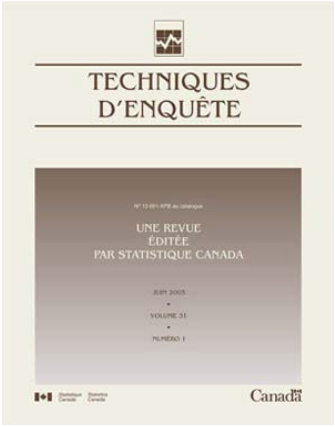
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

# Linéarisation contre Bootstrap pour estimer la variance de l'évolution de l'indice de Gini




par Guillaume Chauvet et Camelia Goga

Date de diffusion : le 21 juin 2018

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) |

Techniques d'enquête, juin 2018
Vol. 44, nº 1, pp. 19-44
Statistique Canada, nº 12-001-X au catalogue


# Linéarisation contre Bootstrap pour estimer la variance de l'évolution de l'indice de Gini

**Guillaume Chauvet et Camelia Goga**[1]

Résumé

Le présent article traite de l'estimation de la variance par linéarisation ou par bootstrap pour l'indice de Gini, et pour l'évolution de cet indice entre deux périodes. Dans le cas d'un seul échantillon, nous adoptons l'approche de linéarisation par la fonction d'influence proposée par Deville (1999), la méthode du bootstrap sans remise proposée par Gross (1980) pour l'échantillonnage aléatoire simple sans remise, et la méthode de tirage avec remise des unités primaires d'écrite dans Rao et Wu (1988) pour l'échantillonnage à plusieurs degrés. Pour obtenir un estimateur de variance dans le cas de deux échantillons, nous utilisons la technique de linéarisation au moyen de fonctions d'influence partielles (Goga, Deville et Ruiz-Gazen, 2009). Nous élaborons aussi une extension des procédures bootstrap étudiées à l'échantillonnage bidimensionnel. Les deux approches sont comparées sur des données simulées.

**Mots-clés :** Estimateur composite; estimateur de Horvitz-Thompson; fonction d'influence; estimateur « intersection »; poids de rééchantillonnage; sondage à deux échantillons; plan d'échantillonnage bidimensionnel; estimateur « union »; estimation de la variance.

## 1 Introduction

L'indice de Gini (Gini, 1914) est l'une des mesures de concentration les plus connues dont l'utilisation est fréquente dans les études économiques. Si $\mathcal{Y}_1$ désigne une variable quantitative positive, telle que le revenu, et $F_1(\cdot)$, sa fonction de répartition définie sur $]-\infty, \infty[$, l'indice de Gini est donné par

$$G_1 = \frac{1}{2}\frac{\iint |v-u|\,dF(u)\,dF(v)}{\int u dF(u)},$$

en supposant que $\int u dF(u) \neq 0$. L'indice de Gini mesure la dispersion d'une variable quantitative positive à l'intérieur d'une population. Habituellement, les instituts de statistique se servent de l'indice de Gini pour évaluer les inégalités de revenu dans un pays à différentes périodes, ou entre différents pays à la même période. Au cours des dernières décennies, l'indice de Gini a également été utilisé dans les domaines de l'économie et de la sociodémographie (voir, par exemple, Navarro, Muntaner, Borrell, Benach, Quiroga, Rodriguez-Sanz, Vergès et Pasarin, 2006; Bhattacharya, 2007; Lai, Huang, Risser et Kapadia, 2008; Barrett et Donald, 2009), de la biologie (Graczyk, 2007), de l'environnement (Druckman et Jackson, 2008; Groves-Kirkby, Denman et Phillips, 2009) ou de l'astrophysique (Lisker, 2008).

Il existe une abondante littérature sur l'estimation de la variance de l'indice de Gini pour des observations tirées de données de sondage; pour une revue, voir Langel et Tillé (2013). Glasser (1962) et Sandström, Wretman et Waldèn (1985) ont étudié le cas de l'échantillonnage aléatoire simple. Sandström, Wretman et Waldèn (1988) ont dressé la liste d'estimateurs de variance possibles pour un plan d'échantillonnage général, y compris un estimateur de variance par le jackknife. Cette dernière approche a été étudiée plus en

1. Guillaume Chauvet, Université Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France. Courriel : chauvet@ensai.fr; Camelia Goga, Laboratoire des Mathématiques de Besançon, Université de Bourgogne Franche-Comté, UMR 6623, Besançon Cedex, France. Courriel : camelia.goga@univ-fcomte.fr.

détail par Yitzhaki (1991), Karagiannis et Kovačević (2000), et Berger (2008). Kovačević et Binder (1997) ont exploré l'estimation de la variance par linéarisation, et Berger (2008) a démontré l'équivalence entre la linéarisation et une technique de jackknife généralisée que Campbell (1980) avait été le premier à suggérer. Qin, Rao et Wu (2010) ont proposé des intervalles de confiance fondés sur le bootstrap et sur la vraisemblance empirique pour l'indice de Gini. Ils ont procédé à l'étude théorique ainsi qu'empirique de ces méthodes dans le cas particulier de l'échantillonnage aléatoire simple avec remise stratifié. En revanche, pour l'évolution de l'indice de Gini, l'estimation de la variance par bootstrap n'a pas été comparée à d'autres méthodes.

Dans le présent article, nous examinons la linéarisation comparativement au bootstrap pour estimer l'évolution de l'indice de Gini. La présentation de l'article est la suivante. À la section 2, nous examinons d'abord l'estimation de l'indice de Gini dans le cas d'un seul échantillon. La notation est définie à la section 2.1, et l'estimateur par substitution de l'indice de Gini est présenté à la section 2.2. L'estimateur de variance par linéarisation est donné à la section 2.3, avec une application à un plan d'échantillonnage aléatoire simple (SI) et à un plan d'échantillonnage à plusieurs degrés. Les principes fondamentaux du bootstrap pondéré sont passés brièvement en revue au début de la section 2.4, et le bootstrap sans remise (BWO pour *Bootstrap Without Replacement*) approprié pour l'échantillonnage SI est présenté à la section 2.4.1, tandis que le bootstrap avec remise des unités primaires d'échantillonnage (BWR pour *Bootstrap With Replacement*) approprié pour l'échantillonnage à plusieurs degrés est présenté à la section 2.4.2. À la section 3, nous considérons l'estimation de l'évolution de l'indice de Gini dans le cas de deux échantillons. La notation est définie à la section 3.1, et les principes de l'estimation composite qui est appliquée sont examinés brièvement à la section 3.1.1 pour le plan SI bidimensionnel (SI2), et à la section 3.1.2, pour un plan d'échantillonnage à deux degrés bidimensionnel (MULT2). L'estimateur composite de la variation de l'indice de Gini est présenté à la section 3.2. L'estimateur de variance par linéarisation au moyen des fonctions d'influence partielles est donné à la section 3.3, avec application au plan SI2 et au plan MULT2. Les extensions du BWO pour le plan SI2 et du BWR pour le plan MULT2 sont présentées à la section 3.4. La méthode de linéarisation et les méthodes bootstrap proposées sont comparées à la section 4 au moyen d'une étude en simulation. Les conclusions sont présentées à la section 5.

# 2 Le cas d'un seul échantillon

## 2.1 Notation

Soit $U$ une population finie de taille $N$ dont les unités peuvent être identifiées par les étiquettes $k = 1, \ldots, N$. Supposons que la variable $\mathcal{Y}_1$ est mesurée sur la population $U$, et soit $y_{11}, \ldots, y_{1N}$ les valeurs prises par $\mathcal{Y}_1$ sur les unités de la population. Soit $M_1 = \sum_{k \in U} \delta_{y_{1k}}$ la mesure discrète prenant une masse unitaire en tout point $y_{1k}$ de la population et 0 ailleurs, où $\delta_{y_{1k}}$ est la masse de Dirac au point $y_{1k}$. La plupart des paramètres d'intérêt $\theta_1$ étudiés dans les sondages peuvent s'écrire sous la forme d'une fonction $T$ de $M_1$, à savoir $\theta_1 = T(M_1)$. Par exemple, le total $t_{y1} = \sum_{k \in U} y_{1k}$ est égal à $\int \mathcal{Y}_1 dM_1$. En pratique, un échantillon $s$ (avec ou sans répétition) est sélectionné selon un plan d'échantillonnage $p(\cdot)$, et nous observons les valeurs $y_{1k}$ pour $k \in s$ uniquement. Un principe de substitution est utilisé pour l'estimation

(voir Deville, 1999, et Goga, Deville et Ruiz-Gazen, 2009). Soit $\pi_k$ l'espérance du nombre de tirages pour l'unité $k$ dans l'échantillon; dans le cas d'un échantillonnage sans remise, il s'agit de la probabilité que l'unité $k$ soit sélectionnée dans l'échantillon. Soit $\hat{M}_1 = \sum_{k\in s} w_k \delta_{y_{1k}}$ la mesure discrète prenant la masse $w_k$ en tout point dans l'échantillon et 0 ailleurs, où $w_k = \pi_k^{-1}$ est le poids d'échantillonnage. La substitution de $\hat{M}_1$ dans $\theta_1$ donne l'estimateur $\hat{\theta}_1 = T(\hat{M}_1)$.

Pour un plan d'échantillonnage sans remise, l'estimateur d'un total par substitution correspond à l'estimateur de Horvitz-Thompson (HT) $\hat{t}_{y1}^{\mathrm{HT}} = \sum_{k\in s} w_k y_{1k}$. L'estimateur de variance Horvitz-Thompson est

$$v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{1k}}{\pi_k} \frac{y_{1l}}{\pi_l}, \tag{2.1}$$

où $\pi_{kl} = \Pr(k, l \in s)$ désigne la probabilité que les unités $k$ et $l$ soient sélectionnées conjointement dans l'échantillon, et $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Dans le cas particulier de l'échantillonnage aléatoire simple sans remise (SI) de taille $n$, nous avons $\hat{t}_{y1}^{\mathrm{HT}} = N\bar{y}_{1,s}$, avec $\bar{y}_{1,s} = n^{-1}\sum_{k\in s} y_{1k}$, et la formule (2.1) donne

$$v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{y_1,s}^2 \quad \text{où} \quad S_{y_1,s}^2 = \frac{1}{n-1}\sum_{k\in s}(y_{1k} - \bar{y}_{1,s})^2. \tag{2.2}$$

Pour un plan d'échantillonnage avec remise, l'estimateur par substitution d'un total correspond à l'estimateur de Hansen-Hurwitz (HH) $\hat{t}_{y1}^{\mathrm{HH}} = \sum_{k\in s} w_k y_{1k}$. Nous considérons le cas important de l'échantillonnage à plusieurs degrés, où les $N$ unités sont groupées à l'intérieur de $N_I$ unités primaires d'échantillonnage (UPE) disjointes $U_1, \ldots, U_{N_I}$, et où un échantillon $s_I$ de taille $m$ est sélectionné avec remise au premier degré. Soit $\pi_{Ii}$ l'espérance du nombre de tirages pour l'UPE $U_i$ dans $s_I$. Un échantillon de deuxième degré $s_i$ est ensuite sélectionné à l'intérieur de toute unité $i \in s_I$ selon un plan d'échantillonnage $p_i(\cdot)$. Soit $\pi_{k|i}$ l'espérance du nombre de tirages pour l'unité $k$ dans $s_i$. La mesure estimée est alors $\hat{M}_1 = \sum_{i\in s_I}\sum_{k\in s_i} \pi_{Ii}^{-1}\pi_{k|i}^{-1}\delta_{y_{1k}}$. Nous avons $\hat{t}_{y1}^{\mathrm{HH}} = \sum_{i\in s_I} \pi_{Ii}^{-1}\hat{Y}_i$ où $\hat{Y}_i = \sum_{k\in s_i} \pi_{k|i}^{-1} y_{1k}$, et un estimateur de variance sans biais pour $\hat{t}_{y1}^{\mathrm{HH}}$ est donné par

$$v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right) = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\hat{Y}_i}{\pi_{Ii}} - \frac{\hat{t}_{y1}^{\mathrm{HH}}}{m}\right)^2. \tag{2.3}$$

## 2.2 Estimation de l'indice de Gini

Si la variable $\mathcal{Y}_1$ est mesurée sur la population $U$, l'indice de Gini est donné par

$$G_1 = \frac{1}{2}\frac{\sum_{k\in U}\sum_{l\in U}|y_{1k} - y_{1l}|}{N\sum_{k\in U} y_{1k}};$$

voir, par exemple, Nygård et Sandström (1985). Il s'ensuit que $G_1$ est nul si $\mathcal{Y}_1$ est constante sur la population, ce qui a lieu quand le total de $\mathcal{Y}_1$ est réparti uniformément entre toutes les unités de la population. Dans le cas opposé, quand un seul individu possède le montant total de $\mathcal{Y}_1$, $G_1$ est maximisé et vaut $1 - 1/N$ : le total de $\mathcal{Y}_1$ est alors concentré en un seul point, ce qui signifie que l'inégalité entre les membres de la population est maximale.

