, nous avons pris la population $MU284$ de municipalités suédoises de Särndal et coll. (1992) (taille de la population $N = 284$) et considéré le logarithme naturel de $RMT85 =$ Revenus de l'imposition municipale de 1985 (en millions de couronnes) comme variable étudiée Y, et le logarithme naturel de $P85 =$ population de 1985 (en milliers) ou de $REV84 =$ valeurs immobilières selon les évaluations de 1984 (en millions de couronnes) comme variable auxiliaire X. Les populations réelles sont présentées à la figure 5.4.

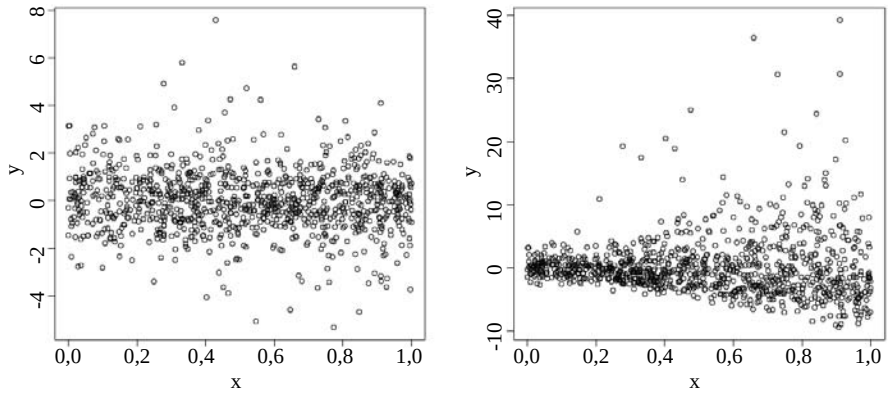

Figure 5.1 Populations générées à partir de $y_i = \varepsilon_i$, où $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$ (à gauche) et $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$ (à droite).

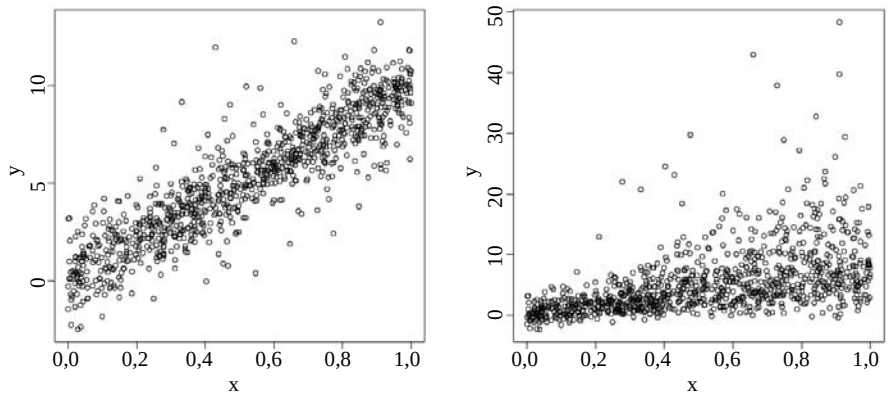

Figure 5.2 Populations générées à partir de $y_i = 10x_i + \varepsilon_i$, où $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$ (à gauche) et $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$ (à droite).

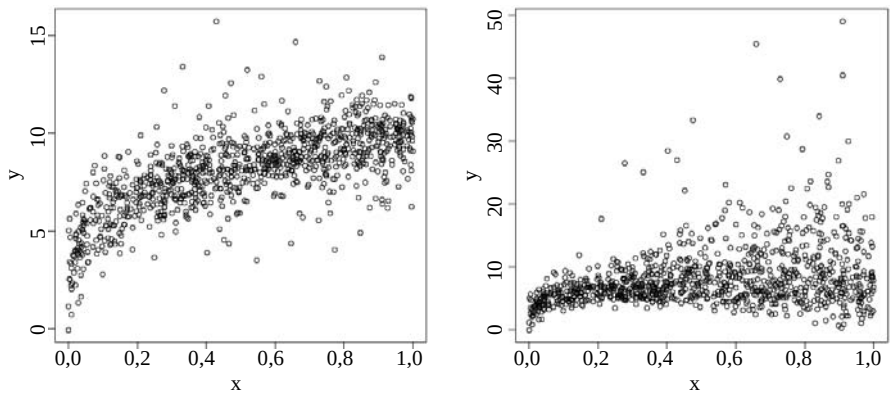

Figure 5.3 Populations générées à partir de $y_i = 10x_i^{1/4} + \varepsilon_i$, où $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$ (à gauche) et $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$ (à droite).

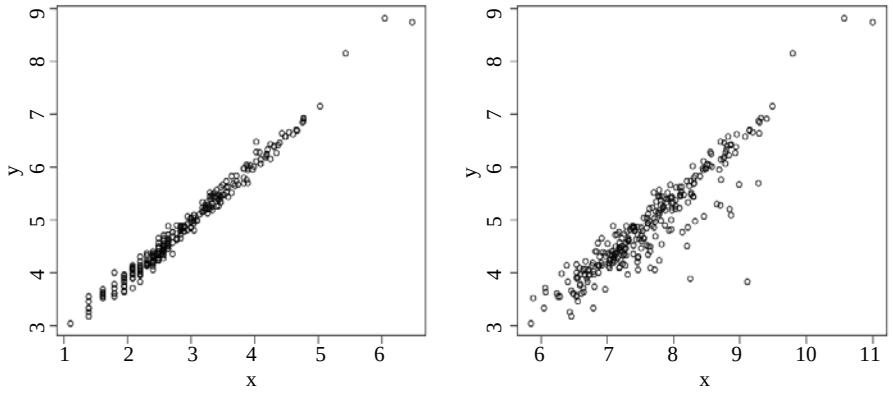

Figure 5.4 Population MU284 de municipalités suédoises de Särndal et coll. (1992). $y_i = \ln RMT85_i$ pour la i^e municipalité, et $x_i = \ln P85_i$ (à gauche) ou $x_i = \ln REV84_i$ (à droite).

Pour chaque population, nous avons sélectionné indépendamment $B = 1\,000$ échantillons. Pour le tirage d'échantillons à partir des populations artificielles, en cas d'échantillonnage aléatoire simple sans remise, nous avons fixé la taille d'échantillon à $n = 100,$ et en cas d'échantillonnage de Poisson, nous avons fixé la taille d'échantillon espérée à $n^* = 100$ et fait en sorte que les probabilités d'inclusion dans l'échantillon soient proportionnelles aux écarts-types des lois t de Student non centrales décalées susmentionnées. Pour le tirage d'échantillons dans les populations réelles, nous avons fixé la taille d'échantillon à $n = 30$ en cas d'échantillonnage aléatoire simple sans remise. Pour l'échantillonnage de Poisson, nous avons fixé la taille d'échantillon espérée à $n^* = 30$ et fait en sorte que les probabilités d'inclusion dans l'échantillon soient proportionnelles aux valeurs absolues des résidus des régressions linéaires par les moindres carrés des valeurs y_i de la population sur les valeurs x_i de la population.

Comme pour la définition des estimateurs non paramétriques, nous avons utilisé la fonction noyau d'Epanechnikov $K(u) := 0{,}75\left(1-u^2\right)$ avec $\lambda = 0{,}15$ ou $\lambda = 0{,}3$ pour les échantillons tirés des populations artificielles et la fonction noyau gaussienne $K(u) := 1/\sqrt{2\pi}e^{-(1/2)u^2}$ avec $\lambda = 1$ ou $\lambda = 2$ pour les échantillons tirés des populations réelles. Dans les tableaux présentant les résultats des simulations, les estimateurs non paramétriques correspondant aux petites et aux grandes valeurs de fenêtre de lissage sont désignés par un s (pour *small*) ou par un l (pour *large*), respectivement, dans l'indice inférieur. Nous avons recouru à la fonction noyau gaussienne pour les échantillons tirés des populations réelles afin d'éviter les problèmes de singularité qui se posent en cas de vides dans le jeu de valeurs x_i échantillonnées. De tels vides sont nettement plus susceptibles d'exister dans le cas des populations réelles que dans celui des populations artificielles, parce que les lois des variables auxiliaires sont asymétriques dans les premières. En fait, dans les populations artificielles, les estimateurs non paramétriques étaient bien définis pour chacun des $B = 1\,000$ échantillons sélectionnés selon le plan d'échantillonnage aléatoire simple sans remise. Pour le plan d'échantillonnage de Poisson, par contre, 47 des $B = 1\,000$ échantillons simulés étaient tels que les estimateurs non paramétriques avec la petite valeur de fenêtre de lissage n'ont pas pu être calculés et seulement un de ces échantillons était tel que les estimateurs non paramétriques avec la grande valeur de fenêtre de lissage étaient indéfinis. Les résultats des simulations s'appliquant aux estimateurs non paramétriques dans les tableaux 5.2 et 5.5 tiennent compte uniquement des échantillons pour lesquels les estimateurs étaient bien définis et sont donc fondés sur un peu moins que les $B = 1\,000$ réalisations.

