Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Estimateurs bayésiens fondés sur le plan de sondage et assistés par un modèle de la taille, du total et de la moyenne d'une population difficile à joindre depuis un échantillon par dépistage de liens avec un échantillon initial en grappes

par Martín H. Félix-Medina

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Estimateurs bayésiens fondés sur le plan de sondage et assistés par un modèle de la taille, du total et de la moyenne d'une population difficile à joindre depuis un échantillon par dépistage de liens avec un échantillon initial en grappes

Martín H. Félix-Medina[1]

Résumé

Nous présentons des estimateurs de type Horvitz-Thompson et de type multiplicité fondés sur le plan de sondage de la taille de la population, ainsi que du total et de la moyenne d'une variable de réponse associée aux éléments d'une population cachée à utiliser avec la variante d'échantillonnage par dépistage de liens proposée par Félix-Medina et Thompson (2004). Étant donné que le calcul des estimateurs nécessite de connaître les probabilités d'inclusion des personnes échantillonnées, mais qu'elles sont inconnues, nous proposons un modèle bayésien qui nous permet de les estimer et, par conséquent, de calculer les estimateurs des paramètres de population. Les résultats d'une petite étude numérique indiquent que les performances des estimateurs proposés sont acceptables.

Mots clés : échantillonnage de Gibbs; population cachée; estimateurs de Horvitz-Thompson; échantillonnage en boule de neige.

1. Introduction

Les méthodes d'échantillonnage classiques ne conviennent pas à l'échantillonnage de populations cachées ou difficiles à détecter, comme les personnes consommant de la drogue, les travailleuses et travailleurs du sexe et les personnes itinérantes, en raison de facteurs comme l'absence de bases de sondage appropriées, la rareté de ces populations et la complexité de l'échantillonnage de leurs membres. Voir Tourangeau (2014) pour une discussion sur ceux-ci ainsi que sur d'autres facteurs. C'est pourquoi plusieurs méthodes d'échantillonnage de ce type de population ont été proposées. L'une d'elles est l'échantillonnage par dépistage de liens (EDL). Cette méthode se fonde sur l'idée de sélectionner un échantillon initial de la population cachée, puis de demander aux personnes de l'échantillon initial de donner le nom d'autres membres de la population qu'elles connaissent. Les personnes nommées qui ne font pas partie de l'échantillon initial sont incluses dans l'échantillon, puis il pourrait leur être demandé de nommer à leur tour les membres de la population qu'elles connaissent. Ce processus peut se poursuivre de cette façon jusqu'à ce que la règle d'arrêt soit satisfaite.

Félix-Medina et Thompson (2004) ont proposé une variante d'échantillonnage par dépistage de liens (EDL) pour estimer la taille d'une population difficile à joindre. Dans le plan de sondage proposé, une partie de la population est couverte par une base de sondage de lieux, comme des bars, des parcs et des rues de pâté de maisons, où les membres de la population ont tendance à se rassembler. Un échantillon aléatoire simple sans remise des lieux est sélectionné à partir de la base de sondage et les membres de la population qui appartiennent à l'un des lieux échantillonnés sont inclus dans l'échantillon. Ensuite, dans chaque lieu échantillonné, on demande aux personnes sélectionnées de nommer leurs contacts qui appartiennent aussi à cette population.

On trouve des estimateurs de la taille de la population, ainsi que des estimateurs du total de la population et de la moyenne d'une variable d'intérêt, comme les dépenses hebdomadaires en drogue d'une personne consommant de la drogue et le nombre hebdomadaire de clients d'un travailleur du sexe, qui ont été proposés aux fins d'utilisation avec

[1]Martín H. Félix-Medina, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacan Sinaloa, Mexique, 80013 (mhfelix@uas.edu.mx)

ce plan de sondage et calculés selon différents modèles dans Félix-Medina et Thompson (2004), Félix-Medina et Monjardin (2006), Félix-Medina et Monjardin (2010), Félix-Medina et coll. (2015) et Félix-Medina (2021). Dans cet article, nous présentons des estimateurs de la taille de la population, du total et de la moyenne, calculés selon une approche fondée sur le plan de sondage, mais assistés par un modèle bayésien. C'est pourquoi nous nous attendons à ce que les estimateurs proposés se comportent de façon acceptable dans un large éventail de conditions.