Si les individus $k \neq l$ possèdent des valeurs différentes de la variable $y_1$, l'indice de Gini $G_1$ égale

$$G_1 = \frac{\sum_{k=1}^{N} y_{1(k)} (2k/N - 1)}{t_{y1}} - \frac{1}{N} = \frac{\sum_{k \in U} y_{1k} \{2F_{1N}(y_{1k}) - 1\}}{t_{y1}} - \frac{1}{N} \tag{2.4}$$

avec $y_{1(1)} \leq \cdots \leq y_{1(N)}$ les valeurs ordonnées et $F_{1N}(\cdot) = N^{-1} \sum_{k \in U} 1_{\{y_{1k} \leq \cdot\}}$ la fonction de répartition en population finie; voir Sandström, Wretman et Waldèn (1988) et Deville (1997) pour des renseignements plus détaillés sur l'obtention de (2.4). Nygård et Sandström (1985) appellent le terme $-1/N$ la correction en population finie de l'indice de Gini et donnent plusieurs raisons d'apporter cette correction, dont la non-négativité de la borne inférieure de $G_1$. Comme cela se fait fréquemment dans la littérature (voir, par exemple, Glasser, 1962), nous ignorons cette correction dans la suite de l'exposé. Nous redéfinissons l'indice de Gini sous la forme

$$G_1 = \frac{\sum_{k \in U} y_{1k} \{2F_{1N}(y_{1k}) - 1\}}{t_{y1}} = \frac{\int \{2F_{1N}(y) - 1\} y dM_1(y)}{\int y dM_1(y)} \tag{2.5}$$

où la fonction de répartition en population finie $F_{1N}(\cdot)$ est une famille de fonctionnelles

$$F_{1N}(y) = \frac{1}{\int dM_1(y)} \int 1_{\{\xi \leq y\}} dM_1(\xi) \tag{2.6}$$

indexée par $y$. L'introduction de $\hat{M}_1$ par substitution dans (2.5) et (2.6) donne l'estimateur

$$\hat{G}_1 = \frac{\int \{2\hat{F}_{1N}(y) - 1\} y d\hat{M}_1(y)}{\int y d\hat{M}_1(y)} = \frac{\sum_{k \in s} w_k \{2\hat{F}_{1N}(y_{1k}) - 1\} y_{1k}}{\sum_{k \in s} w_k y_{1k}}, \tag{2.7}$$

où

$$\hat{F}_{1N}(y) = \frac{1}{\int d\hat{M}_1(y)} \int 1_{\{\xi \leq y\}} d\hat{M}_1(\xi) = \frac{1}{\sum_{k \in s} w_k} \sum_{k \in s} w_k 1_{\{y_{1k} \leq y\}} \tag{2.8}$$

est l'estimateur par substitution de la fonction de répartition $F_{1N}$.

## 2.3 Estimation de la variance par linéarisation

Nous exposons brièvement ici la linéarisation par la fonction d'influence (LFI) (Deville, 1999), qui consiste à donner un développement d'ordre un de l'estimateur par substitution $\hat{\theta}_1 = T(\hat{M}_1)$ autour de la valeur réelle $\theta_1 = T(M_1)$, afin approximer l'erreur par un estimateur linéaire d'une *variable linéarisée* artificielle. Plus précisément, les dérivées premières de $T$ par rapport à $M_1$ sont les fonctions d'influence

$$\mathrm{IT}(M_1; y) = \lim_{h \to 0} \frac{T(M_1 + h\delta_y) - T(M_1)}{h},$$

et $u_{1k} = \mathrm{IT}(M_1; y_{1k})$ est la variable linéarisée pour tout $k \in U$. Supposons que $T(\cdot)$ est homogène, c'est-à-dire qu'il existe un nombre positif $\beta$ dépendant de $T$ tel que $T(rM_1) = r^{\beta} T(M_1)$ pour tout réel $r > 0$. Supposons aussi que $\lim_{N \to \infty} N^{-\beta} T(M_1) < \infty$. Sous certaines hypothèses de régularité supplémentaires concernant $T(\cdot)$ et le plan d'échantillonnage (par exemple, Goga et Ruiz-Gazen, 2014), Deville (1999) établit que

$$\hat{\theta}_1 - \theta_1 = \left( \sum_{k \in s} w_k u_{1k} - \sum_{k \in U} u_{1k} \right) + o_p \left( N^{\beta} n^{-1/2} \right),$$

de sorte que l'erreur $\hat{\theta}_1 - \theta_1$ peut être approximée par l'erreur de l'estimateur HT pour le total de la variable linéarisée $u_{1k}$. Pour un plan d'échantillonnage sans remise, l'utilisation d'un estimateur fondé sur l'échantillon $\hat{u}_{1k}$ de la variable linéarisée $u_{1k}$ dans l'estimateur HT de variance donne l'estimateur de variance

$$v_{\text{LIN}}^{\text{HT}} \left( \hat{\theta}_1 \right) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_{1k}}{\pi_k} \frac{\hat{u}_{1l}}{\pi_l}, \tag{2.9}$$

où $\pi_{kl} = \Pr(k, l \in s)$ désigne la probabilité que les unités $k$ et $l$ soient sélectionnées conjointement dans l'échantillon, et $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Plusieurs résultats de normalité asymptotique ont été prouvés pour des plans d'échantillonnage particuliers; voir Hájek (1960, 1961, 1964), Rosén (1972), Sen (1980), Krewski et Rao (1981), Gordon (1983), Ohlsson (1986, 1989), Chen et Rao (2007), Brändén et Jonasson (2012), Saegusa et Wellner (2013) et Chauvet (2015), entre autres. Si le plan d'échantillonnage est tel que l'estimateur par substitution $\hat{\theta}_1$ satisfait un théorème central limite, un intervalle de confiance de niveau approximatif $(1 - 2\alpha)$ % est $\left[ \hat{\theta}_1 - z_\alpha \sqrt{v_{\text{lin}}\left(\hat{\theta}_1\right)}, \hat{\theta}_1 + z_\alpha \sqrt{v_{\text{lin}}\left(\hat{\theta}_1\right)} \right]$, où $z_\alpha$ est le quantile d'ordre $\alpha$ pour la loi normale centrée réduite.

Dans le cas de l'indice de Gini, nous avons $\beta = 0$ et la variable linéarisée est

$$u_{1k} = 2F_{1N}(y_{1k}) \frac{y_{1k} - \overline{y}_{1k,U<}}{t_{y1}} - y_{1k} \frac{G_1 + 1}{t_{y1}} + \frac{1 - G_1}{N}, \tag{2.10}$$

où $\overline{y}_{1k,U<} = \left( \sum_{l \in U} 1_{\{y_{1l} < y_{1k}\}} \right)^{-1} \sum_{j \in U} y_{1j} 1_{\{y_{1j} < y_{1k}\}}$ désigne la moyenne des $y_{1j}$ inférieurs à $y_{1k}$, voir Deville (1999). Kovačević et Binder (1997) ont obtenu la même expression en appliquant la méthode de linéarisation par les équations estimantes; l'approche de linéarisation de Demnati et Rao (2004) mène aussi au même résultat. La variable linéarisée estimée est

$$\hat{u}_{1k} = 2\hat{F}_{1N}(y_{1k}) \frac{y_{1k} - \overline{y}_{1k,s<}}{\hat{t}_{y1}} - y_{1k} \frac{\hat{G}_1 + 1}{\hat{t}_{y1}} + \frac{1 - \hat{G}_1}{\hat{N}} \tag{2.11}$$

où $\overline{y}_{1k,s<} = \left( \sum_{l \in s} w_l 1_{\{y_{1l} < y_{1k}\}} \right)^{-1} \sum_{j \in s} w_j y_{1j} 1_{\{y_{1j} < y_{1k}\}}$.

Dans le cas particulier de l'échantillonnage SI, l'estimateur de variance par linéarisation pour l'indice de Gini est

$$v_{\text{LIN}}^{\text{HT}} \left( \hat{G}_1 \right) = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) S_{\hat{u}_1,s}^2 \quad \text{où} \quad S_{\hat{u}_1,s}^2 = \frac{1}{n-1} \sum_{k \in s} \left( \hat{u}_{1k} - \overline{\hat{u}}_{1,s} \right)^2, \tag{2.12}$$

et où $\overline{\hat{u}}_{1,s} = n^{-1} \sum_{k \in s} \hat{u}_{1k}$. Dans le cas particulier de l'échantillonnage à plusieurs degrés et de l'échantillonnage avec remise des UPE, l'estimateur de variance par linéarisation pour l'indice de Gini est

$$v_{\text{LIN}}^{\text{HH}} \left( \hat{G}_1 \right) = \frac{m}{m-1} \sum_{i \in s_I} \left( \frac{\hat{U}_{1i}}{\pi_{Ii}} - \frac{\hat{t}_{\hat{u}_1}^{\text{HH}}}{m} \right)^2 \quad \text{où} \quad \hat{U}_{1i} = \sum_{k \in s_i} \pi_{k|i}^{-1} \hat{u}_{1k}. \tag{2.13}$$

## 2.4 Estimation de la variance par bootstrap

L'utilisation de techniques bootstrap dans les sondages a fait l'objet d'un très grand nombre de publications. Les principales techniques bootstrap peuvent être vues comme des cas particuliers du bootstrap pondéré (Bertail et Combris, 1997; Antal et Tillé, 2011; Beaumont et Patak, 2012); consulter aussi Shao et Tu (1995, chapitre 6), Davison et Hinkley (1997, section 3.7), ainsi que Davison et Sardy (2007) pour des études détaillées. Sous une procédure de bootstrap pondéré, la mesure $\hat{M}_1 = \sum_s w_k \delta_{y_k}$ est estimée, conditionnellement à l'échantillon $s$, par la mesure bootstrap

$$\hat{M}_1^* = \sum_{k \in s} w_k D_k \delta_{y_k} \tag{2.14}$$

où $D = \{D_k\}_{k \in s}$ désigne un vecteur (aléatoire) de poids de rééchantillonnage. Nous notons $E_*$ et $V_*$ l'espérance et la variance par rapport au plan de rééchantillonnage. Dans le cas de l'échantillonnage sans remise, le vecteur $D$ est généré de telle façon que

$$E_*\left(\sum_s w_k D_k y_k\right) \simeq \hat{t}_{y1}^{\mathrm{HT}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) \simeq v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) \tag{2.15}$$

correspondent approximativement aux deux premiers moments de l'estimateur HT. Dans le cas de l'échantillonnage avec remise, le vecteur $D$ est généré de telle manière que

$$E_*\left(\sum_s w_k D_k y_k\right) \simeq \hat{t}_{y1}^{\mathrm{HH}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) \simeq v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right) \tag{2.16}$$

correspondent approximativement aux deux premiers moments de l'estimateur HH.

