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

Article

Comparaison de méthodes de restriction de l'ensemble d'échantillons

par Jason C. Legg et Cindy L. Yu

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 75-87
Statistique Canada, N° 12-001-X au catalogue

Comparaison de méthodes de restriction de l'ensemble d'échantillons

Jason C. Legg et Cindy L. Yu [1]

Résumé

Dans le cas de nombreux plans de sondage, la probabilité de sélectionner un échantillon qui produira de mauvaises estimations pour des quantités connues n'est pas nulle. L'échantillonnage aléatoire stratifié permet de réduire l'ensemble de ces échantillons éventuels en fixant la taille de l'échantillon dans chaque strate. Cependant, l'obtention d'échantillons indésirables demeure possible après la stratification. L'échantillonnage réjectif permet d'éliminer les échantillons donnant de mauvais résultats en ne retenant un échantillon que si des fonctions spécifiées des estimations sont comprises entre des limites de tolérance par rapport aux valeurs connues. Les échantillons résultant sont souvent dits équilibrés sur la fonction des variables utilisées dans la méthode de rejet. Nous présentons des modifications de la méthode de rejet de Fuller (2009a) qui donnent plus de souplesse aux règles de rejet. Au moyen de simulations, nous comparons les propriétés des estimations obtenues en suivant une méthode d'échantillonnage réjectif, d'une part, et une procédure d'échantillonnage par la méthode du cube, d'autre part.

Mots clés : Échantillonnage réjectif ; échantillonnage par la méthode du cube ; stratification; échantillonnage équilibré.

1. Introduction

En échantillonnage, une pratique courante consiste à utiliser l'information de population connue au sujet des variables auxiliaires pour améliorer les estimateurs des moyennes et des totaux des caractéristiques d'intérêt. Lorsque l'on dispose de moyennes ou de totaux de population de contrôle pour une variable auxiliaire, on se sert souvent d'estimateurs par la régression et d'autres estimateurs par calage. Soit $(\boldsymbol{x}_i, y_i, p_i)$, $i = 1, 2, ..., N,$ une suite de vecteurs réels, où chaque $\boldsymbol{x}_i$ est un vecteur de dimension k, et un échantillon A, tiré de $F_N = [(\boldsymbol{x}_1, y_1, p_1), ..., (\boldsymbol{x}_N, y_N, p_N)]$ en utilisant un plan de sondage avec probabilités d'inclusion p_i et probabilités d'inclusion conjointe p_{ij}. Supposons que la moyenne de population de $\boldsymbol{x}_i$, $\overline{\boldsymbol{x}}_N$, est connue. Considérons l'estimateur par la régression de la moyenne de population de la forme

$$\overline{y}_{\text{reg}} = \overline{\boldsymbol{z}}_N' \hat{\boldsymbol{\beta}}, \tag{1}$$

où $\boldsymbol{z}_i$ contient les variables du plan de sondage et $\boldsymbol{x}_i$, $\overline{\boldsymbol{z}}_N$ est la moyenne de population de $\boldsymbol{z}_i$ et $\hat{\boldsymbol{\beta}}$ est un estimateur des coefficients de régression. Pour de nombreux plans de sondage, l'estimateur $\hat{\boldsymbol{\beta}}$ de la forme

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in A} \boldsymbol{z}_i \, \phi_i \, p_i^{-2} \, \boldsymbol{z}_i' \right)^{-1} \sum_{i \in A} \boldsymbol{z}_i \, \phi_i \, p_i^{-2} \, y_i, \tag{2}$$

où les valeurs de ϕ_i sont des constantes déterminées par le plan, sera asymptotiquement efficace. Certains exemples de choix de ϕ_i sont $\phi_i = (1 - p_i)$ sous échantillonnage de Poisson et $\phi_{hi} = (N_h - 1)^{-1}(N_h - n_h)$ pour l'élément i dans la strate h sous échantillonnage aléatoire stratifié. Si nous supposons qu'il existe un vecteur $\boldsymbol{d}$ tel que

$$\phi_i \, p_i^{-2} \, \boldsymbol{z}_i' \boldsymbol{d} = p_i^{-1} \tag{3}$$

pour tout i, l'estimateur (1) est convergent sous le plan (Fuller 2002). L'estimateur des coefficients de régression (2) converge avec

$$\boldsymbol{\beta}_N = \left(\sum_{i=1}^{N} \boldsymbol{z}_i \, \phi_i \, p_i^{-1} \, \boldsymbol{z}_i' \right)^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i \, \phi_i \, p_i^{-1} \, y_i.$$

Pour illustrer l'application de l'équation (3), supposons que nous prévoyons sélectionner un échantillon de Poisson et que nous voulons calculer la régression sur une seule covariable x_{1i} passant par l'origine. Si nous ajoutons $(1 - p_i)^{-1} p_i$ dans $\boldsymbol{z}_i$ pour faire $\boldsymbol{z}_i' = (x_{1i}, [1 - p_i]^{-1} p_i)$, l'estimateur (1) sera convergent sous le plan pour $\overline{y}_N$ puisque l'expression (3) est satisfaite en prenant $\boldsymbol{d}' = (0, 1)$. Si nous supposons en outre qu'une colonne de valeurs 1 se trouve dans l'espace colonnes des variables de régression $\boldsymbol{z}_i$, alors pour ces valeurs de ϕ_i, l'estimateur (1) atteint presque la variance asymptotique minimale pour des estimateurs par la régression convergents sous le plan sous certaines conditions de régularité (Rao 1994). Une alternative à la construction d'un estimateur par la régression consiste à partir d'un estimateur convergent sous le plan, tel que l'estimateur par la régression généralisée de Särndal (1980) et à déterminer le meilleur coefficient sachant cette forme de l'estimateur. En débutant avec une forme convergente sous le plan, il n'est plus nécessaire de satisfaire l'expression (3). La condition (3) permet d'exprimer l'estimateur (1) sous la

1. Cindy L. Yu est professeure adjointe au Département de statistique et au Center for Survey Statistics and Methodology à la Iowa State University, Ames, IA 50010. Courriel : cindyyu@iastate.edu ; Jason C. Legg est un chercheur postdoctoral au Center for Survey Statistics and Methodology à la Iowa State University, Ames, IA 50010. Courriel : jason-legg@hotmail.com.

forme d'un estimateur par la régression généralisée (Fuller 2009b, pages 116-117).

Si l'on dispose d'information auxiliaire au niveau de l'unité, on peut également l'intégrer dans le plan d'échantillonnage. Par exemple, dans un cas classique, on suppose que le modèle donné par

$$y_i = \beta_0 + \beta_1 x_i + x_i \varepsilon_i, \qquad (4)$$

$\varepsilon_i \sim \text{ind}(0, \sigma^2)$ et $\text{cov}(\varepsilon_i, x_i) = 0$ s'applique à la population F_N. Selon Isaki et Fuller (1982), les probabilités d'inclusion optimales pour l'estimateur par la régression sont celles qui sont proportionnelles à la racine carrée des variances sous le plan, c'est-à-dire $p_i \propto x_i$ dans le cas qui nous occupe. Une méthode d'échantillonnage possible est l'échantillonnage de Poisson avec les probabilités d'inclusion

$$p_i = \left(\sum_{i=1}^{N} x_i \right)^{-1} n_N x_i, \qquad (5)$$

où $n_N = \sum_{i=1}^{N} p_i$ est une taille d'échantillon cible spécifiée. Un deuxième plan d'échantillonnage fréquent, si l'on émet l'hypothèse du modèle (4), consiste à stratifier la population en se basant sur x. Les strates sont déterminées en fixant leurs limites de sorte que la somme des valeurs ordonnées de x_i créées dans chacune soit à peu près la même dans toutes les strates. Un nombre égal d'unités est sélectionné dans chaque strate. Ce plan de stratification, sous lequel les probabilités d'inclusion sont proches de (5), s'est avéré donner une variance anticipée proche de la meilleure variance sous le modèle d'échantillonnage par choix raisonné dans le cas de deux unités par strate (Fuller 1981).

Un autre moyen d'intégrer dans le plan de sondage l'information issue d'une variable auxiliaire est l'équilibrage. Un échantillon A est équilibré pour la variable $\boldsymbol{z}$ si

$$\overline{\boldsymbol{z}}_{\text{HT}} = N^{-1} \sum_{i \in A} p_i^{-1} \boldsymbol{z}_i = N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i = \overline{\boldsymbol{z}}_N. \qquad (6)$$

Un plan de sondage est équilibré pour $\boldsymbol{z}$ si chaque échantillon dont la probabilité est positive est équilibré pour $\boldsymbol{z}$. L'équilibrage peut être considéré comme un calage par conception. Pour illustrer l'effet de l'équilibrage, considérons un plan avec probabilités d'inclusion égales et $\boldsymbol{z}_i = (1, x_i)'$. La variance de prédiction conditionnelle de $\overline{y}_{\text{reg}}$ sous le modèle (4) est donnée par

$$V(\overline{y}_{\text{reg}} - \overline{y}_N \mid \boldsymbol{x}, \overline{x}_{\text{HT}}) = E\{V(\overline{u}_{\text{HT}} \mid F_N) \mid \boldsymbol{x}, \overline{x}_{\text{HT}}\} + (\overline{x}_N - \overline{x}_{\text{HT}})^2 V(\hat{\beta}_1 \mid \boldsymbol{x}, \overline{x}_{\text{HT}}), \qquad (7)$$

où $u_i = x_i \varepsilon_i$. Pour un plan équilibré, le deuxième terme de (7) est 0, ce qui laisse entendre que nous pourrions améliorer l'estimateur en équilibrant sur x. En pratique, une combinaison d'équilibrage et de calage donnera souvent de meilleurs résultats que l'une ou l'autre technique utilisée seule.

