Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale

par Eric Graf et Yves Tillé

Juin 2014

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 infostats@statcan.gc.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

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

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 « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2014

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/licence-fra.html).

This publication is also available in English.

Note de reconnaissance

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

Signes conventionnels

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)

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 69-88
Statistique Canada, N° 12-001-X au catalogue

Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale

Eric Graf et Yves Tillé[1]

Résumé

Nous avons mis en œuvre la technique de linéarisation généralisée reposant sur le concept de fonction d'influence tout comme l'a fait Osier (2009) pour estimer la variance de statistiques complexes telles que les indices de Laeken. Des simulations réalisées avec le langage R montrent que, pour les cas où l'on a recours à une estimation par noyau gaussien de la fonction de densité des revenus considérés, on obtient un fort biais pour la valeur estimée de la variance. On propose deux autres méthodes pour estimer la densité qui diminuent fortement le biais constaté. L'une de ces méthodes a déjà été esquissée par Deville (2000). Les résultats publiés ici permettront une amélioration substantielle de la qualité des informations sur la précision de certains indices de Laeken diffusées et comparées internationalement.

Mots-clés : Fonction d'influence; enquête SILC; statistiques non linéaires; indices de pauvreté et d'inégalité.

1 Introduction

Deville (2000) a proposé d'estimer la précision des statistiques non linéaires dans des plans de sondage au moyen de la technique de linéarisation généralisée. Celle-ci repose sur le concept de fonction d'influence proposée par Hampel (1974) dans le domaine de la statistique robuste. Osier (2009) a appliqué ces théories pour estimer la variance de statistiques complexes telles que les indices de Laeken (Eurostat 2005) dans l'enquête européenne sur le revenu et les conditions de vie (EU-SILC). Goga, Deville et Ruiz-Gazen (2009) étendent la théorie de Deville (2000) à des enquêtes basées sur deux échantillons. Verma et Betti (2011) fournissent une liste exhaustive des indicateurs de pauvreté conventionnels et des variables linéarisées associées, ils comparent également les performances de la technique de linéarisation à la méthode du Jackknife (*Jackknife repeated replication*). Dans cet article, nous nous restreignons à des indicateurs de pauvreté publiés dans le cadre de l'enquête SILC, en focalisant sur la façon d'estimer la fonction de densité des revenus en différents points de leur distribution.

Dans la section 2, on rappelle les fondements théoriques nécessaires ainsi que les expressions des indices de pauvreté et d'inégalité étudiés, de même que celle de leur linéarisée. Certaines linéarisées dépendent de la fonction de densité de la variable d'intérêt qui est généralement estimée par noyau gaussien. Deux méthodes alternatives au noyau gaussien sont présentées dans la section 3. Les simulations réalisées en langage R sont décrites et commentées dans la section 4. On y montre que la méthode par noyau gaussien peut engendrer un fort biais pour la valeur estimée de la variance des indices pour les cas où l'on a recours à une estimation de la fonction de densité des revenus considérés. On voit aussi que les deux autres méthodes pour estimer la densité proposées dans la section 3 diminuent le biais constaté, ce qui fait aussi l'objet des conclusions dans la dernière partie de ce texte.

1. Eric Graf et Yves Tillé, Institut de Statistique, Faculté des sciences économiques, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Suisse. Courriel : eric.graf@unine.ch et yves.tille@unine.ch.

2 Rappels sur les indices de pauvreté traités ainsi que leur linéarisée

Soit une population finie U constituée de N unités identifiables $u_1,...,u_k,...,u_N$. Pour simplifier l'écriture on désigne par la suite l'unité u_k par son indice k. En pratique la population U est une base de sondage avec un taux de couverture acceptable de la population d'intérêt pour laquelle on désire faire des inférences. On associe à chaque unité k une valeur y_k d'une caractéristique d'intérêt (ici un revenu). Sans nuire à la généralité et pour alléger les notations, on suppose que les y_k sont tous distincts et triés par ordre de grandeur, donc $y_k = y_{[k]}$. Dans les données issues d'enquêtes par échantillonnage, il arrive fréquemment qu'il y ait des doublons, c'est-à-dire plusieurs unités ayant la même valeur y, que ce soit dû à des arrondis ou à des questions-fourchettes. Dans ces cas et pour cette étude, il suffit d'ajouter un montant assez petit (c'est-à-dire négligeable) de la devise tiré au hasard selon une loi uniforme pour que les données soient triables sans équivoque.

Soit un échantillon aléatoire S de taille n obtenu par un plan d'échantillonnage $p(s) = \mathrm{P}(S = s)$, pour tout $s \subset U$. Soit également $\pi_k = \mathrm{P}(k \in s) > 0$ la probabilité d'inclusion dans l'échantillon de l'unité k de U. Soit aussi $d_k = 1/\pi_k$ le poids d'échantillonnage et $w_k = w_k(s)$ un poids d'estimation qui peut être égal à d_k mais qui peut aussi être plus raffiné. Par exemple, w_k peut avoir été obtenu suite à un calage (Deville et Särndal 1992) et refléter ainsi également une correction de non-réponse.

Les estimateurs des indices de pauvreté et d'inégalité sont des statistiques non linéaires qui ne peuvent pas s'exprimer comme des fonctions régulières (c'est-à-dire continûment différentiables jusqu'à l'ordre deux) de totaux. En effet, il s'agit de statistiques de rangs pour l'indice de Gini et de quantiles pour les autres. Comme le relève Osier (2009), leur variance ne peut donc pas être estimée par une linéarisation de Taylor mais nécessite le recours à la méthode de linéarisation généralisée (Deville 2000; Demnati et Rao 2004; Osier 2009). Une alternative pour estimer la variance serait d'utiliser des techniques de rééchantillonnage du genre bootstrap, mais dans le cadre des données de l'enquête SILC, une préférence a été donnée à la technique de linéarisation, du moins pour un certain nombre de pays participants. En effet, les méthodes de rééchantillonnage nécessitent souvent davantage de ressources humaines et machine. De plus, Eurostat collaborant avec une trentaine de pays ayant des plans de sondage différents suivis d'éventuels corrections pour la non-réponse et de calages sur des sources externes, il a semblé plus adéquat d'opter pour une solution analytique pour estimer la variance. Par ailleurs, certains pays pouvaient utiliser le logiciel SAS déjà existant POULPE (Ardilly et Osier 2007) pour produire les estimations nécessaires. Ce fut le cas pour les premiers tests avec les données SILC suisses. On applique ici un mode opératoire qui, comme le relèvent Antal, Langel et Tilllé (2011), concilie l'approche introduite par Deville (2000) à celle de Demnati et Rao (2004). Les deux approches utilisent la notion de *fonction d'influence* qui fut développée initialement dans le domaine des statistiques robustes (Hampel 1974). Antal et coll. (2011) mentionnent aussi que l'on peut retrouver les mêmes linéarisées en appliquant la méthode proposée par Graf (2011, 2013) qui construit une variable linéarisée basée sur un développement en série de Taylor par rapport aux indicatrices d'inclusion dans l'échantillon. Citons aussi le travail de Kovačević et Binder (1997) où une approche de linéarisation par les équations estimantes est développée.