Sous toute technique de bootstrap pondéré, l'estimateur par substitution (*plug-in*) de $\theta_1 = T(M_1)$ est $\hat{\theta}_1^* = T\left(\hat{M}_1^*\right)$, et la variance de $\hat{\theta}_1 = T\left(\hat{M}_1\right)$ est estimée par

$$V_*\left(\hat{\theta}_1^*\right) = E_*\left\{\hat{\theta}_1^* - E_*\left(\hat{\theta}_1^*\right)\right\}^2. \tag{2.17}$$

Puisque l'estimateur de variance (2.17) peut être difficile à calculer exactement, on peut le remplacer par un estimateur de variance basé sur des simulations. Plus précisément, nous générons $C$ réalisations indépendantes $D_1, \ldots, D_C$ du vecteur $D$ et nous désignons par $\hat{\theta}_{1c}^* = T\left(\hat{M}_{1c}^*\right)$ avec $\hat{M}_{1c}^*$ la mesure bootstrap associée au vecteur $D_c$. Alors, $V\left(\hat{\theta}_1\right)$ est estimée par

$$v_B\left(\hat{\theta}_1\right) = \frac{1}{C-1}\sum_{c=1}^{C}\left\{\hat{\theta}_{1c}^* - \frac{1}{C}\sum_{c'=1}^{C}\hat{\theta}_{1c'}^*\right\}^2. \tag{2.18}$$

Deux types d'intervalles de confiance sont habituellement calculés. La méthode des percentiles s'appuie sur les estimations bootstrap ordonnées $\hat{\theta}_{(1c)}^*$, $c = 1, \ldots, C$ pour former un intervalle de confiance à $(1 - 2\alpha)$ % noté $[\hat{\theta}_{(1L)}^*, \hat{\theta}_{(1U)}^*]$ avec $L = \alpha C$ et $U = (1-\alpha)C$. Le bootstrap $-t$ repose sur l'estimation de la statistique pivot $t = (\hat{\theta}_1 - \theta_1)\Big/\sqrt{v_{\mathrm{BWO}}(\hat{\theta}_1)}$ par son analogue bootstrap $t^* = (\hat{\theta}_1^* - \hat{\theta}_1)\Big/\sqrt{v_{\mathrm{BWO}}^*(\hat{\theta}_1^*)}$, où $v_{\mathrm{BWO}}^*(\hat{\theta}_1^*)$ s'obtient en appliquant la procédure bootstrap à la réplique de l'échantillon $s^*$. Le bootstrap $-t$ demande énormément de calcul puisqu'un double bootstrap est nécessaire, et est donc moins intéressant pour

l'utilisateur des données. Par conséquent, nous ne poursuivons pas cette approche et nous nous concentrons sur la méthode des percentiles.

Les méthodes de linéarisation fournissent des formules de variance applicables à des plans d'échantillonnage généraux, mais comprennent des calculs de dérivées qui peuvent être ardus pour les paramètres d'intérêt complexes, tels que l'indice de Gini. Contrairement à la linéarisation, le bootstrap évite le travail théorique grâce au calcul répété du système d'estimations existant. Les poids de rééchantillonnage sont fournis avec l'ensemble de données et peuvent être utilisés facilement pour produire des estimations de variance pour une grande gamme de statistiques. Cependant, les techniques bootstrap ne conviennent habituellement pas pour les plans d'échantillonnage généraux. Autrement dit, un plan d'échantillonnage particulier requiert habituellement un schéma de rééchantillonnage taillé sur mesure. Dans le présent article, nous nous concentrons sur deux techniques bootstrap particulières, qui seront généralisées à la section 3 au contexte de deux échantillons.

### 2.4.1 Bootstrap sans remise pour l'échantillonnage SI

Quand l'échantillon $s$ est sélectionné par échantillonnage SI, nous considérons le bootstrap sans remise (BWO) introduit par Gross (1980). L'approche s'étend facilement à l'échantillonnage aléatoire simple stratifié (STSI) avec un nombre fini de strates. Supposons que $N/n$ est un entier. Alors, le vecteur $D$ s'obtient en créant d'abord une pseudopopulation $U^*$ de taille $N$ en reproduisant $N/n$ fois chaque unité $k$ dans l'échantillon original $s$, puis en sélectionnant une réplique d'échantillon SI $s^*$ de taille $n$ dans $U^*$.

La mesure bootstrap est donnée par (2.14), où le poids de rééchantillonnage $D_k$ est le nombre de fois que l'unité $k \in s$ est sélectionnée dans $s^*$. La construction de $U^*$ peut être évitée en notant que, sous la procédure BWO, le vecteur $D$ suit une loi hypergéométrique multivariée. Par conséquent, les poids de rééchantillonnage peuvent être produits directement. On peut montrer que la procédure BWO mène à

$$E_*\left(\sum_s w_k D_k y_k\right) = \hat{t}_{y1}^{\mathrm{HT}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) = \frac{1-n^{-1}}{1-N^{-1}} v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right), \tag{2.19}$$

où $v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right)$ est donné dans (2.2), de sorte que l'on obtient approximativement une concordance avec l'équation (2.15) pour une grande taille d'échantillon.

Plusieurs solutions ont été proposées pour traiter le cas où $N/n$ n'est pas un entier; voir Chao et Lo (1985), Bickel et Freedman (1984), Sitter (1992b), Booth, Butler et Hall (1994), Presnell et Booth (1994), entre autres. La généralisation de l'estimation de la variance par la procédure BWO pour des plans d'échantillonnage avec probabilités inégales est examinée dans Särndal, Swensson et Wretman (1992) et Chauvet (2007).

### 2.4.2 Bootstrap avec remise pour l'échantillonnage à plusieurs degrés

Quand l'échantillon $s$ est tiré par échantillonnage à plusieurs degrés et par échantillonnage à probabilités inégales avec remise des UPE, nous considérons le bootstrap des UPE (BWR) introduit par Rao et Wu (1988). Une réplique d'échantillon avec remise $s_I^*$ de taille $m-1$ est sélectionnée par échantillonnage aléatoire simple avec remise (SIR) dans l'échantillon de premier degré original $s_I$. La mesure bootstrap est

$$\hat{M}_1^* = \frac{m}{m-1}\sum_{i\in s_1^*}\sum_{k\in s_i}\pi_{Ii}^{-1}\pi_{k|i}^{-1}\delta_{y1k} = \sum_{k\in s} w_k D_k \delta_{yk}, \tag{2.20}$$

où le poids de rééchantillonnage $D_k$ égale $m(m-1)^{-1}$ multiplié par le nombre de fois que l'UPE contenant $k$ est sélectionnée dans $s_1^*$.

La taille de rééchantillonnage $m-1$ est utilisée pour reproduire l'estimateur de variance sans biais habituel dans le cas linéaire (voir Rao et Wu, 1988). On peut montrer que la procédure BWR mène à

$$E_*\left(\sum_s w_k D_k y_k\right) = \hat{t}_{y1}^{\mathrm{HH}} \quad \text{et} \quad V_*\left(\sum_s w_k D_k y_k\right) = v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right), \tag{2.21}$$

où $v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right)$ est donné dans (2.3), de sorte que l'équation (2.16) est exactement reproduite. La procédure BWR est particulièrement simple, puisqu'elle comporte un rééchantillonnage pour le premier degré d'échantillonnage uniquement, et que les sous-échantillons d'unités d'échantillonnage secondaires (USE) ne sont pas modifiés à l'intérieur des UPE rééchantillonnées.

# 3 Le cas de deux échantillons

## 3.1 Notation et estimation composite

Supposons maintenant que deux variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ sont mesurées sur la population $U$, et soit $y_{d1},\ldots,y_{dN}$ les valeurs prises par $\mathcal{Y}_d, d=1,2,$ sur les unités de la population. Les variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ peuvent par exemple correspondre à une caractéristique d'intérêt observée à deux périodes différentes $\tau_1$ et $\tau_2$. Nous considérons l'estimation de paramètres $\Delta\theta$ qui peuvent s'écrire sous la forme d'une fonctionnelle $\Delta\theta = T(M_1, M_2)$, où $M_d = \sum_{k\in U}\delta_{\{y_{dk}\}}$. Par exemple, le cas linéaire $\Delta t = t_{y2} - t_{y1}$ correspond à la différence entre les totaux $t_{y2} = \sum_{k\in U} y_{2k}$ et $t_{y1} = \sum_{k\in U} y_{1k}$.

Soit $s_1$ et $s_2$ deux échantillons de tailles $n_1$ et $n_2$, respectivement, tirés de la même population $U$ selon un plan d'échantillonnage bidimensionnel $p(\cdot,\cdot)$ (voir Goga, 2003). La variable $\mathcal{Y}_1$ est mesurée sur $s_1$, tandis que la variable $\mathcal{Y}_2$ est mesurée sur $s_2$. L'insertion des estimateurs fondés sur l'échantillon $\hat{M}_d$ dans $\Delta\theta$ donne l'estimateur par substitution $\widehat{\Delta\theta} = T\left(\hat{M}_1, \hat{M}_2\right)$. Contrairement au cas d'un seul échantillon, plusieurs estimateurs $\hat{M}_d$ sont possibles. Dans la suite de l'exposé, nous nous concentrons sur la classe générale d'*estimateurs composites* introduite par Goga, Deville et Ruiz-Gazen (2009). Nous notons $s_{1\bullet} = s_1 \setminus s_2$, $s_3 = s_1 \cap s_2$ et $s_{2\bullet} = s_2 \setminus s_1$. Pour $\lozenge \in \{1\bullet, 3, 2\bullet\}$, nous notons $\pi_{\lozenge,k}$ le nombre prévu de tirages de l'unité $k$ dans $s_\lozenge$ et $\hat{M}_{d,\lozenge} = \sum_{k\in s_\lozenge} w_{\lozenge,k}\delta_{y_{dk}}$, où $w_{\lozenge,k} = \pi_{\lozenge,k}^{-1}$. Les estimateurs composites de $M_1$ et $M_2$ sont

$$\hat{M}_1^{\mathrm{co}}(a) = a\,\hat{M}_{1,1\bullet} + (1-a)\,\hat{M}_{1,3} \quad \text{et} \quad \hat{M}_2^{\mathrm{co}}(b) = b\,\hat{M}_{2,2\bullet} + (1-b)\,\hat{M}_{2,3}, \tag{3.1}$$

où $a$ et $b$ sont des constantes connues. Le choix $a = b = 0$ mène à l'*estimateur « intersection »* avec $\hat{M}_1^{\mathrm{int}} = \hat{M}_{1,3}$ et $\hat{M}_2^{\mathrm{int}} = \hat{M}_{2,3}$, où seul est utilisé l'échantillon « intersection » $s_3$ (correspondant à l'intersection).

Si l'on estime le paramètre $\Delta t = t_{y2} - t_{y1}$, l'estimateur composite est donné par

$$\widehat{\Delta t}^{\,\mathrm{co}}(a,b) \;=\; \hat{t}_{y_2}^{\,\mathrm{co}} - \hat{t}_{y_1}^{\,\mathrm{co}}, \tag{3.2}$$

où $\hat{t}_{y_1}^{\,\mathrm{co}} = \int y d\hat{M}_1^{\,\mathrm{co}}(y)$ et $\hat{t}_{y_2}^{\,\mathrm{co}} = \int y d\hat{M}_2^{\,\mathrm{co}}(y)$. Il peut se réécrire sous la forme

$$\widehat{\Delta t}^{\,\mathrm{co}}(a,b) \;=\; b\left(\hat{t}_{y_2,s_{2\bullet}} - \hat{t}_{y_2,s_3}\right) - a\left(\hat{t}_{y_1,s_{1\bullet}} - \hat{t}_{y_1,s_3}\right) + \left(\hat{t}_{y_2,s_3} - \hat{t}_{y_1,s_3}\right), \tag{3.3}$$

où $\hat{t}_{y_d,s_\diamond} = \sum_{k\in s_\diamond} w_{\diamond,k}\, y_{dk}$. La variance de l'estimateur composite est

$$V\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\} \;=\; (b,-a,1)\, V\left\{\left(\hat{t}_{y_2,s_{2\bullet}} - \hat{t}_{y_2,s_3}, \hat{t}_{y_1,s_{1\bullet}} - \hat{t}_{y_1,s_3}, \hat{t}_{y_2,s_3} - \hat{t}_{y_1,s_3}\right)^{\top}\right\}(b,-a,1)^{\top}. \tag{3.4}$$

Trouver le vecteur $(a_{\mathrm{opt}}, b_{\mathrm{opt}})^{\top}$ qui minimise la variance en (3.4) mène à l'*estimateur composite optimal* (Goga, Deville et Ruiz-Gazen, 2009, section 3.6). Notons qu'il ne s'agit pas d'un estimateur proprement dit, puisqu'il dépend de quantités inconnues qui doivent être estimées en pratique. Cependant, il représente une référence utile que nous utiliserons pour évaluer des estimateurs composites plus simples.

