Published in Techniques d'enquête 28, issue 1, 83-91, 2002
which should be used for any reference to this work

Estimation sans biais par calage sur la répartition dans les plans simples sans remise

Yves Tillé
Groupe de Statistique, Université de Neuchâtel
Espace de l'Europe 4, Case postale 1825, 2002 Neuchâtel , Suisse
e-mail : yves.tille@unine.ch

12 mars 2002

Résumé

L'estimateur post-stratifié a parfois des post-strates vides. Pour pallier ce problème, on construit un estimateur post-stratifié dont les tailles des post-strates sont fixées dans l'échantillon. Les tailles des post-strates sont alors aléatoires dans la population. Ensuite, on construit un estimateur lissé en effectuant une moyenne mobile des estimateurs post-stratifiés. Cette technique permet de construire une théorie exacte du calage sur la répartition. L'estimateur obtenu est non seulement calé sur la répartition, il est linéaire, et exactement sans biais. On compare ensuite l'estimateur calé à l'estimateur par la régression. On propose enfin un estimateur approché de la variance validé au moyen de simulations.

1 Introduction

Lors d'une enquête par sondage, il arrive que l'on connaisse les valeurs d'un caractère auxiliaire pour toutes les unités de la population. Cette information peut être disponible quand les unités sont sélectionnées dans une base de données qui contient d'autres variables d'intérêt. On est alors tenté de caler les résultats d'une enquête sur cette information auxiliaire. Soit, on ne retient de cette variable auxiliaire que certaines fonctions (moments, effectifs) en vue d'utiliser une méthode de calage (voir par exemple Deville et Särndal, 1992 ou Estevao, Hidiroglou, et Särndal, 1995), soit on peut découper cette variable en classes en vue d'utiliser un estimateur post-stratifié.

Si l'on opte pour l'estimateur post-stratifié, le découpage en strates s'avère délicat. Théoriquement, les strates doivent être définies avant la sélection de l'échantillon. Où faut-il placer les bornes des post-strates ? De quelles tailles doivent être les post-strates ? Cette dernière question est la plus embarrassante, car le problème principal de la post-stratification est la possibilité d'obtenir des post-strates vides. Les tailles des post-strates doivent donc être suffisamment grandes pour que la probabilité d'obtenir une post-strate vide soit négligeable.

Ces problèmes ne se limitent pas aux estimateurs post-stratifiés, En effet, les estimateurs par la régression ou calés peuvent également ne pas exister pour certains échantillons.

Notre objectif est de définir une nouvelle méthode permettant d'utiliser l'information auxiliaire dans la population. Cette méthode est basée sur la définition de post-strates pour lesquelles le nombre d'unités est fixé dans l'échantillon, et non dans la population. On peut ainsi importer dans l'estimateur une information auxiliaire complexe issue de la connaissance de toutes les valeurs prises par la variable auxiliaire, tout en évitant à la fois le problème de la définition des bornes des post-strates et le problème des post-strates vides.

Cet article est organisé comme suit : En Section 2, la notation est définie, en Section 3, on donne le principe du conditionnement par les rangs, qui permet de définir des estimateurs sans biais en Section 4. Ensuite, en Section 5, on définit l'estimateur lissé, et un cas particulier est examiné en détail en Section 6. En Section 7, on donne une application à l'estimation d'une fonction de répartition. En section 8, on compare ce nouvel estimateur à l'estimateur par la régression et l'estimateur du plan simple sans remise. Le calcul de variance est abordé en Section 9. Suite à l'impossibilité de donner une solution exacte, on donne une approximation en Section 10, qui est testée par des simulations en Section 11. Enfin, des conclusions générales sont données en Section 12.

2 Notation

On suppose que l'on a une population composée de N unités d'observation dont les étiquettes des observations sont notées $\{1, ..., k, ..., N\}$. Dans cette population, on s'intéresse à un caractère d'intérêt $Y_k, k \in U$. L'objectif consiste à estimer le total $Y = \sum_{k \in U} Y_k$. On sélectionne un échantillon aléatoire S de taille fixe n au moyen d'un plan aléatoire simple sans remise. On note I_k la variable aléatoire indicatrice qui prend la valeur 1 si l'unité k est dans l'échantillon et 0 sinon. Les probabilités d'inclusion d'ordre un sont donc définies par $Pr(k \in S) = \pi_k = n/N, k \in U$, et les probabilités d'inclusion d'ordre deux par $Pr(k, \ell \in S) = \pi_{k\ell} = n(n-1)/(N(N-1)), k \neq \ell \in U$.

On s'intéressera à la classe des estimateurs linéaires de Y qui s'écrit

$$\widehat{Y}_w = \sum_{k \in S} w_k Y_k,$$

où les poids w_k peuvent dépendre de l'échantillon S et donc être aléatoires.

Une des possibilités consiste à prendre $w_k = 1/\pi_k = n/N$, ce qui donne l'estimateur de Horvitz-Thompson,

$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{Y_k}{\pi_k} = \frac{N}{n} \sum_{k \in S} Y_k,$$

qui est sans biais.

Nous allons cependant nous intéresser à la classe plus générale des estimateurs pondérés conditionnellement (Tillé, 1998, 1999b) où les unités sont pondérées par des inverses de probabilités d'inclusion conditionnelles. Si Z est une statistique quelconque, alors l'estimateur pondéré conditionnellement

$$\widehat{Y}_Z = \sum_{k \in S} \frac{Y_k}{E(I_k|Z)} \tag{1}$$

est strictement sans biais si et seulement si $E(I_k|Z) > 0$, pour tout $k \in U$. En effet,

$$E\left(\widehat{Y}|Z\right) = \sum_{k \in U} \frac{E(I_k|Z)Y_k}{E(I_k|Z)} = Y.$$

Comme l'estimateur est conditionnellement sans biais, il est également non-conditionnellement sans biais. L'estimateur (1) généralise, selon le choix de la statistique Z utilisée, l'estimateur stratifié, mais aussi (à une approximation près) l'estimateur par la régression (voir Tillé, 1998).