Les tableaux 5.1 à 5.4 donnent le biais simulé (BIAIS) et la racine carrée de l'erreur quadratique moyenne simulée (REQM) pour chaque estimateur de la fonction de répartition à différents niveaux de t auxquels $F_N(t)$ a été estimée : en se basant, par exemple, sur les valeurs $\tilde{F}_b(t)$, $b = 1,2,\ldots,B,$ tirées de l'estimateur $\tilde{F}(t)$,

$$\text{BIAIS} := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right) \times 10\,000$$

et

$$\text{REQM} := \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2} \times 10\,000.$$

La REQM montre que les estimateurs fondés sur les valeurs prédites modifiées sont habituellement plus efficaces. Dans le cas de l'échantillonnage dans les populations réelles, l'augmentation des REQM est parfois assez grande. Comme prévu, les estimateurs fondés sur le modèle ont tendance à être plus efficaces que les estimateurs par la différence généralisée sous échantillonnage aléatoire simple sans remise quand les deux types d'estimateurs sont approximativement sans biais. Sous échantillonnage de Poisson, le BIAIS des estimateurs fondés sur le modèle augmente, mais demeure néanmoins concurrentiel. Une plus grande variabilité des probabilités d'inclusion dans l'échantillon modifierait certainement ce résultat, car elle augmenterait le BIAIS des estimateurs fondés sur le modèle. Les résultats des simulations ne doivent donc pas être considérés comme contredisant Johnson, Breidt et Opsomer (2008) qui se prononcent en faveur des estimateurs par la différence généralisée (appelés estimateurs assistés par modèle dans leur article), soutenant qu'il s'agit d'« un bon choix global pour les estimateurs de la fonction de répartition ».

Tableau 5.1
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\hat{F}_s(t)$	6	216	-3	433	31	512	23	434	12	207
$\hat{F}_l(t)$	15	219	10	430	0	502	-10	429	3	213
$\hat{F}_s^*(t)$	6	209	-30	411	22	484	22	414	3	200
$\hat{F}_l^*(t)$	15	214	-9	409	10	477	1	407	-10	207
$\tilde{F}_s(t)$	6	213	8	425	24	504	-4	430	8	207
$\tilde{F}_l(t)$	6	210	10	417	22	494	-8	422	6	206
$\tilde{F}_s^*(t)$	8	213	9	426	25	503	-5	432	5	206
$\tilde{F}_l^*(t)$	7	210	10	417	23	494	-6	424	4	206
$\tilde{F}_\pi(t)$	7	208	11	411	19	489	-5	417	6	200
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	26	225	33	376	8	477	26	419	33	209
$\hat{F}_l(t)$	52	236	23	374	-5	475	38	421	29	213
$\hat{F}_s^*(t)$	20	195	-29	351	-89	471	11	407	30	202
$\hat{F}_l^*(t)$	36	201	-11	357	-94	473	28	410	21	204
$\tilde{F}_s(t)$	8	211	11	370	-7	473	4	415	16	211
$\tilde{F}_l(t)$	5	208	8	367	-5	468	5	411	16	212
$\tilde{F}_s^*(t)$	11	210	11	372	-11	475	4	416	15	210
$\tilde{F}_l^*(t)$	7	208	11	368	-7	468	8	412	15	211
$\tilde{F}_\pi(t)$	1	211	1	391	-6	477	8	399	18	210
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	32	201	25	275	13	250	-14	264	-36	217
$\hat{F}_l(t)$	114	250	152	304	12	236	-180	312	-86	242
$\hat{F}_s^*(t)$	-50	165	12	226	51	216	26	230	13	172
$\hat{F}_l^*(t)$	-46	155	-14	199	69	195	23	211	17	156
$\tilde{F}_s(t)$	-5	186	4	275	15	248	11	269	-2	201
$\tilde{F}_l(t)$	-5	184	7	274	17	250	5	269	-2	196
$\tilde{F}_s^*(t)$	-10	180	5	275	16	245	14	266	-1	200
$\tilde{F}_l^*(t)$	-9	176	3	272	15	242	13	262	-1	194
$\tilde{F}_\pi(t)$	-7	203	14	413	37	472	17	405	1	206
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	24	204	23	351	27	403	26	382	29	208
$\hat{F}_l(t)$	94	242	135	372	51	392	13	380	15	212
$\hat{F}_s^*(t)$	55	182	-9	301	-18	368	-23	359	37	202
$\hat{F}_l^*(t)$	124	210	-31	278	-63	363	-8	356	48	200
$\tilde{F}_s(t)$	-2	194	-4	349	11	401	18	377	13	208
$\tilde{F}_l(t)$	-2	190	-5	345	12	398	17	374	11	209
$\tilde{F}_s^*(t)$	0	191	-5	352	14	401	20	376	13	207
$\tilde{F}_l^*(t)$	-1	189	-6	344	13	397	18	375	12	209
$\tilde{F}_\pi(t)$	-4	205	-5	401	21	470	24	401	14	207
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	81	207	44	316	17	384	-2	376	23	203
$\hat{F}_l(t)$	138	258	183	356	35	367	-50	374	8	208
$\hat{F}_s^*(t)$	7	146	-14	274	16	352	-8	358	15	197
$\hat{F}_l^*(t)$	9	144	10	246	-2	323	-18	339	24	186
$\tilde{F}_s(t)$	3	175	3	319	10	383	17	374	10	203
$\tilde{F}_l(t)$	0	178	5	316	11	380	17	370	8	202
$\tilde{F}_s^*(t)$	1	167	5	320	12	383	17	374	9	203
$\tilde{F}_l^*(t)$	-1	164	6	316	13	379	20	368	8	201
$\tilde{F}_\pi(t)$	4	209	11	412	25	477	27	422	10	200

Tableau 5.1 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	59	234	95	402	66	455	51	395	26	208
$\hat{F}_l(t)$	94	259	190	441	147	467	98	400	16	212
$\hat{F}_s^*(t)$	30	184	33	343	-123	435	-34	385	40	203
$\hat{F}_l^*(t)$	57	201	58	331	-148	437	2	382	34	203
$\tilde{F}_s(t)$	1	205	7	386	12	449	17	392	13	208
$\tilde{F}_l(t)$	-1	204	0	385	9	445	20	389	11	209
$\tilde{F}_s^*(t)$	3	201	8	389	7	449	13	392	14	207
$\tilde{F}_l^*(t)$	0	198	6	383	9	446	19	390	13	208
$\tilde{F}_\pi(t)$	0	205	-2	399	9	463	25	398	14	208

Tableau 5.2
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrales avec $\nu = 5$ dl et avec paramètres de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\hat{F}_s(t)$	-10	252	-11	593	-22	738	-20	743	6	357
$\hat{F}_l(t)$	-1	237	9	543	-15	621	-5	590	11	302
$\hat{F}_s^*(t)$	22	244	-29	485	-3	555	9	515	-17	297
$\hat{F}_l^*(t)$	14	238	-10	492	-5	564	14	524	-1	283
$\tilde{F}_s(t)$	-6	247	0	579	-27	724	-40	736	3	349
$\tilde{F}_l(t)$	-2	231	11	526	-1	598	-10	566	7	285
$\tilde{F}_s^*(t)$	23	248	23	505	-4	562	-27	531	-20	304
$\tilde{F}_l^*(t)$	12	240	20	504	1	573	-13	538	-6	287
$\tilde{F}_\pi(t)$	-6	220	-7	543	-37	741	-44	929	-48	1 058
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	17	164	30	411	4	749	14	590	15	190
$\hat{F}_l(t)$	47	173	19	383	-1	602	57	498	15	187
$\hat{F}_s^*(t)$	21	175	-7	378	-89	554	-11	473	3	192
$\hat{F}_l^*(t)$	29	152	-3	367	-99	555	27	481	3	184
$\tilde{F}_s(t)$	1	159	10	406	-11	737	-5	579	-2	194
$\tilde{F}_l(t)$	1	158	9	388	-5	586	14	482	-1	192
$\tilde{F}_s^*(t)$	14	186	27	409	-3	562	-17	487	-10	200
$\tilde{F}_l^*(t)$	3	160	22	399	-11	566	-5	482	-2	193
$\tilde{F}_\pi(t)$	-3	162	-7	451	-31	738	-29	980	-55	1 067
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	8	461	21	561	-12	259	-18	218	-30	164
$\hat{F}_l(t)$	78	429	183	451	2	248	-161	261	-79	189
$\hat{F}_s^*(t)$	-69	306	12	340	10	267	15	199	6	143
$\hat{F}_l^*(t)$	-59	294	4	302	56	205	15	172	17	124
$\tilde{F}_s(t)$	-25	441	4	560	-10	257	9	219	5	153
$\tilde{F}_l(t)$	-14	372	35	410	-10	262	4	219	5	151
$\tilde{F}_s^*(t)$	-31	333	-2	386	-29	294	4	227	-1	161
$\tilde{F}_l^*(t)$	-20	339	15	372	-10	259	11	215	4	151
$\tilde{F}_\pi(t)$	-15	385	3	746	-37	917	-35	1 004	-48	1 070

Tableau 5.2 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrales avec $\nu = 5$ dl et avec paramètres de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	-4	516	30	671	7	453	11	344	6	182
$\hat{F}_l(t)$	63	409	129	539	61	421	9	341	1	180
$\hat{F}_s^*(t)$	44	300	-29	433	-45	422	-47	345	12	180
$\hat{F}_l^*(t)$	107	314	-41	420	-60	397	-22	323	31	171
$\tilde{F}_s(t)$	-27	502	8	667	-8	450	0	344	-8	185
$\tilde{F}_l(t)$	-10	364	16	510	11	425	-2	345	-7	182
$\tilde{F}_s^*(t)$	-6	325	-9	479	-25	447	-14	356	-10	187
$\tilde{F}_l^*(t)$	-7	332	-9	489	-5	426	-3	344	-6	182
$\tilde{F}_\pi(t)$	-16	349	-2	705	-21	886	-42	1 013	-61	1 069
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\hat{F}_s(t)$	36	497	47	629	9	418	-11	320	15	191
$\hat{F}_l(t)$	56	393	186	490	43	383	-48	308	13	184
$\hat{F}_s^*(t)$	-29	276	-19	383	-18	380	-43	335	-1	204
$\hat{F}_l^*(t)$	-29	274	10	355	7	336	-29	290	23	179
$\tilde{F}_s(t)$	-30	475	12	630	4	421	7	317	6	191
$\tilde{F}_l(t)$	-42	336	31	452	11	390	8	312	8	186
$\tilde{F}_s^*(t)$	-31	306	5	429	-18	406	-14	344	-8	210
$\tilde{F}_l^*(t)$	-28	308	14	424	7	387	5	315	7	191
$\tilde{F}_\pi(t)$	-15	380	10	739	-23	891	-37	993	-47	1 064
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\hat{F}_s(t)$	24	308	69	687	53	690	38	406	2	188
$\hat{F}_l(t)$	47	301	131	553	139	561	91	393	-2	186
$\hat{F}_s^*(t)$	15	237	2	435	-135	513	-59	411	12	186
$\hat{F}_l^*(t)$	27	235	18	435	-149	506	-5	374	13	179
$\tilde{F}_s(t)$	-28	274	-8	673	4	688	3	403	-10	191
$\tilde{F}_l(t)$	-29	251	-12	512	17	541	7	395	-9	188
$\tilde{F}_s^*(t)$	-3	255	-12	481	-7	536	-20	422	-12	196
$\tilde{F}_l^*(t)$	-12	251	-16	489	2	538	-4	399	-9	189
$\tilde{F}_\pi(t)$	-10	267	-8	608	-4	860	-38	1 009	-63	1 066