Un estimateur de variance s'obtient en substituant dans (3.4) un estimateur de la matrice de variance-covariance. L'obtention des estimateurs de variance est décrite en détail aux sections 3.1.1 et 3.1.2 pour deux exemples de plans d'échantillonnage bidimensionnels.

### 3.1.1 Plan SI bidimensionnel

Le plan SI bidimensionnel (SI2) de taille fixée $(n_{1\bullet}, n_3, n_{2\bullet})$ attribue des probabilités égales à tous les $s = (s_1, s_2)$ pour lesquels les sous-échantillons associés $s_{1\bullet}$, $s_3$ et $s_{2\bullet}$ possèdent les tailles requises $n_{1\bullet}$, $n_3$ et $n_{2\bullet}$, voir Goga (2003) ainsi que Qualité et Tillé (2008). Le plan SI2 a pour propriété intéressante que les échantillons marginaux $s_{1\bullet}$, $s_3$ et $s_{2\bullet}$ sont des échantillons SI provenant de la population $U$. De même, $s_1$ est un échantillon SI de taille $n_1 = n_{1\bullet} + n_3$, et $s_2$ est un échantillon SI de taille $n_2 = n_{2\bullet} + n_3$. Pour le plan d'échantillonnage SI2, l'estimateur composite en (3.3) donne

$$\widehat{\Delta t}^{\,\mathrm{co}}(a,b) \;=\; Nb\left(\bar{y}_{2,s_{2\bullet}} - \bar{y}_{2,s_3}\right) - Na\left(\bar{y}_{1,s_{1\bullet}} - \bar{y}_{1,s_3}\right) + N\left(\bar{y}_{2,s_3} - \bar{y}_{1,s_3}\right), \tag{3.5}$$

et la variance de l'estimateur composite s'exprime par

$$V\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\} \;=\; N^2\left\{c_1(a)\, S_{y_1,U}^2 - 2c_{12}(a,b)\, S_{y_1y_2,U} + c_2(b)\, S_{y_2,U}^2\right\}, \tag{3.6}$$

avec

$$\begin{aligned} c_1(a) &= \frac{(1-a)^2}{n_3} + \frac{a^2}{n_1 - n_3} - \frac{1}{N}, \\ c_2(b) &= \frac{(1-b)^2}{n_3} + \frac{b^2}{n_2 - n_3} - \frac{1}{N}, \\ c_{12}(a,b) &= \frac{(1-a)(1-b)}{n_3} - \frac{1}{N}, \end{aligned}$$

voir l'annexe pour une preuve.

Nous considérons deux exemples. Le choix $a = b = 0$ mène à l'estimateur « intersection »

$$\widehat{\Delta t}^{\,\text{int}} = \widehat{\Delta t}^{\,\text{co}}(0,0) = \frac{N}{n_3}\sum_{k\in s_3}(y_{2k} - y_{1k}), \tag{3.7}$$

et l'expression de la variance se simplifie en

$$V\left\{\widehat{\Delta t}^{\,\text{int}}\right\} = N^2\left(\frac{1}{n_3} - \frac{1}{N}\right)S^2_{y_2-y_1,U}. \tag{3.8}$$

Le choix $a = n_1^{-1}n_{1\bullet}$ et $b = n_2^{-1}n_{2\bullet}$ mène à l'*estimateur « union »*

$$\widehat{\Delta t}^{\,\text{uni}} = \widehat{\Delta t}^{\,\text{co}}(n_1^{-1}n_{1\bullet}, n_2^{-1}n_{2\bullet}) = \frac{N}{n_2}\sum_{k\in s_2} y_{2k} - \frac{N}{n_1}\sum_{k\in s_1} y_{1k} \tag{3.9}$$

où les échantillons complets sont utilisés, et la variance peut s'écrire sous la forme

$$V\left\{\widehat{\Delta t}^{\,\text{uni}}\right\} = N^2\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S^2_{y_1,U} - 2\left(\frac{n_3}{n_1\,n_2} - \frac{1}{N}\right)S_{y_1y_2,U} + \left(\frac{1}{n_2} - \frac{1}{N}\right)S^2_{y_2,U}\right\}. \tag{3.10}$$

Les variances de l'estimateur « union » et de l'estimateur « intersection » ont été établies par Qualité et Tillé (2008), voir aussi Tam (1984).

Le choix de $a$ et $b$ revêt une importance pratique si l'on veut obtenir un estimateur composite efficace. Après un peu de calcul, le vecteur $(a_{\text{opt}}, b_{\text{opt}})^\top$ qui minimise la variance de $\widehat{\Delta t}^{\,\text{co}}(a,b)$ est donné par

$$(a_{\text{opt}}, b_{\text{opt}})^\top = A^{-1}X \tag{3.11}$$

avec

$$A = \begin{pmatrix} \dfrac{n_1}{n_1 - n_3} & -\dfrac{S_{y_1y_2,U}}{S^2_{y_1,U}} \\ -\dfrac{S_{y_1y_2,U}}{S^2_{y_2,U}} & \dfrac{n_2}{n_2 - n_3} \end{pmatrix} \quad \text{et} \quad X = \left(1 - \frac{S_{y_1y_2,U}}{S^2_{y_1,U}}, 1 - \frac{S_{y_1y_2,U}}{S^2_{y_2,U}}\right)^\top. \tag{3.12}$$

Pour deux variables $\mathcal{Y}_1$ et $\mathcal{Y}_2$ se rapportant à une même caractéristique observée à deux périodes différentes, $S_{y_1y_2,U}$ doit, en principe, être proche de $S^2_{y_1,U}$ et $S^2_{y_2,U}$. Le vecteur $X$ dans (3.12) est, à son tour, proche du vecteur nul, et si la taille de l'échantillon « intersection » $s_3$ est comparable à celles de $s_{1\bullet}$ et $s_{2\bullet}$, nous obtenons $a_{\text{opt}} \simeq 0$ et $b_{\text{opt}} \simeq 0$. Par conséquent, l'utilisation de l'estimateur « intersection » où $a = b = 0$ paraît raisonnable en pratique. Au contraire, l'estimateur « union » peut être très inefficace; voir la section 4.2 pour un exemple. Ces conclusions concordent avec celles de Qualité et Tillé (2008), section 2.2.2.

Plusieurs estimateurs de variance peuvent être utilisés pour l'estimateur composite. L'estimation des dispersions sur l'échantillon « intersection » uniquement donne l'estimateur de variance sans biais

$$v_{\text{int}}^{\text{HT}}\left\{\widehat{\Delta t}^{\,\text{co}}(a,b)\right\} = N^2\left\{c_1(a)\,S^2_{y_1,s_3} - 2c_{12}(a,b)\,S_{y_1y_2,s_3} + c_2(b)\,S^2_{y_2,s_3}\right\}, \tag{3.13}$$

tandis qu'une estimation sur les échantillons entiers donne

$$v_{\text{uni}}^{\text{HT}}\left\{\widehat{\Delta t}^{\,\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S^2_{y_1,s_1} - 2c_{12}(a,b)S_{y_1y_2,s_3} + c_2(b)S^2_{y_2,s_2}\right\}. \tag{3.14}$$

Berger (2004) a considéré l'estimation de la variance pour l'estimateur « union » sous un plan d'échantillonnage rotatif à entropie maximale en estimant séparément les trois composantes dans (3.6).

### 3.1.2 Plan à plusieurs degrés bidimensionnel

Considérons maintenant un plan d'échantillonnage à deux degrés bidimensionnel (MULT2). Nous supposons qu'un échantillon de premier degré $s_I$ de taille $m$ est d'abord sélectionné avec remise parmi les UPE $U_1,\ldots,U_{N_I}$. À l'intérieur de chaque UPE $i \in s_I$, on sélectionne ensuite un échantillon SI2 de taille $(n^i_{1\bullet}, n^i_{3}, n^i_{2\bullet})$. Ce type de plan d'échantillonnage se dégage en particulier dans le cas d'un plan à deux degrés autopondéré en deux vagues, avec à la deuxième vague un remplacement partiel des USE sélectionnées à la première vague. L'estimateur composite en (3.3) donne

$$\widehat{\Delta t}^{\,\text{co}}(a,b) = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\text{co}}(a,b) \tag{3.15}$$

où

$$\widehat{\Delta t}^{\,i,\text{co}}(a,b) = N_i b\left(\overline{y}_{2,s^i_{2\bullet}} - \overline{y}_{2,s^i_3}\right) - N_i a\left(\overline{y}_{1,s^i_{1\bullet}} - \overline{y}_{1,s^i_3}\right) + N_i\left(\overline{y}_{2,s^i_3} - \overline{y}_{1,s^i_3}\right), \tag{3.16}$$

où $\overline{y}_{d,s^i_\lozenge} = (n^i_\lozenge)^{-1}\sum_{k\in s^i_\lozenge} y_{\lozenge k}$, où $s^i_\lozenge = s_\lozenge \cap U_i$, et où $N_i$ désigne le nombre d'USE dans l'UPE $u_i$.

Par exemple, en utilisant uniquement les échantillons communs à l'intérieur des UPE, on obtient l'estimateur « intersection »

$$\widehat{\Delta t}^{\,\text{int}} = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\text{int}} \qquad \text{avec} \qquad \widehat{\Delta t}^{\,i,\text{int}} = N_i\left(\overline{y}_{2,s^i_3} - \overline{y}_{1,s^i_3}\right). \tag{3.17}$$

En utilisant les échantillons complets à l'intérieur des UPE, on obtient l'estimateur « union »

$$\widehat{\Delta t}^{\,\text{uni}} = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\text{uni}} \qquad \text{avec} \qquad \widehat{\Delta t}^{\,i,\text{uni}} = N_i\left(\overline{y}_{2,s^i_2} - \overline{y}_{1,s^i_1}\right). \tag{3.18}$$

Nous notons que, pour tout vecteur de valeurs $(a,b)^\top$, la variance due au premier degré d'échantillonnage pour $\widehat{\Delta t}^{\,\text{co}}(a,b)$ est la même. Les estimateurs composites possibles diffèrent donc en ce qui concerne la variance de second degré uniquement. Compte tenu de la discussion de la section 3.1.1, nous nous attendons par conséquent à ce que l'estimateur « intersection » soit proche de l'estimateur composite optimal; voir la section 4.2 pour un exemple. Un estimateur de variance sans biais pour $\widehat{\Delta t}^{\,\text{co}}(a,b)$ est donné par

$$v^{\text{HH}}\left\{\widehat{\Delta t}^{\,\text{co}}(a,b)\right\} = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\widehat{\Delta t}^{\,i,\text{co}}(a,b)}{\pi_{Ii}} - \frac{\widehat{\Delta t}^{\,\text{co}}(a,b)}{m}\right)^2. \tag{3.19}$$

## 3.2 Estimation de l'évolution de l'indice de Gini

L'évolution de l'indice de Gini $\Delta G = G_2 - G_1$ peut s'écrire sous la forme

$$\Delta G = \frac{\int \{2F_{2N}(y) - 1\}\, y dM_2(y)}{\int y dM_2(y)} - \frac{\int \{2F_{1N}(y) - 1\}\, y dM_1(y)}{\int y dM_1(y)} \tag{3.20}$$

où $F_{dN}(y) = N^{-1}\sum_{k\in U} 1_{\{y_{dk}\le y\}}$, $d = 1, 2$. L'utilisation de l'estimation composite mène à

$$\widehat{\Delta G}^{\text{co}}(a,b) = \frac{\int \left\{2\hat{F}_{2N}^{\text{co}}(y) - 1\right\} y d\hat{M}_2^{\text{co}}(y)}{\int y d\hat{M}_2^{\text{co}}(y)} - \frac{\int \left\{2\hat{F}_{1N}^{\text{co}}(y) - 1\right\} y d\hat{M}_1^{\text{co}}(y)}{\int y d\hat{M}_1^{\text{co}}(y)} \tag{3.21}$$

où $\hat{F}_{dN}^{\text{co}}(y) = \left\{\int d\hat{M}_d^{\text{co}}(y)\right\}^{-1} \int 1_{\{\xi\le y\}}\, d\hat{M}_d^{\text{co}}(\xi)$.