3 Conditionnement sur des rangs

Supposons maintenant que les N valeurs $X_1, ..., X_k, ..., X_N$ d'un caractère auxiliaire x soient connues sur les N unités de la population. Dans un premier temps, on suppose que tous les X_k prennent des valeurs distinctes, cette hypothèse sera ensuite levée en section 5. Le rang R_k de l'unité k où

$$R_k = \#\{\ell \in U | X_\ell \leq X_k\}.$$

On note par ailleurs $r_j, j = 1, ...n$, les rangs de la population ordonnés des n unités sélectionnées dans l'échantillon donc $r_1 < r_2 < ... < r_{n-1} < r_n$. Les r_j sont des variables aléatoires ayant une distribution hypergéométrique négative (voir Tillé, 1999b).

La statistique utilisée pour définir les probabilités d'inclusion conditionnelles est un sous-ensemble de $\{r_1, ..., r_j, ..., r_n\}$. On définit d'abord

- un entier q tel que $2 \leq q \leq n$, définissant la période,
- un entier b tel que $2 \leq b$, définissant la bordure,
- un entier ℓ tel que $b \leq \ell \leq b + q - 1$, définissant le décalage.

Les quantité q, b, et ℓ servent à définir un sous-ensemble d'indices :

$$E_\ell = \{r_\ell, r_{\ell+q}, r_{\ell+2q}, ..., r_{\ell+hq}, ..., r_{\ell+Hq}\}, \text{ pour } \ell = b, ..., b + q - 1.$$

Par exemple, si $n = 18, q = 4, b = 3$, alors

$$E_3 = \{r_3, r_7, r_{11}, r_{15}\}, E_4 = \{r_4, r_8, r_{12}, r_{16}\}, E_5 = \{r_5, r_9, r_{13}\}, E_6 = \{r_6, r_{10}, r_{14}\}.$$

La probabilité d'inclusion conditionnelle est calculée par rapport à l'un des E_ℓ.

La valeur de H est définie de manière à ce que $\ell + Hq \leq n - b + 1$ et donc H est le plus grand entier tel que $H \leq (n - b - \ell + 1)/q$. Il est donc clair que H dépend de ℓ.

Ensuite, on peut calculer les probabilités d'inclusion :

$$E\left(I_k|E_\ell\right) = \begin{cases} 1 & \text{si } k \in E_\ell \\ \dfrac{q-1}{r_{\ell+hq} - r_{\ell+(h-1)q} - 1} & \text{si } r_{\ell+(h-1)q} < k < r_{\ell+hq}, h = 1, ..., H \\ \dfrac{\ell - 1}{r_\ell - 1} & \text{si } k < r_\ell \\ \dfrac{n - (\ell + Hq)}{N - r_{\ell+Hq}} & \text{si } k > r_{\ell+Hq}. \end{cases}$$

Ces probabilités d'inclusion sont donc assez contrastées. Cependant elles sont toutes positives, y compris sur les bords. Il est important de prendre une bordure $b \geq 2$ afin que la première et la dernière post-strate ne soit pas vide.

4 Une classe d'estimateurs sans biais

Comme $E\left(I_k|E_\ell\right) > 0$, on peut construire un estimateur qui est sans biais et même conditionnellement sans biais par rapport à E_ℓ. En notant $y_1, ..., y_j, ..., y_n$ les n valeurs prises par les unités dans l'échantillon ordonnées selon les R_k, on obtient

$$\begin{aligned} \widehat{Y}_\ell &= \sum_{k \in S} \frac{Y_k}{E(I_k|E_\ell)} \\ &= \frac{r_\ell - 1}{\ell - 1} \sum_{j=1}^{\ell-1} y_j + y_\ell \\ &\quad + \sum_{h=1}^{H} \left(\frac{r_{\ell+hq} - r_{\ell+(h-1)q} - 1}{q-1} \sum_{j=1}^{q-1} y_{\ell+(h-1)q+j} + y_{\ell+hq} \right) \\ &\quad + \frac{N - r_{\ell+Hq}}{n - (\ell + Hq)} \sum_{j=\ell+Hq+1}^{n} y_j \\ &= N_{0|\ell} \hat{\bar{y}}_{0|\ell} + y_\ell + \sum_{h=1}^{H} \left(N_{h|\ell} \hat{\bar{y}}_{h|\ell} + y_{\ell+hq} \right) + N_{H+1|\ell} \hat{\bar{y}}_{H+1|\ell} \end{aligned}$$

où

$$N_{0|\ell} = r_\ell - 1,$$
$$N_{h|\ell} = r_{\ell+hq} - r_{\ell+(h-1)q} - 1, h = 1, ..., H,$$
$$N_{H+1|\ell} = N - r_{\ell+Hq},$$
$$\hat{\bar{y}}_{0|\ell} = \frac{1}{\ell - 1} \sum_{j=1}^{\ell-1} y_j,$$

$$\hat{\bar{y}}_{h|\ell} = \frac{1}{q-1} \sum_{j=1}^{q-1} y_{\ell+(h-1)q+j}, h = 1, ..., H,$$

et

$$\hat{\bar{y}}_{H+1|\ell} = \frac{1}{n-(\ell+Hq)} \sum_{j=\ell+Hq+1}^{n} y_j.$$

Cet estimateur est en réalité un estimateur post-stratifié où les tailles des post-strates sont fixées dans l'échantillon. Comme $E(I_k|E_\ell) > 0$, $\widehat{Y}_\ell$ est strictement sans biais non-conditionnellement et conditionnellement à E_ℓ, ce qui n'est évidemment pas le cas de l'estimateur post-stratifié classique, car ce dernier a une probabilité non-nulle d'avoir une post-strate vide. En fixant la taille des post-strates dans l'échantillon, la constitution de post-strates vides devient impossible. La taille correspondante de la post-strate dans la population est une variable aléatoire $N_{h|\ell}$.