Tableau 5.3
Populations réelles (taille de population $N = 284$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAISR**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\hat{F}_s(t)$	133	421	339	625	180	529	-265	490	-187	439
$\hat{F}_l(t)$	52	380	67	588	45	555	-63	469	-87	370
$\hat{F}_s^*(t)$	8	81	-154	203	90	130	62	123	6	54
$\hat{F}_l^*(t)$	28	66	-170	212	69	112	57	109	2	50
$\tilde{F}_s(t)$	-28	300	-24	497	8	483	-48	421	-38	319
$\tilde{F}_l(t)$	-28	326	-96	569	-52	544	3	466	1	319
$\tilde{F}_s^*(t)$	26	177	-11	302	0	244	1	308	-18	102
$\tilde{F}_l^*(t)$	29	179	-10	302	-2	243	-1	308	-21	104
$\tilde{F}_\pi(t)$	22	388	-10	771	9	864	5	731	-43	394

Tableau 5.3 (suite)
Populations réelles (taille de population $N = 284$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAISR**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\hat{F}_s(t)$	143	449	303	643	138	554	-217	543	-166	446
$\hat{F}_l(t)$	62	395	62	611	36	582	-49	519	-71	376
$\hat{F}_s^*(t)$	-11	204	-32	300	-101	328	42	285	31	155
$\hat{F}_l^*(t)$	36	183	-40	288	-149	345	6	261	34	122
$\tilde{F}_s(t)$	5	340	-22	548	4	557	-30	498	-23	332
$\tilde{F}_l(t)$	-2	349	-78	599	-36	588	10	522	8	331
$\tilde{F}_s^*(t)$	24	303	7	446	-6	494	2	439	-13	209
$\tilde{F}_l^*(t)$	29	304	4	443	-6	495	-1	432	-18	192
$\tilde{F}_\pi(t)$	34	395	1	766	16	880	9	744	-37	398

Tableau 5.4
Populations réelles (taille de population $N = 284$). BIAIS et REQM des estimateurs de la fonction de répartition sous échantillonnage de Poisson avec probabilités d'inclusion proportionnelles à la valeur absolue des résidus de la régression linéaire des valeurs y_i de la population sur les valeurs x_i de la population. Taille espérée $n^* = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAIS	**REQM**	**BIAIS**	**REQM**	**BIAISR**	**REQM**	**BIAIS**	**REQM**	**BIAIS**	**REQM**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\hat{F}_s(t)$	204	420	485	668	239	519	-412	626	-90	317
$\hat{F}_l(t)$	180	424	417	684	319	614	-239	548	-148	348
$\hat{F}_s^*(t)$	-41	97	-118	199	132	178	40	140	-71	104
$\hat{F}_l^*(t)$	11	70	-147	211	63	128	-25	122	-85	106
$\tilde{F}_s(t)$	24	360	30	649	0	675	-68	614	58	368
$\tilde{F}_l(t)$	9	390	-63	737	-64	774	-7	682	75	414
$\tilde{F}_s^*(t)$	16	184	-14	307	36	283	16	323	-11	103
$\tilde{F}_l^*(t)$	25	187	-15	312	30	286	14	328	-11	112
$\tilde{F}_\pi(t)$	40	445	73	1 983	12	2 498	-43	3 094	-49	3 341
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\hat{F}_s(t)$	349	660	1 185	1 373	890	1 059	458	654	-32	270
$\hat{F}_l(t)$	287	601	1 003	1 236	771	989	484	695	42	263
$\hat{F}_s^*(t)$	317	453	739	866	761	879	624	701	159	207
$\hat{F}_l^*(t)$	364	471	720	842	718	824	572	647	96	158
$\tilde{F}_s(t)$	35	488	82	818	-31	772	7	634	-8	326
$\tilde{F}_l(t)$	22	500	3	878	-98	852	40	704	27	354
$\tilde{F}_s^*(t)$	37	317	32	498	-13	513	32	412	7	157
$\tilde{F}_l^*(t)$	51	313	30	498	-30	518	12	411	-10	149
$\tilde{F}_\pi(t)$	32	671	19	1 658	-172	2 354	-173	2 787	-191	2 935

Considérons enfin les résultats des simulations concernant les estimateurs de variance de la section 4. Les tableaux 5.5 à 5.8 donnent le biais relatif (BIAISR) et la racine carrée de l'erreur quadratique moyenne relative (REQMR) pour chacun d'eux. Par exemple, selon les estimations de variance $\tilde{V}_b\left(\tilde{F}(t)\right)$, $b = 1, 2, \ldots, B$, obtenues au moyen de l'estimateur $\tilde{V}\left(\tilde{F}(t)\right)$,

$$\text{BIAISR} := \frac{1}{B}\sum_{b=1}^{B}\frac{\tilde{V}_b\left(\tilde{F}(t)\right) - V_B\left(\tilde{F}(t)\right)}{V_B\left(\tilde{F}(t)\right)} \times 10\,000$$

et

$$\text{REQMR} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{V}_b\left(\tilde{F}(t)\right) - V_B\left(\tilde{F}(t)\right)\right)^2}}{V_B\left(\tilde{F}(t)\right)} \times 10\,000$$

où

$$V_B\left(\tilde{F}(t)\right) := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2.$$

À titre de référence, nous donnons également les BIAISR et REQMR de l'estimateur

$$\tilde{V}\left(\tilde{F}_\pi(t)\right) := \frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j} - \pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j} I(y_i \le t) I\left(y_j \le t\right)$$

pour la variance de l'estimateur de Horvitz-Thompson.

Tableau 5.5
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-1 092	32 442	-1 249	3 895	-1 714	3 077	-1 536	3 828	-824	34 601
$\tilde{V}(\tilde{F}_l(t))$	-576	31 726	-603	3 838	-1 122	3 374	-951	3 758	-441	33 055
$\tilde{V}(\tilde{F}_s^*(t))$	-1 091	32 579	-1 292	3 914	-1 708	3 085	-1 640	3 828	-802	34 809
$\tilde{V}(\tilde{F}_l^*(t))$	-556	31 881	-622	3 857	-1 148	3 361	-1 025	3 749	-425	33 184
$\tilde{V}(\tilde{F}_\pi(t))$	42	30 952	57	3 928	-592	3 776	-287	3 825	551	33 462
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1 900	29 622	50	4 707	-917	3 557	-998	3 695	-1 480	29 417
$\tilde{V}(\tilde{F}_l(t))$	-1 359	29 623	535	4 572	-395	3 881	-527	3 736	-1 277	28 267
$\tilde{V}(\tilde{F}_s^*(t))$	-1 832	30 119	-101	4 710	-991	3 530	-1 077	3 704	-1 398	29 927
$\tilde{V}(\tilde{F}_l^*(t))$	-1 362	29 713	465	4 559	-420	3 865	-591	3 718	-1 236	28 489
$\tilde{V}(\tilde{F}_\pi(t))$	-351	29 132	1 096	4 215	-78	4 074	574	4 067	-638	29 507
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2 170	11 624	-1 027	2 480	-816	3 274	-1 424	2 583	-1 946	8 681
$\tilde{V}(\tilde{F}_l(t))$	-1 534	11 605	-529	2 632	-148	2 975	-859	2 590	-1 151	9 015
$\tilde{V}(\tilde{F}_s^*(t))$	-1 765	12 107	-1 108	2 529	-714	3 366	-1 318	2 660	-1 905	8 658
$\tilde{V}(\tilde{F}_l^*(t))$	-1 062	11 948	-671	2 735	-212	3 291	-762	2 785	-1 048	8 590
$\tilde{V}(\tilde{F}_\pi(t))$	254	31 545	-52	3 726	136	4 152	267	3 992	35	30 264
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1 642	25 809	-855	3 541	-1 076	3 038	-1 081	3 030	-1 361	21 157
$\tilde{V}(\tilde{F}_l(t))$	-950	25 692	-323	3 509	-597	3 312	-617	3 164	-1 124	20 231
$\tilde{V}(\tilde{F}_s^*(t))$	-1 385	26 406	-997	3 505	-1 089	3 045	-1 096	3 033	-1 310	21 393
$\tilde{V}(\tilde{F}_l^*(t))$	-832	26 212	-292	3 556	-614	3 317	-716	3 154	-1 135	20 286
$\tilde{V}(\tilde{F}_\pi(t))$	105	29 621	507	3 857	209	4 244	425	3 910	-337	29 082