Habituellement, dans un cadre d'échantillonnage temporel, les échantillons $s_1$ et $s_2$ ne sont pas indépendants. Par conséquent, nos conditions diffèrent de l'estimation usuelle des fonctionnelles dépendantes des fonctions de répartition estimées sur des échantillons indépendants; voir, par exemple, Pires et Branco (2002) et Reid (1981), qui donnent le développement d'ordre un d'une fonctionnelle pour deux échantillons utilisant les fonctions d'influence partielles. Davison et Hinkley (1997, page 71) donnent des méthodes bootstrap sous un cadre similaire. Sous un plan d'échantillonnage bidimensionnel général $p(\cdot,\cdot)$, Goga, Deville et Ruiz-Gazen (2009) donnent une technique de linéarisation pour deux échantillons de fonctionnelles bivariées que nous utiliserons dans la suite de l'exposé.

## 3.3 Estimation de la variance par linéarisation

Pour obtenir la variance asymptotique de $\widehat{\Delta\theta}^{\text{co}}(a,b)$, nous adoptons le cadre asymptotique introduit par Goga, Deville et Ruiz-Gazen (2009), qui est une extension du cas à deux échantillons du cadre asymptotique d'Isaki et Fuller (1982). Définissons, quand elles existent, les *fonctions d'influence partielles* d'une fonctionnelle $T(M_1, M_2)$ au point $y$ par

$$I_1T(M_1, M_2; y) = \lim_{h\to 0} \frac{T(M_1 + h\delta_y, M_2) - T(M_1, M_2)}{h},$$
$$I_2T(M_1, M_2; y) = \lim_{h\to 0} \frac{T(M_1, M_2 + h\delta_y) - T(M_1, M_2)}{h}.$$

Nous définissons les *variables linéarisées* $u_{dk} = I_dT(M_1, M_2; y_{dk})$ pour $d = 1, 2$ comme étant les fonctions d'influence partielles de $T$ pour $(M_1, M_2)$ et $y = y_{dk}$. Pour l'évolution de l'indice de Gini $\Delta G$, nous pouvons calculer les variables linéarisées $u_{dk}$ en utilisant (2.10), à savoir

$$u_{dk} = 2F_{dN}(y_{dk})\frac{y_{dk} - \bar{y}_{dk,U<}}{t_{y_d}} - y_{dk}\frac{G_d + 1}{t_{y_d}} + \frac{1 - G_d}{N}, \tag{3.22}$$

où $\bar{y}_{dk,U<} = \left(\sum_{l\in U} 1_{\{y_{dl}<y_{dk}\}}\right)^{-1} \sum_{j\in U} y_{dj} 1_{\{y_{dj}<y_{dk}\}}$. La variable linéarisée estimée est

$$\hat{u}_{dk} = 2\hat{F}_{dN}^{\text{co}}(y_{dk})\frac{y_{dk} - \overline{y}_{dk,s<}^{\text{co}}}{\hat{t}_{y1}^{\text{co}}} - y_{dk}\frac{\hat{G}_{d}^{\text{co}} + 1}{\hat{t}_{y1}^{\text{co}}} + \frac{1 - \hat{G}_{d}^{\text{co}}}{\hat{N}}. \tag{3.23}$$

### 3.3.1 Plan SI bidimensionnel

Dans le cas du plan SI2 présenté à la section 3.1.1, l'insertion des variables $u_{dk}$ calculées en (3.22) dans la formule de variance (3.6) donne l'approximation de la variance

$$V\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} \simeq N^2\left\{c_1(a)S_{u_1,U}^2 - 2c_{12}(a,b)S_{u_1u_2,U} + c_2(b)S_{u_2,U}^2\right\},$$

voir le théorème 1 dans Goga, Deville et Ruiz-Gazen (2009). Pour obtenir un estimateur de variance, les variables linéarisées peuvent être estimées de plusieurs façons. Si l'on utilise seulement l'échantillon « intersection » $s_3$, les variables linéarisées estimées $\hat{u}_d$ s'obtiennent au moyen de (3.23) en prenant $\hat{M}_1^{\text{co}} = \hat{M}_{1,3}$ et $\hat{M}_2^{\text{co}} = \hat{M}_{2,3}$. Un estimateur de variance s'obtient alors en insérant ces variables linéarisées dans (3.13). Cela donne

$$v_{\text{int}}^{\text{HT}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S_{\hat{u}_1,s_3}^2 - 2c_{12}(a,b)S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b)S_{\hat{u}_2,s_3}^2\right\}. \tag{3.24}$$

Si les deux échantillons $s_1$ et $s_2$ sont utilisés, les variables linéarisées estimées $\hat{u}_d$ s'obtiennent au moyen de (3.23) en prenant $\hat{M}_1^{\text{co}} = \hat{M}_{1,1}$ et $\hat{M}_2^{\text{co}} = \hat{M}_{2,2}$. Un estimateur de variance s'obtient alors en insérant ces variables linéarisées dans (3.14). Cela donne

$$v_{\text{uni}}^{\text{HT}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S_{\hat{u}_1,s_1}^2 - 2c_{12}(a,b)S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b)S_{\hat{u}_2,s_2}^2\right\}. \tag{3.25}$$

### 3.3.2 Plan à plusieurs degrés bidimensionnel

Dans le cas du plan MULT2 présenté à la section 3.1.2, les variables linéarisées peuvent également être estimées de plusieurs façons. Pour simplifier, nous considérons l'utilisation de l'échantillon « intersection » $s_3$ seulement, de sorte que les variables linéarisées estimées $\hat{u}_d$ s'obtiennent au moyen de (3.23) en prenant $\hat{M}_1^{\text{co}} = \hat{M}_{1,3}$ et $\hat{M}_2^{\text{co}} = \hat{M}_{2,3}$. Un estimateur de variance s'obtient alors en insérant ces variables linéarisées dans (3.19). Cela donne

$$v^{\text{HH}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\widehat{\Delta u}^{i,\text{co}}(a,b)}{\pi_{Ii}} - \frac{\widehat{\Delta u}^{\text{co}}(a,b)}{m}\right)^2, \tag{3.26}$$

où $\widehat{\Delta u}^{\text{co}}(a,b)$ et $\widehat{\Delta u}^{i,\text{co}}(a,b)$ s'obtiennent à partir de (3.15) et (3.16), respectivement, en remplaçant $y_{dk}$ par $\hat{u}_{dk}$.

## 3.4 Estimation de la variance par bootstrap

Les méthodes bootstrap n'ont pas encore été étudiées dans le cas de l'évolution de l'indice de Gini. Les principes des techniques de bootstrap pondéré peuvent être étendus au contexte de deux échantillons,

c'est-à-dire que chaque mesure $\hat{M}_{d,\Diamond}$ avec $d = 1, 2$ et $\Diamond \in \{1\bullet, 3, 2\bullet\}$ est estimée, conditionnellement aux échantillons sélectionnés au départ, par une mesure bootstrap pondérée $\hat{M}^*_{d,\Diamond}$ qui permet de reproduire, au moins approximativement, les deux premiers moments d'un estimateur sans biais dans le cas linéaire. À la section 3.4.1, nous examinons une généralisation du bootstrap sans remise (BWO) au plan SI2. À la section 3.4.2, nous proposons une généralisation du bootstrap avec remise (BWR) au plan MULT2.

### 3.4.1 Une généralisation du bootstrap sans remise au plan SI2

Nous considérons d'abord le plan SI2. La construction d'une pseudopopulation $U^*$ est plus complexe dans le cas de deux échantillons, puisque les variables d'intérêt mesurées aux vagues $\tau_1$ et $\tau_2$ doivent être disponibles pour chaque unité dans $U^*$. Nous décrivons donc un algorithme bootstrap où seul l'échantillon « intersection » $s_3$ est utilisé pour construire la pseudopopulation $U^*$, dans l'esprit de l'estimateur de variance « intersection » en (3.24).

Supposons que $N/n_3$ est un entier. Les vecteurs $D_\Diamond$ s'obtiennent en créant d'abord une pseudopopulation $U^*$ de taille $N$ en dupliquant $N/n_3$ fois chaque unité $k$ de l'échantillon original $s_3$. Une réplique d'échantillon SI2 $s^* = (s^*_{1\bullet}, s^*_3, s^*_{\bullet 2})$ de taille $(n_{1\bullet}, n_3, n_{2\bullet})$ est ensuite sélectionnée dans $U^*$. Les mesures bootstrap sont alors

$$\hat{M}^*_{d,\Diamond} = \sum_{k \in s_3} w_{\Diamond,k} D_{\Diamond,k} \delta_{y_{dk}}, \tag{3.27}$$

avec $D_{\Diamond,k}$ le nombre de fois que l'unité $k$ est sélectionnée dans la réplique d'échantillon $s^*_\Diamond$. Dans le cas linéaire, l'estimateur bootstrap du paramètre $\Delta t$ est alors donné par

$$\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b) = b\left(\hat{t}_{y_2,s^*_{2\bullet}} - \hat{t}_{y_2,s^*_3}\right) - a\left(\hat{t}_{y_1,s^*_{1\bullet}} - \hat{t}_{y_1,s^*_3}\right) + \left(\hat{t}_{y_2,s^*_3} - \hat{t}_{y_1,s^*_3}\right), \tag{3.28}$$

où $\hat{t}_{y_d,s^*_\Diamond} = \sum_{k \in s_3} w_{\Diamond,k} D_{\Diamond,k} y_{dk}$. Après un peu de calcul, nous obtenons

$$E_*\left\{\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b)\right\} = \widehat{\Delta t}^{\,\mathrm{int}} \quad \text{et} \quad V_*\left\{\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b)\right\} = \frac{1 - n_3^{-1}}{1 - N^{-1}} v_{\mathrm{int}}^{\mathrm{HT}}\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\}, \tag{3.29}$$

où $\widehat{\Delta t}^{\,\mathrm{int}}$ est donné en (3.7), et $v_{\mathrm{int}}^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right)$ est donné en (3.13). La généralisation du bootstrap sans remise (BWO) permet donc de reproduire exactement l'estimateur « intersection » du premier moment et de reproduire approximativement l'estimateur « intersection » du deuxième moment pour une grande valeur de $n_3$.

La construction de $U^*$ peut être évitée en notant que, sous la procédure BWO, chaque vecteur $D_\Diamond$ suit une loi hypergéométrique multivariée. Par conséquent, les poids de rééchantillonnage peuvent être produits directement. L'algorithme peut être adapté au cas général où $N/n_3$ n'est pas un entier en appliquant n'importe laquelle des techniques mentionnées à la section 2.4.

### 3.4.2 Une généralisation du bootstrap avec remise pour le plan à plusieurs degrés bidimensionnel

Nous considérons maintenant le plan d'échantillonnage à deux degrés bidimensionnel avec un échantillon de premier degré commun $s_I$ présenté à la section 3.1.2. La procédure bootstrap proposée est similaire à celle décrite dans Rao et Wu (1988). Une réplique d'échantillon $s_I^*$ de taille $m-1$ est tirée par échantillonnage aléatoire simple avec remise (SIR) dans l'échantillon de premier degré original $s_I$. Les mesures bootstrap sont alors

$$\hat{M}_{d,\Diamond}^{*} = \frac{m}{m-1}\sum_{i\in s_I^*}\sum_{k\in s_\Diamond^i}\pi_{Ii}^{-1}\pi_{\Diamond k|i}^{-1}\delta_{y_{dk}} \quad \text{où} \quad \pi_{\Diamond k|i} = \frac{n_\Diamond^i}{N_i}. \tag{3.30}$$

Celle-ci peut se réécrire sous la forme

$$\hat{M}_{d,\Diamond}^{*} = \sum_{k\in s_\Diamond} w_{\Diamond,k} D_{\Diamond,k}\delta_{y_{dk}}, \tag{3.31}$$

où $s_\Diamond$ est l'union des échantillons $s_\Diamond^i$ pour $i \in s_I$, et où le poids de rééchantillonnage $D_{\Diamond,k}$ est égal à $m(m-1)^{-1}$ multiplié par le nombre de fois que l'UPE contenant $k$ est sélectionnée dans $s_I^*$.