Les plans de sondage équilibrés ont une certaine valeur pratique supplémentaire. Dans le cas de nombreux plans, il existe une probabilité non nulle de sélectionner un échantillon contenant des valeurs indésirables pour les variables auxiliaires. Ainsi, un échantillon indésirable pourrait être un échantillon dont la répartition est insuffisante pour les domaines ou un échantillon présentant un grand nombre de valeurs extrêmes pour les variables auxiliaires. Même si les plans stratifiés réduisent l'ensemble d'échantillons éventuels de ce genre en fixant la taille d'échantillon dans chaque strate, l'obtention d'échantillons indésirables demeure possible. Par exemple, certains échantillons stratifiés pourraient être associés à certains poids négatifs à cause de l'utilisation d'estimateurs par la régression. L'équilibrage peut éliminer les échantillons donnant de mauvais résultats en retenant uniquement ceux qui produisent des estimations proches des quantités connues et ne possédant que des poids positifs pour les estimateurs par la régression.

L'échantillonnage équilibré a été proposé par Royall et Cumberland (1981) comme moyen de réduire le biais sous le modèle causé par la spécification incorrecte des modèles de superpopulation polynomiaux. Valliant, Dorfman et Royall (2000) discutent des incidences de l'équilibrage sous l'angle d'une approche prédictive de l'échantillonnage. Deville et Tillé (2004) ont étudié des méthodes de sélection d'échantillons équilibrés dans le cadre fondé sur le plan de sondage décrit plus haut. Le lecteur est invité à consulter également Tillé (2006, chapitre 8) pour un traitement détaillé de l'équilibrage. En pratique, il est parfois impossible de trouver un plan parfaitement équilibré. Un équilibrage très strict peut produire un plan présentant certaines probabilités d'inclusion conjointes extrêmes, y compris des probabilités d'inclusion nulles. Par conséquent, en pratique, on procède à un équilibrage partiel.

Dans le présent article, nous comparons, au moyen d'études par simulation, les propriétés des plans obtenus en appliquant deux méthodes d'équilibrage, à savoir l'échantillonnage réjectif de Fuller (2009a) et l'échantillonnage par la méthode du cube de Tillé (2006). Nous présentons aussi des modifications de la méthode d'échantillonnage réjectif de Fuller qui donnent plus de souplesse à l'équilibrage. À la section 2, nous décrivons l'échantillonnage réjectif et l'échantillonnage par la méthode du cube. À la section 3, nous comparons les propriétés des probabilités d'inclusion des deux méthodes d'équilibrage. À la section 4, nous présentons certains résultats de simulation obtenus en utilisant les échantillons équilibrés. À la section 5, nous donnons les corrections apportées à la méthode réjective. Enfin, à la section 6, nous présentons nos conclusions.

2. Méthodes d'échantillonnage équilibré

L'échantillonnage réjectif comprend le rejet de tout échantillon qui ne satisfait pas une tolérance d'équilibrage spécifiée. Fuller (2009a) donne une condition pour le rejet d'un échantillon, tandis que Royall et Herson (1973) en présentent une autre. Dans la méthode de Fuller avec le vecteur de variables d'équilibrage $\boldsymbol{z}$, un échantillon tiré sous un plan de sondage initial est retenu si

$$(\overline{\boldsymbol{z}}_{\mathrm{HT}} - \overline{\boldsymbol{z}}_N)' \, [V(\overline{\boldsymbol{z}}_{\mathrm{HT}} \mid F_N)]^{-1} \, (\overline{\boldsymbol{z}}_{\mathrm{HT}} - \overline{\boldsymbol{z}}_N) < \gamma \qquad (8)$$

pour une constante donnée $\gamma > 0$, où $\overline{\boldsymbol{z}}_{\mathrm{HT}}$ est l'estimateur de Horvitz-Thompson de la moyenne pour la variable $\boldsymbol{z}$, F_N est la population finie donnée,

$$V(\overline{\boldsymbol{z}}_{\mathrm{HT}} \mid F_N) = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} (p_{ij} - p_i \, p_j) \, \boldsymbol{z}_i \, \boldsymbol{z}_i' \, p_i^{-1} p_j^{-1},$$

p_i est la probabilité d'inclusion de l'unité i et p_{ij} est la probabilité d'inclusion conjointe de l'unité i et de l'unité j sous le plan initial. Autrement, l'échantillon est rejeté, un nouvel échantillon est tiré sous le plan de sondage initial et le respect de la condition (8) est vérifié pour le nouvel échantillon. Si le plan de sondage original obéit au théorème de la limite centrale, le premier membre de (8) est asymptotiquement une variable aléatoire χ^2 dont le nombre de degrés de liberté est égal au nombre de variables auxiliaires. Un taux de rejet approximatif peut être fixé en utilisant les quantiles d'une loi du χ^2 pour γ. Le choix du taux de rejet dépendra des objectifs particuliers de chaque enquête. Un faible taux de rejet pourrait ne pas réduire fortement la variance, mais donner à un chercheur un degré de certitude suffisant de ne pas sélectionner un échantillon très médiocre. Par ailleurs, un taux de rejet élevé pourrait réduire considérablement la variance, mais produire des tailles d'échantillon insuffisantes pour l'exécution d'analyses de domaine non planifiées. Par exemple, si un chercheur décide de procéder à une analyse de domaine sur la queue de la distribution d'une variable d'équilibrage, les probabilités d'inclusion conjointe pourraient être faibles, si bien que, pour de nombreux échantillons, le domaine ne contiendra que quelques unités.

La méthode du cube a été élaborée par Tillé et Deville, et décrite dans Tillé (2006). Elle a pour objectif de tirer un échantillon équilibré en se servant de probabilités d'inclusion de premier ordre prédéterminées. Si le vecteur des probabilités d'inclusion de premier ordre ne produit pas de plan équilibré, une étape supplémentaire consistant à minimiser une contrainte de coût est utilisée. Contrairement à la méthode réjective, les probabilités d'inclusion initiales d'ordre plus élevé ne sont pas spécifiées préalablement. L'étape de minimisation du coût assure le maintien des probabilités d'inclusion de premier ordre initiales spécifiées.

Pour faciliter la compréhension de la méthode du cube, Tillé (2006) décrit l'échantillonnage géométriquement. L'ensemble de tous les échantillons possibles est défini comme étant l'ensemble des vecteurs correspondant aux sommets d'un hypercube unité N-dimensionnel, ou N-cube. Par exemple, si $N = 3$, le sommet $(0, 1, 1)$ désigne un échantillon contenant la deuxième et la troisième unité. Un plan d'équilibrage est créé en utilisant l'équation d'équilibrage (6) et la probabilité d'inclusion souhaitée p_i pour $i = 1, ..., N$. Tout échantillon situé à l'intersection du plan d'équilibrage et d'un sommet de l'hypercube unité N-dimensionnel est équilibré. Le plan de sondage est équilibré si chaque point de l'intersection entre le plan d'équilibrage et l'hypercube unité est un sommet de ce cube. La procédure d'échantillonnage par la méthode du cube débute par la sélection d'un vecteur dans le plan d'équilibrage ; cette étape est suivie par une marche aléatoire du point initial jusqu'à une arête de l'hypercube unité. Tillé donne à l'étape de la marche aléatoire le nom de phase de vol. Si le point rencontré sur l'arête à la fin de la marche aléatoire est un sommet de l'hypercube unité, l'échantillon est sélectionné. Sinon, une méthode de minimisation des coûts est utilisée pour convertir les composantes fractionnaires du vecteur d'arêtes en nombres entiers. Les composantes entières du vecteur d'arêtes ne sont pas modifiées durant l'étape de la minimisation du coût. Tillé donne à l'étape de minimisation du coût le nom de phase d'atterrissage. L'échantillonnage réjectif avec taux de rejet élevé produit des résultats semblables à l'échantillonnage par la méthode du cube.

D'autres méthodes que l'échantillonnage réjectif et l'échantillonnage par la méthode du cube peuvent être utilisées pour obtenir des échantillons presque équilibrés. Par exemple, la stratification où les limites des strates sont déterminées par les variables $\boldsymbol{x}$ peut également avoir certains effets équilibrants sur les échantillons (Fuller 1981). Le processus suivi pour décider du nombre de variables qu'il convient d'utiliser dans les méthodes d'échantillonnage réjectif et d'échantillonnage par la méthode du cube est essentiellement le même que celui utilisé pour décider du nombre de variables à inclure dans un estimateur par la régression.