Tableau 5.5 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2 465	30 612	-1 121	4 594	-1 512	3 183	-1 958	3 076	-863	19 720
$\tilde{V}(\tilde{F}_l(t))$	-1 780	28 103	-663	4 420	-1 092	3 319	-1 491	3 140	-439	18 985
$\tilde{V}(\tilde{F}_s^*(t))$	-2 052	33 980	-1 150	4 619	-1 537	3 217	-1 948	3 127	-954	19 637
$\tilde{V}(\tilde{F}_l^*(t))$	-1 194	33 573	-691	4 472	-1 124	3 368	-1 438	3 228	-357	19 245
$\tilde{V}(\tilde{F}_\pi(t))$	-81	30 001	9	3 756	-110	3 996	-598	3 661	440	32 455
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1 873	29 437	-758	3 759	-621	3 476	-709	3 599	-1 298	27 679
$\tilde{V}(\tilde{F}_l(t))$	-1 267	28 511	-284	3 661	-131	3 758	-321	3 552	-1 075	26 790
$\tilde{V}(\tilde{F}_s^*(t))$	-1 710	30 670	-928	3 741	-628	3 510	-777	3 603	-1 245	27 972
$\tilde{V}(\tilde{F}_l^*(t))$	-939	30 486	-270	3 764	-171	3 803	-375	3 581	-1 014	26 926
$\tilde{V}(\tilde{F}_\pi(t))$	178	29 640	599	3 816	533	4 324	590	3 874	-404	28 917

Tableau 5.6
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrale avec $\nu = 5$ dl et avec paramètre de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student centrale avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-3 306	65 777	-4 248	8 032	-5 093	4 242	-6 258	4 844	-5 652	32 037
$\tilde{V}(\tilde{F}_l(t))$	-2 048	47 035	-2 656	4 705	-2 434	3 116	-3 310	3 939	-3 092	29 380
$\tilde{V}(\tilde{F}_s^*(t))$	-3 362	36 855	-2 488	4 409	-1 910	3 147	-2 869	3 910	-4 329	23 247
$\tilde{V}(\tilde{F}_l^*(t))$	-2 696	39 509	-2 076	4 450	-1 768	3 163	-2 648	3 811	-3 244	26 343
$\tilde{V}(\tilde{F}_\pi(t))$	113	129 637	259	15 120	618	6 327	193	5 429	273	6 097
$y_i = \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-740	125 975	-2 522	14 864	-5 466	3 658	-4 896	6 691	-1 551	83 262
$\tilde{V}(\tilde{F}_l(t))$	-391	83 047	-1 503	8 946	-2 428	4 099	-2 228	5 526	-1 154	54 680
$\tilde{V}(\tilde{F}_s^*(t))$	-3 260	58 072	-2 649	7 661	-2 260	3 936	-2 795	5 011	-2 116	48 739
$\tilde{V}(\tilde{F}_l^*(t))$	-716	77 935	-2 000	7 979	-1 934	4 235	-2 279	5 243	-1 243	52 531
$\tilde{V}(\tilde{F}_\pi(t))$	666	251 134	-564	26 553	-87	7 344	-2	6 029	407	6 610
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-6 801	7 898	-6 470	4 281	-1 059	22 596	-398	32 401	-1 650	72 632
$\tilde{V}(\tilde{F}_l(t))$	-4 978	5 826	-2 898	4 473	-603	9 530	206	15 226	-1 157	40 466
$\tilde{V}(\tilde{F}_s^*(t))$	-4 520	6 691	-2 710	4 213	-3 245	6 723	-1 156	12 681	-2 458	32 907
$\tilde{V}(\tilde{F}_l^*(t))$	-4 226	6 206	-1 674	5 062	-978	7 874	55	12 781	-1 283	33 737
$\tilde{V}(\tilde{F}_\pi(t))$	-707	47 550	118	7 214	609	4 409	743	4 628	435	4 800

Tableau 5.6 (suite)
Populations artificielles (taille de population $N = 1\,000$). BIAISR et REQMR des estimateurs de variance sous échantillonnage de Poisson avec probabilités d'inclusion dans l'échantillon π_i proportionnelles aux écarts-types des lois t de Student non centrale avec $\nu = 5$ dl et avec paramètre de non-centralité $\mu = 15x_i$. Taille espérée d'échantillon $n^* = 100$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
$y_i = 10x_i + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-7 398	8 847	-6 235	3 667	-2 493	8 171	-1 051	16 299	-1 440	71 943
$\tilde{V}(\tilde{F}_l(t))$	-4 548	9 463	-3 136	3 282	-1 187	4 246	-832	7 638	-982	45 182
$\tilde{V}(\tilde{F}_s^*(t))$	-3 902	11 727	-2 808	3 409	-2 411	3 501	-1 721	6 737	-1 671	41 389
$\tilde{V}(\tilde{F}_l^*(t))$	-3 598	10 771	-2 610	3 462	-1 284	3 988	-852	7 008	-972	43 017
$\tilde{V}(\tilde{F}_\pi(t))$	146	57 044	-42	8 708	520	4 784	214	4 686	390	5 085
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ i.i.d. t de Student avec $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-7 731	8 568	-6 597	3 484	-2 442	7 775	-903	16 067	-1 967	56 480
$\tilde{V}(\tilde{F}_l(t))$	-4 611	9 378	-2 990	3 252	-874	4 119	-347	7 420	-1 310	35 051
$\tilde{V}(\tilde{F}_s^*(t))$	-4 747	11 909	-2 679	3 298	-1 896	3 272	-2 248	5 747	-3 382	27 222
$\tilde{V}(\tilde{F}_l^*(t))$	-4 223	10 380	-2 100	3 494	-788	3 731	-550	5 975	-1 795	29 856
$\tilde{V}(\tilde{F}_\pi(t))$	-428	47 038	-206	7 350	641	4 504	738	4 708	487	4 943
$y_i = 10x_i^{1/4} + \varepsilon_i$, avec $\varepsilon_i \sim$ indép. t de Student non centrale avec $\nu = 5$ et $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-4 936	40 696	-6 111	4 579	-5 549	4 035	-1 864	14 381	-1 509	84 892
$\tilde{V}(\tilde{F}_l(t))$	-3 004	29 404	-2 764	3 962	-2 436	3 606	-1 234	7 357	-1 103	53 875
$\tilde{V}(\tilde{F}_s^*(t))$	-4 328	27 704	-2 516	4 235	-2 671	3 332	-2 586	5 955	-1 939	47 601
$\tilde{V}(\tilde{F}_l^*(t))$	-3 454	28 267	-2 263	4 160	-2 329	3 574	-1 433	6 682	-1 171	50 985
$\tilde{V}(\tilde{F}_\pi(t))$	152	98 607	663	12 879	15	5 376	20	5 080	429	5 619

Tableau 5.7
Populations réelles (taille de population $N = 284$). BIAISR et REQMR des estimateurs de variance sous échantillonnage aléatoire simple sans remise. Taille d'échantillon $n = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\tilde{V}(\tilde{F}_s(t))$	-2 853	16 809	-1 700	3 037	-1 554	2 984	-1 100	4 633	-5 503	16 257
$\tilde{V}(\tilde{F}_l(t))$	-1 110	16 374	-1 827	2 760	-1 683	2 847	-927	4 387	-3 016	18 685
$\tilde{V}(\tilde{F}_s^*(t))$	-1 043	19 081	-91	7 728	-448	9 120	-484	7 715	-1 877	65 298
$\tilde{V}(\tilde{F}_l^*(t))$	-424	18 971	104	7 819	-382	9 110	-301	7 799	-1 058	62 968
$\tilde{V}(\tilde{F}_\pi(t))$	-186	29 720	-603	3 901	31	3 971	500	4 383	-74	28 418
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\tilde{V}(\tilde{F}_s(t))$	-2 283	16 303	-1 450	3 538	-945	3 526	-1 071	4 300	-4 832	19 401
$\tilde{V}(\tilde{F}_l(t))$	-1 095	16 755	-1 427	3 181	-938	3 390	-780	4 051	-2 753	20 551
$\tilde{V}(\tilde{F}_s^*(t))$	-1 737	14 642	-298	5 648	-546	5 282	-736	5 679	-3 564	38 344
$\tilde{V}(\tilde{F}_l^*(t))$	-1 174	14 111	-27	5 856	-422	5 452	-228	5 974	-1 433	43 923
$\tilde{V}(\tilde{F}_\pi(t))$	-307	28 421	-460	3 963	-344	3 850	112	4 235	-401	27 987

Tableau 5.8
Populations réelles (taille de population $N = 284$). BIAISR et REQMR des estimateurs de variance sous échantillonnage de Poisson avec probabilités d'inclusion proportionnelles à la valeur absolue des résidus de la régression linéaire des valeurs y_i de la population sur les valeurs x_i de la population. Taille espérée $n^* = 30$