Selon Deville (2000), on cerne l'influence de l'unité k sur un paramètre d'intérêt θ au niveau de la population par une variation infinitésimale de l'importance attribuée à cette unité. On exprime ledit paramètre comme une fonctionnelle $\theta = T(M)$, où M est une mesure allouant une masse unité,

$M(k) = M_k = 1$, uniquement aux points du continuum correspondant aux unités $k \in U$. La spécialisation de la mesure générale M en une mesure discrète fait passer la fonctionnelle T, définie *a priori* sur un continuum, en une fonctionnelle discrète, tout comme le total Y est défini par la somme des y_k sur notre population finie. La *fonction d'influence* de T, ou la *variable linéarisée*, est définie par

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{ pour tout } k \in U,$$

où δ_k est la mesure de Dirac pour l'unité k ($\delta_k(i) = 1$ si $i = k$ et 0 sinon). En pratique on ne dispose que des données connues sur un échantillon S et Deville (2000) obtient une linéarisée $\hat{z}_k$ ou *fonction d'influence empirique*, en : 1) évaluant la limite ci-dessus par calcul différentiel, 2) remplaçant dans l'évaluation les quantités inconnues par les quantités correspondantes estimées à partir de l'échantillon. Il justifie ce procédé en montrant que :

$$T(\hat{M}) - T(M) \approx \left(\sum_{k \in S} w_k z_k - \sum_{k \in U} z_k \right).$$

Le résultat central est que, sous des conditions asymptotiques décrites dans Deville (2000), qui sont en principe satisfaites lorsque l'échantillon est « assez grand », la variance du total estimé de la variable $\hat{z}_k$ est une approximation de la variance de la statistique (complexe) $\hat{\theta}$:

$$\text{var}\left[\sum_{k \in s} \hat{z}_k w_k \right] \approx \text{var}(\hat{\theta}).$$

Le point de départ de l'approche de Deville est donc le paramètre de population et non l'estimateur qu'on se propose d'utiliser pour l'évaluation à partir de l'échantillon. Dans les cas où l'estimateur utilisé découle naturellement de l'expression du paramètre de population (comme par exemple le total Y approché par l'estimateur de Horvitz-Thompson), le procédé ne présente pas d'ambiguïté. Mais des imprécisions surviennent si l'on estime le même total Y en ayant recours à l'estimateur par le quotient grâce à une variable auxiliaire x. En effet, dans ce cas, l'approche de Deville ne précisant pas la forme de l'estimateur du total à utiliser fournira plutôt une fonction d'influence constante égale à 1, au lieu de faire intervenir le quotient inconnu d'intérêt.

Une alternative qui évite ces problèmes est celle de Demnati-Rao, lorsque rattachée au préalable au cadre de Deville comme cela est fait dans Antal et coll. (2011). Ces auteurs présentent l'approche de Demnati-Rao comme résultant du cadre de Deville lorsque la mesure M employée n'est pas la mesure discrète définie sur U présentée précédemment, mais plutôt la mesure suivante définie sur S, l'échantillon :

$$\hat{M}(k) = w_k, k \in S$$

où w_k est un poids. En définissant la mesure sur S, on se donne dans les faits comme point de départ l'estimateur et non le paramètre; c'est le paramètre qui se retrouve exprimé au départ sous la forme d'une fonctionnelle et non le paramètre de population à estimer. Autrement dit, on se donne comme fonctionnelle celle qui correspond à l'estimateur pour lequel on souhaite avoir une estimation de la variance par linéarisation généralisée. On obtient ensuite la linéarisée en fonction de cette fonctionnelle de la façon suivante :

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_k) - T(\hat{M})}{t}, \text{ pour tout } k \in S.$$

Antal et coll. (2011) observent que, dans la mesure où la fonctionnelle apparaissant dans cette limite s'exprime comme une fonction explicite des variables que sont les poids assignés par la mesure $\hat{M}$ aux observations, cette linéarisée est en fait une fonction des dérivées partielles par rapport aux poids :

$$I[T(\hat{M})]_k = \frac{\partial T(\hat{M})}{\partial w_k}.$$

Antal et coll. (2011) relèvent que, les linérarisées que nous rappelons dans la suite, peuvent être obtenues par les deux approches. En effet, l'évaluation de la limite à-la-Demnati-Rao ne mène pas nécessairement à l'estimation de la variance suggérée par Deville (2000). L'approche pratique utilisée dans le cadre de cet article pourrait donc être nommée comme étant celle de Deville-Demnati-Rao en reconnaissance au cadre théorique fourni par Deville (2000) et l'algorithmique pratique de ce cadre qu'apportent Demnati et Rao (2004).

Par cette méthode, on peut estimer la variance de $\hat{\theta}$ quel que soit le plan d'échantillonnage, et donc obtenir un intervalle de confiance, en substituant la variable linéarisée dans la formule de variance pour un total correspondant au plan choisi. Sous un plan aléatoire simple sans remise, l'estimateur de la variance d'un indice d'inégalité $\hat{\theta}$ est donné par

$$\widehat{\text{var}}_{\text{lin}}[\hat{\theta}] = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{k\in S}(\hat{z}_k - \overline{z})^2, \tag{2.1}$$

avec

$$\overline{z} = n^{-1}\sum_{k\in S}\hat{z}_k.$$

Dans la suite, dans le cadre de la mesure du revenu de la population, nous rappelons les définitions empiriques des indices d'inégalité considérés ainsi que l'expression de leurs linéarisées telles que nous les avons mises en œuvre.

2.1 L'indice de Gini

L'indice de Gini, G, est compris entre 0 (en cas d'égalité totale, tous gagnent le même montant) et 1 (en cas d'inégalité totale, c'est-à-dire un individu gagne tout et les autres rien). L'indice G s'exprime en fonction des revenus cumulés d'une certaine proportion d'individus les plus pauvres. Si $\mathcal{Y}$ est la variable aléatoire représentant les revenus, $f(y)$ sa fonction de densité et $F(y)$ sa fonction de répartition, alors la *courbe de Lorenz* (Lorenz 1905) est définie par

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)\,dy}{\int_0^{\infty} yf(y)\,dy} = \frac{1}{\mathrm{E}(\mathcal{Y})}\int_0^{\alpha} F^{-1}(u)\,du.$$

L'indice de Gini représente deux fois la surface comprise entre la courbe de Lorenz et la ligne (diagonale $f_{eg}(x) = x$) de l'égalité parfaite (Figure 2.1). Il est donc défini par :