Un logiciel a été développé pour le tirage d'échantillons par la méthode du cube. Dans le cas de l'échantillonnage réjectif, des progiciels standard peuvent être utilisés pour tirer un échantillon et calculer (8). Une boucle doit être rédigée pour achever la procédure. Des programmes de tirage d'échantillons par la méthode du cube ont été écrits pour SAS et R. Voir Rousseau et Tardieu (2004) pour SAS, et Matei et Tillé (2005) pour R, ainsi que Deville et Tillé (2004), pour renseignements détaillés sur les procédures implémentées. Le programme R disponible dans la bibliothèque *sampling* a été utilisé pour les simulations décrites

dans le présent article. Comme l'étape de minimisation du coût de l'échantillonnage par la méthode du cube requiert d'importantes ressources informatiques si l'on traite plus de 20 variables d'équilibrage, nous recommandons d'ajouter une étape de suppression de variables à la phase d'atterrissage dans les programmes.

3. Probabilités d'inclusion

Soit π_i la probabilité d'inclusion de premier ordre de l'unité i et π_{ij}, la probabilité d'inclusion conjointe des unités i et j sous un plan de sondage équilibré. Les probabilités d'inclusion de premier ordre initiales sont des données d'entrée requises tant pour l'échantillonnage réjectif que pour l'échantillonnage par la méthode du cube. Dans le cas de l'échantillonnage réjectif, les probabilités d'inclusion de premier ordre diffèrent des valeurs initiales, les unités proches de la moyenne de population ayant, sous cette méthode, une probabilité d'inclusion légèrement plus élevée que les unités éloignées de la moyenne. Par contre, dans l'échantillonnage par la méthode du cube, les probabilités d'inclusion de premier ordre demeurent celles de la spécification initiale. Autrement dit, pour l'échantillonnage par la méthode du cube, $\pi_i = p_i$. Bien que, pour l'échantillonnage réjectif, $\pi_i \neq p_i$, en général, les estimateurs pris en considération utiliseront p_i plutôt que π_i.

Afin d'illustrer les différences entre les probabilités d'inclusion initiales et finales, nous avons simulé des échantillons de taille 20 tirés d'une population de 100 unités. La population de valeurs de x a été générée sous forme de variables aléatoires tirées d'une loi normale standard. La méthode de rejet s'appuyait sur l'échantillonnage aléatoire simple comme plan de sondage initial et était équilibrée sur x. Pour l'échantillonnage par la méthode du cube, nous avons utilisé un vecteur d'équilibrage de $\boldsymbol{z}_i = (p_i, x_i)'$, où $p_i = 20/100$ pour tout i. L'inclusion de p_i dans le vecteur d'équilibrage pour l'échantillonnage par la méthode du cube avait pour but de contrôler la taille d'échantillon afin que le plan de sondage résultant soit comparable à l'utilisation de l'échantillonnage aléatoire simple comme plan initial dans la simulation de l'échantillonnage réjectif. Nous avons estimé les probabilités d'inclusion de premier ordre en utilisant une simulation Monte Carlo de taille 100 000 (figure 1). La courbe a été obtenue par un ajustement non paramétrique. Pour l'échantillonnage réjectif, nous avons utilisé un taux de rejet d'environ 90 %. En vertu de la théorie de l'échantillonnage réjectif, les probabilités d'inclusion de premier ordre correspondent approximativement à une fonction quadratique de la distance $x_i - \bar{x}_N$ pour un plan de sondage initial à probabilités égales (Fuller 2009a). Le graphique donne à penser que toutes les probabilités d'inclusion de premier ordre sont égales à 0,2 pour le plan d'échantillonnage obtenu par la méthode du cube. Comme prévu, la figure 1 indique que la méthode du cube maintient les probabilités d'inclusion de premier ordre spécifiées, mais que l'échantillonnage réjectif ne le fait pas. Par conséquent, l'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion initiales (p_i) et les échantillons réjectifs est biaisé.

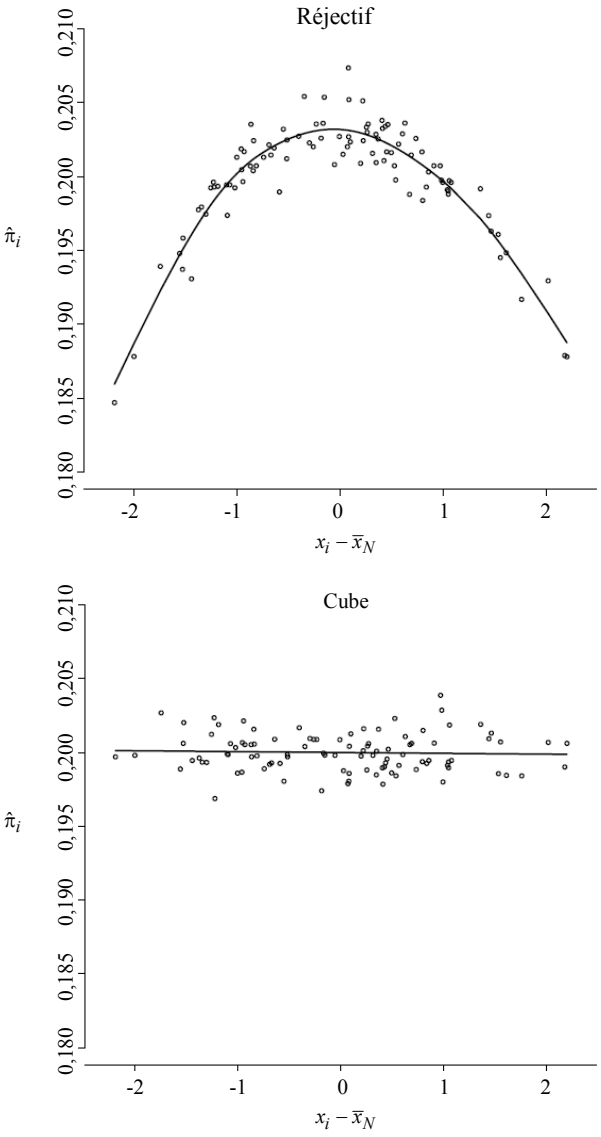

Figure 1 Probabilités d'inclusion de premier ordre simulées. Pour la méthode réjective, la variable d'équilibrage est $z_i = x_i$ et pour la méthode du cube, $z_i = (p_i, x_i)'$, où $p_i = 20/100$

Dans la procédure d'échantillonnage réjectif, les probabilités d'inclusion conjointe diffèrent de celles du plan de sondage initial. Une paire d'unités i et j sont peu susceptibles d'avoir une probabilité d'inclusion conjointe élevée si $x_i + x_j - 2\bar{x}_N$ est proche de zéro pour un plan de sondage initial à probabilités égales. Nous avons estimé les probabilités d'inclusion conjointe pour des échantillons simulés

de taille 20 tirés d'une population de 100 (figure 2). Pour l'échantillonnage aléatoire simple, la probabilité d'inclusion conjointe est 0,038. Sous échantillonnage réjectif, les probabilités d'inclusion conjointe sont approximativement égales à une fonction quadratique de $x_i + x_j$. Le tracé des probabilités d'inclusion conjointe sous échantillonnage par la méthode du cube en fonction de $x_i + x_j$ semble présenter des angles plus aigus que celui des probabilités d'inclusion conjointe sous échantillonnage réjectif. Les probabilités d'inclusion conjointe élevées observées pour la méthode du cube sont associées à des paires d'unités situées sur les côtés opposés éloignés de $\bar{x}_N$. Autrement dit, pour la valeur d'échantillon de $x_i + x_j$, les paires dont la valeur de $|x_i| + |x_j|$ est grande ont une grande probabilité d'inclusion (figure 3).