L'estimateur $\widehat{Y}_\ell$ possède une autre propriété intéressante. En utilisant la définition des $E(I_k|E_\ell)$, on montre assez facilement que

$$\sum_{k \in S} \frac{1}{E(I_k|E_\ell)} = N.$$

L'estimateur est donc calé sur la taille de la population. Cette propriété, que possède également l'estimateur de Horvitz-Thompson dans les plans simples, n'est donc pas perdue. Les unités dont les rangs sont dans E_ℓ sont appelées unités pivots, et sont affectées d'un poids égal à 1, ce qui rend les poids très inégaux. On peut donc reprocher à $\widehat{Y}_\ell$ d'utiliser des poids fortement dispersés. Ce problème peut être résolu en réalisant un lissage des estimateurs.

5 Lissage des estimateurs

Pour résoudre le problème de la dispersion des poids, on réalise une moyenne mobile d'estimateurs de la manière suivante :

$$\widehat{Y}_c = \frac{1}{q} \sum_{\ell=b}^{b+q-1} \widehat{Y}_\ell.$$

$\widehat{Y}_c$ garde toutes les propriétés des $\widehat{Y}_\ell$. Il est donc sans biais, calé sur N et linéaire, il peut donc s'écrire sous la forme

$$\widehat{Y}_c = \sum_{j=1}^{n} w_j y_j,$$

où

$$w_j = \begin{cases} \frac{1}{q}\sum_{\ell=b}^{b+q-1} \frac{r_\ell - 1}{\ell - 1} & j < b \\ \frac{1}{q}\left(\sum_{\ell=b}^{b+q-1} \frac{r_{j+\ell-b} - r_{m^-(j+\ell-b-q)} - 1}{j+\ell-b-m^-(j+\ell-b-q)-1} + 1\right) & b \le j < b+q-1 \\ \frac{1}{q}\left(\sum_{\ell=b}^{b+q-1} \frac{r_{j+\ell-b} - r_{j+\ell-b-q} - 1}{q-1} + 1\right) & b+q-1 \le j \le n-b+2-q \\ \frac{1}{q}\left(\sum_{\ell=b}^{b+q-1} \frac{r_{m^+(j+\ell-b)} - r_{j+\ell-b-q} - 1}{m^+(j+\ell-b)-j+\ell-b-q-1} + 1\right) & n-b+2-q < j \le n-b+1 \\ \frac{1}{q}\sum_{\ell=b}^{b+q-1} \frac{N+1-r_{n+1-\ell}-1}{n+1-(n+1-\ell)-1} = \frac{1}{q}\sum_{\ell=b}^{b+q-1} \frac{N-r_{n+1-\ell}}{\ell-1} & n-b+1 < j, \end{cases} \tag{2}$$

$$m^-(x) = \begin{cases} 0 & \text{si } x < b \\ x & \text{sinon} \end{cases}, m^+(x) = \begin{cases} n+1 & \text{si } x > n-b+1 \\ x & \text{sinon} \end{cases}$$

$r_0 = 0$, et $r_{n+1} = N + 1$.

Sous l'apparente complexité due au traitement particulier des bords, le système de pondération est relativement simple. Dans le cas où l'on ne se trouve pas trop près des bords, il vaut alors

$$w_j = \frac{1}{q}\left(\sum_{\ell=b}^{b+q-1} \frac{r_{j+\ell-b} - r_{j+\ell-b-q} - 1}{q-1} + 1\right) = \frac{1}{q(q-1)}\sum_{\alpha=0}^{q-1}\left(r_{j+\alpha} - r_{j+\alpha-q}\right).$$

Si toutes les valeurs de la variable auxiliaire ne sont pas distinctes, on peut affecter les rangs unités ayant des valeurs communes au hasard. Par exemple, si on a $X_1 = 2, X_2 = 5, X_3 = 5, X_4 = 7, X_5 = 8$, on choisit avec une probabilité $1/2$, soit les rangs $R_1 = 1, R_2 = 2, R_3 = 3, R_4 = 4, R_5 = 5$, soit $R_1 = 1, R_2 = 3, R_3 = 2, R_4 = 4, R_5 = 5$. On calcule alors l'estimateur lissé pour chaque permutation, et on fait leur moyenne. Cette méthode a l'avantage de conserver un estimateur sans biais. En effet, pour chacune des permutations possibles, l'estimateur est sans biais. En pratique, il est n'est pas nécessaire de calculer les estimateurs pour toutes les permutations. On peut calculer l'estimateur d'une permutation, et ensuite, on égalise simplement les poids des unités ayant les mêmes valeurs pour la variable x.

6 Cas où $q = 2$, $b = 2$

Quand $q = 2$, et $b = 2$, on obtient après quelques calculs

$$\begin{aligned} \widehat{Y}_c &= \frac{1}{2}\left\{\sum_{j=3}^{n-2} y_j\left(r_{j+1} - r_{j-1}\right)\right. \\ &\quad + \frac{r_3 + 2r_2 - 3}{2} y_1 + \frac{r_3 + 1}{2} y_2 \\ &\quad \left. + \frac{r_{n+1} - r_{n-2} + 1}{2} y_{n-1} + \frac{3r_{n+1} - 2r_{n-1} - r_{n-2} - 3}{2} y_n\right\} \\ &= \frac{1}{2}\left\{\sum_{j=1}^{n} y_j\left(r_{j+1} - r_{j-1}\right)\right. \end{aligned}$$

$$+y_1\frac{r_3-3}{2}+y_2\frac{2r_1+1-r_3}{2}$$
$$\left.+y_{n-1}\,\frac{r_{n+1}+r_{n-2}+1-2r_n}{2}+y_n\frac{r_{n+1}-r_{n-2}-3}{2}\right\},$$

où $r_0=0$ et $r_{n+1}=N+1$. On retrouve un estimateur proposé par Ren (2000, p. 140) et obtenu avec un argument de calage. Seule la gestion des bords est légèrement différente.