$$G = 2\int_0^1 [\alpha - L(\alpha)]\,d\alpha.$$

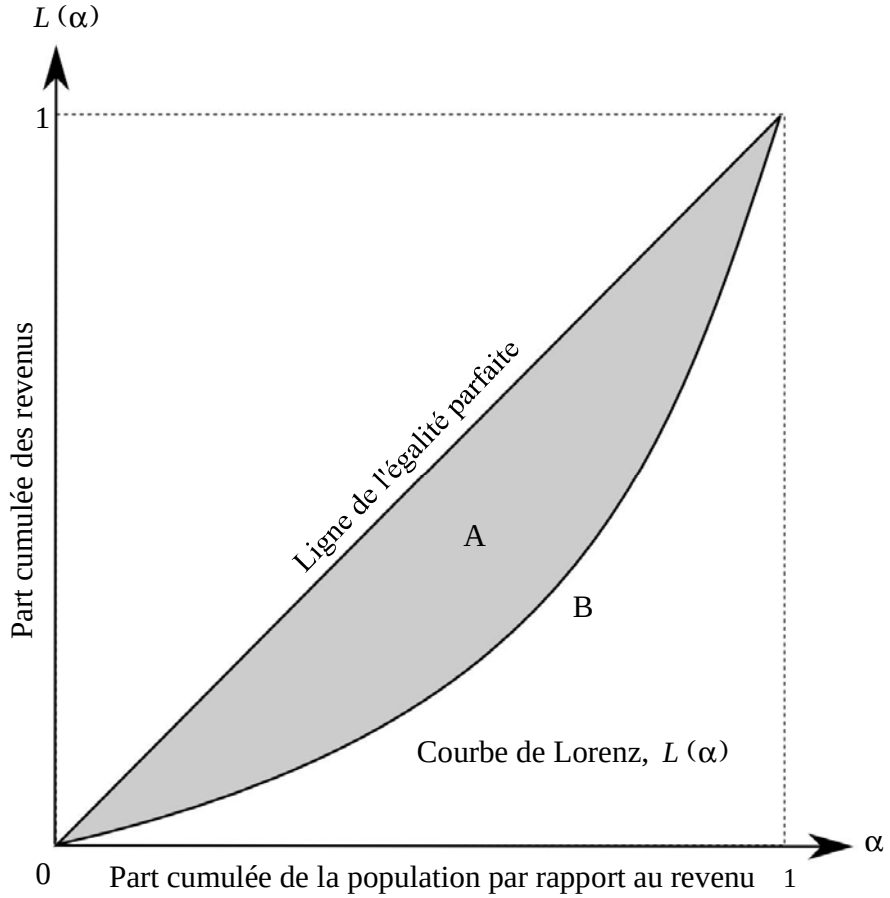

Figure 2.1 Indice de Gini G et courbe de Lorenz $L(\alpha)$. $G = 2A$, $A + B = 1/2$

Dans le cas d'une population finie, les y_k ne sont pas aléatoires et l'indice de Gini est défini sur la population U par :

$$G = \frac{2\sum_{k\in U} k y_k}{N\sum_{k\in U} y_k} - \frac{N+1}{N},$$

où les y_k ont été préalablement triés par leur rang. Calculé sur un échantillon, on l'estime par :

$$\begin{aligned}\hat{G} &= \frac{2}{\hat{N}\hat{Y}}\sum_{k\in S} w_k \hat{N}_k y_k - \left(1 + \frac{1}{\hat{N}\hat{Y}}\sum_{k\in S} w_k^2 y_k\right) \\ &= \frac{\sum_{k\in S}\sum_{\ell\in S} w_k w_\ell \left|y_k - y_\ell\right|}{2\hat{N}\hat{Y}},\end{aligned}$$

où $\hat{N}_k = \sum_{\ell\in S} w_\ell \mathbf{1}_{[y_\ell \le y_k]}$ est la somme cumulée des poids w_k, $\hat{Y} = \sum_{k\in S} w_k y_k$ étant le revenu total estimé pour la population et $\hat{N} = \sum_{k\in S} w_k$ la taille estimée de cette population. L'expression peut être simplifiée si tous les poids sont égaux et valent tous N/n :

$$\hat{G} = \frac{2\sum_{k\in S} k y_k}{n\sum_{k\in S} y_k} - \frac{n+1}{n}.$$

Notons que la définition peut varier d'un facteur $n/(n-1)$ selon les auteurs (Osier 2009; Eurostat 2004b), mais cette subtilité est négligeable dès que la taille de l'échantillon est assez grande.

Langel et Tillé (2012) ont fait une synthèse des différentes approches permettant d'obtenir la même variable linéarisée de l'indice de Gini estimée sur l'échantillon :

$$\hat{z}_k^{\text{GINI}} = \frac{1}{\hat{N}\hat{Y}}\left[2\hat{N}_k(y_k - \hat{\bar{Y}}_k) + \hat{Y} - \hat{N}y_k - \hat{G}\left(\hat{Y} + y_k\hat{N}\right)\right],$$

où $\hat{\bar{Y}}_k = \sum_{\ell=1}^{k} w_\ell y_\ell / \hat{N}_k$, les y_ℓ étant triés et distincts.

2.2 Le Quintile Share Ratio (QSR ou S_{80}/S_{20})

On trouve un bon tour d'horizon sur cet indice dans Langel et Tillé (2012). Soient q_{80} et q_{20} les 80^{e} et 20^{e} percentiles de la fonction de répartition $\mathrm{F}(y)$. Le QSR est le ratio de la somme des revenus des 20 % les plus riches sur les 20 % les plus pauvres. Dans le cas continu, on peut le définir ainsi :

$$\text{QSR} = \frac{\mathrm{E}(\mathcal{Y} \mid \mathcal{Y} > q_{80})}{\mathrm{E}(\mathcal{Y} \mid \mathcal{Y} < q_{20})} = \frac{1 - L(0{,}8)}{L(0{,}2)},$$

où $\mathcal{Y}$ serait une variable aléatoire représentant les revenus. Dans le cas de populations finies, le QSR peut être exprimé et estimé au niveau de l'échantillon en fonction des sommes partielles

$$\widehat{\text{QSR}} = \frac{\hat{Y} - \hat{Y}_{0,8}}{\hat{Y}_{0,2}},$$

où, suite aux résultats obtenus par Langel et Tillé (2011), nous utiliserons la définition suivante de la somme partielle, ce qui diffère très légèrement de la définition officielle d'Eurostat (2004a),

$$\hat{Y}_\alpha = \sum_{k\in S} w_k y_k H\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right), \tag{2.2}$$

avec

$$H(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \le x < 1 \\ 1 & \text{si } x \ge 1. \end{cases}$$

Pour obtenir la linéarisée du QSR, il faut d'abord calculer la linéarisée de la somme partielle (2.2) qui est donnée par :

$$I(Y_\alpha)_k = y_k H(\alpha N - k + 1) + \left[\alpha - \mathbf{1}_{[y_k < Q_\alpha]}\right] Q_\alpha,$$

où $Q_\alpha = y_i$, avec $\hat{N}_{i-1} < \alpha\hat{N} \le \hat{N}_i$, correspond à la première définition du quantile d'une population finie dans l'article de Hyndman et Fan (1996). Osier (2009) obtient une linéarisée qui dépend de la densité de la variable $\mathcal{Y}$. Langel et Tillé (2011) ont cependant montré qu'une simplification permet d'éluder le

problème de l'estimation de cette densité pour le QSR et qu'il n'est donc pas nécessaire de faire une approximation par noyau de la densité des revenus comme le propose Osier (2009).

La fonction d'influence dépend de celles des sommes partielles :

$$I\,(\mathrm{QSR})_k \;=\; z_k^{\mathrm{QSR}} \;=\; \frac{y_k \;-\; I\,(Y_{0,8})}{Y_{0,2}} \;-\; \frac{(Y \;-\; Y_{0,8})\,I\,(Y_{0,2})}{Y_{0,2}^2}.$$

En faisant les substitutions nécessaires, on trouve que la linéarisée estimée sur la base de l'échantillon vaut

$$\begin{aligned}\hat{z}_k^{\mathrm{QSR}} \;=\; & \frac{y_k \;-\; \left\{ y_k H\left(\frac{0{,}8\hat{N} \;-\; \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0,8}\left[0{,}8 \;-\; \mathbf{1}_{\left[y_k \,<\, \hat{Q}_{0,8}\right]}\right]\right\}}{\hat{Y}_{0,2}} \\ & -\; \frac{\left(\hat{Y} \;-\; \hat{Y}_{0,8}\right)\left\{ y_k H\left(\frac{0{,}2\hat{N} \;-\; \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0,2}\left[0{,}2 \;-\; \mathbf{1}_{\left[y_k \,<\, \hat{Q}_{0,2}\right]}\right]\right\}}{\hat{Y}_{0,2}^2}.\end{aligned} \tag{2.3}$$