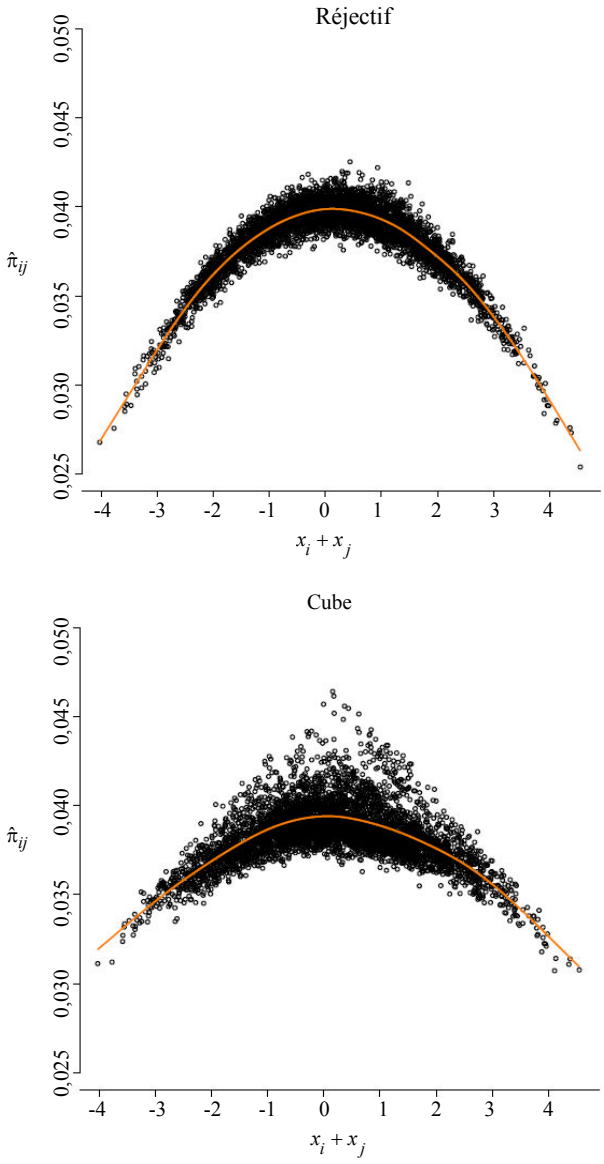

Figure 2 Probabilités d'inclusion de deuxième ordre simulées. Pour la méthode réjective, la variable d'équilibrage est $z_i = x_i$ et pour la méthode du cube, $z_i = (p_i, x_i)'$, où $p_i = 20/100$

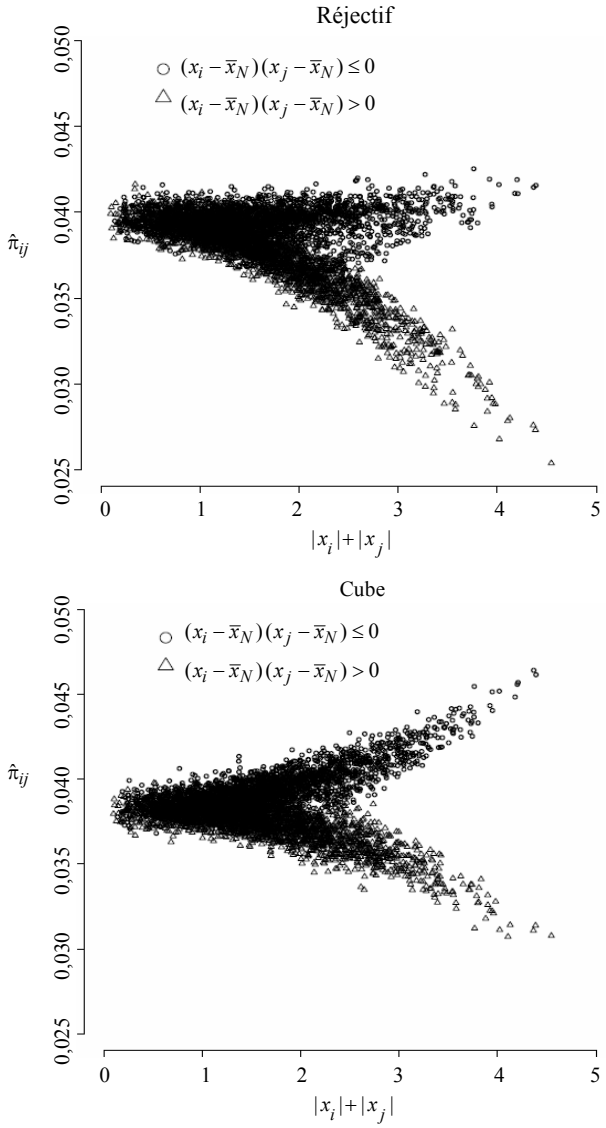

Figure 3 Probabilités d'inclusion de deuxième ordre simulées avec les sommes absolues de x. Pour la méthode réjective, la variable d'équilibrage est $z_i = x_i$ et pour la méthode du cube, $z_i = (p_i, x_i)'$, où $p_i = 20/100$

L'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion initiales présente un biais d'ordre $O_p(n^{-1})$ sous échantillonnage réjectif, tandis qu'il est sans biais sous échantillonnage par la méthode du cube. L'estimateur de variance de Horvitz-Thompson standard présente un biais sous les deux procédures. En utilisant des méthodes Monte Carlo, il est possible d'estimer des probabilités d'inclusion permettant d'utiliser des estimateurs de Horvitz-Thompson presque sans biais. Cependant, pour une grande population, il est difficile de simuler un nombre d'échantillons suffisant pour obtenir une estimation précise de la probabilité d'inclusion conjointe pour une paire d'unités. Au lieu d'estimer la variance, on peut utiliser un estimateur par la régression et l'estimateur de variance pour ce dernier. Cette approche est intuitivement séduisante,

parce que l'équilibrage est, par conception, semblable à la régression. Lorsque l'on utilise l'estimateur par la régression, le biais de ce dernier est du même ordre sous la méthode du cube que sous la méthode réjective. Pour l'échantillonnage réjectif, Fuller (2009a) donne les conditions de convergence de l'estimateur de variance de l'estimateur par la régression. Pour l'échantillonnage par la méthode du cube, Deville et Tillé (2005), ainsi que Tillé (2006) laissent entendre qu'en utilisant l'estimateur de variance pour un estimateur par la régression, on obtient une bonne approximation de la variance de l'estimateur de Horvitz-Thompson. Les estimateurs de variance proposés par Deville et Tillé (2005) donnent de bons résultats quand les probabilités d'inclusion conjointe du plan de sondage obtenu par la méthode du cube sont approximativement égales aux probabilités d'inclusion conjointe d'un plan de sondage de Poisson. Dans les études par simulation de la section 4, nous évaluons les estimateurs de variance proposés dans Fuller (2009a) et dans Deville et Tillé (2005).

4. Simulation de l'estimateur par la régression

Une population de taille 100 a été générée à partir du modèle

$$y_i = x_i + 0,55x_i^2 + x_i\,\varepsilon_i \qquad (9)$$

$\varepsilon_i \sim$ iid $N(0,\ 0,4)$, où les x_i sont des valeurs fixées dans l'intervalle de 0 à 4 (figure 4). Soixante-douze des valeurs de x étaient des valeurs inférieures à 1,15 simulées aléatoirement d'après une loi exponentielle standard. Les 28 autres valeurs, variant de 0,18 à 4,0, ont été ajoutées de manière déterministe pour former l'ensemble de données de x. Les valeurs fixes de x ont été sélectionnées de manière que leur distribution soit assez étalée vers la droite afin que des grandes et des petites strates soient produites lors de la stratification de la population sur x de façon que les sommes intra-strate des valeurs x_i triées soient approximativement égales. La population a été maintenue fixe après la sélection initiale. Nous avons choisi le modèle (9), qui contient un terme quadratique, pour simuler la performance de la stratégie de sondage et d'estimation sous l'hypothèse du modèle (4) pour le plan de sondage et l'estimation.

Nous prenons pour plans de sondage initiaux l'échantillonnage de Poisson et l'échantillonnage aléatoire stratifié avec deux unités par strate. Nous avons déterminé les strates en fixant les limites de manière que la somme intra-strate des valeurs triées de x_i soit à peu près la même dans toutes les strates. Nous avons fixé la taille d'échantillon à 20 et formé dix strates. Les tailles de strate étaient 35, 15, 11, 9, 8, 7, 5, 4, 3 et 3. Pour la méthode réjective, nous avons utilisé un échantillonnage stratifié de deux unités par strate avec probabilités d'inclusion égales dans la strate. Nous avons choisi les limites des strates de cette façon pour que la probabilité d'inclusion de l'unité i soit presque proportionnelle à x_i, ce qui est la probabilité d'inclusion optimale sous le modèle (9) (Ikasi et Fuller 1982). Ce plan de stratification peut aussi être équilibré partiellement sur x par la voie d'un plan de sondage standard. Dans le plan d'échantillonnage aléatoire stratifié, l'équilibre est atteint en utilisant une fonction escalier pour approximer une droite. Le plan stratifié sera également partiellement équilibré sur x^2. Le plan d'échantillonnage aléatoire stratifié est destiné à illustrer dans quelle mesure un équilibrage supplémentaire peut être avantageux. Nous avons tiré deux unités par strate afin d'obtenir le nombre maximal de strates tout en permettant une estimation de variance sans biais. Fuller (1981) a montré que, dans le cas de deux unités par strate, ce plan d'échantillonnage stratifié possède une variance anticipée proche de la variance du meilleur modèle d'échantillonnage à choix raisonné sous (4). Pour l'échantillonnage de Poisson, nous avons pris, pour la taille d'échantillon prévue de 20, des probabilités d'inclusion initiales égales aux probabilités d'inclusion initiales du plan d'échantillonnage stratifié.