Exemple 1 *Soit une population de taille $N=20$. Supposons que les valeurs de la variable d'intérêt se trouvent dans la Table 1. On suppose, en outre que*

TAB. 1 – Exemple d'une population de taille $N=20$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_k	9	71	72	35	91	14	3	36	64	38	81	52	78	62	86	16	20	59	84	55
R_k	2	14	15	6	20	3	1	7	13	8	17	9	16	12	19	4	5	11	18	10

l'échantillon de taille $n=7$ est composé des unités de rangs $\{3,7,8,11,12,15,17\}$. Si on prend $q=2$, $\ell=2$, $b=2$ on obtient $E_2=\{r_2,r_4,r_6\}=\{7,11,15\}$. Ensuite on peut calculer $E(I_k|E_2=\{7,11,15\})$. Les probabilités d'inclusion conditionnelles sont les suivantes :

$$E(I_3|E_2=\{7,11,15\})=1/6,$$
$$E(I_7|E_2=\{7,11,15\})=1,$$
$$E(I_8|E_2=\{7,11,15\})=1/3,$$
$$E(I_{11}|E_2=\{7,11,15\})=1,$$
$$E(I_{12}|E_2=\{7,11,15\})=1/3,$$
$$E(I_{15}|E_2=\{7,11,15\})=1,$$
$$E(I_{17}|E_2=\{7,11,15\})=1/5.$$

L'estimateur

$$\widehat{Y}_0=\sum\frac{y_k}{E(I_k|E_2=\{7,11,15\})}$$

est donc sans biais et conditionnellement sans biais. De plus, il est linéaire et

$$\sum_{k\in S}\frac{1}{E(I_k|E_2=\{7,11,15\})}=N.$$

D'autre part, si on prend $q=2$, $\ell=3$, $b=2$, on obtient $E_3=\{r_3,r_5\}=\{8,12\}$. Alors, avec le même exemple, on calcule $E(I_k|E_3=\{8,12\})$, et on obtient

$$E(I_3|E_3=\{8,12\})=2/7,$$

$$E(I_7|E_3 = \{8, 12\}) = 2/7,$$
$$E(I_8|E_3 = \{8, 12\}) = 1,$$
$$E(I_{11}|E_3 = \{8, 12\}) = 1/3,$$
$$E(I_{12}|E_3 = \{8, 12\}) = 1,$$
$$E(I_{15}|E_3 = \{8, 12\}) = 2/8 = 1/4,$$
$$E(I_{17}|E_3 = \{8, 12\}) = 2/8 = 1/4.$$

L'estimateur

$$\widehat{Y}_1 = \sum \frac{y_k}{E(I_k|E_3 = \{8, 12\})}$$

est également sans biais et linéaire.

Enfin, on calcule la moyenne des deux estimateurs :

$$\widehat{Y}_c = \frac{\widehat{Y}_0 + \widehat{Y}_1}{2}.$$

Les poids s'obtiennent simplement en faisant la moyenne des poids des estimateurs $\widehat{Y}_0$ *et* $\widehat{Y}_1$*, et valent*

$$w_3 = (6 + 7/2)/2 = 19/4,$$
$$w_7 = (1 + 7/2)/2 = 9/4,$$
$$w_8 = (3 + 1)/2 = 2,$$
$$w_{11} = (1 + 3)/2 = 2,$$
$$w_{12} = (3 + 1)/2 = 2,$$
$$w_{15} = (1 + 4)/2 = 5/2,$$
$$w_{17} = (5 + 4)/2 = 9/2.$$

L'estimateur $\widehat{Y}_c$ *est linéaire et strictement sans biais.*

7 Application à l'estimation de la répartition

Il existe de multiples méthodes permettant d'utiliser de manière appropriée de l'information auxiliaire pour estimer une fonction de répartition. On peut trouver une exposé de ces techniques dans Ren (2000) et dans Wu et Sitter (2001). La méthode que nous proposons permet également d'estimer la répartition. La répartition dans la population est définie par

$$F_1(y) = \frac{1}{N} \sum_{k \in U} I\{y_k \leq y\},$$

et peut être estimée par

$$\widehat{F}_1(y) = \frac{\sum_{k\in S} w_k I\{y_k \leq y\}}{\sum_{k\in S} w_k},$$

où $I\{y \leq y_k\}$ est la fonction indicatrice, et les w_k sont les poids associés aux unités k qui valent $1/\pi_k = N/n$ pour l'estimateur de Horvitz-Thompson, et qui sont donnés en (2) pour l'estimateur calé.

Notons que les deux fonctions sont discrètes, mais que les sauts sont beaucoup moins nombreux sur S que sur U. Pour atténuer les différences des répartitions entre l'échantillon et la population, nous avons lissé les fonctions de répartition, en utilisant comme Deville (1995) une interpolation linéaire des centres des contremarches, ce qui consiste à définir $F_2(y)$ en reliant les points

$$\frac{1}{2}\{F_1(y_k) - F_1(y_k - \epsilon)\},$$

pour $k \in U$, où ϵ est un réel strictement positif arbitrairement petit. Ensuite, on définit $\widehat{F}_2(y)$ en reliant les points

$$\frac{1}{2}\left\{\widehat{F}_1(y_k) - \widehat{F}_1(y_k - \epsilon)\right\},$$

pour l'échantillon.

Exemple 2 *Une population de taille $N = 1000$ a été générée au moyen de variables logarithmico-normales indépendantes et équidistribuées. Un échantillon de taille $n = 16$ a ensuite été sélectionné et on a fixé $h = 5$. La figure 1 donne $F_2(x)$ dans la population.*

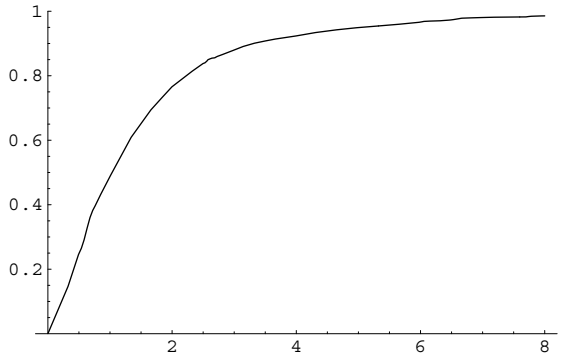

FIG. 1 – Fonction de répartition dans la population

Ensuite, la figure 2 montre $F_2(x)$ et la répartition estimée par l'estimateur de Horvitz-Thompson. Enfin, la figure 3 montre $F_2(x)$ et la répartition estimée par l'estimateur calé.