2.3 Linéarisée d'un quantile

Avant de traiter les indices de pauvreté, il convient de donner quelques détails sur la linéarisée d'un quantile d'ordre α. Celle-ci s'estime par :

$$\hat{z}_k^{Q_\alpha} \;=\; -\frac{1}{f\left(\hat{Q}_\alpha\right)}\frac{1}{\hat{N}}\left[\mathbf{1}_{\left[y_k \,\le\, \hat{Q}_\alpha\right]} \;-\; \alpha\right],$$

où le quantile pondéré peut être défini de manière similaire à la somme partielle (2.2) et $f\,(\cdot)$ est la fonction de densité des revenus qui est discutée en détails à la section 3. Notons qu'Eurostat (2004a) préconise la deuxième définition de Hyndman et Fan (1996). On pourrait discuter de la définition d'Eurostat et utiliser une autre définition du quantile, par exemple $Q_\alpha \;=\; y_{k-1} + (y_k \;-\; y_{k-1})[\alpha N \;-\; (k-1)]$ où $\alpha N \;<\; k \;\le\; \alpha N + 1,$ ce qui correspond à la quatrième définition selon Hyndman et Fan (1996). On estime alors le quantile sur l'échantillon par

$$\hat{Q}_\alpha \;=\; y_{k-1} + (y_k \;-\; y_{k-1})\left(\frac{\alpha\hat{N} \;-\; \hat{N}_{k-1}}{w_k}\right).$$

La linéarisée du quantile dépend de la valeur de la fonction de densité des revenus en ce quantile. Or, la vraie densité des revenus n'est pas connue et doit donc aussi être estimée à partir de l'échantillon. Deville (2000) et Osier (2009) proposent de l'estimer par noyau gaussien. On revient plus en détails sur le problème de l'estimation de f dans la section 3.

En plus du problème de l'estimation de la fonction de densité des revenus, Croux (1998) montre que la fonction d'influence empirique de la médiane n'est pas un estimateur convergent de la fonction d'influence (théorique) correspondante. Pour une variable positive (cas des revenus), la fonction d'influence empirique de la médiane (cas que Croux traite dans son article) converge vers une distribution exponentielle dont l'espérance est la fonction d'influence. Elle résiste mal aux valeurs extrêmes s'il y en a une trop grande proportion. On dira qu'elle manque de robustesse, dans le sens que la valeur de l'estimateur sur la base de l'échantillon peut s'éloigner fortement de la vraie valeur sur la population en

raison d'observations extrêmes (c'est-à-dire très grandes par rapport aux autres) présentes dans l'échantillon (voir Hampel (1974) pour les idées de base sur la robustesse en population infinie, et Beaumont, Haziza et Ruiz-Gazen (2013) pour des réflexions récentes à ce sujet dans le cadre d'échantillonnage en population finie).

2.4 La médiane et le seuil de risque de pauvreté (ARPT)

Soit $\hat{m} = \hat{Q}_{0,5}$ la médiane estimée sur l'échantillon, le seuil de pauvreté, noté ARPT (*At Risk of Poverty Threshold*), est défini comme étant égal à 60 % de la médiane :

$$\text{ARPT} = 0{,}6F^{-1}(0{,}5)$$
$$\widehat{\text{ARPT}} = 0{,}6\hat{Q}_{0,5} = 0{,}6\hat{m}.$$

Il s'agit d'une mesure absolue qui dépend de l'échelle. La linéarisée de l'ARPT est proportionnelle à celle de la médiane :

$$\hat{z}_k^{\text{ARPT}} = I(\text{ARPT})_k = 0{,}6I(\text{MED})_k = -\frac{0{,}6}{f(\hat{m})}\frac{1}{\hat{N}}\left[\mathbf{1}_{[y_k \le \hat{m}]} - 0{,}5\right].$$

2.5 Le taux de risque de pauvreté (ARPR)

Le taux de risque de pauvreté (*At Risk of Poverty Rate*), $\text{ARPR} \in [0,1]$, correspond à la proportion de la population au-dessous du seuil de risque de pauvreté, $\text{ARPR} = F(\text{ARPT})$. Il est indépendant de l'échelle tout comme l'indice de Gini, le QSR et le RMPG (voir section 2.7). La définition officielle d'Eurostat (2004a) de son estimation à partir de l'échantillon est

$$\widehat{\text{ARPR}} = \frac{\sum_{y_k < \widehat{\text{ARPT}}} w_k}{\hat{N}}.$$

La linéarisée de l'ARPR est donnée par Osier (2009) :

$$\begin{aligned}\hat{z}_k^{\text{ARPR}} &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \widehat{\text{ARPT}}]} - \widehat{\text{ARPR}}\right) - \frac{f(\widehat{\text{ARPT}})}{f(\hat{m})}\frac{0{,}6}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}]} - 0{,}5\right) \\ &= \frac{1}{\hat{N}}\left(\mathbf{1}_{\left[y_k \le \widehat{\text{ARPT}}\right]} - \widehat{\text{ARPR}}\right) + f(\widehat{\text{ARPT}})\,\hat{z}_k^{\text{ARPT}}.\end{aligned}$$

Ici, la fonction de densité des revenus doit être estimée en deux points : en la médiane et en l'ARPT.

2.6 La médiane des pauvres

Il s'agit du revenu médian des personnes en-dessous du seuil de risque de pauvreté : $m_p = F^{-1}(1/2\,F(\text{ARPT}))$. On l'estime selon la même procédure que tout autre quantile dont la définition exacte peut varier. La linéarisée de m_p (Osier 2009) dépend de celle de l'ARPR :

$$\hat{z}_k^{m_p} = \frac{1}{f(\hat{m}_p)}\frac{\hat{z}_k^{\text{ARPR}}}{2} - \frac{1}{\hat{N}}\left(\mathbf{1}_{\left[y_k \le \hat{m}_p\right]} - F(\hat{m}_p)\right).$$

L'estimation de la densité des revenus intervient donc trois fois : en la médiane et en l'ARPT dans $\hat{z}_k^{\text{ARPR}}$ et en la médiane des pauvres m_p.

2.7 Le Relative Median Poverty Gap (RMPG)

Il s'agit de la différence relative entre le seuil de risque de pauvreté et la médiane des pauvres. On définit $\text{RMPG} = 0$ si tous les « pauvres » gagnent un montant égal au seuil et $\text{RMPG} = 1$ si les pauvres ne gagnent rien du tout. C'est une mesure de « combien pauvre sont les pauvres » :

$$\text{RMPG} = \frac{\text{ARPT} - m_p}{\text{ARPT}}.$$

Son estimation à partir de l'échantillon est donc déjà décrite. L'influence de chaque observation sur le RMPG est donnée par Osier (2009) :

$$\hat{z}_k^{\text{RMPG}} = \frac{\hat{m}_p \hat{z}_k^{\text{ARPT}} - \widehat{\text{ARPT}}\, \hat{z}_k^{m_p}}{\widehat{\text{ARPT}}^2}.$$

L'estimation de la densité de la distribution des revenus intervient à quatre reprises : une fois dans le calcul de $\hat{z}_k^{\text{ARPT}}$ et trois fois dans celui de $\hat{z}_k^{m_p}$.