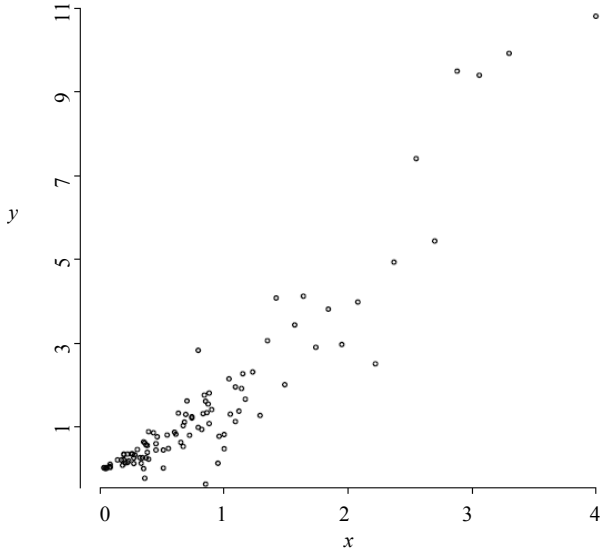

Figure 4 Population simulée sous le modèle (9)

L'estimateur par la régression étudié dans le présent article est de la forme (1) avec $\hat{\boldsymbol{\beta}}$ défini en (2). La variable de régression $\boldsymbol{z}$ est un vecteur de variables auxiliaires qui contient les variables du plan de sondage et x. Pour les plans de Poisson, nous avons utilisé $\boldsymbol{z}_i = (1,\ p_i,\ x_i,\ (1-p_i)^{-1}p_i)'$ comme vecteur des variables d'équilibrage et comme vecteur des variables de régression. La première variable fournit un contrôle pour la taille de population, la deuxième est un contrôle pour la taille d'échantillon, la

troisième fournit l'équilibre sur x et la quatrième garantit que l'estimateur par la régression est convergent sous le plan. Voir la condition (3) pour la convergence sous le plan de $\bar{y}_{\text{reg}}$ et l'ensemble $\boldsymbol{d} = (0, 0, 0, 1)'$. Pour les échantillons stratifiés avec deux unités par strate, le vecteur de variables d'équilibrage est $(x_i, I_{1i}, I_{2i}, ..., I_{10i})$ pour l'échantillonnage par la méthode du cube, où les I_{hi} sont les variables indicatrices de strate définies comme étant

$$I_{hi} = \begin{cases} 1 & \text{unité } i \text{ dans la strate } h \\ 0 & \text{sinon} \end{cases}$$

pour $h = 1, 2, ..., 10$. Seule la variable x est incluse dans la méthode d'équilibrage réjectif puisque l'échantillon tiré de ce plan de sondage initial est automatiquement équilibré sur les variables indicatrices de strate. Pour les deux méthodes d'équilibrage, le vecteur de variables de régression est $\boldsymbol{z}_i = (x_i, I_{1i}, ..., I_{10i})'$.

Pour les plans de sondage initiaux, les estimateurs de variance pour $\bar{y}_{\text{reg}}$ sont les estimateurs de variance de la moyenne de $e_i = y_i - \boldsymbol{z}_i'\boldsymbol{\beta}_N$ calculés avec $\hat{e}_i$, où $\hat{e}_i = y_i - \boldsymbol{z}_i'\hat{\boldsymbol{\beta}}$. Pour l'échantillonnage de Poisson, l'estimateur de variance est

$$\hat{V}(\bar{y}_{\text{reg}}) = (n-s)^{-1} n \bar{\boldsymbol{z}}_N' \hat{\boldsymbol{M}}_{zz}^{-1} \sum_{i \in A} \boldsymbol{z}_i \, p_i^{-4} \times (1-p_i)^3 \, \hat{e}_i^2 \, \boldsymbol{z}_i' \hat{\boldsymbol{M}}_{zz}^{-1} \bar{\boldsymbol{z}}_N, \quad (10)$$

où

$$\hat{\boldsymbol{M}}_{zz} = N^{-1} \sum_{i \in A} \boldsymbol{z}_i \, p_i^{-2} (1-p_i) \boldsymbol{z}_i',$$

et s est le nombre de variables dans $\boldsymbol{z}$. L'obtention de (10) est décrite en annexe.

Dans le cas de l'échantillonnage aléatoire stratifié avec deux unités par strate, l'estimateur de variance de $\bar{y}_{\text{reg}}$ est

$$\hat{V}(\bar{y}_{\text{reg}}) = (H-1)^{-1} H \sum_{h=1}^{H} [(1-W_h)^{1/2} \{0{,}5 W_h + (\bar{\boldsymbol{z}}_N - \bar{\boldsymbol{z}}) \hat{\boldsymbol{M}}_{zz,h}^{-1} \phi_h W_h^2 (\boldsymbol{z}_{h1} - \boldsymbol{z}_{h2})\} \times (\hat{e}_{h1} - \hat{e}_{h2})]^2, \quad (11)$$

où

$$\hat{\boldsymbol{M}}_{zz,h} = N_h^{-1} \sum_{i \in A_h} \boldsymbol{z}_i \, p_i^{-2} \, \phi_h \, \boldsymbol{z}_i',$$

A_h est l'ensemble d'échantillons dans la strate h, $W_h = n_h/N_h$, $\phi_h = (N_h - 1)^{-1}(N_h - 2)$ pour les unités dans la strate h, $\boldsymbol{z}_{hi}$ est le vecteur de variables auxiliaires $\boldsymbol{z}_i$ dans la strate h,

$$\hat{e}_{hi} = y_{hi} - \bar{y}_h - (\boldsymbol{z}_{hi} - \bar{\boldsymbol{z}}_h)'\hat{\boldsymbol{\beta}},$$

$\bar{y}_h$ et $\bar{\boldsymbol{z}}_h$ sont les moyennes de strate de y_{hi} et $\boldsymbol{z}_{hi}$, respectivement, et $H = 10$ est le nombre de strates. Les calculs menant à l'expression (11) sont les mêmes que ceux décrits en annexe et ne sont donc pas présentés.

Pour l'échantillonnage réjectif, nous avons utilisé les mêmes estimateurs de variance (10) et (11) en nous servant des probabilités d'inclusion du plan initial pour calculer l'estimateur de variance de $\bar{y}_{\text{reg}}$ pour les échantillons obtenus. Fuller (2009a) a prouvé que les propriétés en grand échantillon de l'estimateur par la régression sous échantillonnage réjectif sont les mêmes que celles de l'estimateur par la régression pour la procédure d'inclusion originale sous certaines conditions de régularité. Dans le cas de l'échantillonnage par la méthode du cube, nous avons évalué un estimateur de variance proposé par Deville et Tillé (2005) pour $\bar{y}_{\text{reg}}$ en utilisant les échantillons donnés par cette méthode.

Soit $p(\cdot)$ le plan de sondage initial et $\pi(\cdot)$, le scénario résultant après équilibrage. Le nombre d'échantillons sélectionnés était de 30 000 pour chaque simulation Monte Carlo sous les plans de sondage initiaux, l'échantillonnage par la méthode du cube et l'échantillonnage réjectif avec des taux de rejet de 90 % et 95 %. Nous avons construit l'estimateur de Horvitz-Thompson $\bar{y}_{\text{HT}}$ et l'estimateur par la régression $\bar{y}_{\text{reg}}$ en utilisant les probabilités d'inclusion initiales p_i. Soulignons que, pour l'échantillonnage réjectif, l'estimateur de Horvitz-Thompson utilisant les probabilités d'inclusion initiales n'est pas l'estimateur de Horvitz-Thompson sous le plan équilibré. Pour chaque plan initial, nous avons calculé les quantités qui suivent dans les études par simulation.

- $V_p(\bar{y}_{\text{HT}})$ (ou $V_p(\bar{y}_{\text{reg}})$) : Variance Monte Carlo de l'estimateur de Horvitz-Thompson (ou de l'estimateur par la régression) en utilisant les échantillons provenant des plans de sondage initiaux.
- $V_\pi(\bar{y}_{\text{HT}})$ (ou $V_\pi(\bar{y}_{\text{reg}})$) : Variance Monte Carlo de l'estimateur de Horvitz-Thompson (ou de l'estimateur par la régression) pour les échantillons équilibrés.
- $\text{biais}_\pi(\bar{y}_{\text{HT}})$ (ou $\text{biais}_\pi(\bar{y}_{\text{reg}})$) : Biais Monte Carlo de l'estimateur de Horvitz-Thompson (ou de l'estimateur par la régression), en utilisant les échantillons équilibrés.

Pour les échantillons obtenus par la méthode du cube,

- $\hat{V}_{DT}(\bar{y}_{\text{reg}})$: variance estimée de l'estimateur par la régression en utilisant les estimateurs de variance de Deville et Tillé (2005) et chaque échantillon obtenu par la méthode du cube.
- $\text{moy}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))$: moyenne Monte Carlo de $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ en utilisant tous les échantillons obtenus par la méthode du cube.

Deville et Tillé (2005) recommandent plusieurs estimateurs de variance basés sur une approximation de l'échantillonnage

de Poisson avec des corrections pour les contraintes connues dans la variance sous le plan. Les trois premiers estimateurs décrits dans Deville et Tillé (2005) ne différant que légèrement, nous n'avons utilisé que le deuxième dans les études par simulation. Deville et Tillé (2005) proposent aussi un quatrième estimateur, mais celui-ci requiert la résolution d'un système d'équations non linéaires dont l'ajout à la simulation aurait demandé beaucoup de ressources informatiques. Cependant, leur quatrième estimateur pourrait donner de meilleurs résultats que les autres pour les plans d'échantillonnage stratifiés, puisqu'il reproduit la variance d'un échantillon aléatoire stratifié quand le vecteur de variables d'équilibrage contient des indicateurs de strate.