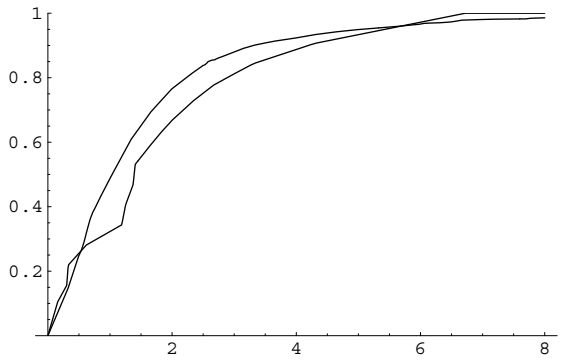

FIG. 2 – Fonction de répartition dans la population et estimateur de Horvitz-Thompson de la répartition

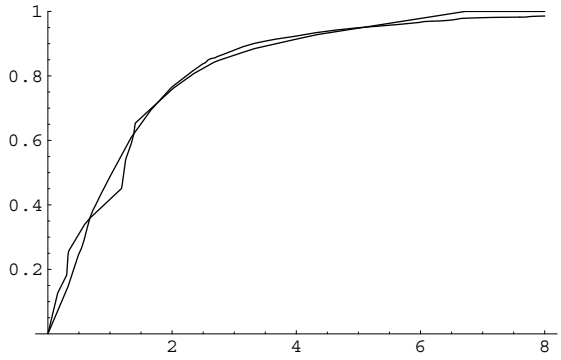

FIG. 3 – Fonction de répartition dans la population et estimateur calé de la répartition

8 Comparaison à l'estimateur par la régression

Afin de comparer les qualités de l'estimateur proposé, un ensemble de simulations a été réalisé pour comparer l'estimateur calé sur la répartition avec l'estimateur de Horvitz-Thompson et l'estimateur par la régression. Trois populations de taille 1000 on été générées selon les modèles suivants.

- *Modèle A Dépendance linéaire :* La population est générée selon le modèle $X_k \sim N(0,1)$ et $Y_k = X_k + 1.33333 \times \epsilon_k$ où $\epsilon_k \sim N(0,1)$. Le coefficient de corrélation obtenu dans la population est $\rho = 0.616154$.
- *Modèle B Dépendance non linéaire 1 :* La population est générée selon le modèle $X_k \sim N(0,1)$ et $Y_k = (0.2 + X_k)^2 + 1.33333 \times \epsilon_k$ où $\epsilon_k \sim N(0,1)$. Le coefficient de corrélation obtenu dans la population est $\rho = 0.28975$.
- *Modèle C Dépendance non linéaire 2 :* La population est générée selon le modèle $X_k \sim N(0,1)$ et $Y_k = (0.4 + X_k)^2 + 1.33333 \times \epsilon_k$ où $\epsilon_k \sim N(0,1)$.

Le coefficient de corrélation obtenu dans la population est $\rho = 0.476158$.

Dans chaque population 100 000 échantillons de taille 100 ont été sélectionnés. Pour chaque échantillon trois systèmes de pondération ont été calculés.

1. les poids associés au plan simple $w_k = N/n$,
2. les poids de l'estimateur calé sur la répartition donné en (2) en prenant la fenêtre $q = 10$ et la bordure $b = 6$,
3. les poids de l'estimateur par la régression donnés par

$$w_k = \frac{N}{n} + (X - \widehat{X}_{HT}) \frac{(X_k - \widehat{\overline{X}})}{\sum_{k \in S} (X_k - \widehat{\overline{X}})^2},$$

où X est le total des X_k sur la population, $\widehat{X}_{HT}$ est l'estimateur de Horvitz-Thompson de X, et $\widehat{\overline{X}} = \widehat{X}_{HT}/N$.

Au moyen de ces poids, l'estimateur de la moyenne et des neuf déciles ont été calculés pour chaque échantillon. Ensuite, on estime la variance de ces estimateurs au moyen des simulations.

TAB. 2 – Modèle A : variance des estimateurs (référence : Horvitz-Thompson=1)

Paramètre	Calage répartition	Estim. Régression
moyenne	0.674422	0.632608
1er décile	0.905273	0.893876
2ème décile	0.815403	0.802082
3ème décile	0.842681	0.815071
4ème décile	0.806749	0.768283
5ème décile	0.783731	0.740765
6ème décile	0.818051	0.782549
7ème décile	0.794411	0.773794
8ème décile	0.857114	0.844874
9ème décile	0.884424	0.884032

Les résultats sont donnés dans les tableaux 2, 3 et 4. Les variances ont été ramenées à 1 pour le plan simple. Pour le modèle linéaire, l'estimateur par la régression est légèrement préférable. Cependant, pour le cas non-linéaire, le gain de précision de l'estimateur calé sur la répartition est très important sur la moyenne comme sur les quantiles. L'estimateur que nous proposons se comporte donc de manière robuste en cas de relation non-linéaire entre la variable auxiliaire et la variable d'intérêt.