	$t = F_N^{-1}(0{,}05)$		$t = F_N^{-1}(0{,}25)$		$t = F_N^{-1}(0{,}50)$		$t = F_N^{-1}(0{,}75)$		$t = F_N^{-1}(0{,}95)$	
	BIAISR	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**	**BIAISR**	**REQMR**
Population MU284 avec $Y = \ln RMT85$ et $X = \ln P85$										
$\tilde{V}(\tilde{F}_s(t))$	-3 502	26 342	-1 841	14 037	-2 691	12 087	-3 415	9 674	-5 932	26 823
$\tilde{V}(\tilde{F}_l(t))$	-2 159	27 610	-1 782	14 010	-2 840	12 002	-3 186	10 177	-4 455	26 802
$\tilde{V}(\tilde{F}_s^*(t))$	-434	22 455	515	15 503	-506	31 296	-1 460	23 496	-2 649	78 527
$\tilde{V}(\tilde{F}_l^*(t))$	-80	22 921	677	15 575	-280	33 294	-1 283	26 612	-1 597	72 166
$\tilde{V}(\tilde{F}_\pi(t))$	-294	361 991	522	75 891	43	48 764	-241	36 354	90	32 354
Population MU284 avec $Y = \ln RMT85$ et $X = \ln REV84$										
$\tilde{V}(\tilde{F}_s(t))$	-5 220	18 699	-3 667	8 749	-3 222	7 537	-3 018	9 279	-4 955	44 597
$\tilde{V}(\tilde{F}_l(t))$	-4 254	20 765	-3 100	9 180	-3 435	7 231	-3 196	8 540	-3 461	43 206
$\tilde{V}(\tilde{F}_s^*(t))$	-2 938	18 922	-1 110	11 828	-1 265	8 726	-1 040	10 963	-3 682	89 262
$\tilde{V}(\tilde{F}_l^*(t))$	-1 938	19 997	-699	12 641	-1 003	9 305	-599	11 545	-1 558	98 798
$\tilde{V}(\tilde{F}_\pi(t))$	-143	128 401	493	33 934	-255	18 473	-91	17 904	327	16 463

Comme le montrent les résultats des simulations, les estimateurs de variance souffrent d'une grande variabilité. Ce problème touche aussi l'estimateur de variance pour l'estimateur de Horvitz-Thompson qui, à l'occasion, présente de très grandes REQMR. Il est en outre intéressant de noter que, si le BIAISR des estimateurs de variance pour les estimateurs par la différence généralisée est presque toujours négatif et parfois assez grand en valeur absolue, celui de l'estimateur de variance pour l'estimateur de Horvitz-Thompson est positif dans la plupart des cas considérés.

Remerciements

La présente étude a été financée en partie par la subvention FAR 2014-ATE-0200 octroyée par *University of Milano-Bicocca*.

Annexe

Soit β une suite de nombres réels. Tout au long de la présente annexe, nous désignerons par $O_{i_1,i_2,\ldots,i_k}(\beta)$ les termes de reste qui peuvent dépendre de $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ et qui sont de même ordre que la suite β uniformément pour $i_1, i_2, \ldots, i_k \in U$. Formellement, $R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) = O_{i_1,i_2,\ldots,i_k}(\beta)$ si

$$\sup_{i_1,i_2,\ldots,i_k \in U} \left| R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) \right| = O(\beta).$$

En outre, pour simplifier la notation, nous écrirons m_i à la place de $m(x_i)$ et σ_i^2 à la place de $\sigma^2(x_i)$.

Biais de l'estimateur fondé sur le modèle de Kuo

$$
\begin{aligned}
E\left(\hat{F}(t)-F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[I\left(\varepsilon_j \le t-m_j\right)-I\left(\varepsilon_i \le t-m_i\right)\right]\right)\\
&= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t-m_j|x_j\right)-G\left(t-m_i|x_i\right)\right]\\
&= \frac{1}{2N}\sum_{i\notin s}\Big[G^{(2,0)}\left(t-m_i|x_i\right)\left(m_i'\right)^2-G^{(1,0)}\left(t-m_i|x_i\right)m_i''\\
&\qquad -2G^{(1,1)}\left(t-m_i|x_i\right)m_i'+G^{(0,2)}\left(t-m_i|x_i\right)\Big]\sum_{j\in s} w_{i,j}\left(x_j-x_i\right)^2+o\left(\lambda^2\right)\\
&= \lambda^2\frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b\Big[G^{(2,0)}\left(t-m(x)|x\right)\left(m'(x)\right)^2-G^{(1,0)}\left(t-m(x)|x\right)m''(x)\\
&\qquad -2G^{(1,1)}\left(t-m(x)|x\right)m'(x)+G^{(0,2)}\left(t-m(x)|x\right)\Big]h_{\bar{s}}(x)\,dx+o\left(\lambda^2\right).
\end{aligned}
$$

Biais de l'estimateur par la différence généralisée de Kuo

Écrivons

$$
\begin{aligned}
\tilde{F}(t)-F_N(t) &= \frac{1}{N}\Bigg\{\sum_{i\notin s}\sum_{j\in s}\tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t-m_j\right)-I\left(\varepsilon_i \le t-m_i\right)\right]\\
&\qquad +\sum_{i\in s}\left(1-\frac{1}{\pi_i}\right)\sum_{j\in s}\tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t-m_j\right)-I\left(\varepsilon_i \le t-m_i\right)\right]\Bigg\}.
\end{aligned}
$$

Des étapes similaires à celles suivies pour $\hat{F}(t)$ montrent que

$$
\begin{aligned}
E\left(\tilde{F}(t)-F_N(t)\right) &= \lambda^2\frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b\Big[G^{(2,0)}\left(t-m(x)|x\right)\left(m'(x)\right)^2-G^{(1,0)}\left(t-m(x)|x\right)m''(x)\\
&\qquad -2G^{(1,1)}\left(t-m(x)|x\right)m'(x)+G^{(0,2)}\left(t-m(x)|x\right)\Big]h(x)\,dx+o\left(\lambda^2\right),
\end{aligned}
$$

où

$$
h(x) := h_{\bar{s}}(x)+\left(1-\pi^{-1}(x)\right)h_s(x).
$$

Variance de l'estimateur fondé sur le modèle de Kuo

$$
\begin{aligned}
\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}I\left(\varepsilon_j \le t-m_j\right)-\frac{1}{N}\sum_{i\notin s}I\left(y_i \le t\right)\right)\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\\
&\quad +\frac{1}{N^2}\sum_{i\notin s}\left[G\left(t-m_i|x_i\right)-G^2\left(t-m_i|x_i\right)\right]\\
&= A_1+A_2,
\end{aligned}
$$

où

$$\begin{aligned} A_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right] \\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\left(\sum_{i\notin s} w_{i,j}\right)^2 \\ &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right)-G^2\left(t-m(x)|x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\ &\quad + O\left((n\lambda)^{-1}\alpha\right) \end{aligned}$$

et

$$\begin{aligned} A_2 &:= \frac{1}{N^2}\sum_{i\notin s}\left[G\left(t-m_i|x_i\right)-G^2\left(t-m_i|x_i\right)\right] \\ &= \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|x\right)-G^2\left(t-m(x)|x\right)\right]h_{\bar{s}}(x)\,dx+O\left(n^{-1}\alpha\right). \end{aligned}$$

Donc,

$$\begin{aligned} \operatorname{var}\left(\hat{F}(t)-F_N(t)\right) &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|\,x\right)-G^2\left(t-m(x)|\,x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\ &\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)|\,x\right)-G^2\left(t-m(x)|\,x\right)\right]h_{\bar{s}}(x)\,dx+O\left((n\lambda)^{-1}\alpha\right). \end{aligned}$$

Variance de l'estimateur par la différence généralisée de Kuo

Notons que,

$$\tilde{F}(t)-F_N(t)=\frac{1}{N}\left\{\sum_{j\in s} I\left(y_j\le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)+\left(\pi_j^{-1}-1\right)\right]-\sum_{i\notin s} I\left(y_i\le t\right)\right\}$$

de sorte que

$$\begin{aligned} \operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j\le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right) \\ &\quad + \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s} I\left(y_i\le t\right)\right) \\ &= B_1+A_2, \end{aligned}$$

où le terme A_2 est le même que dans la variance de $\hat{F}(t)$, et où

$$
\begin{aligned}
B_1 &:= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j \le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j} + \left(\pi_j^{-1}-1\right) - \sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right) \\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j \mid x_j\right) - G^2\left(t-m_j \mid x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j} + \left(\pi_j^{-1}-1\right) - \sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]^2 \\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j \mid x_j\right) - G^2\left(t-m_j \mid x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j} + \left(\pi_j^{-1}-1\right)\left(1-\sum_{i\in s}\tilde{w}_{i,j}\right)\right]^2 + O\left(\lambda n^{-1}\right) \\
&= \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G\left(t-m(x) \mid x\right) - G^2\left(t-m(x) \mid x\right)\right]\left[h_{\tilde{s}}(x)/h_s(x)\right]h_{\tilde{s}}(x)\,dx \\
&\quad + O\left((n\lambda)^{-1}\alpha + \lambda n^{-1}\right) \\
&= A_1 + O\left((n\lambda)^{-1}\alpha + \lambda n^{-1}\right).
\end{aligned}
$$

Donc,

$$
\operatorname{var}\left(\tilde{F}(t) - F_N(t)\right) = \operatorname{var}\left(\hat{F}(t) - F_N(t)\right) + O\left((n\lambda)^{-1}\alpha + \lambda n^{-1}\right).
$$

Biais de l'estimateur fondé sur le modèle avec valeurs prédites modifiées

Soit $\hat{\tilde{m}}_i := \sum_{k\in s} w_{i,k} m_k$, $c_{i,j} := 1 - w_{j,j} + w_{i,j}$ et

$$
d_{i,j} := \frac{1}{c_{i,j}}\left[\left(1-c_{i,j}\right)\left(t-m_i\right) + \left(\hat{\tilde{m}}_j - m_j\right) - \left(\hat{\tilde{m}}_i - m_i\right) + \sum_{k\in s, k\ne j}\left(w_{j,k} - w_{i,k}\right)\varepsilon_k\right].
$$

Observons que $w_{i,j} = O_{i,j}\left((n\lambda)^{-1}\right)$ d'où

$$
y_j - \hat{\tilde{m}}_j \le t - \hat{\tilde{m}}_i
$$

est (asymptotiquement, aussitôt que $c_{i,j} > 0$) équivalent à

$$
\varepsilon_j \le t - m_i + d_{i,j}.
$$

Comme $d_{i,j}$ ne dépend pas de ε_j, il s'ensuit que

$$
\begin{aligned}
E\left(I\left(y_j - \hat{\tilde{m}}_j \le t - \hat{\tilde{m}}_i\right)\right) &= E\left(I\left(\varepsilon_j \le t - m_i + d_{i,j}\right)\right) \\
&= E\left(E\left(I\left(\varepsilon_j \le t - m_i + d_{i,j}\right) \mid \varepsilon_k, k \ne j\right)\right) \qquad \text{(A.1)} \\
&= E\left(G\left(t - m_i + d_{i,j} \mid x_j\right)\right).
\end{aligned}
$$