Pour les échantillons obtenus par la méthode réjective,

- $\hat{V}(\bar{y}_{\text{reg}})$: variance estimée de l'estimateur par la régression en utilisant l'équation (10) (ou (11)) pour le plan de sondage initial de Poisson (ou stratifié à deux unités par strate) et chaque échantillon équilibré.
- $\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))$: moyenne Monte Carlo de $\hat{V}(\bar{y}_{\text{reg}})$ en utilisant les échantillons équilibrés.

Dans les simulations, nous avons également calculé $\hat{V}(\bar{y}_{\text{reg}})$ pour les échantillons obtenus par la méthode du cube aux fins de comparaison.

Le tableau 1 donne les estimations pour le plan d'échantillonnage de Poisson. La variance de la moyenne de Horvitz-Thompson sous l'échantillonnage de Poisson initial avec taille de l'échantillon prévue de 20 et sans équilibrage est $V_p(\bar{y}_{\text{HT}}) = 0{,}08$. Dans le tableau 1, les variances sont normalisées par $V_p(\bar{y}_{\text{HT}})$, et les biais sont normalisés par $\sqrt{V_p(\bar{y}_{\text{HT}})}$. L'estimateur de Horvitz-Thompson est sans biais sous les plans obtenus par la méthode du cube, parce que l'échantillonnage par la méthode du cube retient les probabilités d'inclusion de premier ordre. L'estimateur de Horvitz-Thompson en utilisant les probabilités d'inclusion du plan initial est biaisé sous l'échantillonnage réjectif, parce que les probabilités d'inclusion diffèrent des probabilités d'inclusion du plan initial, comme l'indique la figure 1. Le biais de l'estimateur par la régression sous échantillonnage réjectif est plus faible que celui de l'estimateur de Horvitz-Thompson calculé avec les probabilités d'inclusion du plan initial. Le biais de $\bar{y}_{\text{reg}}$ est du même ordre sous la méthode du cube et la méthode réjective. L'accroissement du taux de rejet augmente le biais de $\bar{y}_{\text{reg}}$ pour les plans obtenus par la méthode réjective. Cependant, sous les deux méthodes d'équilibrage et les deux taux de rejet, les biais de $\bar{y}_{\text{reg}}$ sont négligeables comparativement aux variances Monte Carlo. Pour l'estimateur de Horvitz-Thompson en utilisant les probabilités d'inclusion du plan initial, le gain réalisé en utilisant l'échantillon équilibré est important pour la méthode du cube ainsi que la méthode réjective. Les erreurs quadratiques moyennes sont réduites encore davantage en utilisant l'estimateur par la régression avec l'une ou l'autre méthode d'équilibrage. Le gain dû à l'utilisation de l'estimateur par la régression est plus important pour l'échantillonnage réjectif que pour l'échantillonnage par la méthode du cube, vraisemblablement parce que cette dernière produit un équilibre plus strict que la méthode réjective. Les deux méthodes donnent lieu à des variances similaires pour l'estimateur par la régression. La variance de l'estimateur par la régression sous le plan initial de Poisson est $V_p(\bar{y}_{\text{reg}}) = 0{,}249$ (relativement à $V_p(\bar{y}_{\text{HT}})$). En comparant la valeur de 0,249 à la quatrième ligne du tableau 1, nous voyons que, pour l'estimateur par la régression, le gain résultant de l'utilisation d'échantillons équilibrés est modéré. Ce résultat est en harmonie avec la constatation de Fuller (2009a) selon laquelle la réduction de la variance de $\bar{y}_{\text{reg}}$ en utilisant des échantillons obtenus par la méthode réjective est due à une correction de deuxième ordre. L'estimateur de variance de $\bar{y}_{\text{reg}}$ en utilisant (10) présente un petit biais tant pour les échantillons obtenus par la méthode du cube que pour ceux produits par la méthode réjective ($\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))$ dans le tableau 1). L'estimateur de variance $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ proposé dans Deville et Tillé (2005) donne des résultats comparables à $\hat{V}(\bar{y}_{\text{reg}})$ dans (10) puisque le deuxième estimateur de variance de Deville et Tillé (2005) est très proche de (10) pour l'échantillonnage de Poisson. Ce résultat appuie l'allégation selon laquelle, dans les estimateurs de variance de Deville et Tillé (2005), l'hypothèse de l'approximation de Poisson est satisfaite pour le cas du plan d'échantillonnage de Poisson.

Tableau 1
Propriétés des échantillons de taille prévue de 20 basés sur l'échantillonnage de Poisson. $V_p(\bar{y}_{\text{HT}}) = 0{,}08$ et $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0{,}249$

	Cube	Réj. 90 %	Réj. 95 %
$\text{biais}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0,002	-0,016	-0,007
$\text{biais}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0,002	0,002	0,005
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0,142	0,270	0,220
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0,131	0,136	0,129
$\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,122	0,123	0,121
$\text{moy}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,120	-	-

Au tableau 2, nous présentons les estimations sous le plan d'échantillonnage stratifié à deux unités par strate initial. La variance de la moyenne de Horvitz-Thompson sous le plan de stratification initial est $V_p(\bar{y}_{\text{HT}}) = 0{,}011$ et toutes les estimations sont normalisées au moyen de cette valeur. Puisque, dans ce plan initial, la stratification contrôle la plupart de l'effet de x sur y, l'estimateur par la régression n'offre pas d'amélioration importante par rapport

à l'estimateur de Horvitz-Thompson calculé en utilisant les probabilités d'inclusion du plan initial. Le biais et la variance de $\bar{y}_{\text{HT}}$ sont proches de ceux de $\bar{y}_{\text{reg}}$ sous la méthode du cube ainsi que la méthode réjective. Le biais estimé plus grand de $\bar{y}_{\text{HT}}$ sous l'échantillonnage par la méthode du cube est dû à l'erreur de Monte Carlo. L'amélioration due à l'équilibrage sur x n'est pas importante comparativement à celle obtenue pour l'exemple d'échantillonnage de Poisson. Cependant, sous ce plan stratifié initial fortement contrôlé, dans lequel les échantillons initiaux sont déjà partiellement équilibrés sur x, un équilibrage supplémentaire et l'utilisation des estimateurs $\bar{y}_{\text{reg}}$ peuvent encore offrir un avantage modéré. Ce résultat peut être constaté pour $\bar{y}_{\text{reg}}$ en comparant la quatrième ligne du tableau 2 à la variance de $\bar{y}_{\text{reg}}$ sous le plan initial $V_p(\bar{y}_{\text{reg}}) = 0{,}987$. Par conséquent, dans ce cas, une bonne stratégie consiste à combiner la stratification, l'équilibrage et la régression, conclusion qui est semblable à celle tirée par Deville et Tillé (2004). L'estimateur de variance $\hat{V}(\bar{y}_{\text{reg}})$ obtenu en utilisant (11) donne des estimations qui, en moyenne, pour les variances de l'estimateur par la régression sous la méthode du cube ainsi que la méthode réjective, sont proches des variances réelles. Cependant, l'estimateur de variance $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ proposé par Deville et Tillé (2005) donne de médiocres résultats pour l'échantillonnage par la méthode du cube. Cela pourrait être dû au fait que l'approximation de l'échantillonnage de Poisson dans le deuxième estimateur de variance de Deville et Tillé (2005) repose sur l'hypothèse que les probabilités d'inclusion conjointe sont éloignées des probabilités d'inclusion conjointe réelles dans les petites strates. Or, les probabilités d'inclusion conjointe dans les petites strates sont plus proches de celles de l'échantillonnage aléatoire stratifié que de celles de l'échantillonnage de Poisson. Ce problème pourrait expliquer pourquoi $\hat{V}(\bar{y}_{\text{reg}})$ donné par (11) en utilisant les probabilités initiales d'inclusion pour deux unités par strate est moins biaisé que $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ dans ce cas.

Tableau 2
Propriétés des échantillons de taille prévue de 20 basés sur l'échantillonnage stratifié. $V_p(\bar{y}_{\text{HT}}) = 0{,}011$ et $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0{,}987$

	Cube	Réj. 90 %	Réj. 95 %
$\text{biais}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0,028	0,014	0,010
$\text{biais}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0,013	0,014	0,010
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0,910	0,866	0,813
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0,929	0,865	0,813
$\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,907	0,881	0,775
$\text{moy}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,792	-	-

Afin d'évaluer les propriétés en grand échantillon des méthodes d'équilibrage, nous avons quadruplé la taille de la simulation de Poisson. Nous avons répété quatre fois la population et sélectionné un échantillon de taille prévue de 80. La variance d'une moyenne de Horvitz-Thompson sous l'échantillonnage de Poisson est $V_p(\bar{y}_{\text{HT}}) = 0{,}020$ et la variance de l'estimateur par la régression est $V_p(\bar{y}_{\text{reg}}) = 0{,}132$. Les variances et biais relatifs résultants sont proches des résultats pour les échantillons de taille 20 (tableau 3). Les résultats des simulations corroborent le résultat théorique de Fuller (2009a) selon lequel l'estimateur par la régression est un estimateur d'ordre $O_p(n^{-1/2})$ après un rejet du type utilisé dans le présent article. Bien que cela ne soit pas prouvé ici, l'estimateur par la régression utilisé après l'échantillonnage par la méthode du cube semble posséder des propriétés semblables à l'estimateur par la régression lorsque l'on utilise l'échantillonnage réjectif.