TAB. 3 – Modèle B : variance des estimateurs (référence : Horvitz-Thompson=1)

Paramètre	Calage répartition	Estim. Régression
moyenne	0.429689	0.953025
1er décile	0.913598	0.958656
2ème décile	0.919394	1.009270
3ème décile	0.829860	0.987950
4ème décile	0.792094	0.989114
5ème décile	0.703908	0.992023
6ème décile	0.622705	1.009830
7ème décile	0.550028	0.981249
8ème décile	0.443828	1.010340
9ème décile	0.549615	1.029120

TAB. 4 – Modèle C : variance des estimateurs (référence : Horvitz-Thompson=1)

Paramètre	Calage répartition	Estim. Régression
moyenne	0.30768	0.808114
1er décile	0.95560	0.983582
2ème décile	0.85920	0.970913
3ème décile	0.73854	0.930401
4ème décile	0.65728	0.950651
5ème décile	0.60500	0.956807
6ème décile	0.52139	0.930514
7ème décile	0.45709	0.907537
8ème décile	0.40752	0.903593
9ème décile	0.39820	0.860050

9 Variance et estimation de variance

Pour calculer la variance de $\widehat{Y}_c$, on commence par calculer la variance de $\widehat{Y}_\ell$. Comme $\widehat{Y}_\ell$, est sans biais conditionnellement à E_ℓ, on a

$$V\left(\widehat{Y}_\ell\right) = EV\left(\widehat{Y}_\ell | E_\ell\right).$$

Comme dans chacune des post-strates, conditionnellement à E_ℓ le plan est simple sans remise de taille fixe, on a

$$\begin{aligned} V\left(\widehat{Y}_\ell | E_\ell\right) &= \sum_{h=0}^{H+1} N_{h|\ell}^2 V\left(\hat{\bar{y}}_{h|\ell}\right) \\ &= \sum_{h=0}^{H+1} N_{h|\ell}^2 \frac{N_{h|\ell} - n_{h|\ell}}{N_{h|\ell}} \frac{S_{h|\ell}^2}{n_{h|\ell}}, \end{aligned} \tag{3}$$

où

$$n_{0|\ell} = \ell - 1,$$

$$n_{h|\ell} = q - 1, h = 1, ..., H,$$

$$n_{H+1|\ell} = n - (\ell + Hq),$$

$$\bar{Y}_{0|\ell} = \frac{1}{N_{0|\ell}} \sum_{k=1}^{r_{\ell-1}} Y_{(k)},$$

$$\bar{Y}_{h|\ell} = \frac{1}{N_{h|\ell}} \sum_{k=r_{\ell+(h-1)q}+1}^{r_{\ell+hq}-1} Y_{(k)}, h = 1, ..., H,$$

$$\bar{Y}_{H+1|\ell} = \frac{1}{N_{H|\ell}} \sum_{k=N-r_{\ell+Hq}+1}^{N} Y_{(k)},$$

$$S^2_{0|\ell} = \frac{1}{N_{0|\ell} - 1} \sum_{k=1}^{r_{\ell-1}} \left(Y_{(k)} - \bar{Y}_{0|\ell}\right)^2,$$

$$S^2_{h|\ell} = \frac{1}{N_{h|\ell} - 1} \sum_{k=r_{\ell+(h-1)q}+1}^{r_{\ell+hq}-1} \left(Y_{(k)} - \bar{Y}_{h|\ell}\right)^2, , h = 1, ..., H,$$

et

$$S^2_{H+1|\ell} = \frac{1}{N_{H+1|\ell} - 1} \sum_{k=N-r_{\ell+Hq}+1}^{N} \left(Y_{(k)} - \bar{Y}_{H+1|0}\right)^2,$$

où les $Y_{(k)}$ représentent les valeurs de Y_k triées selon l'ordre croissant des X_k.

On remarque qu'il est très difficile de calculer la variance non-conditionnelle de $\widehat{Y}_\ell$, c'est-à-dire l'espérance de $V\left(\widehat{Y}_\ell | E_\ell\right)$. En effet, $N_{h|\ell}$ et $S^2_{h|\ell}$ sont aléatoires. Cependant, on peut estimer simplement $V\left(\widehat{Y}_\ell | E_\ell\right)$ et obtenir un estimateur sans biais de la variance conditionnelle (et donc de la variance) en estimant simplement (3), par

$$\widehat{V}\left(\widehat{Y}_\ell | E_\ell\right) = \sum_{h=0}^{H+1} N^2_{h|\ell} \frac{N_{h|\ell} - n_{h|\ell}}{N_{h|\ell} n_{h|\ell}} s^2_{h|\ell}, \quad (4)$$

où

$$s^2_{0|\ell} = \frac{1}{n_{0|\ell} - 1} \sum_{j=1}^{\ell-1} \left(y_j - \hat{\bar{y}}_{0|\ell}\right)^2,$$

$$s^2_{h|\ell} = \frac{1}{n_{h|\ell} - 1} \sum_{j=1}^{q-1} \left(y_{\ell+(h-1)q+j} - \hat{\bar{y}}_{h|\ell}\right)^2, h = 1, ..., H,$$

et

$$s^2_{H+1|\ell} = \frac{1}{n_{h|\ell} - 1} \sum_{j=\ell+Hq+1}^{n} \left(y_j - \hat{\hat{y}}_{H+1|\ell} \right)^2 .$$

L'estimateur $\widehat{V}\left(\widehat{Y}_\ell | E_\ell\right)$ est non seulement sans biais pour $V\left(\widehat{Y}_\ell | E_\ell\right)$ mais aussi pour $V\left(\widehat{Y}_\ell\right)$.

10 Approximations pour le calcul de la variance

Malheureusement, le calcul de la variance de $\widehat{Y}_c$ devient plus complexe à cause des covariances. En effet, on a

$$V\left(\widehat{Y}_c\right) = \frac{1}{q^2} \sum_{\ell=b}^{b+q-1} \sum_{i=b}^{b+q-1} Cov\left(\widehat{Y}_\ell, \widehat{Y}_i\right).$$

Quand $\ell = i$, le problème consiste à estimer $V\left(\widehat{Y}_i\right)$, ce qui se fait sans difficulté. Quand $\ell \neq i$, il faut calculer

$$Cov\left(\widehat{Y}_\ell, \widehat{Y}_i\right) = ECov\left(\widehat{Y}_\ell, \widehat{Y}_i | E_\ell\right) + Cov\left(E(\widehat{Y}_\ell | E_\ell), E(\widehat{Y}_i | E_\ell)\right).$$

Comme $E(\widehat{Y}_\ell | E_\ell) = Y$, on obtient

$$Cov\left(\widehat{Y}_\ell, \widehat{Y}_i\right) = ECov\left(\widehat{Y}_\ell, \widehat{Y}_i | E_\ell\right) = EE\left(\widehat{Y}_\ell \widehat{Y}_i | E_\ell\right) - Y^2.$$

Le calcul $E\left(\widehat{Y}_\ell, \widehat{Y}_i | E_\ell\right)$ semble malheureusement inextricable, il faut donc chercher une approximation.