Or, en utilisant le fait que

$$d_{i,j} = (1-c_{i,j})(t-m_i) + (\hat{\hat{m}}_j - m_j) - (\hat{\hat{m}}_i - m_i) + \sum_{k\in s, k\neq j} (w_{j,k} - w_{i,k})\varepsilon_k + R(d_{i,j}), \tag{A.2}$$

où

$$E^{1/4}\left(|R(d_{i,j})|^4\right) = O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right), \tag{A.3}$$

on voit en examinant (A.1) que

$$\begin{aligned} E\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right)\right) &= E\left(G\left(t - m_i + d_{i,j}\right)\middle| x_j\right) \\ &= G\left(t - m_i \middle| x_j\right) + G^{(1,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}\right) \\ &\quad + \frac{1}{2} G^{(2,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}^2\right) + o_{i,j}\left(\lambda^4 + (n\lambda)^{-1}\right). \end{aligned} \tag{A.4}$$

Donc,

$$\begin{aligned} E\left(\hat{F}^*(t) - F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left(I\left(y_j - \hat{m}_j \le t - \hat{m}_i\right) - I\left(y_i \le t\right)\right)\right) \\ &= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_i \middle| x_j\right) - G\left(t - m_i \middle| x_i\right)\right] \\ &\quad + \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}\right) \\ &\quad + \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(2,0)}\left(t - m_i \middle| x_j\right) E\left(d_{i,j}^2\right) + o\left(\lambda^4 + (n\lambda)^{-1}\right) \\ &:= C_1 + C_2 + C_3 + o\left(\lambda^4 + (n\lambda)^{-1}\right). \end{aligned} \tag{A.5}$$

Considérons d'abord C_1 et notons que

$$\begin{aligned} C_1 &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_i \middle| x_j\right) - G\left(t - m_i \middle| x_i\right)\right] \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(0,2)}\left(t - m_i \middle| x_i\right)\sum_{j\in s} w_{i,j}\left(x_j - x_i\right)^2 + o\left(\lambda^2\right) \\ &= \lambda^2 \frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}\left(t - m(x) \middle| x\right) h_{\bar{s}}(x)\,dx + o\left(\lambda^2\right). \end{aligned}$$

Considérons ensuite C_2. (A.2) et (A.3) impliquent que

$$\begin{aligned} E\left(d_{i,j}\right) &= (1-c_{i,j})(t-m_i) + (\hat{\hat{m}}_j - m_j) - (\hat{\hat{m}}_i - m_i) + O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right) \\ &= (w_{j,j} - w_{i,j})(t-m_i) + m_j'' \sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - m_i'' \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2 \\ &\quad + o_{i,j}\left(\lambda^2\right) + O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right) \\ &= (w_{j,j} - w_{i,j})(t-m_i) + (m_j'' - m_i'')\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 \\ &\quad + m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2\right) \\ &\quad + o_{i,j}\left(\lambda^2\right) + O_{i,j}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right) \end{aligned}$$

de sorte que

$$C_2 = C_{2,a} + C_{2,b} + C_{2,c} + o\left(\lambda^2\right) + O\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right),$$

où

$$\begin{aligned} C_{2,a} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \mid x_j\right)\left(w_{j,j} - w_{i,j}\right)(t - m_i) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t - m_i \mid x_i\right)(t - m_i)\sum_{j\in s} w_{i,j}\left(w_{j,j} - w_{i,j}\right) + O\left(n^{-1}\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}\left(t - m(x) \mid x\right)(t - m(x))\left[h_{\bar{s}}(x)/h_s(x)\right]dx \\ &\quad + O\left((n\lambda)^{-1}\lambda^{-1}\alpha + n^{-1}\right) \end{aligned}$$

avec $\kappa := \int_{-1}^{1} K^2(u)\,du$,

$$\begin{aligned} C_{2,b} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \mid x_j\right)\left(m_j'' - m_i''\right)\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 \\ &= o\left(\lambda^2\right) \end{aligned}$$

et

$$\begin{aligned} C_{2,c} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t - m_i \mid x_j\right) m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2\right) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t - m_i \mid x_i\right) m_i''\left(\sum_{j\in s} w_{i,j}\sum_{k\in s} w_{j,k}\left(x_k - x_j\right)^2 - \sum_{k\in s} w_{i,k}\left(x_k - x_i\right)^2\right) + o\left(\lambda^2\right) \\ &= o\left(\lambda^2\right). \end{aligned}$$

Considérons enfin C_3. Notons que, d'après (A.2) et (A.3),

$$E\left(d_{i,j}^2\right) = \sum_{k\in s}\left(w_{j,k} - w_{i,k}\right)^2 \sigma_k^2 + O_{i,j}\left(\lambda^4 + (n\lambda)^{-2}\right) \tag{A.6}$$

d'où

$$\begin{aligned} C_3 &= \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(2,0)}\left(t - m_i \mid x_j\right)\sum_{k\in s}\left(w_{j,k} - w_{i,k}\right)^2 \sigma_k^2 + O\left(\lambda^4 + (n\lambda)^{-2}\right) \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(2,0)}\left(t - m_i \mid x_i\right)\sigma_i^2 \sum_{j\in s} w_{i,j}\sum_{k\in s}\left(w_{j,k} - w_{i,k}\right)^2 + o\left((n\lambda)^{-1}\right) + O\left(\lambda^4\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}\left(t - m(x) \mid x\right)\sigma^2(x)\left[h_{\bar{s}}(x)/h_s(x)\right]dx + o\left((n\lambda)^{-1}\right) + O\left(\lambda^4\right) \end{aligned}$$

avec $\theta := \int_{-1}^{1} K(v) \int_{-1}^{1} K(u+v) K(u)\, du dv$.

En substituant les développements susmentionnés à C_1, C_2 et C_3 dans (A.5), on obtient finalement

$$\begin{aligned} E\left(\hat{F}^*(t) - F_N(t)\right) &= \lambda^2 \frac{N-n}{N} \frac{\mu_2}{\mu_0} \int_a^b G^{(0,2)}(t - m(x) \mid x) h_{\bar{s}}(x)\, dx \\ &\quad + \frac{1}{n\lambda} \frac{N-n}{N} \left[\frac{K(0) - \kappa}{\mu_0} \int_a^b G^{(1,0)}(t - m(x) \mid x)(t - m(x)) h_s^{-1}(x) h_{\bar{s}}(x)\, dx \right. \\ &\quad \left. + \frac{\kappa - \theta}{\mu_0^2} \int_a^b G^{(2,0)}(t - m(x) \mid x) \sigma^2(x) h_s^{-1}(x) h_{\bar{s}}(x)\, dx \right] \\ &\quad + o\left(\lambda^2 + (n\lambda)^{-1}\right). \end{aligned}$$

Biais de l'estimateur par la différence généralisée avec valeurs prédites modifiées

Soit $\tilde{d}_{i,j}$ l'équivalent pondéré selon le plan de sondage de $d_{i,j}$ et observons que

$$\begin{aligned} \tilde{F}^*(t) - F_N(t) &= \frac{1}{N} \left[\sum_{i \notin s} \sum_{j \in s} \tilde{w}_{i,j} \left(I\left(\varepsilon_j \le t - m_i + \tilde{d}_{i,j}\right) - I(y_i \le t) \right) \right. \\ &\quad \left. + \sum_{i \in s} \left(1 - \pi_i^{-1}\right) \sum_{j \in s} \tilde{w}_{i,j} \left(I\left(\varepsilon_j \le t - m_i + \tilde{d}_{i,j}\right) - I(y_i \le t) \right) \right]. \end{aligned} \tag{A.7}$$

En adaptant la preuve qui mène à (A.4), on voit que le développement asymptotique en (A.4) est également vérifié en prenant $\tilde{d}_{i,j}$ à la place de $d_{i,j}$. L'adaptation de la partie restante de la preuve mène en bout de ligne à

$$\begin{aligned} E\left(\tilde{F}^*(t) - F_N(t)\right) &= \lambda^2 \frac{N-n}{N} \frac{\mu_2}{\mu_0} \int_a^b G^{(0,2)}(t - m(x) \mid x) h(x)\, dx \\ &\quad + \frac{1}{n\lambda} \frac{N-n}{N} \left[\frac{K(0) - \kappa}{\mu_0} \int_a^b G^{(1,0)}(t - m(x) \mid x)(t - m(x)) h_s^{-1}(x) h(x)\, dx \right. \\ &\quad \left. + \frac{\kappa - \theta}{\mu_0^2} \int_a^b G^{(2,0)}(t - m(x) \mid x) \sigma^2(x) h_s^{-1}(x) h(x)\, dx \right] \\ &\quad + o\left(\lambda^2 + (n\lambda)^{-1}\right), \end{aligned}$$

où

$$h(x) := h_{\bar{s}}(x) + \left(1 - \pi^{-1}(x)\right) h_s(x).$$

Variance de l'estimateur fondé sur le modèle avec valeurs prédites modifiées

Écrivons

$$\hat{F}^*(t) - F_N(t) = \frac{1}{N}\left(\sum_{i \notin s} \sum_{j \in s} w_{i,j} I\left(\varepsilon_j \le t - m_i + d_{i,j}\right) - \sum_{i \notin s} I\left(\varepsilon_i \le t - m_i\right) \right)$$

et observons que

$$\text{var}\left(\hat{F}^*(t) - F_N(t)\right) = D_1 + D_2 + D_3,$$

où

$$D_1 := \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j \in s} w_{i_1,j} w_{i_2,j} \,\text{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + d_{i_1,j}\right), I\left(\varepsilon_j \le t - m_{i_2} + d_{i_2,j}\right)\right),$$

$$D_2 := \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s, j_2 \ne j_1} w_{i_1,j_1} w_{i_2,j_2} \times \text{cov}\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right)$$

et où $D_3 := A_2$ provenant de la variance de l'estimateur fondé sur le modèle de Kuo.