Tableau 3
Propriétés des échantillons de taille prévue de 80 basés sur l'échantillonnage de Poisson. $V_p(\bar{y}_{\text{HT}}) = 0{,}02$ et $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0{,}132$

	Cube	Réj. 90 %	Réj. 95 %
$\text{biais}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0,002	-0,006	-0,007
$\text{biais}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0,002	0,000	-0,001
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0,127	0,267	0,224
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0,122	0,124	0,123
$\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,121	0,121	0,121
$\text{moy}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,121	-	-

5. Corrections apportées à la méthode de rejet

Dans la méthode d'échantillonnage réjectif de Fuller, la même importance est accordée à toutes les variables d'équilibrage. Pour un grand nombre de celles-ci, on ne peut s'attendre à un équilibre exact sur toutes les variables et l'approximation pourrait être médiocre pour certaines variables importantes. Par conséquent, un praticien pourrait souhaiter obtenir un équilibre plus strict sur un sous-ensemble de variables d'équilibrage. Par exemple, un chercheur pourrait vouloir utiliser l'échantillonnage de Poisson par souci de simplicité, mais exercer aussi un certain contrôle sur la taille aléatoire de l'échantillon. Une taille d'échantillon aléatoire peut compliquer la planification de l'étude et contribuer fortement à la variance des estimateurs. L'échantillonnage équilibré permet de réduire la variation des tailles d'échantillon en équilibrant sur la variable p_i, qui est la probabilité d'inclusion de premier ordre initiale. Dans la méthode d'échantillonnage réjectif de Fuller, la variance de la taille d'échantillon s'accroît quand le nombre de variables d'équilibrage augmente et que le taux de rejet est maintenu

constant. Il est possible de modifier la méthode d'échantillonnage réjectif de manière que l'équilibre sur p_i soit plus strict que sur d'autres variables.

Un moyen d'accroître l'équilibrage sur un sous-ensemble de variables consiste à modifier la fonction de test de rejet. Dans Fuller (2009a), l'ordre de l'approximation des probabilités d'inclusion de premier et de deuxième ordres demeure le même quand, dans la forme quadratique de rejet, la matrice de variance est remplacée par une matrice définie positive symétrique de même ordre.

Pour déterminer les pondérations pour l'échantillonnage réjectif pondéré, il est commode de transformer les variables d'équilibrage de façon que $V(\overline{z}_{\mathrm{HT}} \mid F_N)$ soit une matrice diagonale. Pour l'échantillonnage réjectif pondéré, la statistique de test est

$$\sum_{q=1}^{m} c_q V(\overline{z}_{\mathrm{HT},\,q} \mid F_N)^{-1}(\overline{z}_{\mathrm{HT},\,q} - \overline{z}_{N,\,q})^2, \qquad (12)$$

où m est le nombre de variables d'équilibrage, z_q est la q^{e} variable d'équilibrage et c_q représente les poids sélectionnés. Nous pouvons donner au poids appliqué à la première variable $z_{1i} = p_i$ une grande valeur relativement à celle des poids appliqués aux autres variables afin de réduire la variation de la taille de l'échantillon. La transformation utilisée est celle de Gramm-Schmidt en se servant de la variance sous le plan de sondage initial. L'équilibrage est effectué sur les variables transformées, mais la première variable n'est pas transformée. Les estimateurs de Horvitz-Thompson des variables transformées ne sont pas corrélés. L'équilibrage sur les variables transformées produira encore un équilibre sur les variables originales, puisque chaque variable transformée est le résidu d'une opération de régression sur les variables précédentes.

Un parallèle peut être établi entre l'équation (12) et le terme de pénalité de la fonction de distance qui sous-tend le calage par la régression Ridge. Voir Rao et Singh (1997), Beaumont et Bocci (2008), ainsi que Chambers (1996). En particulier, le choix des poids c_q est semblable au problème de sélection des coûts appropriés dans le calage par la régression Ridge. Donc, l'échantillonnage réjectif en utilisant la statistique de test (12) peut être considéré comme l'intégration du calage par la régression Ridge à l'étape de l'élaboration du plan de sondage.

Un deuxième moyen de produire un équilibre plus strict sur un sous-ensemble de variables consiste à effectuer le rejet séparément pour divers sous-ensembles. Une statistique de test est produite pour chaque sous-ensemble et, pour être accepté, il faut qu'un échantillon soit accepté à tous les tests. Dans le cas de l'échantillonnage de Poisson, une statistique de test pourrait donner lieu au rejet si la taille d'échantillon n'est pas comprise entre les limites de tolérance spécifiées pour la taille d'échantillon prévue. Cette deuxième approche nécessite certaines hypothèses de plus que celles de Fuller (2009a), mais un argument similaire peut être utilisé pour justifier la procédure.

Afin de prouver les propriétés de convergence de la méthode de rejet à tests multiples, il est commode de considérer deux sous-ensembles de variables d'équilibrage et d'imaginer que le rejet est effectué séquentiellement sur chaque sous-ensemble. Nous donnons à la méthode de rejet sur deux sous-ensembles le nom de méthode d'échantillonnage réjectif en deux étapes. Supposons que $\boldsymbol{z}_i' = (\boldsymbol{z}_{1i}', \boldsymbol{z}_{2i}')$ est le vecteur de variables d'équilibrage et que le plan de sondage original est désigné par $p(\cdot)$. La procédure est la suivante.

Étape 1 : Sélectionner un échantillon en utilisant $p(\cdot)$ et rejeter les échantillons en appliquant la condition d'équilibrage (8) au premier sous-ensemble $\boldsymbol{z}_1$,

$$Q_1 = (\overline{z}_{\mathrm{HT},\,1} - \overline{z}_{N,\,1})' V(\overline{z}_{\mathrm{HT},\,1} \mid F_N)^{-1}(\overline{z}_{\mathrm{HT},\,1} - \overline{z}_{N,\,1}) < \gamma_1.$$

Étape 2 : Utiliser l'échantillon accepté à l'étape 1 pour vérifier la condition d'équilibrage (8) sur le deuxième sous-échantillon $\boldsymbol{z}_2$,

$$Q_2 = (\overline{z}_{\mathrm{HT},\,2} - \overline{z}_{N,\,2})' V(\overline{z}_{\mathrm{HT},\,2} \mid F_N)^{-1}(\overline{z}_{\mathrm{HT},\,2} - \overline{z}_{N,\,2}) < \gamma_2.$$

Rejeter l'échantillon si la condition n'est pas satisfaite et répéter l'étape 1.

En pratique, aussi bien pour l'échantillonnage réjectif pondéré que pour l'échantillonnage réjectif en deux étapes, des tâtonnements sont vraisemblablement nécessaires pour choisir les valeurs de γ. Dans le cas de la méthode pondérée, la forme quadratique devient une somme de multiples des variables aléatoires χ^2, ce qui rend le choix de γ plus difficile que dans le cas non pondéré. Nous avons utilisé des approximations par appariement des moments pour choisir les valeurs de γ qui fournissent des taux de rejet proches de ceux souhaités, mais nous avons ensuite procédé à de petites simulations pour déterminer le taux de rejet sous forme d'une fonction de γ. Pour la méthode en deux étapes, nous avons utilisé une approximation par une loi du χ^2 pour sélectionner une valeur de γ_1 donnant approximativement le taux de rejet souhaité à la première étape et nous avons utilisé une deuxième approximation par le χ^2 pour sélectionner une valeur initiale de γ_2 donnant approximativement le taux de rejet souhaité à la deuxième étape. Nous avons ajusté le deuxième paramètre γ_2 afin d'obtenir le taux de rejet global cible. Le choix des valeurs de γ dans la méthode en deux étapes est subjectif, car de nombreuses combinaisons de γ_1 et γ_2 peuvent produire le même taux global. En pratique, un praticien imposera vraisemblablement un bornage strict pour le premier sous-ensemble de variables et des bornes lâches sur les variables d'équilibrage restantes.

La moyenne et la variance en grand échantillon de l'estimateur par la régression sous l'échantillonnage réjectif en deux étapes sont les même que celles de l'estimateur par la régression sous le plan de sondage original. En outre, l'estimateur habituel de la variance sous le plan de sondage original pour l'estimateur par la régression convient pour l'échantillonnage réjectif en deux étapes. La preuve de cette déclaration, qui est une extension de la preuve donnée dans Fuller (2009a), peut être obtenue sur demande.