Dans le cas linéaire, l'estimateur bootstrap du paramètre $\Delta t$ est alors

$$\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b) = \frac{m}{m-1}\sum_{i\in s_I^*}\pi_{Ii}^{-1}\widehat{\Delta t}^{\,i,\mathrm{co}}(a,b) \tag{3.32}$$

où $\widehat{\Delta t}^{\,i,\mathrm{co}}(a,b)$ est défini en (3.16). Après un peu de calcul, nous obtenons

$$E_*\left\{\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b)\right\} = \widehat{\Delta t}^{\,\mathrm{co}}(a,b) \quad \text{et} \quad V_*\left\{\widehat{\Delta t}^{\,\mathrm{co}^*}(a,b)\right\} = v^{\mathrm{HH}}\left\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\right\}, \tag{3.33}$$

où $\widehat{\Delta t}^{\,\mathrm{co}}(a,b)$ est donné en (3.15), et $v^{\mathrm{HH}}\{\widehat{\Delta t}^{\,\mathrm{co}}(a,b)\}$ est donné en (3.19). La généralisation proposée du bootstrap avec remise permet donc de reproduire exactement l'estimateur composite du premier moment et l'estimateur associé au deuxième moment.

# 4 Étude par simulations

Pour commencer, cinq populations artificielles sont générées comme il est décrit à la section 4.1. À la section 4.2, l'estimateur « union » est comparé à l'estimateur « intersection » en ce qui concerne la variance asymptotique. Une expérience Monte Carlo est ensuite présentée à la section 4.3, et les performances de la linéarisation et du bootstrap sont comparées dans le cas d'un plan d'échantillonnage SI2. Une comparaison similaire est faite à la section 4.4 dans le cas du plan d'échantillonnage à deux degrés bidimensionnel.

## 4.1 Configuration de la simulation

Nous avons généré cinq populations finies de taille $N$ = 40 000, contenant chacune deux variables étudiées $y_1$ et $y_2$. Les valeurs $y_{1k}$ et les valeurs $y_{2k}$ ont été produites conformément au modèle lognormal

$$y_{dk} = \exp(\alpha_d \varepsilon_k). \tag{4.1}$$

Les résidus $\varepsilon_k$ ont été générés selon une loi normale centrée réduite. Les valeurs des indices de Gini pour les cinq populations sont présentées au tableau 4.1.

**Tableau 4.1**
**Indices de Gini pour 5 populations**

| Population | Pop. 1 | Pop. 2 | Pop. 3 | Pop. 4 | Pop. 5 |
|---|---|---|---|---|---|
| $G_1$ | 0,249 | 0,298 | 0,348 | 0,397 | 0,447 |
| $G_2$ | 0,259 | 0,318 | 0,378 | 0,437 | 0,496 |
| $\Delta G$ | 0,010 | 0,020 | 0,030 | 0,040 | 0,049 |

Dans chacune des cinq populations, les unités ont été groupées en $M = 500$ grappes de taille égale $N_0 = 80$. Les grappes ont été construites de façon que le coefficient de corrélation intragrappe par rapport à la variable $y_1$ soit approximativement égal à 0,20 dans chaque population.

## 4.2 Comparaison des estimateurs « union » et « intersection »

À la présente section, nous comparons l'estimateur « union » à l'estimateur « intersection » pour la variation de l'indice de Gini en ce qui concerne la variance asymptotique. Nous considérons deux plans d'échantillonnage, à savoir le plan SI2 présenté à la section 3.1.1 avec $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 1 000; 1 000), (1 000; 2 000; 1 000) ou (1 000; 4 000; 1 000); et le plan MULT2 présenté à la section 3.1.2 avec $m = 300$ et $(n^i_{1\bullet}, n^i_3, n^i_{2\bullet})$ = (10; 10; 10), (10; 20; 10) ou (10; 40; 10).

Pour chaque population, nous calculons la variance asymptotique $V_{\text{lin}}(\widehat{\Delta G}^{\text{uni}})$ de l'estimateur « union », et la variance asymptotique $V_{\text{lin}}(\widehat{\Delta G}^{\text{int}})$ de l'estimateur « intersection ». Afin de comparer ces variances, nous calculons l'efficacité relative définie comme étant

$$\text{ER}\left\{\widehat{\Delta G}^{\cdot}\right\} = \frac{V_{\text{lin}}\left\{\widehat{\Delta G}^{(\cdot)}\right\}}{V_{\text{lin}}\left\{\widehat{\Delta G}^{\text{opt}}\right\}}, \tag{4.2}$$

où $\widehat{\Delta G}^{\text{opt}}$ est l'estimateur optimal.

Les résultats sont présentés au tableau 4.2. L'estimateur « union » est très inefficace. Sa variance asymptotique est de 15 à 244 fois plus élevée que celle de l'estimateur « intersection » pour SI2, et de 2 à 44 fois plus élevée que celle de l'estimateur « intersection » pour MULT2. La différence entre les deux estimateurs a tendance à diminuer quand la taille de l'échantillon commun augmente ou quand $\Delta G$ augmente. Par ailleurs, l'estimateur « intersection » est légèrement moins efficace que l'estimateur optimal pour SI2, la valeur de ER variant de 1,33 à 2,46, et approximativement aussi efficace que l'estimateur optimal pour MULT2, la valeur de ER variant de 1,02 à 1,12. Ces constatations appuient le raisonnement heuristique de la section 3.1.1. Étant donné la médiocre performance de l'estimateur « union », et la bonne performance de l'estimateur « intersection », nous examinons uniquement ce dernier dans la suite de l'étude par simulations.

**Tableau 4.2**
**Efficacité relative des estimateurs de variance « union » et « intersection » pour 5 populations**

| Plan | Taille d'échantillon | Pop. 1 | | Pop. 2 | | Pop. 3 | | Pop. 4 | | Pop. 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\widehat{\Delta G}^{\text{uni}}$ | $\widehat{\Delta G}^{\text{int}}$ | $\widehat{\Delta G}^{\text{uni}}$ | $\widehat{\Delta G}^{\text{int}}$ | $\widehat{\Delta G}^{\text{uni}}$ | $\widehat{\Delta G}^{\text{int}}$ | $\widehat{\Delta G}^{\text{uni}}$ | $\widehat{\Delta G}^{\text{int}}$ | $\widehat{\Delta G}^{\text{uni}}$ | $\widehat{\Delta G}^{\text{int}}$ |
| SI2 | $n_3$ = 1 000 | 600,22 | 2,46 | 200,23 | 2,27 | 96,72 | 2,10 | 58,73 | 1,96 | 39,35 | 1,85 |
| | $n_3$ = 2 000 | 410,23 | 1,84 | 141,71 | 1,76 | 70,71 | 1,68 | 44,18 | 1,61 | 30,33 | 1,54 |
| | $n_3$ = 4 000 | 250,02 | 1,47 | 88,40 | 1,43 | 45,17 | 1,40 | 28,86 | 1,36 | 20,23 | 1,33 |
| MULT2 | $n_3^i$ = 10 | 49,10 | 1,12 | 19,89 | 1,13 | 11,83 | 1,14 | 8,84 | 1,15 | 7,28 | 1,16 |
| | $n_3^i$ = 20 | 23,08 | 1,05 | 9,75 | 1,05 | 6,08 | 1,05 | 4,73 | 1,06 | 4,04 | 1,07 |
| | $n_3^i$ = 40 | 9,15 | 1,02 | 4,25 | 1,02 | 2,90 | 1,02 | 2,41 | 1,02 | 2,16 | 1,02 |

## 4.3 Comparaison de la linéarisation et du bootstrap pour le plan SI2

À la présente section, nous comparons la linéarisation et le bootstrap pour l'estimation de la variance et pour la production des intervalles de confiance, dans le cas de l'estimateur « intersection », pour la variation de l'indice de Gini sous le plan d'échantillonnage SI2. Pour chaque population, nous avons tiré $B = 10\,000$ échantillons bidimensionnels selon le plan SI2 indexé par $(n_{1\bullet}, n_3, n_{2\bullet}) = (1\,000; 1\,000; 1\,000)$, $(n_{1\bullet}, n_3, n_{2\bullet}) = (1\,000; 2\,000; 1\,000)$ ou $(n_{1\bullet}, n_3, n_{2\bullet}) = (1\,000; 4\,000; 1\,000)$. Dans chaque échantillon, nous avons calculé l'estimateur « intersection » $\widehat{\Delta G}^{\text{int}}$ de la variation de l'indice de Gini. Pour cet estimateur, nous avons calculé i) l'estimateur de variance par linéarisation $v_{\text{int}}(\widehat{\Delta G}^{\text{int}})$ donné en (3.24), et ii) l'estimateur de variance par bootstrap $v_{\text{BWO}}(\widehat{\Delta G})$, en suivant la procédure bootstrap décrite à la section 3.4.1.

Pour mesurer le biais d'un estimateur de variance $v(\widehat{\Delta G})$, nous avons utilisé le biais relatif de Monte Carlo en pourcentage

$$\text{BR}\left\{v\left(\widehat{\Delta G}\right)\right\} = 100 \times \frac{B^{-1}\sum_{b=1}^{B} v\left(\widehat{\Delta G}_b\right) - \text{EQM}\left(\widehat{\Delta G}\right)}{\text{EQM}\left(\widehat{\Delta G}\right)}, \tag{4.3}$$

où $v(\widehat{\Delta G}_b)$ désigne l'estimateur $v(\widehat{\Delta G})$ dans le $b^{\text{e}}$ échantillon, et $\text{EQM}(\widehat{\Delta G})$ est une approximation fondée sur la simulation de l'erreur quadratique moyenne réelle de $\widehat{\Delta G}$, obtenue d'après une exécution indépendante de 100 000 simulations. Comme mesure de stabilité de $v(\widehat{\Delta G})$, nous utilisons la stabilité relative

$$\text{SR}\left\{v\left(\widehat{\Delta G}\right)\right\} = \frac{\left[B^{-1}\sum_{b=1}^{B}\left\{v\left(\widehat{\Delta G}\right) - \text{EQM}\left(\widehat{\Delta G}\right)\right\}^2\right]^{1/2}}{\text{EQM}\left(\widehat{\Delta G}\right)}. \tag{4.4}$$

Enfin, nous avons comparé les taux de couverture de i) l'intervalle de confiance fondé sur l'hypothèse de normalité avec utilisation de l'estimateur de variance par linéarisation et ii) l'intervalle de confiance bootstrap par la méthode des percentiles. Les estimateurs de variance bootstrap et les intervalles de confiance bootstrap sont fondés sur $C = 1\,000$ répliques bootstrap. Les taux d'erreur des intervalles de confiance (avec un taux d'erreur nominal unilatéral de 2,5 % dans chaque queue) sont comparés. La comparaison avec le taux d'erreur nominal de 5 % n'a donné aucune différence qualitative et est donc omise.

Les résultats sont présentés au tableau 4.3. Les estimateurs de variance présentent tous deux un biais négatif. Ce biais est modéré (moins de 5 %) dans la plupart des cas, sauf pour la plus petite taille d'échantillon $n = 1\,000$, et pour la population $U_5$ possédant la valeur la plus élevée de $\Delta G$. L'estimateur de variance par bootstrap présente systématiquement un biais légèrement plus important que l'estimateur de variance par linéarisation, mais la différence diminue à mesure que la taille d'échantillon augmente. Pour les deux estimateurs de variance, l'instabilité augmente avec $\Delta G$. L'estimateur de variance par bootstrap est légèrement plus stable pour la taille d'échantillon la plus petite $n = 1\,000$, mais la situation est inversée quand la taille d'échantillon augmente. Pour ce qui est de la couverture des intervalles de confiance, les deux méthodes aboutissent à une sous-couverture qui concorde avec le biais négatif des deux estimateurs de variance. Les intervalles de confiance fondés sur l'hypothèse de normalité donnent une couverture un peu meilleure que les intervalles de confiance bootstrap par la méthode des percentiles. Pour les deux catégories d'intervalles de confiance, la sous-couverture est plus importante quand $\Delta G$ augmente, et diminue quand la taille d'échantillon augmente.