3 L'estimation de la fonction de densité des revenus

Dans une approche basée sur le plan (*design based*) en population finie, l'inférence se fait par rapport au plan de sondage $\text{P}(S)$ utilisé pour sélectionner l'échantillon S dans la population U de taille finie N. Dans cette approche, seules les indicatrices d'inclusion dans l'échantillon sont aléatoires, toutes les autres grandeurs sont fixes. La fonction de répartition des revenus au niveau de la population est alors une fonction en escaliers : $F_y(x) = \sum_{k \in U} \mathbf{1}_{y_k \le x} \big/ N$ et sa dérivée, la fonction de densité, n'existe pas à cause des discontinuités. Si l'on ne veut pas se placer dans une approche basée sur le modèle (*model based*) avec un modèle de super population pour justifier le terme de fonction de densité des revenus, il faut artificiellement lisser la fonction de répartition pour qu'elle devienne dérivable. C'est donc par abus de langage que nous nous autorisons ici à parler de fonction de densité. Avec cette volonté de lissage, Deville (2000) et Osier (2009) proposent d'estimer la fonction de densité des revenus par noyau gaussien :

$$\begin{aligned} K(u) &= \frac{1}{h\sqrt{2\pi}} e^{-u^2/2}, \qquad u = \frac{x - y_k}{h} \\ \hat{f}_1(x) &= \frac{1}{\hat{N}} \sum_{k \in S} w_k K\left(\frac{x - y_k}{h}\right) \\ &= \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \exp\left[-\frac{(x - y_k)^2}{2h^2}\right] \end{aligned} \tag{3.1}$$

où h est la largeur de la bande qu'Osier estime par $\hat{h} = \hat{\sigma}\hat{N}^{-0,2}$ et $\hat{\sigma}$ est l'écart-type estimé de la distribution empirique des revenus :

$$\hat{\sigma} = \sqrt{\frac{\sum_{k\in S} w_k y_k^2}{\hat{N}} - \left(\frac{\sum_{k\in S} w_k y_k}{\hat{N}}\right)^2} = \sqrt{\frac{\sum_{k\in S} w_k y_k^2}{\hat{N}} - \bar{y}_w^2}.$$

Notons que cette estimation de σ n'est pas robuste étant très sensible aux valeurs extrêmes de y. Des données de revenus ont très souvent une queue de distribution étendue vers la droite avec des valeurs très élevées possibles, on parle de valeurs aberrantes représentatives (*representative outliers*) au sens de Chambers (1986) et Hulliger (1999). Comme le montrent nos simulations dans la section 4, cela peut fortement biaiser nos estimations de variance. Verma et Betti (2011) procèdent également par noyau rappelant que, selon Silverman (1986), le choix du noyau n'est pas crucial pour assurer la convergence de $\hat{f}(y)$ vers $f(y)$, alors que celui de la largeur de bande l'est. Ils utilisent une valeur conseillée par Silverman dans le cas de distributions avec un coefficient d'asymétrie positif, $h = 0{,}79(\hat{Q}_{75} - \hat{Q}_{25})\hat{N}^{-0,2}$. Dans leurs conclusions, ils relèvent que la méthode par linéarisation peut être problématique en raison d'irrégularités de la fonction de densité empirique. On ajoutera que ces problèmes sont d'autant plus préoccupants qu'il est fréquent, dans les données issues d'enquêtes, d'avoir des agglomérats d'observations à certaines valeurs (dues à des arrondis ou des questions-fourchettes), ce qui peut compliquer l'estimation de la densité. La suite de l'article décrit des solutions que nous avançons pour réduire le biais de la variance estimée.

3.1 Passer par le logarithme

Une solution qui, comme on le verra plus loin, donne de très bons résultats est de simplement passer par le logarithme pour estimer la densité en x. Si l'on pose $v = \log(x + a)$, où x est le revenu et a un nombre réel positif par exemple égal à $(|\min_k(y_k)| + 1)$ dans le cas où l'on aurait des revenus négatifs ou nuls (en négligeant le fait que a serait estimé), on a que

$$F_v(v) = \mathrm{P}(\mathcal{V} \leq v) = \mathrm{P}(\log(\mathcal{Y} + a) \leq v) = \mathrm{P}(\mathcal{Y} \leq e^v - a) = F_y(e^v - a),$$

où $\mathcal{V}$ et $\mathcal{Y}$ seraient de variables aléatoires. Donc,

$$f_v(v) = \frac{dF_v(v)}{dv} = \frac{dF_y(e^v - a)}{dv} = f_y(e^v - a)e^v.$$

Autrement dit $f_v(v) = f_y(x)(x + a)$, ce qui nous donne l'estimateur suivant de la densité en x :

$$\hat{f}_2(x) = \frac{\hat{f}_v(v)}{x + a} = \frac{\hat{f}_y(\log(x + a))}{x + a}. \tag{3.2}$$

L'estimation de la densité en x de $\mathcal{Y}$ peut donc s'évaluer en estimant celle du logarithme de la variable, divisée par la valeur de la variable au point qui nous intéresse. La propriété reste valable en population finie. Le fait de passer par le logarithme a l'avantage de diminuer l'effet levier exercé par les grandes valeurs des revenus dans le calcul de l'approximation de la densité par noyau. Les simulations montrent que cette méthode très simple réduit fortement le biais.

3.2 Plus proches voisins avec largeur de bande minimale

Deville (2000) esquisse une autre manière du type « plus proches voisins » (voir Silverman 1986) d'estimer la densité en utilisant le noyau

$$K_D(u) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \le u < b \\ 0 & \text{sinon,} \end{cases},$$

avec $u = y_k$ et où le choix de a et b, vérifiant $x \in [a,b]$, reste à déterminer et pourrait dépendre de x. La distance $(b-a)$ représente la largeur de bande h. L'estimation de la densité vaudrait alors

$$\begin{aligned} \hat{f}_D(x,a,b) &= \frac{1}{\hat{N}} \sum_{k \in S} K_D(y_k) \\ &= \frac{1}{\hat{N}} \sum_{k \in S} w_k \frac{1}{b-a} \mathbf{1}_{y_k \in [a,b[} \\ &= \frac{\hat{F}_y(b) - \hat{F}_y(a)}{b-a}, x \in [a,b[\end{aligned} \tag{3.3}$$

avec $\hat{F}_y(x) = \sum_{k \in S} w_k \mathbf{1}_{y_k \le x} \Big/ \hat{N}$.

Notons que l'estimation de la densité (3.3) n'est pas une fonction continue et qu'elle ne serait pas très adaptée pour estimer des valeurs de densité à l'extrémité des queues de la distribution. Puisque nos travaux ne reposent pas trop sur les queues de la distribution, nous considérons cette approche comme une option.

Notre deuxième proposition d'estimation de la densité en x s'inspire de l'idée ci-dessus. Elle est du type « plus proches voisins », mais impose aussi une largeur de bande minimale : notre méthode impose d'utiliser au minimum les p plus proches observations du point x tout en imposant une largeur de bande minimale $h(p) \ge h_{\text{opt}}$ où

$$h_{\text{opt}} = \frac{0{,}9 \min(\hat{\sigma}, \hat{Q}_{75} - \hat{Q}_{25})}{1{,}34 \sqrt[5]{\hat{N}}}$$

est la règle empirique (*rule of thumb*) de Silverman (1986) pour déterminer la largeur de la bande. Cette valeur est aussi utilisée par défaut par la fonction R `density` pour la largeur de la bande si rien n'est spécifié. Cette solution est plus robuste que (3.1) et évite les problèmes que l'on rencontre lorsque plusieurs valeurs y_k sont très proches les unes des autres, ce qui arrive fréquemment parce que les personnes interrogées ont tendance à arrondir leur revenu.