Une première voie consiste à chercher un majorant de la variance, comme

$$Cov\left(\widehat{Y}_\ell, \widehat{Y}_i\right) \leq \sqrt{V\left(\widehat{Y}_\ell\right) V\left(\widehat{Y}_i\right)},$$

on a un majorant donné par

$$V\left(\widehat{Y}_c\right) \leq \frac{1}{q^2} \sum_{\ell=b}^{b+q-1} \sum_{i=b}^{b+q-1} \sqrt{V\left(\widehat{Y}_\ell\right) V\left(\widehat{Y}_i\right)} = \frac{1}{q^2} \left(\sum_{\ell=b}^{b+q-1} \sqrt{V\left(\widehat{Y}_\ell\right)} \right)^2,$$

ce qui peut être estimé par

$$\widehat{V}_1\left(\widehat{Y}_c\right) = \frac{1}{q^2} \left(\sum_{\ell=b}^{b+q-1} \sqrt{\widehat{V}\left(\widehat{Y}_\ell | E_\ell\right)} \right)^2,$$

ce qui revient à estimer l'écart-type des moyennes par la moyenne des écart-types.

Enfin, une seconde voie peut-être donnée par une technique de résidus. En effet, de manière générale, quand on redresse un estimateur au moyen d'une technique de calage, on estime la variance au moyen d'une technique de résidus (voir à ce sujet Deville et Särndal, 1992, et Deville 1999). Dans le cas du calcul de la variance de $\widehat{Y}_\ell$, on peut utiliser une technique de résidus pour obtenir la variance exacte. En effet, considérons la variable

$$v_k(\ell) = \begin{cases} \left(\frac{N^2(N-n)}{Nn(N-1)}\right)^{-1/2} \left(\frac{N_{h|\ell}^2(N_{h|\ell}-n_{h|\ell})}{N_{h|\ell}n_{h|\ell}(N_{h|\ell}-1)}\right)^{1/2} \left(Y_k - \bar{Y}_{h|\ell}\right) \\ \qquad\qquad \text{si } k = r_{\ell+(h-1)q+1}, ..., r_{\ell+hq-1} \\ 0 \text{ si } k = r_{\ell+(h-1)q} \text{ ou } k = r_{\ell+hq} \end{cases}$$

qui peut apparaître comme un résidu associé à l'estimateur $\widehat{Y}_\ell$. La variable $v_k(\ell)$ injectée dans l'expression classique du plan simple sans remise de taille fixe est exactement égale à la variance conditionnelle $\widehat{Y}_\ell$ donnée en (3). En effet,

$$N^2 \frac{N-n}{nN} \frac{1}{N-1} \sum_{k \in U} \left(v_k - \frac{\sum_{k \in U} v_k}{N}\right)^2 = V\left(\widehat{Y}_\ell | E_\ell\right).$$

Cette variable dépend cependant des $\bar{Y}_{h|\ell}$ qui sont inconnus. On peut estimer $v_k(\ell)$ par

$$\hat{v}_j(\ell) = \begin{cases} \left(\frac{N^2(N-n)}{Nn(n-1)}\right)^{-1/2} \left(\frac{N_{h|\ell}^2(N_{h|\ell}-n_{h|\ell})}{N_{h|\ell}n_{h|\ell}(n_{h|\ell}-1)}\right)^{1/2} \left(y_j - \widehat{\bar{y}}_{h|\ell}\right) \\ \qquad\qquad \text{si } j = \ell + (h-1)q + 1, ..., \ell + hq - 1 \\ 0 \text{ si } j = \ell + (h-1)q \text{ ou } j = \ell + hq \end{cases}$$

Si on injecte $\hat{v}_k(\ell)$ dans l'estimateur de la variance du plan simple sans remise, on obtient un estimateur sans biais de la variance conditionnelle, et donc de la variance.

$$N^2 \frac{N-n}{nN} \frac{1}{n-1} \sum_{j=1}^{n} \left(\hat{v}_j - \frac{\sum_{j=1}^{n} \hat{v}_j}{n}\right)^2 = \widehat{V}\left(\widehat{Y}_\ell | E_\ell\right).$$

Deville (1999) montre que la variance d'une somme d'estimateurs peut se calculer en faisant la somme des résidus associés à ces estimateurs, les résidus étant calculés par linéarisation. Pour estimer la variance de $\widehat{Y}_c$, on pourrait donc simplement prendre la moyenne des résidus $\hat{v}_k(\ell)$, ce qui s'écrit

$$\hat{v}_k = \frac{1}{q} \sum_{\ell=b}^{b+q-1} \hat{v}_k(\ell).$$

On pourrait ainsi estimer la variance par

$$\widehat{V}_2\left(\widehat{Y}_c\right) = \frac{N^2(N-n)}{nN} \frac{1}{n-1} \sum_{k \in S} \left(\hat{v}_k - \frac{\sum_{k \in S} \hat{v}_k}{n}\right)^2.$$

Ces deux estimateurs de variances sont à évaluer par des simulations.