Considérons D_1. Observons que

$$\begin{aligned} \text{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + d_{i_1,j}\right), I\left(\varepsilon_j \le t - m_{i_2} + d_{i_2,j}\right)\right) &= E\left(G\left(t - m_{i_1} + d_{i_1,j} \wedge t - m_{i_2} + d_{i_2,j} \middle| x_j\right)\right) \\ &\quad - E\left(G\left(t - m_{i_1} + d_{i_1,j} \middle| x_j\right)\right) E\left(G\left(t - m_{i_2} + d_{i_2,j} \middle| x_j\right)\right). \end{aligned} \tag{A.8}$$

Puisque

$$\left|\left(t - m_{i_1} + d_{i_1,j} \wedge t - m_{i_2} + d_{i_2,j}\right) - \left(t - m_{i_1} \wedge t - m_{i_2}\right)\right| \le \left|d_{i_1,j}\right| + \left|d_{i_2,j}\right|,$$

il découle de (A.6) que

$$E\left(G\left(t - m_{i_1} + d_{i_1,j} \wedge t - m_{i_2} + d_{i_2,j} \middle| x_j\right)\right) = G\left(t - m_{i_1} \wedge t - m_{i_2} \middle| x_j\right) + O_{i_1,i_2,j}\left(\lambda^2 + (n\lambda)^{-1/2}\right). \tag{A.9}$$

En outre, de (A.1), (A.4) et (A.6), il découle que

$$E\left(G\left(t - m_i + d_{i,j} \middle| x_j\right)\right) = G\left(t - m_i \middle| x_j\right) + O_{i,j}\left(\lambda^2 + (n\lambda)^{-1/2}\right). \tag{A.10}$$

En utilisant (A.9) et (A.10) pour obtenir un développement asymptotique pour la covariance en (A.8) et en introduisant par substitution le résultat dans la définition de D_1, on obtient

$$
\begin{aligned}
D_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+d_{i_1,j}\right), I\left(\varepsilon_j \le t-m_{i_2}+d_{i_2,j}\right)\right)\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\,\middle|\, x_j\right)\right)\right.\\
&\qquad\qquad \left. -E\left(G\left(t-m_{i_1}+d_{i_1,j}\,\middle|\, x_j\right)\right)E\left(G\left(t-m_{i_2}+d_{i_2,j}\,\middle|\, x_j\right)\right)\right]\\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\left[G\left(t-m_{i_1}\wedge t-m_{i_2}\,\middle|\, x_j\right)-G\left(t-m_{i_1}\,\middle|\, x_j\right)G\left(t-m_{i_2}\,\middle|\, x_j\right)\right]\\
&\quad + O\left(\lambda^2 n^{-1}+(n\lambda)^{-1/2}\,n^{-1}\right)\\
&= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j\,\middle|\, x_j\right)-G^2\left(t-m_j\,\middle|\, x_j\right)\right]\left(\sum_{i\notin s} w_{i,j}\right)^2 + O\left(\lambda n^{-1}+(n\lambda)^{-1/2}\,n^{-1}\right)\\
&= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\, x\right)-G^2\left(t-m(x)\,|\, x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx\\
&\quad + O\left((n\lambda)^{-1}\alpha+n^{-1}\lambda+n^{-1}(n\lambda)^{-1/2}\right).
\end{aligned}
\tag{A.11}
$$

Considérons ensuite

$$
D_2 := \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}\times\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right), I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right).
$$

Puisque

$$
\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right), I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)=0
$$

si $\left|x_{i_1}-x_{i_2}\right|>2\lambda,$ il s'ensuit que les termes de reste R_{i_1,j_1,i_2,j_2}, dont la contribution à la covariance susmentionnée est d'ordre $O_{i_1,j_1,i_2,j_2}(\beta)$ pour une suite β qui tend vers zéro, apportent à D_2 un terme d'ordre $O(\lambda\beta)$. Or, soit

$$
\begin{aligned}
b_{i,j_1,j_2} &:= c_{i,j_1}^{-1}\left(w_{j_1,j_2}-w_{i,j_2}\right),\\
a_{i,j_1,j_2} &:= t-m_i+d_{i,j_1}-b_{i,j_1,j_2}\varepsilon_{j_2}
\end{aligned}
$$

et notons que

$$
t-m_i+d_{i,j_1}=a_{i,j_1,j_2}+b_{i,j_1,j_2}\varepsilon_{j_2}.
$$

Puisque a_{i,j_1,j_2} ne dépend pas de ε_{j_1} ni de ε_{j_2}, il s'ensuit que

$$
\begin{aligned}
&E\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right)I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)\\
&\quad = E\left(E\left(I\left(\varepsilon_{j_1}\le a_{i_1,j_1,j_2}+b_{i_1,j_1,j_2}\varepsilon_{j_2}\right)I\left(\varepsilon_{j_2}\le a_{i_2,j_2,j_1}+b_{i_2,j_2,j_1}\varepsilon_{j_1}\right)\,\middle|\,\varepsilon_k, k\neq j_1, j_2\right)\right)\\
&\quad = E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1}+b_{i_2,j_2,j_1}\varepsilon\,\middle|\, x_{j_2}\right)dG\left(\varepsilon\,|\, x_{j_1}\right)\right)\\
&\quad + E\left(\int_{-\infty}^{\varepsilon^*_{i_2,i_1,j_2,j_1}} G\left(a_{i_1,j_1,j_2}+b_{i_1,j_1,j_2}\varepsilon\,\middle|\, x_{j_1}\right)dG\left(\varepsilon\,|\, x_{j_2}\right)\right)\\
&\quad - E\left(G\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\,\middle|\, x_{j_1}\right)G\left(\varepsilon^*_{i_2,i_1,j_2,j_1}\,\middle|\, x_{j_2}\right)\right),
\end{aligned}
\tag{A.12}
$$

où

$$\varepsilon^*_{i_1,i_2,j_1,j_2} := \frac{a_{i_1,j_1,j_2} + a_{i_2,j_2,j_1} b_{i_1,j_1,j_2}}{1 - b_{i_1,j_1,j_2} b_{i_2,j_2,j_1}}.$$

Notons que les deux espérances aux troisième et quatrième lignes de (A.12) sont les mêmes si i_1 et j_1 sont remplacés par i_2 et j_2, respectivement. Donc, il suffit d'analyser la première espérance. Étant donné que

$$\varepsilon^*_{i_1,i_2,j_1,j_2} = t - m_{i_1} + d_{i_1,j_1} + b_{i_1,j_1,j_2}\left(t - m_{i_2} - \varepsilon_{j_2}\right) + R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right),$$

où

$$E^{1/4}\left(\left|R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right)\right|^4\right) = O_{i_1,i_2,j_1,j_2}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right),$$

on voit que

$$\begin{aligned}
&E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon \,\middle|\, x_{j_2}\right) dG\left(\varepsilon \,\middle|\, x_{j_1}\right)\right) \\
&= G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\
&\quad + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d_{i_2,j_2}\right) + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) b_{i_2,j_2,j_1} \int_{-\infty}^{t - m_{i_1}} \varepsilon dG\left(\varepsilon \,\middle|\, x_{j_1}\right) \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned} \tag{A.13}$$

et que

$$\begin{aligned}
&E\left(G\left(\varepsilon^*_{i_1,i_2,j_1,j_2} \,\middle|\, x_{j_1}\right) G\left(\varepsilon^*_{i_2,i_1,j_2,j_1} \,\middle|\, x_{j_2}\right)\right) \\
&= G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\
&\quad + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left[E\left(d_{i_2,j_2}\right) + b_{i_2,j_2,j_1}\left(t - m_{i_1}\right)\right] \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) \\
&\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right).
\end{aligned} \tag{A.14}$$

En utilisant les développements asymptotiques en (A.4), (A.13) et (A.14), on obtient

$$\begin{aligned}
&\operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right) \\
&= G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) b_{i_2,j_2,j_1} \gamma_{i_1,j_1} + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) b_{i_1,j_1,j_2} \gamma_{i_2,j_2} \\
&\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \operatorname{cov}\left(d_{i_1,j_1}, d_{i_2,j_2}\right) \\
&\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned} \tag{A.15}$$

où

$$\gamma_{i,j} := \int_{-\infty}^{t-m_i} \varepsilon dG\left(\varepsilon \,|\, x_j\right).$$