Les valeurs y_k, $k = 1,...,n$, étant supposées ordonnées par leur rang, la largeur $h(p)$ de la fenêtre autour de x est initialement déterminée par les p plus proches observations, avec $p \ll n$. Pour les simulations exposées dans la section suivante, après différents essais, le p initial a été fixé à 30. On impute comme densité en x la densité estimée au point observé y_j le plus proche inférieur ou égal à x, c'est-à-dire $j = \max(k \mid y_k \le x)$, $k = 1,...,n$. La largeur de la bande en x dépendra en fait des p_j plus proches observations autour de y_j, avec $p_j \ge p$. On la désignera par $h(p_j)$ dans la suite pour le

rappeler. La densité n'est donc estimée qu'en des points observés sans qu'un lissage ou une interpolation soient menés entre les $\hat{f}(y_j)$. L'algorithme pour estimer $\hat{f}(y_j)$ est le suivant (voir aussi Figure 3.1) :

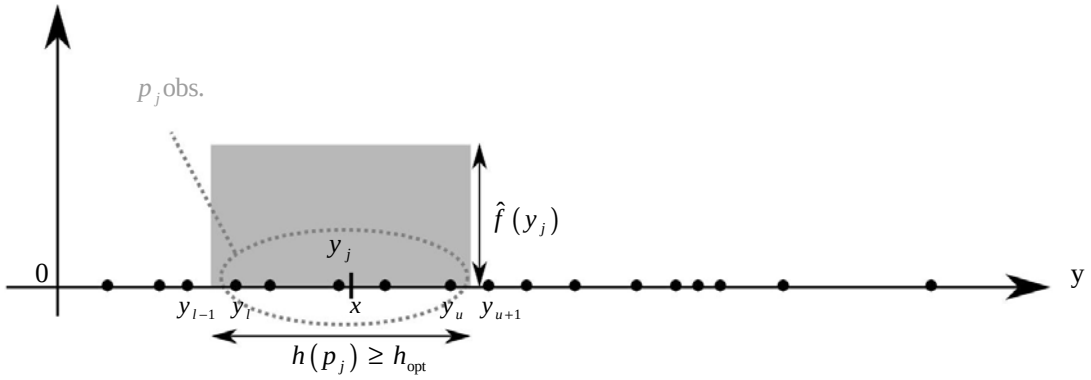

Figure 3.1 Largeur de la fenêtre $h(p_j)$

1. La largeur initiale de la fenêtre autour du point y_j, avec $p_j = p$, est définie par :

$$h(p_j) = \frac{y_u + y_{u+1}}{2} - \frac{y_\ell + y_{\ell-1}}{2}; \quad u = \begin{cases} j + p_j/2 - 1 & \text{si } p_j \text{ est pair} \\ j + \lfloor p_j/2 \rfloor & \text{si } p_j \text{ est impair} \end{cases}$$
$$\ell = j - \lfloor p_j/2 \rfloor.$$

2. Si la largeur de fenêtre $h(p_j)$ ainsi obtenue est inférieure à h_{opt}, on incrémente les deux bornes :
   borne supérieure : $u \to u + 1$, tant que $u < n$,
   borne inférieure : $l \to l - 1$, tant que $l > 1$,
   ce qui implique $p_j \to p_j + 2$, sauf si $u = n$ ou $l = 1$, on n'a alors plus le même nombre de points à gauche et à droite de y_j.

3. Répéter 2 jusqu'à ce que $h(p_j) \geq h_{\text{opt}}$.

4. La densité estimée en x est alors donnée par

$$\hat{f}(x) = \hat{f}(y_j) = \begin{cases} \dfrac{p_j}{nh(p_j)} & \text{sans pondération,} \\ \dfrac{\sum_{p_j \text{ plus proche s de } y_j} w_j^{\text{std}}}{nh(p_j)} & \text{avec pondération,} \end{cases}$$

   avec les poids standardisés $w_k^{\text{std}} = w_k / \overline{w}$, $k = 1, \ldots, n$.

Le nombre d'observations p_j prises en compte pour le calcul peut varier et dépend de la courbure locale de la fonction de répartition empirique. La condition $h(p_j) \geq h_{opt}$ garantit une largeur de fenêtre minimale aux endroits où beaucoup d'observations seraient concentrées sur un petit intervalle. On rend la procédure encore plus solide en combinant cette approche avec la précédente, c'est-à-dire en estimant la densité du logarithme de la variable divisée par sa valeur (non logarithmisée) :

$$\hat{f}_3(x) = \frac{\hat{f}(\log(x+a))}{x+a}. \tag{3.4}$$

3.3 Robustesse de la linéarisée

Comme mentionné plus haut, dans le cas de la médiane ou pour les autres quantiles, Croux (1998) relève que la fonction d'influence empirique ou linéarisée estimée à partir de l'échantillon n'est pas aussi robuste qu'il n'y paraît, même si l'on connaît la fonction de densité. Nous avons vérifié cela pour les données SILC utilisées dans les simulations modélisées avec une loi Bêta Généralisée de seconde espèce (GB2) grâce à la fonction `profml.gb2` de R (Graf et Nedyalkova 2011). Sur de petits échantillons $(n \leq 100)$, le biais potentiel de la linéarisée engendré par un trop grand nombre de valeurs extrêmes peut aussi biaiser l'estimation de la variance calculée à partir de cette dernière. Pour de plus grands échantillons $(n \geq 1\,000)$, un biais relatif maximal dans la variance estimée à l'aide de la linéarisée empirique *vs.* théorique peut atteindre jusqu'à 5 %. Il est cependant en-dessous du pourcent en valeur absolue dans les trois quarts des cas.

4 Résultats

Des simulations ont été menées sur trois jeux de données réelles afin de comparer et évaluer les différentes méthodes d'estimation de la fonction de densité : $\hat{f}_1(x)$ voir (3.1), $\hat{f}_2(x)$ voir (3.2) et $\hat{f}_3(x)$ voir (3.4). Ces dernières sont nécessaires à l'estimation de la variance de certains indices de pauvreté et d'inégalité.

1. Le premier jeu de données correspond aux revenus équivalents des ménages de l'enquête SILC suisse menée par l'Office Fédéral Suisse de la Statistique en 2009. On y dispose de 17 534 individus avec un revenu équivalent non nul.
2. Le deuxième jeu de données est aussi issu de l'enquête SILC 2009 mais se restreint aux personnes salariées. Il s'agit des salaires du registre de la Centrale de Compensation qui a été couplé avec les répondants de l'enquête. Nous n'avons donc pas de problème de non-réponse et disposons de 7 922 individus avec un salaire non nul.
3. Le troisième fichier de test, nommé *Ilocos*, est celui à disposition dans le package `ineq` de R (Zeileis 2012). Il contient 632 observations. Il s'agit du revenu des ménages d'une des 16 régions des Philippines appelée Ilocos. Les données sont issues de deux enquêtes de l'Institut National de Statistiques des Philippines en 1997 et 1998.

Les trois jeux de données ont un coefficient d'asymétrie positif, ce qui est typique des distributions de revenus. Chaque jeu de données étant considéré comme une population, on a sélectionné dans un premier temps 10 000 échantillons aléatoires simples sans remise de différentes tailles. Les valeurs des différents indices sont calculées pour chacun des échantillons, ce qui nous procure une estimation Monte Carlo de leur variance, $\text{var}_{\text{sim}}(\hat{\theta})$, pour un indice de pauvreté ou d'inégalité θ. L'estimateur par linéarisation de la variance est noté $\widehat{\text{var}}_{\text{lin}}(\hat{\theta})$ et est calculé en utilisant la linéarisée $\hat{z}^{\hat{\theta}}$ estimée pour chaque échantillon :

$$\widehat{\text{var}}_{\text{lin}}(\hat{\theta}) = \frac{N(N-n)}{n}\text{var}\left(\hat{z}_S^{\hat{\theta}}\right),$$

où n est la taille d'échantillon utilisée pour les simulations et

$$\text{var}\left(\hat{z}_S^{\hat{\theta}}\right) = \frac{1}{n-1}\sum_{k\in S}\left(\hat{z}_{S,k}^{\hat{\theta}} - \overline{z}_S^{\hat{\theta}}\right)$$

avec $\overline{z}_S^{\hat{\theta}} = n^{-1}\sum_S \hat{z}_{S,k}^{\hat{\theta}}$, voir (2.1).