Afin d'examiner certaines propriétés des deux méthodes, nous avons répété les simulations Monte Carlo pour le plan d'échantillonnage de Poisson initial avec la variable p_i séparée des trois autres variables. Nous avons transformé le vecteur de variables d'équilibrage de façon que la matrice de variance des estimateurs de Horvitz-Thompson des totaux soit diagonale. Pour la méthode avec pondération, nous avons fixé le poids appliqué à la composante p_i de forme quadratique à 1,5, ceux appliqués aux autres composantes, à 1 et nous avons donné à γ la valeur 0,627. Cette méthode de pondération limitait les échantillons à ceux dont la taille variait de 18 à 22. Pour la méthode en deux étapes, tout échantillon dont la taille se situait en dehors de la fourchette de 18 à 22 a été rejeté à la première étape, puis la forme quadratique pour les trois variables restantes a été vérifiée en utilisant une valeur de γ de 0,63 pour la deuxième étape. Étant donné les bonnes propriétés de l'estimateur de variance $\hat{V}(\bar{y}_{\text{reg}})$ donné par (10), le tableau 4 ne contient que ses valeurs Monte Carlo moyennes $\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))$.

Tableau 4
Propriétés des échantillons de taille prévue de 20 obtenus par la méthode réjective avec corrections basées sur l'échantillonnage de Poisson, et taux de rejet de 95 %

	Pondération	Deux étapes
$\text{biais}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0,005	-0,014
$\text{biais}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0,003	0,002
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0,210	0,217
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0,132	0,132
$\text{moy}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0,121	0,121
$V_\pi(n)$	1,237	1,902

Les résultats obtenus pour une taille d'échantillon prévue de 20 et un taux de rejet proche de 95 % étaient semblables pour les deux méthodes de correction (tableau 4). L'estimateur de Horvitz-Thompson calculé pour la méthode pondérée donnait d'un peu meilleurs résultats que celui calculé pour la méthode en deux étapes. L'une des raisons de cette différence est que la variation des tailles d'échantillon était nettement plus faible dans le cas de la méthode pondérée ($V_\pi(n)$ à la dernière ligne du tableau 4). Des simulations supplémentaires en utilisant des tailles d'échantillon prévues plus grandes ont donné des variances relatives similaires. L'efficacité de l'estimateur par la régression était à peu près la même pour les deux méthodes. Les estimateurs de Horvitz-Thompson calculés en utilisant les probabilités d'inclusion du plan de sondage initial pour ces deux méthodes de correction avaient d'un peu meilleures propriétés que l'estimateur de Horvitz-Thompson obtenu pour la méthode réjective sans contrôle supplémentaire de la taille de l'échantillon.

6. Discussion

L'échantillonnage réjectif et l'échantillonnage par la méthode du cube produisent des estimateurs par la régression dont la performance est à peu près la même. L'équilibrage produit des gains importants quand le plan de sondage initial donne peu de contrôle sur les valeurs auxiliaires qui entrent dans les échantillons. Un plan d'échantillonnage bien stratifié offre nombre des avantages de l'équilibrage sur une variable continue. Toutefois, la poursuite de l'équilibrage après la stratification peut encore donner lieu à de petites améliorations de l'erreur quadratique moyenne pour les estimateurs par la régression. En outre, l'équilibrage pourrait être utilisé pour prévenir les poids négatifs que produisent les estimateurs par la régression (Fuller 2009a).

Pour les simulations, nous avons fixé le taux de rejet à 90 % pour la population la plus grande. Quand les tailles de la population et de l'échantillon augmentent, il est possible d'accroître le taux de rejet tout en maintenant un grand ensemble d'échantillons possibles. Des simulations supplémentaires ont été exécutées avec des taux de rejet proches de 99 %, mais les données ne sont pas présentées car les différences entre les résultats obtenus avec un taux de 95 % et un taux de 99 % étaient très faibles et le biais de $\bar{y}_{\text{reg}}$ demeurait négligeable. La légère réduction de la variance due à l'équilibrage diminue à mesure que les conditions d'équilibrage sont rendues plus rigoureuses.

Dans certains cas particulier, un chercheur pourrait souhaiter effectuer un équilibrage strict sur certaines variables et plus lâche sur d'autres. Des améliorations peuvent être obtenues en choisissant des poids différents pour les diverses variables ou en répartissant les variables entre des ensembles de test distincts. Les méthodes d'échantillonnage réjectif pondéré et en deux étapes donnent des résultats comparables, si bien que le choix entre ces méthodes dépendra en grande partie de la facilité de mise en œuvre.

Remerciements

Les présents travaux ont été financés aux termes de l'accord de coopération n° 68-3A75-4-122 conclu entre le Natural Resources Conservation Service du USDA et le

Center for Survey Statistics and Methodology de la Iowa State University. Les auteurs remercient le rédacteur associé Wayne A. Fuller et deux examinateurs anonymes de leurs commentaires constructifs qui leur ont permis d'améliorer l'article.

Annexe

Partons de

$$V(\bar{y}_{\text{reg}} \mid F_N) = V(\bar{y}_{\text{reg}} - \bar{y}_N \mid F_N).$$

Posons que

$$\bar{y}_N = \bar{\boldsymbol{z}}_N' \boldsymbol{\beta}_N$$

et notons que

$$y_i = \boldsymbol{z}_i' \boldsymbol{\beta}_N + e_{Ni}$$

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} \boldsymbol{z}_i'\right]^{-1} \sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} (\boldsymbol{z}_i' \boldsymbol{\beta}_N + e_i)$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + \left[N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} \boldsymbol{z}_i'\right]^{-1} N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} e_i. \qquad (13)$$

Sous les hypothèses (hypothèses classiques de convergence sous le plan de sondage)

$$N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} \boldsymbol{z}_i' = N^{-1}\sum_{i\in U} \boldsymbol{z}_i \phi_i p_i^{-1} \boldsymbol{z}_i' + O_p(n^{-1/2}).$$

Écrivons

$$N^{-1}\sum_{i\in U} \boldsymbol{z}_i \phi_i p_i^{-1} \boldsymbol{z}_i' = \boldsymbol{M}_{zz,N}.$$

Utilisons le même argument pour développer le terme $N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} e_i$. Alors, le développement de (13) est

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + \boldsymbol{M}_{zz,N}^{-1} N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} e_i + O_p(n^{-1}).$$

Pour construire les intervalles de confiance pour $\bar{y}_N$, il est suffisant de prendre en considération la variance du terme linéarisé. Par conséquent, considérons, en utilisant la notation de Särndal, Swensson et Wretman (1992),

$$\text{AV}(\bar{y}_{\text{reg}}) = \bar{\boldsymbol{z}}_N' \boldsymbol{M}_{zz,N}^{-1} V(\bar{\boldsymbol{b}}_{\text{HT}} \mid F_N) \boldsymbol{M}_{zz,N}^{-1} \bar{\boldsymbol{z}}_N$$

où

$$\boldsymbol{b}_i = \boldsymbol{z}_i \phi_i p_i^{-1} e_i.$$

La variance de l'estimateur HT pour la moyenne de b_i sous échantillonnage de Poisson est

$$\sum_{i\in U} (1 - p_i) p_i^{-1} \boldsymbol{b}_i \boldsymbol{b}_i'.$$

Ensuite, appliquons que $\phi = 1 - p_i$ pour obtenir l'approximation asymptotique de variance pour la partie linéarisée de $\bar{y}_{\text{reg}}$

$$\text{AV}(\bar{y}_{\text{reg}}) = \bar{\boldsymbol{z}}_N' \boldsymbol{M}_{zz,N}^{-1} \sum_{i\in U} (1 - p_i)^3 p_i^{-3} \boldsymbol{z}_i e_i^2 \boldsymbol{z}_i' \boldsymbol{M}_{zz,N}^{-1} \bar{\boldsymbol{z}}_N.$$

Nous obtenons l'estimateur de variance en remplaçant les totaux de population par les estimateurs HT sous échantillonnage de Poisson et en intégrant une correction du nombre de degrés de liberté devant $n/(n-s)$ à cause de la petite taille d'échantillon.

Bibliographie

Beaumont, J.-F., et Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66, 1, 5-20.

Chambers, R.L. (1996). Robust Case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Fuller, W.A. (1981). An empirical Study of the ratio estimator and estimators of its variance: Comment. *Journal of the American Statistical Association*, 76, 78-80.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. À venir *Biometrika*.

Fuller, W.A. (2009b). *Sampling Statistics*. Hoboken, New Jersey : John Wiley & Sons, Inc.

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

Matei, A., et Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 543-570.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Rao, J.N.K., et Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 57-65.

Rousseau, S., et Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Rapport technique, INSEE, Paris.

Royall, R.M., et Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Royall, R.M., et Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

Särndal, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

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

Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer Science+ Business Media, Inc.

Tillé, Y., et Matei, A. (2005). The R package Sampling. *The Comprehensive R Archive Network*, *http://cran.r-project.org/*, *Manual of the Contributed Packages*.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York : John Wiley & Sons, Inc.