**Tableau 4.3**
**Biais relatif, stabilité relative et taux d'erreur nominaux unilatéraux pour l'estimation de la variance par linéarisation et par bootstrap de l'estimateur « intersection » de la variation de l'indice de Gini pour 5 populations sous le plan d'échantillonnage SI2**

| Pop. | Linéarisation | | | | | Bootstrap | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BR | SR | I | S | I+S | BR | SR | I | S | I+S |
| Taille d'échantillon $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 1 000; 1 000) | | | | | | | | | | |
| Pop. 1 | -1,41 | 24,6 | 1,8 | 4,5 | 6,3 | -1,83 | 24,6 | 1,8 | 4,9 | 6,7 |
| Pop. 2 | -1,98 | 32,4 | 1,6 | 5,2 | 6,8 | -2,64 | 32,1 | 1,7 | 5,9 | 7,6 |
| Pop. 3 | -2,80 | 41,9 | 1,3 | 6,3 | 7,7 | -3,83 | 40,9 | 1,3 | 7,0 | 8,3 |
| Pop. 4 | -4,00 | 52,5 | 1,0 | 7,7 | 8,7 | -5,57 | 50,6 | 1,1 | 8,2 | 9,3 |
| Pop. 5 | -5,80 | 64,0 | 1,0 | 9,2 | 10,1 | -8,11 | 60,6 | 0,8 | 9,9 | 10,7 |
| Taille d'échantillon $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 2 000; 1 000) | | | | | | | | | | |
| Pop. 1 | -1,38 | 17,3 | 1,6 | 3,7 | 5,3 | -1,67 | 17,8 | 1,8 | 4,1 | 5,9 |
| Pop. 2 | -1,64 | 23,0 | 1,4 | 4,3 | 5,8 | -2,05 | 23,2 | 1,4 | 4,7 | 6,1 |
| Pop. 3 | -1,99 | 30,1 | 1,2 | 5,0 | 6,2 | -2,58 | 30,0 | 1,1 | 5,3 | 6,4 |
| Pop. 4 | -2,50 | 38,4 | 1,0 | 6,0 | 6,9 | -3,38 | 37,9 | 1,0 | 6,3 | 7,3 |
| Pop. 5 | -3,30 | 47,9 | 0,7 | 7,2 | 7,9 | -4,62 | 46,7 | 0,7 | 7,5 | 8,2 |
| Taille d'échantillon $(n_{1\bullet}, n_3, n_{2\bullet})$ = (1 000; 4 000; 1 000) | | | | | | | | | | |
| Pop. 1 | -0,60 | 11,9 | 2,0 | 3,4 | 5,3 | -0,68 | 12,8 | 2,1 | 3,4 | 5,5 |
| Pop. 2 | -0,67 | 15,9 | 1,8 | 3,7 | 5,6 | -0,80 | 16,5 | 2,0 | 3,9 | 5,9 |
| Pop. 3 | -0,83 | 20,8 | 1,8 | 4,4 | 6,2 | -1,03 | 21,3 | 1,9 | 4,4 | 6,3 |
| Pop. 4 | -1,13 | 26,7 | 1,5 | 5,0 | 6,6 | -1,46 | 26,9 | 1,6 | 5,0 | 6,6 |
| Pop. 5 | -1,64 | 33,4 | 1,4 | 5,8 | 7,1 | -2,18 | 33,5 | 1,4 | 5,8 | 7,1 |

## 4.4 Comparaison de la linéarisation et du bootstrap pour le plan MULT2

À la présente section, nous comparons la linéarisation et le bootstrap pour l'estimation de la variance et la production des intervalles de confiance, dans le cas de l'estimateur « intersection » pour la variation de l'indice de Gini sous le plan d'échantillonnage MULT2 présenté à la section 3.1.2. Pour chaque population, nous avons tiré $B = 10\,000$ échantillons à deux degrés bidimensionnels selon le plan MULT2 indexé par $m = 300$ et $(n_{1\bullet}^i, n_3^i, n_{2\bullet}^i) = (10; 10; 10)$, (10; 20; 10) ou (10; 40; 10). Dans chaque échantillon, nous avons calculé l'estimateur « intersection » $\widehat{\Delta G}^{\text{int}}$ de la variation de l'indice de Gini. Pour cet estimateur, nous avons calculé i) l'estimateur de variance par linéarisation $v^{\text{HH}}\{\widehat{\Delta G}^{\text{co}}(a,b)\}$ donné en (3.26), et

ii) l'estimateur de variance par bootstrap $v_{\mathrm{BWR}}(\widehat{\Delta G}^{\mathrm{int}})$, obtenu selon la procédure bootstrap décrite à la section 3.4.2.

Pour mesurer le biais d'un estimateur de variance $v(\widehat{\Delta G})$, nous avons utilisé le biais relatif Monte Carlo en pourcentage défini par l'équation (4.3), et la stabilité relative définie par l'équation (4.4). L'erreur quadratique moyenne réelle de $\widehat{\Delta G}$ a été obtenue par une exécution indépendante de 100 000 simulations. En outre, nous avons comparé les taux de couverture de i) l'intervalle de confiance fondé sur l'hypothèse de normalité avec utilisation de l'estimateur de variance par linéarisation et ii) l'intervalle de confiance bootstrap par la méthode des percentiles. Les estimateurs de variance par bootstrap et les intervalles de confiance par bootstrap sont fondés sur $C = 1\,000$ répliques bootstrap. Les taux d'erreur des intervalles de confiance (avec un taux d'erreur nominal unilatéral de 2,5 % dans chaque queue) sont comparés. La comparaison avec le taux d'erreur nominal de 5 % n'a produit aucune différence qualitative et est donc omise.

Les résultats sont présentés au tableau 4.4. Les estimateurs de variance sont tous deux approximativement sans biais pour les petites valeurs de $\Delta G$, mais présentent un biais négatif modéré qui augmente avec $\Delta G$. L'estimateur de variance par bootstrap présente un biais plus important que l'estimateur de variance par linéarisation. Pour les deux estimateurs de variance, l'instabilité augmente avec $\Delta G$. L'estimateur de variance par bootstrap est légèrement plus stable que l'estimateur de variance par linéarisation. Les deux méthodes donnent lieu à une sous-couverture qui concorde avec le biais négatif des deux estimateurs de variance. Les intervalles de confiance fondés sur l'hypothèse de normalité donnent des résultats légèrement meilleurs. Pour les deux intervalles de confiance, la sous-couverture est plus importante quand $\Delta G$ augmente, et diminue quand la taille d'échantillon augmente.

**Tableau 4.4**
**Biais relatif, stabilité relative et taux d'erreur nominaux unilatéraux pour l'estimation de la variance par linéarisation et par bootstrap de l'estimateur « intersection » de la variation de l'indice de Gini pour 5 populations sous le plan d'échantillonnage MULT2**

| Pop. | Linéarisation | | | | | Bootstrap | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BR | SR | I | S | I+S | BR | SR | I | S | I+S |
| Tailles d'échantillon $m = 300$ et $(n_{1\bullet}^i, n_{3}^i, n_{2\bullet}^i) = (10; 10; 10)$ | | | | | | | | | | |
| Pop. 1 | 1,23 | 33,8 | 0,6 | 4,9 | 5,5 | 1,09 | 33,2 | 0,6 | 6,0 | 6,6 |
| Pop. 2 | 0,64 | 41,1 | 0,8 | 5,5 | 6,3 | -0,20 | 39,7 | 0,6 | 6,5 | 7,1 |
| Pop. 3 | -0,42 | 48,7 | 0,7 | 7,1 | 7,8 | -2,05 | 46,6 | 0,7 | 8,4 | 9,1 |
| Pop. 4 | -2,07 | 56,4 | 0,8 | 8,4 | 9,2 | -4,47 | 53,3 | 0,6 | 9,6 | 10,2 |
| Pop. 5 | -4,44 | 63,7 | 0,9 | 9,2 | 10,1 | -7,56 | 59,5 | 0,4 | 10,3 | 10,7 |
| Tailles d'échantillon $m = 300$ et $(n_{1\bullet}^i, n_{3}^i, n_{2\bullet}^i) = (10; 20; 10)$ | | | | | | | | | | |
| Pop. 1 | 1,70 | 32,6 | 1,5 | 4,9 | 6,4 | -1,70 | 32,3 | 1,5 | 6,0 | 7,5 |
| Pop. 2 | 1,10 | 39,0 | 1,4 | 5,4 | 6,8 | -1,91 | 38,3 | 1,5 | 6,9 | 8,4 |
| Pop. 3 | 0,17 | 45,6 | 1,2 | 7,4 | 8,6 | -2,49 | 44,4 | 1,1 | 7,7 | 8,8 |
| Pop. 4 | -1,17 | 52,0 | 1,0 | 9,0 | 10,0 | -3,58 | 50,3 | 0,8 | 9,7 | 10,5 |
| Pop. 5 | -3,03 | 57,9 | 0,9 | 10,4 | 11,3 | -5,35 | 55,4 | 0,7 | 11,0 | 11,7 |
| Tailles d'échantillon $m = 300$ et $(n_{1\bullet}^i, n_{3}^i, n_{2\bullet}^i) = (10; 40; 10)$ | | | | | | | | | | |
| Pop. 1 | -0,99 | 32,1 | 1,2 | 6,1 | 7,3 | -3,21 | 32,2 | 1,7 | 6,7 | 8,4 |
| Pop. 2 | -1,68 | 38,3 | 1,4 | 6,7 | 8,1 | -3,70 | 38,3 | 1,4 | 7,6 | 9,0 |
| Pop. 3 | -2,58 | 44,6 | 1,3 | 7,5 | 8,8 | -4,40 | 44,5 | 1,2 | 8,9 | 10,1 |
| Pop. 4 | -3,78 | 50,6 | 1,1 | 8,9 | 10,0 | -5,50 | 50,1 | 0,9 | 10,6 | 11,5 |
| Pop. 5 | -5,39 | 55,9 | 0,8 | 10,9 | 11,7 | -7,16 | 54,8 | 0,6 | 12,8 | 13,4 |

# 5 Conclusion

Dans le présent article, nous considérons l'estimation de l'évolution de l'indice de Gini. Nous avons présenté la classe d'estimateurs composites introduite par Goga et coll. (2009), et étudié plus particulièrement l'estimateur « intersection » qui utilise l'échantillon commun seulement, et l'estimateur « union » qui utilise les échantillons disponibles complets. Nous avons justifié tant de manière heuristique que par une étude en simulation décrite à la section 4.2 que l'estimateur « intersection » peut être proche de l'estimateur optimal, tandis que l'estimateur « union » présente des propriétés médiocres dans tous les scénarios considérés. L'estimateur « intersection » est également facile à calculer, tandis que l'estimateur optimal fait intervenir des quantités inconnues qui doivent être estimées en pratique. Par conséquent, nous recommandons l'utilisation de l'estimateur « intersection » pour estimer la variation de l'indice de Gini.

Nous avons également comparé les méthodes de linéarisation et de bootstrap pour estimer la variance et pour produire les intervalles de confiance. Dans les scénarios que nous avons considérés dans l'étude en simulation, la linéarisation donnait de meilleurs résultats que le bootstrap, les biais relatifs étant habituellement plus faibles pour l'estimateur de variance, et les taux de couverture étant meilleurs pour les intervalles de confiance fondés sur l'hypothèse de normalité que pour ceux fondés sur la méthode des percentiles. Les intervalles de confiance fondés sur le bootstrap $-t$ (non examinés dans l'étude en simulation) seraient une option intéressante, mais étant donné l'importance des calculs requis, ils sont moins attractifs pour un utilisateur des données. La linéarisation présente aussi l'avantage d'offrir une approche unifiée convenant pour tout plan d'échantillonnage, alors qu'un plan d'échantillonnage donné requiert habituellement une procédure bootstrap particulière, comme l'illustre l'application du bootstrap sans remise (BWO) pour l'échantillonnage SI et du bootstrap avec remise (BWR) pour l'échantillonnage à plusieurs degrés.