La qualité de l'estimateur de variance par linéarisation est évaluée en comparant la valeur Monte Carlo espérée de la variance estimée par linéarisation, notée $E_{\text{sim}}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right]$, à la « vraie » variance Monte Carlo $\text{var}_{\text{sim}}(\hat{\theta})$ en terme de biais relatif :

$$\text{RB}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] = \frac{E_{\text{sim}}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] - \text{var}_{\text{sim}}(\hat{\theta})}{\text{var}_{\text{sim}}(\hat{\theta})}. \tag{4.1}$$

Pour le deuxième jeu de données (SILC 2009, revenu des personnes salariées) nous avons, dans un deuxième temps, également sélectionné 10 000 échantillons aléatoires sans remise selon un plan stratifié, puis calé les poids de sondage de manière à respecter huit marges sociodémographiques de totaux connus sur la population de 7 922 individus. Les cinq strates utilisées correspondent à des tranches d'âge des personnes salariées (voir tableau 4.1).

Les huit cellules de calage ont été obtenues par les croisements des trois variables dichotomiques suivantes (variables auxiliaires du calage) :

1. MARIÉ indiquant si la personne est mariée ou pas,
2. CHEF indiquant si la personne a une fonction de chef ou pas dans son emploi,
3. HOMME renseignant sur le sexe de l'individu.

Les totaux sur la population de 7 922 individus pour ces cellules de calage sont indiqués dans le tableau 4.2.

Tableau 4.1
Définition des strates utilisées dans les simulations sur les données SILC 2009 et trois tailles d'échantillons (revenu des personnes salariées, N = 7 922)

strate h	description	N_h	%	n_h		
1	personnes de moins de 25 ans	1 187	15,0	75	112	150
2	personnes de 26 à 35 ans	1 359	17,2	86	129	171
3	personnes de 36 à 45 ans	2 137	27,0	135	202	270
4	personnes de 46 à 55 ans	1 864	23,5	117	177	235
5	personnes de plus de 55 ans	1 375	17,4	87	130	174
	TOTAL	7 922	100,0	500	750	1 000

Tableau 4.2
Marges de calage utilisées dans les simulations sur les données SILC 2009 (revenu des personnes salariées, $N = 7\,922$)

marge	MARIÉ	CHEF	HOMME	Tot. pop.	%
1	0	0	0	1 487	18,8
2	0	0	1	1 208	15,2
3	0	1	0	323	4,1
4	0	1	1	457	5,8
5	1	0	0	1 759	22,2
6	1	0	1	1 278	16,1
7	1	1	0	328	4,1
8	1	1	1	1 082	13,7
			TOTAL	7 922	100,0

Pour chaque échantillon stratifié, un calage sur marge (méthode linéaire) a été effectué de manière à ce que les sommes des poids après calage respectent les huit marges citées ci-dessus. Les estimations ponctuelles des indicateurs ainsi que leur linéarisée ont été calculées pour chaque échantillon à l'aide des poids calés.

Pour estimer la variance, on a appliqué la méthode développée par Deville (2000). Celle-ci consiste à linéariser aussi par rapport au calage en calculant les résidus $e^{\hat{\theta}}$ de la régression (pondérée par les poids de sondage) des linérarisées des indicateurs sur les variables auxiliaires du calage. La variance du total des résidus ainsi calculés sous plan aléatoire stratifié sans remise est alors un estimateur de la variance de l'indicateur estimé, ce qui est la grandeur qui nous intéresse :

$$\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta}) = \sum_{h=1}^{H} \frac{N_h}{n_h}(N_h - n_h)\, s^2_{e^{\hat{\theta}}_h} \tag{4.2}$$

avec

$$s^2_{e^{\hat{\theta}}_h} = \frac{1}{n_h - 1}\sum_{k \in S_h}\left(e_k^{\hat{\theta}} - \bar{e}^{\hat{\theta}}\right)^2$$

La qualité de l'estimateur de variance par linéarisation est évaluée de manière analogue au cas du sondage aléatoire simple, voir (4.1).

Les tableaux 4.3, 4.4 et 4.5 présentent les biais relatifs de la variance obtenus pour les trois jeux de données utilisés et décrits plus haut dans le cas du sondage aléatoire simple. Le tableau 4.6 présente les biais relatifs de la variance obtenus dans le cas du sondage aléatoire stratifié avec poids calés. Dans les parties supérieures des tableaux, on trouve les valeurs pour l'indice de Gini et le QSR. Ces derniers ne nécessitent pas d'estimer la fonction de densité des revenus. L'estimation de leur variance fonctionne bien. On note un problème de sous-estimation de la variance pour l'indice de Gini dans le cas stratifié avec calage (tableau 4.6).

Dans le cas du premier jeu de données considéré, le tableau 4.3 ne révèle pas de très grandes différences si ce n'est que l'estimation de la densité de revenus par $\hat{f}_3(x)$ mène à des résultats plus conservateurs. En effet, le biais relatif reste du même ordre de grandeur mais positif, alors qu'il est négatif pour les deux autres manières d'estimer la densité. Pour le deuxième jeu de donnée, le tableau 4.4 montre qu'il est essentiel de passer par le logarithme ou d'utiliser la méthode par plus proches voisins avec

largeur de bande minimale. Avec cette dernière méthode, tous les biais relatifs tombent en-dessous de 10 % dès que les tailles d'échantillons sont suffisamment grandes (voir la dernière colonne du tableau). Les simulations sur les mêmes données avec un plan de sondage stratifié et un calage renforcent et confirment ces résultats (voir tableau 4.6). Pour le troisième jeu de données, le tableau 4.5 indique les mêmes tendances, bien que les résultats soient plus instables en raison des petites tailles d'échantillon et de population. Cette constatation n'est pas surprenante, le nombre minimal de voisins à prendre en compte ayant été fixé à 30. En l'occurrence, pour le jeu de données Ilocos, des simulations avec un p plus petit fixé à 10 ne changent rien au final, car la condition $h(p_j) \geq h_{\text{opt}}$ le remonte systématiquement au-dessus de 30.

De plus, de manière générale, on constate que plus il faut avoir recours à l'estimation de la densité par noyau gaussien - $\hat{f}_1(x)$ - plus l'erreur augmente. En effet, les biais relatifs de la variance pour la médiane des pauvres et le RMPG sont presque systématiquement plus grands en valeur absolue que ceux pour les autres indicateurs. Pour le RMPG, l'erreur peut se compenser (cas du tableau 4.3) s'il y a suffisamment d'observations car l'estimation de la densité intervient tant au numérateur qu'au dénominateur.

Tableau 4.3
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires simples sans remise sur les données SILC 2009 (revenu équivalent des ménages, $N = 17\,534$)

	Tailles d'échantillons (taux de sondage)								
Indice	$n = 500$ (2,9 %)			$n = 750$ (4,3 %)			$n = 1\,000$ (5,7 %)		
GINI	-0,02			-0,02			-0,02		
QSR	0,01			0,00			0,00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0,08	-0,06	0,04	-0,09	-0,07	0,03	-0,09	-0,07	0,04
ARPR	-0,05	-0,01	-0,00	-0,09	-0,06	-0,05	-0,08	-0,05	-0,03
RMPG	-0,09	-0,07	*0,15*	*-0,10*	-0,07	*0,12*	-0,09	-0,06	*0,14*
MEDP	*-0,16*	*-0,12*	0,09	*-0,19*	*-0,13*	0,05	*-0,18*	*-0,11*	0,07
MED	-0,08	-0,06	0,05	-0,08	-0,06	0,04	-0,08	-0,06	0,04

Tableau 4.4
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires simple sans remise sur les données SILC 2009 (revenu des personnes salariées, $N = 7\,922$)