11 Simulations pour les estimateurs de variance

TAB. 5 – Résultats des simulations

	Corrélation :	0.802			
q	$V_{si}(\widehat{Y}_c)$	$E_{si}\widehat{V}_1(\widehat{Y}_c)$	$E_{si}\widehat{V}_2(\widehat{Y}_c)$	$BR_{si}\widehat{V}_1(\widehat{Y}_c)$	$BR_{si}\widehat{V}_2(\widehat{Y}_c)$
4	0.045	0.065	0.054	0.444	0.200
5	0.045	0.066	0.057	0.467	0.267
6	0.056	0.076	0.070	0.357	0.250
7	0.058	0.079	0.059	0.362	0.017
8	0.063	0.088	0.087	0.397	0.381
	Corrélation :	0.481			
q	$V_{si}(\widehat{Y}_c)$	$E_{si}\widehat{V}_1(\widehat{Y}_c)$	$E_{si}\widehat{V}_2(\widehat{Y}_c)$	$BR_{si}\widehat{V}_1(\widehat{Y}_c)$	$BR_{si}\widehat{V}_2(\widehat{Y}_c)$
4	0.048	0.066	0.059	0.375	0.229
5	0.045	0.060	0.054	0.333	0.200
6	0.044	0.056	0.051	0.273	0.159
7	0.044	0.054	0.051	0.227	0.159
8	0.045	0.052	0.048	0.156	0.067
	Corrélation :	0.111			
q	$V_{si}(\widehat{Y}_c)$	$E_{si}\widehat{V}_1(\widehat{Y}_c)$	$E_{si}\widehat{V}_2(\widehat{Y}_c)$	$BR_{si}\widehat{V}_1(\widehat{Y}_c)$	$BR_{si}\widehat{V}_2(\widehat{Y}_c)$
4	0.281	0.471	0.363	0.676	0.292
5	0.297	0.420	0.356	0.414	0.199
6	0.279	0.363	0.316	0.301	0.133
7	0.267	0.342	0.324	0.281	0.213
8	0.282	0.327	0.281	0.160	−0.004

Les simulations présentées dans le tableau 5 sont basées sur des populations de taille $N = 100$, qui sont générées au moyen de variables aléatoires indépendantes normales. Pour chaque cas étudié, on donne la valeur de q, et le coefficient de corrélation entre la variable d'intérêt Y_k et le rang R_k de la variable auxiliaire X_k. La borne b est définie en prenant la partie entière de $q/2+1$. L'objectif étant de valider l'estimateur de variance, on tire 3000 échantillons de taille $n = 20$ pour chaque simulation, et on compare la variance estimée par les simulations de l'estimateur calé $V_{si}(\widehat{Y}_c)$ aux espérances sous les simulations des deux estimateurs de variance notées $E_{si}(\widehat{V}_\alpha(\widehat{Y}_c)), \alpha = 1, 2$. Les deux dernières colonnes des tableaux présentent le biais relatif défini par

$$BR_{si}\widehat{V}_\alpha(\widehat{Y}_c) = \frac{E_{si}\widehat{V}_\alpha(\widehat{Y}_c) - V_{si}(\widehat{Y}_c)}{V_{si}(\widehat{Y}_c)}, \alpha = 1, 2.$$

Les simulations montrent que les deux estimateurs proposés surestiment la variance. La surestimation semble diminuer quand q croît. L'estimateur $\widehat{V}_2(\widehat{Y}_c)$ a manifestement le plus petit biais. On préconisera donc l'utilisation de $\widehat{V}_2(\widehat{Y}_c)$.

12 Conclusions

L'estimateur que nous proposons est un des rares estimateurs qui soit à la fois sans biais, linéaire, qui utilise de l'information auxiliaire, et qui est calé sur la taille de la population. Il est paramétrable au moyen de la largeur de la fenêtre q. Ce nouvel estimateur est robuste par rapport à l'estimateur par la régression. Il peut prendre en compte une information auxiliaire ayant une relation non-linéaire avec la variable d'intérêt. La plupart des simulations semblent montrer que la largeur de la fenêtre n'a pas beaucoup d'impacts sur la précision d'un estimateur de moyenne. Cependant, il apparaît aussi qu'une petite largeur de fenêtre n'est pas pénalisante, même si il n'y a pas de dépendance entre la variable auxiliaire et la variable d'intérêt. Plus q est petit, plus le calage sera serré, et l'estimateur de la variance sera alors fortement pénalisé, car un degré de liberté est perdu dans chaque post-strate. Le choix de q doit donc tenir compte de ce problème.

Il existe beaucoup d'autres méthodes permettant d'utiliser l'information donnée par une fonction de répartition (voir Ren, 2000) pour améliorer un estimateur. Les résultats que nous avons présentés se réduisent aux plans simples, mais nous pensons qu'ils sont importants, au même titre que la post-stratification est importante comme cas particulier des techniques de calage. En effet, la post-stratification est un des rares exemples où l'on peut montrer avec exactitude que le calage correspond à une approche conditionnelle. De plus, notre approche peut être vue comme un calage sur une fonction de répartition fournissant un estimateur sans biais. Une bonne technique générale de calage sur la répartition devrait donc retrouver dans les plans simples la méthode que nous avons présentée.

Remerciements

Nous remercions Jean-Claude Deville et Anne-Catherine Favre, deux arbitres et un éditeur associé pour leurs commentaires constructifs qui ont permis d'améliorer considérablement cet article.

Références

Deville, J.-C. (1995), *Estimation de la variance du coefficient de Gini mesuré par sondage,* INSEE Méthode, Document de travail, Méthodologie F9510.

Deville, J.-C. (1999), Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation, *Techniques d'Enquête* **25,** 219-230.

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

Estevao, V., Hidiroglou, M.A. et Särndal, C.-E. (1995), Methodological principle for a generalized estimation system in Statistics Canada, *Journal of Official Statistics*, **11**, 181-204.

Ren, R. (2000), *Estimation par calage sur la répartition*, Thèse de Doctorat en préparation, Paris, Université Paris Dauphine.

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

Tillé, Y. (1998), Estimation in surveys using conditional inclusion probabilities : simple random sampling, *International Statistical Review*, **66**, 303-322.

Tillé, Y. (1999a), Sur la détermination a posteriori des bornes des post-strates, in G. Brossier, et A.-M. Dussaix, *Les sondages,* Dunod, pages 202-208.

Tillé, Y. (1999b), Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles : enquêtes à plan d'échantillonnage complexe, *Techniques d'enquêtes*, **25**, 57-66.

Wu, C., et Sitter, R.R. (2001), Variance estimation for the finite population distribution function with complete auxiliary information, *Canadian Journal of Statistics*, **29**, 289-307.

Table des matières