Observons maintenant que

$$b_{i,j_1,j_2} = w_{j_1,j_2} - w_{i,j_2} + O_{i,j_1,j_2}\left((n\lambda)^{-2}\right)$$

et que

$$\begin{aligned}
\operatorname{cov}\left(d_{i_1,j_1}, d_{i_2,j_2}\right) &= \frac{1}{c_{i_1,j_1} c_{i_2,j_2}} \sum_{k \in s; k \ne j_1, j_2} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right)\sigma_k^2 \\
&= \sum_{k \in s} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right)\sigma_k^2 + O_{i_1,i_2,j_1,j_2}\left((n\lambda)^{-2}\right)
\end{aligned}$$

de sorte que

$$D_2 = 2D_{2a} + D_{2b} + o\left(\lambda^5 + n^{-1}\right), \tag{A.16}$$

où

$$\begin{aligned}
D_{2a} &:= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s, j_2 \ne j_1} w_{i_1,j_1} w_{i_2,j_2} G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left(w_{j_1,j_2} - w_{i_1,j_2}\right)\gamma_{i_2,j_2} \\
&= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s} w_{i_1,j_1} w_{i_2,j_2} G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left(w_{j_1,j_2} - w_{i_1,j_2}\right)\gamma_{i_2,j_2} + O\left(n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2} \sum_{j_2 \in s} G^{(1,0)}\left(t - m_{j_2} \,\middle|\, x_{j_2}\right)\gamma_{j_2,j_2} \left[\sum_{j_1 \in s} w_{j_1,j_2} \sum_{i_1 \notin s} w_{i_1,j_1} \sum_{i_2 \notin s} w_{i_2,j_2} - \left(\sum_{i \notin s} w_{i,j_2}\right)^2\right] \\
&\quad + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= O\left((n\lambda)^{-1}\alpha + n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right)
\end{aligned} \tag{A.17}$$

et

$$
\begin{aligned}
D_{2b} \quad &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s, j_2\neq j_1} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,|\,x_{j_2}\right) \\
&\quad \times\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2 \\
&= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s} w_{i_1,j_1}w_{i_2,j_2}G^{(1,0)}\left(t-m_{i_1}\,|\,x_{j_1}\right)G^{(1,0)}\left(t-m_{i_2}\,|\,x_{j_2}\right) \\
&\quad \times\sum_{k\in s}\left(w_{j_1,k}-w_{i_1,k}\right)\left(w_{j_2,k}-w_{i_2,k}\right)\sigma_k^2 + O\left(n^{-1}(n\lambda)^{-1}\right) \qquad \text{(A.18)} \\
&= \frac{1}{N^2}\sum_{k\in s}\sigma_k^2\left[G^{(1,0)}\left(t-m_k\,|\,x_k\right)\right]^2\left(\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left(w_{j,k}-w_{i,k}\right)\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= \frac{1}{N^2}\sum_{k\in s}\sigma_k^2\left[G^{(1,0)}\left(t-m_k\,|\,x_k\right)\right]^2\left(\sum_{j\in s} w_{j,k}\sum_{i\notin s} w_{i,j} - \sum_{i\notin s} w_{i,k}\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\
&= O\left((n\lambda)^{-1}\alpha + n^{-1}\lambda\right).
\end{aligned}
$$

En regroupant tout, on obtient finalement

$$
\begin{aligned}
\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right) \quad &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\,x\right)-G^2\left(t-m(x)\,|\,x\right)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx \\
&\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\,x\right)-G^2\left(t-m(x)\,|\,x\right)\right]h_{\bar{s}}(x)\,dx + o\left(\lambda^5+n^{-1}\right).
\end{aligned}
$$

Variance de l'estimateur par la différence généralisée avec valeurs prédites modifiées

Étant donné (A.7), nous allons montrer que

$$
\operatorname{var}\left(\tilde{F}^*(t)-F_N(t)\right) = \operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right) + o\left(n^{-1}\right) \qquad \text{(A.19)}
$$

en démontrant que

$$
\operatorname{var}\left(\frac{1}{N}\sum_{i\in s}\left(1-\pi_i^{-1}\right)\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j\le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i\le t\right)\right)\right) = o\left(n^{-1}\right). \qquad \text{(A.20)}
$$

Pour prouver (A.20), observons que la variance dans le premier membre peut s'écrire

$$
E_1+E_2+E_3-2E_4-2E_5,
$$

où

$$E_1 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t-m_{i_2}+\tilde{d}_{i_2,j}\right)\right),$$

$$E_2 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j_1\in s}\sum_{j_2\in s, j_2\neq j_1}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j_2}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t-m_{i_1}+\tilde{d}_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t-m_{i_2}+\tilde{d}_{i_2,j_2}\right)\right),$$

$$E_3 := \frac{1}{N^2}\sum_{i\in s}\left(1-\pi_i^{-1}\right)^2 \operatorname{var}\left(I\left(\varepsilon_i \le t-m_i\right)\right),$$

$$E_4 := \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\left(1-\pi_i^{-1}\right)\left(1-\pi_j^{-1}\right)\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_i+\tilde{d}_{i,j}\right), I\left(\varepsilon_j \le t-m_j\right)\right),$$

et finalement

$$E_5 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s, j\neq i_2}\tilde{w}_{i_1,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t-m_{i_2}\right)\right).$$

Pour commencer, considérons E_1 et E_2. Notons que, à part i) le fait que les indices de sommation i_1 et i_2 s'étendent sur s au lieu du complément de s dans U, ii) la présence des facteurs $\left(1-\pi_i^{-1}\right)$ et iii) le fait que les $w_{i,j}$ et les $d_{i,j}$ sont remplacés par leurs équivalents pondérés selon le plan de sondage $\tilde{w}_{i,j}$ et $\tilde{d}_{i,j}$, E_1 et E_2 sont semblables à D_1 et D_2 provenant de $\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right)$, respectivement. L'adaptation des preuves qui mènent aux développements asymptotiques pour D_1 et D_2 montre donc que

$$E_1 = \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]\left[1-\pi^{-1}(x)\right]^2 h_s(x)\,dx + o\left(n^{-1}\right)$$

et que

$$E_2 = o\left(\lambda^5 + n^{-1}\right).$$

Comme pour E_3, on constate immédiatement que

$$E_3 = E_1 + o\left(n^{-1}\right),$$

tandis que, pour traiter E_4 et E_5, on a besoin des développements asymptotiques pour

$$\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t-m_{i_2}\right)\right) \tag{A.21}$$

pour le cas où $j = i_2$ et celui où $j \neq i_2$. Dans le premier cas, nous pouvons faire appel à des arguments similaires à ceux utilisés pour prouver (A.9) et (A.10), ce qui donne

$$\begin{aligned}&\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t-m_j\right)\right)\\ &\quad = G\left(t-m_{i_1}\wedge t-m_j \,\middle|\, x_j\right) - G\left(t-m_{i_1}\,\middle|\, x_j\right)G\left(t-m_j\,\middle|\, x_j\right) + O\left(\lambda^2+(n\lambda)^{-1/2}\right).\end{aligned}$$

Par contre, quand $j \neq i_2$, la covariance dans (A.21) diffère de zéro uniquement si $\left|x_j - x_{i_2}\right| \leq \lambda$ ou $\left|x_{i_1} - x_{i_2}\right| \leq \lambda$, et en adaptant (A.12), on peut montrer que

$$\begin{aligned}
&E\left(I\left(\varepsilon_j \leq t - m_{i_1} + \tilde{d}_{i_1,j}\right) I\left(\varepsilon_{i_2} \leq t - m_{i_2}\right)\right) \\
&\quad = E\left(E\left(I\left(\varepsilon_j \leq \tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon_{i_2}\right) I\left(\varepsilon_{i_2} \leq t - m_{i_2}\right) \middle| \varepsilon_k, k \neq i, j\right)\right) \\
&\quad = E\left(\int_{-\infty}^{t-m_{i_2}} G\left(\tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon \middle| x_j\right) dG\left(\varepsilon \middle| x_{i_2}\right)\right) \\
&\quad = G\left(t - m_{i_1} \middle| x_j\right) G\left(t - m_{i_2} \middle| x_{i_2}\right) + G\left(t - m_{i_2} \middle| x_{i_2}\right) G^{(1,0)}\left(t - m_{i_1} \middle| x_j\right) E\left(d_{i_1,j}\right) \\
&\qquad + G^{(1,0)}\left(t - m_{i_1} \middle| x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + \frac{1}{2} G\left(t - m_{i_2} \middle| x_{i_2}\right) G^{(2,0)}\left(t - m_{i_1} \middle| x_j\right) E\left(d_{i_1,j}^2\right) \\
&\qquad + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned}$$

où $\tilde{a}_{i,j,k}$ et $\tilde{b}_{i,j,k}$ sont les équivalents pondérés selon le plan de sondage de $a_{i,j,k}$ et $b_{i,j,k}$, respectivement. En adaptant également (A.4) pour tenir compte des poids de sondage, on constate que

$$\begin{aligned}
\operatorname{cov}\left(I\left(\varepsilon_j \leq t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \leq t - m_{i_2}\right)\right) &= G^{(1,0)}\left(t - m_{i_1} \middle| x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right) \\
&= G^{(1,0)}\left(t - m_i \middle| x_j\right)\left(\tilde{w}_{j,i_2} - \tilde{w}_{i_1,i_2}\right)\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right)
\end{aligned}$$

de sorte que (voir les étapes qui mènent aux développements asymptotiques des termes D_1 et D_2 dans la variance de l'estimateur en deux étapes fondé sur le modèle)

$$E_4 = E_1 + o\left(n^{-1}\right)$$

et

$$E_5 = o\left(\lambda^5 + n^{-1}\right).$$

Cela achève la preuve de (A.20) et donc (A.19) s'ensuit.

Bibliographie

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals Statistics*, 28(4), 1026-1053.

Chambers, R.L., et Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series 37.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations using non-parametric calibration. *Journal of the American Statistical Association*, 88(421), 268-277.

Chen, J., et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Dorfman, A.H., et Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452-1475.

Fan, J., et Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4), 2008-2036.

Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726-748.

Johnson, A.A., Breidt, F.J. et Opsomer, J.D. (2008). Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice*, 2(3), 419-431.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. Dans les *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, 280-285.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472), 1429-1442.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

Rueda, M., Martínez, S., Martínez, H. et Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

Rueda, M., Sànchez-Borrego, I., Arcos, A. et Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71(1), 33-44.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York : Springer.

Wang, J.C., et Opsomer, J.D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1), 91-106.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90(4), 937-951.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.