	Tailles d'échantillons (taux de sondage)								
Indice	$n = 500$ (6,3 %)			$n = 750$ (9,5 %)			$n = 1\,000$ (12,6 %)		
GINI	-0,03			-0,03			-0,02		
QSR	-0,00			0,00			0,00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	0,07	0,05	*0,13*	0,06	0,04	*0,10*	0,06	0,03	0,08
ARPR	-0,05	-0,04	-0,02	-0,05	-0,04	-0,01	-0,06	-0,05	-0,02
RMPG	*0,61*	*0,12*	*0,15*	*0,60*	*0,11*	0,08	*0,59*	0,09	0,05
MEDP	*0,73*	*0,17*	*0,18*	*0,72*	*0,16*	*0,10*	*0,72*	*0,15*	0,07
MED	0,07	0,04	*0,13*	0,06	0,04	*0,10*	0,05	0,03	0,07

Tableau 4.5
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires simple sans remise sur les données Ilocos (revenu des ménages, $N = 632$)

Indice	Tailles d'échantillons (taux de sondage)					
	$n = 50$ (7,9 %)			$n = 63$ (10,0 %)		
GINI	*-0,16*			*-0,13*		
QSR	0,00			0,00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0,05	-0,06	-0,01	-0,03	-0,03	-0,01
ARPR	*-0,31*	-0,01	*-0,12*	*-0,33*	-0,03	*-0,18*
RMPG	*1,55*	*0,83*	*0,26*	*1,54*	*0,16*	*0,39*
MEDP	*1,02*	*0,28*	*-0,26*	*1,05*	0,07	*-0,11*
MED	0,04	0,03	0,08	0,07	0,07	0,09

Tableau 4.6
Biais relatifs (4.1) de la variance obtenus avec 10 000 échantillons aléatoires stratifiés sans remise, avec poids calés sur huit marges sociodémographiques, sur les données SILC 2009 (revenu des personnes salariées, $N = 7\,922$)

Indice	Tailles d'échantillons (taux de sondage)								
	$n = 500$ (6,3 %)			$n = 750$ (9,5 %)			$n = 1\,000$ (12,6 %)		
GINI	*-0,21*			*-0,20*			*-0,20*		
QSR	-0,06			-0,06			-0,07		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0,07	-0,09	-0,01	-0,08	*-0,10*	-0,04	-0,09	*-0,11*	-0,06
ARPR	*-0,10*	*-0,10*	-0,08	-0,07	-0,06	-0,05	-0,06	-0,06	-0,05
RMPG	*0,63*	*0,13*	*0,13*	*0,61*	*0,11*	0,08	*0,59*	*0,10*	0,04
MEDP	*0,71*	*0,16*	*0,15*	*0,68*	*0,13*	0,09	*0,66*	*0,12*	0,04
MED	-0,07	-0,09	-0,01	-0,08	*-0,10*	-0,04	-0,08	*-0,11*	-0,06

En résumé, on voit que l'on peut tout autant surestimer la variance $\left(\mathrm{RB}[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})] > 0\right)$ que la sous-estimer $\left(\mathrm{RB}[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})] < 0\right)$ selon l'indice et le jeu de données considérés. Le fait de passer au logarithme $(\hat{f}_2(x))$ améliore très significativement la situation. La méthode par « plus proches voisins » $(\hat{f}_3(x))$ fait disparaître tout problème si on a suffisamment de données (cas des tableaux 4.3, 4.4 et 4.6). Cette dernière a un peu plus de difficulté si les échantillons sont petits (cas du tableau 4.5). Les variations illogiques et les biais qui subsistent dans les tableaux peuvent aussi être dus au manque de robustesse des linéarisées sur certains échantillons comme mentionné à la section 3.3.

5 Conclusions

Dans plusieurs pays, des enquêtes par échantillonnage de niveau national publient les valeurs extrapolées des indices de Laeken (Eurostat 2005) car elles constituent des indicateurs importants permettant d'aiguiller les preneurs de décisions aux niveaux politique et social. Il est dès lors essentiel de pouvoir quantifier la précision de ces mesures, ce qui soulève la question de la pertinence de l'estimation

de la précision dont on dispose. Ce travail montre qu'une amélioration substantielle peut être faite dans l'estimation de la précision d'indices de pauvreté et d'inégalité qui nécessitent le recours à une estimation (locale) de la densité des revenus ou de la variable monétaire considérée.

Les simulations menées montrent que la méthode d'estimation de la densité par noyau gaussien actuellement mise en œuvre dans la plupart des cas n'est pas à recommander sans au moins passer par le logarithme comme proposé dans la section 3.1, faute de quoi le biais de la variance estimée peut devenir très important. La méthode par « plus proches voisins » (section 3.2) qui impose aussi une largeur de bande minimale peut donner des résultats encore meilleurs, surtout si on a des agglomérats d'observations en certaines valeurs dans les données étudiées. Elle nécessite cependant de choisir un nombre minimal p de voisins en fonction des données sur lesquelles on travaille. Si l'on a peu d'observations à disposition, il est préférable d'en rester à la variante du logarithme. Dans tous les cas, nous espérons que ce travail contribuera à une prise de conscience du fait qu'il faut être méticuleux lors de l'implémentation du calcul des linéarisées pour tous les indicateurs faisant intervenir des quantiles.

6 Remerciements

Ce travail a été réalisé dans le cadre d'une convention de collaboration entre l'Institut de Statistique de l'Université de Neuchâtel et l'Office Fédéral de la Statistique (OFS). Nous tenons également à remercier en particulier la section *Revenus, consommation et conditions de vie* de l'OFS pour nous avoir mis à disposition les données de la partie suisse de l'enquête européenne sur le revenu et les conditions de vie. Merci aussi à Matti Langel et Anne Massiani pour leur appui au cours de nos investigations.

Bibliographie

Antal, E., Langel, M. et Tilllé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. *Proceedings 58th World Statistical Congress*, Dublin.

Ardilly, P., et Osier, G. (2007). Cross-sectional variance estimation for the French "Labour Force Survey". *Survey Research Methods*, 1, 75-83.

Beaumont, J.-F., Haziza, D. et Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Croux, C. (1998). Limit behaviour of the empirical influence function of the median. *Statistics & Probability Letters*, 37, 331-340.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 1, 17-27.

Deville, J.-C. (2000). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques de résidus. *Techniques d'enquêtes*, 25, 2, 219-230.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Eurostat (2004a). Common cross-sectional eu indicators based on eu-silc; the gender pay gap. Documents de travail et d'études de l'Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-rev/04.

Eurostat (2004b). Theoretical study of the gini index. Documents de travail et d'études de l'Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-A/04.

Eurostat (2005). The continuity of indicators during the transition between ECHP and EU-SILC. Documents de travail et d'études de l'Office for Official Publications of the European Communities, Luxembourg.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96, 691-709.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. Dans *Compositional Data Analysis, Theory and Applications*, (Éds., V. Pawlosky-Glahn et A. Buccianti), Oxford : Wiley, chapitre 9, 114-127.

Graf, M. (2013). A simplified approach to linerarization variance for surveys. *University of Neuchâtel*, document de travail.

Graf, M., et Nedyalkova, D. (2011). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.0.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 54-63.

Hyndman, R.J., et Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50, 361-365.

Kovačević, M.S., et Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

Langel, M., et Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976-2985.

Langel, M., et Tillé, Y. (2012). Variance estimation of the Gini index: Revisiting a result several times published. Dans *Press in Journal of the Royal Statistical Society - Series A.*

Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *American Statistical Association*, 9, 209-219.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Londres.

Verma, V., et Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics*, 38, 1549-1576.

Zeileis, A. (2012). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-10.