L'étude par simulations nous permet de constater que les taux de couverture ne sont parfois bien respectés ni dans le cas de la linéarisation ni dans celui du bootstrap, particulièrement dans le contexte de l'échantillonnage à plusieurs degrés et même si l'on utilise de grandes tailles d'échantillon. Des intervalles de confiance donnant de meilleurs taux de couverture tout en demandant un temps de calcul raisonnable sont nécessaires. Cette question devrait faire l'objet d'une étude plus approfondie.

# Remerciements

Nous remercions Anne Ruiz-Gazen pour des discussions utiles. Nous remercions également deux examinateurs et un rédacteur associé pour leurs commentaires et suggestions constructifs qui nous ont permis d'améliorer considérablement le manuscrit.

# Annexe

## Preuve de l'équation (3.6)

Partant de (3.3), nous avons $\widehat{\Delta t}^{\,\mathrm{co}} = N(A^{\top}X)$, où $X = (\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3})^{\top}$ et $A = (b, -a, 1)^{\top}$. Cela mène à

$$V\left\{\widehat{\Delta t}^{\,\mathrm{co}}\right\} = N^2\{A^\top V(X)A\}. \tag{A.1}$$

Nous calculons les éléments de $V(X)$ séparément. Nous avons

$$\begin{aligned} V(\overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \left(\frac{1}{n_3} - \frac{1}{N}\right) S^2_{y_2-y_1,U} \\ &= \left(\frac{1}{n_3} - \frac{1}{N}\right)\left(S^2_{y_2,U} + S^2_{y_1,U} - 2S_{y_1y_2,U}\right). \end{aligned}$$

En outre, puisque $E(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3} \mid s_2) = 0,$ nous avons

$$\begin{aligned} V(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}) &= \mathrm{EV}(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3} \mid s_2), \\ &= \mathrm{EV}\left(\frac{n_2}{n_{2\bullet}}\overline{y}_{2,s_2} - \frac{n_3}{n_{2\bullet}}\overline{y}_{2,s_3} - \overline{y}_{2,s_3} \mid s_2\right) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)^2 \mathrm{EV}(\overline{y}_{2,s_3} \mid s_2) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)^2 \left(\frac{1}{n_3} - \frac{1}{n_2}\right) S^2_{y_2,U} \\ &= \frac{n_2}{n_3(n_2 - n_3)} S^2_{y_2,U} \end{aligned}$$

et

$$\begin{aligned} \mathrm{Cov}(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \mathrm{ECov}(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3} \mid s_2) \\ &= \mathrm{ECov}\left(\frac{n_2}{n_{2\bullet}}\overline{y}_{2,s_2} - \frac{n_3}{n_{2\bullet}}\overline{y}_{2,s_3} - \overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3} \mid s_2\right) \\ &= -\left(1 + \frac{n_3}{n_{2\bullet}}\right)\mathrm{ECov}(\overline{y}_{2,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3} \mid s_2) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)\left(\frac{1}{n_3} - \frac{1}{n_2}\right)\left(S^2_{y_2,U} - S_{y_1y_2,U}\right) \\ &= -\frac{1}{n_3}\left(S^2_{y_2,U} - S_{y_1y_2,U}\right). \end{aligned}$$

Des arguments similaires mènent à

$$\begin{aligned} V(\overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}) &= \frac{n_1}{n_3(n_1 - n_3)} S^2_{y_1,U}, \\ \mathrm{Cov}(\overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \frac{1}{n_3}\left(S^2_{y_1,U} - S_{y_1y_2,U}\right). \end{aligned}$$

Enfin, nous considérons $\mathrm{Cov}(\bar{y}_{2,s_2\bullet} - \bar{y}_{2,s_3}, \bar{y}_{1,s_1\bullet} - \bar{y}_{1,s_3})$. Nous commençons par calculer $\mathrm{Cov}(\bar{y}_{2,s_2\bullet}, \bar{y}_{1,s_1\bullet})$, qui peut s'écrire sous la forme

$$\begin{aligned}
\mathrm{Cov}(\bar{y}_{2,s_2\bullet}, \bar{y}_{1,s_1\bullet}) &= \mathrm{Cov}(E(\bar{y}_{2,s_2\bullet} \mid s_{1\bullet}), E(\bar{y}_{1,s_1\bullet} \mid s_{1\bullet})) \\
&= \mathrm{Cov}(\bar{y}_{2,U\setminus s_1\bullet}, \bar{y}_{1,s_1\bullet}) \\
&= \mathrm{Cov}\left(\frac{N}{N-n_{1\bullet}}\bar{y}_{2,U} - \frac{n_{1\bullet}}{N-n_{1\bullet}}\bar{y}_{2,s_1\bullet}, \bar{y}_{1,s_1\bullet}\right) \\
&= -\frac{n_{1\bullet}}{N-n_{1\bullet}}\mathrm{Cov}(\bar{y}_{2,s_1\bullet}, \bar{y}_{1,s_1\bullet}) \\
&= -\frac{n_{1\bullet}}{N-n_{1\bullet}}\left(\frac{1}{n_{1\bullet}} - \frac{1}{N}\right)S_{y_1y_2,U} \\
&= -\frac{1}{N}S_{y_1y_2,U}.
\end{aligned}$$

Des arguments similaires mènent à

$$\mathrm{Cov}(\bar{y}_{2,s_2\bullet}, \bar{y}_{1,s_3}) = \mathrm{Cov}(\bar{y}_{2,s_3}, \bar{y}_{1,s_1\bullet}) = -\frac{1}{N}S_{y_1y_2,U}.$$

Nous obtenons

$$\begin{aligned}
\mathrm{Cov}(\bar{y}_{2,s_2\bullet} - \bar{y}_{2,s_3}, \bar{y}_{1,s_1\bullet} - \bar{y}_{1,s_3}) &= \frac{1}{N}S_{y_1y_2,U} + \mathrm{Cov}(\bar{y}_{2,s_3}, \bar{y}_{1,s_3}) \\
&= \frac{1}{N}S_{y_1y_2,U} + \left(\frac{1}{n_3} - \frac{1}{N}\right)S_{y_1y_2,U} \\
&= \frac{1}{n_3}S_{y_1y_2,U}.
\end{aligned}$$

En résumé, nous obtenons

$$V(X) = \begin{pmatrix}
\frac{n_2}{n_3(n_2-n_3)}S^2_{y_2,U} & \frac{1}{n_3}S_{y_1y_2,U} & -\frac{1}{n_3}\left(S^2_{y_2,U} - S_{y_1y_2,U}\right) \\
 & \frac{n_1}{n_3(n_1-n_3)}S^2_{y_1,U} & \frac{1}{n_3}\left(S^2_{y_1,U} - S_{y_1y_2,U}\right) \\
 & & \left(\frac{1}{n_3} - \frac{1}{N}\right)\left(S^2_{y_2,U} + S^2_{y_1,U} - 2S_{y_1y_2,U}\right)
\end{pmatrix}$$

qui, avec (A.1), mène à (3.6).

# Bibliographie

Antal, E., et Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106, 534-543.

Barrett, G.F., et Donald, S.G. (2009). Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *Journal of Business and Economic Statistics*, 27, 1-17.

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Revue Internationale de Statistique*, 80, 127-148.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 451-467.

Berger, Y.G. (2008). A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics*, 24, 541-555.

Bertail, P., et Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Économie et de Statistique*, 46, 49-83.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, 137, 674-707.

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Brändén, P., et Jonasson, J. (2012). Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39, 830-838.

Campbell, C. (1980). A different view of finite population estimation. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 319-324.

Chao, M.-T., et Lo, S.-H. (1985). A Bootstrap method for finite population. *Sankhyā, Series A*, 47, 3, 399-405.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. Thèse de doctorat, Université Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, 43(6), 2484-2506.

Chen, J., et Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17, 1047-1064.

Davison, A.C., et Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Davison, A.C., et Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, 23, 3, 371-386.

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. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2004001/article/6991-fra.pdf.

Deville, J.-C. (1997). Estimation de la variance de l'indice de Gini mesurée par sondage. *Actes des Journées de Méthodologie Statistique*, *Insee Méthodes*.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-fra.pdf.

Druckman, A., et Jackson, T. (2008). Measuring resource inequalities: The concepts and methodology for an area-based Gini coefficient. *Ecological Economics*, 65, 242-252.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze Lettere ed Arti.*

Glasser, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Thèse de doctorat, Université Rennes 2.

Goga, C., et Ruiz-Gazen, A. (2014). Efficient estimation of nonlinear finite population parameters using nonparametrics. *Journal of the Royal Statistical Society B*, 76, 113-140.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96, 691-709.

Gordon, L. (1983). Successive sampling in large finite populations. *Annals of Statistics*, 11, 702-706.

Graczyk, P.P. (2007). Gini coefficient: A new way to express selectivity of kinase inhibitors against a family of Kinases. *Journal of Medicinal Chemistry*, 50, 5773-5779.

Gross, S.T. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

Groves-Kirkby, C.J., Denman, A.R. et Phillips, P.S. (2009). Lorenz Curve and Gini coefficient: Novel tools for analysing seasonal variation of environmental radon gas. *Journal of Environmental Management*, 90, 2480-2487.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Tud. Akad. Mat. Kutatò Int. Közl.*, 5, 361-374.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *Annals of Mathematical Statistics*, 32, 506-523.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Karagiannis, E., et Kovačević, M.S. (2000). A method to calculate the jackknife variance estimator for the Gini coefficient. *Oxford Bulletin of Economics and Statistics*, 62, 119-122.

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

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics,* 9, 1010-1019.

Lai, D., Huang, J., Risser, J.M. et Kapadia, A.S. (2008). Statistical properties of generalized Gini coefficient with application to health inequality measurement. *Social Indicator Research,* 87, 249-258.

Langel, M., et Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society, Series A,* 176, 521-540.

Lisker, T. (2008). Is the Gini coefficient a stable measure on galaxy structure? *The Astrophysical Journal Supplement Series,* 179, 319-325.

Navarro, V., Muntaner, C., Borrell, C., Benach, J., Quiroga, A., Rodríguez-Sanz, M., Vergès, N. et Pasarín, M.I. (2006). Politics and health outcomes. *The Lancet*, 18, 1033-1037.

Nygård, F., et Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 4, 399-412.

Ohlsson, E. (1986). Asymptotic normality of the Rao-Hartley-Cochran estimator: An application of the martingale CLT. *Scandinavian Journal of Statistics*, 13, 17-28.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields*, 81, 341-352.

Pires, A.M., et Branco, J.A. (2002). Partial influence functions. *Journal of Multivariate Analysis,* 83, 451-468.

Presnell, B., et Booth, J.G. (1994). *Resampling Methods for Sample Surveys*. Rapport technique.

Qin, Y., Rao, J.N.K. et Wu, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. *Economic Modelling,* 27, 1429-1435.

Qualité, L., et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée. *Techniques d'enquête*, 34, 2, 193-201. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2008002/article/10758-fra.pdf.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

Reid, N. (1981). Influence functions for censored data. *The Annals of Statistics*, 9, 78-92.

Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Annals of Mathematical Statistics*, 43, 373-397, 748-776.

Saegusa, T., et Wellner, J.A. (2013). Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics*, 41, 269-295.

Sandström, A., Wretman, J.H. et Waldèn, B. (1985). Variance estimators of the Gini coefficient - Simple random sampling. *Metron,* 43, 41-70.

Sandström, A., Wretman, J.H. et Waldèn, B. (1988). Variance estimators of the Gini coefficient - Probability sampling. *Journal of Business and Economic Statistics,* 6, 113-119.

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

Sen, P.K. (1980). Limit theorems for an extended coupon collector's problem and for successive subsampling with varying probabilities. *Calcutta Statistical Association Bulletin,* 29, 113-132.

Shao, J., et Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association,* 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics,* 20, 135-154.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician,* 38, 288-289.

Yitzhaki, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics,* 9, 235-239.