2. Plan de sondage et notation

La variante de l'EDL proposée par Félix-Medina et Thompson (2004) est la suivante. Soit U une population finie de τ éléments. On suppose qu'une partie U_1 de U est couverte par une base de sondage de N lieux $A_1, \dots, A_N$, où les membres de la population ont tendance à se rassembler. Soit m_i le nombre d'éléments de U_1 qui appartiennent à A_i, $i = 1, \dots, N$. Comme dans un échantillonnage en grappes classique, on suppose qu'une personne dans U_1 appartient à un seul lieu. Alors, le nombre de personnes dans U_1 est $\tau_1 = \sum_1^N m_i$. Soit $U_2 = U - U_1$ la partie de U qui n'est pas couverte par la base. Notons que la taille de U_2 est $\tau_2 = \tau - \tau_1$. Un échantillon aléatoire simple sans remise $S_A = \{A_1, \dots, A_n\}$ de n lieux a été sélectionné à partir de la base de sondage. Les m_i éléments de U_1 qui appartiennent à chaque $A_i \in S_A$ sont inclus dans l'échantillon. Soit S_0 l'ensemble de membres de U_1 qui appartiennent aux lieux échantillonnés et soit $m = \sum_1^n m_i$ la taille de S_0. Dans chaque lieu échantillonné, on demande aux personnes de nommer leurs contacts qui appartiennent aussi à la population. On considère qu'une personne nommée est liée à un lieu si l'un des membres de ce lieu a donné son nom. On suppose $x_{ij}^{(k)} = 1$ si la personne $j \in U_k - A_i$ est liée au lieu $A_i \in S_A$, et $x_{ij}^{(k)} = 0$ sinon, $k = 1, 2$. Dans le cas d'une personne échantillonnée $j \in U_k$, les renseignements suivants qui lui sont associés sont enregistrés : la valeur $y_j^{(k)}$ de la variable d'intérêt y; les valeurs $x_{ij}^{(k)}$, $i = 1, \dots, n$, des variables d'indicateur de lien, et le sous-ensemble auquel la personne appartient parmi les sous-ensembles suivants de U : $U_1 - S_0$, $A_i \in S_A$ ou U_2 en particulier. Nous noterons par S_1 et S_2 les ensembles de r_1 et r_2 personnes dans $U_1 - S_0$ et U_2 qui sont liées à au moins un lieu $A_i \in S_A$. Enfin, supposons que $S_1^* = S_0 \cup S_1$ et $S_2^* = S_2$ sont les ensembles de $m + r_1$ et r_2 personnes échantillonnées de respectivement U_1 et U_2.

3. Estimateurs de la taille, du total et de la moyenne

Supposons que $Y_k = \sum_1^{\tau_k} y_j^{(k)}$ et $\bar{Y}_k = Y_k/\tau_k$ sont le total de la population et la moyenne des valeurs y associées aux éléments de U_k, $k = 1, 2$, et que $Y = Y_1 + Y_2$ et $\bar{Y} = Y/\tau$ sont le total et la moyenne correspondants des éléments de U. Notons que si $y_j^{(k)} = 1$ pour chaque $j \in U_k$, alors $Y_k = \tau_k$, $k = 1, 2$ et $Y = \tau$. Les estimateurs de Horvitz-Thompson de τ_k, Y_k, τ et Y sont $\hat{\tau}_k = \sum_{j \in S_k^*} 1/\pi_j^{(k)}$, $\hat{Y}_k = \sum_{j \in S_k^*} y_j^{(k)}/\pi_j^{(k)}$, $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$ et $\hat{Y} = \hat{Y}_1 + \hat{Y}_2$, où $\pi_j^{(k)}$ est la probabilité d'inclusion de l'élément $j \in U_k$, qui sera calculé ultérieurement. Les estimateurs des moyennes $\bar{Y}_k$ et $\bar{Y}$ fondées sur les estimateurs de Horvitz-Thompson sont $\hat{\bar{Y}}_k = \hat{Y}_k/\hat{\tau}_k$, $k = 1, 2$ et $\hat{\bar{Y}} = \hat{Y}/\hat{\tau}$.

Pour calculer la probabilité d'inclusion $\pi_j^{(k)}$, nous définirons les variables $N_j^{(k)} = \sum_{i=1}^N x_{ij}^{(k)}$ et $X_j^{(k)} = \sum_{i=1}^n x_{ij}^{(k)}$, qui représentent le nombre de lieux dans la base de sondage et dans l'échantillon S_A, respectivement, qui sont liés à une personne $j \in U_k$, $k = 1, 2$. Notons que $N_j^{(k)}$ est inconnu, alors que $X_j^{(k)}$ est connu. Puisque l'échantillon S_A est sélectionné par un plan d'échantillonnage aléatoire simple sans remise (EASSR), $X_j^{(2)}$ suit une loi hypergéométrique $\left(N, N_j^{(2)}, n\right)$. Alors, $\pi_j^{(2)} = 1 - \Pr\left(X_j^{(2)} = 0 \middle| N_j^{(2)}\right) = 1 - \binom{N - N_j^{(2)}}{n} / \binom{N}{n}$, $j \in U_2$. Dans le cas d'une personne $j \in U_1$, la distribution conditionnelle de $X_j^{(1)}$ étant donné que $j \notin S_0$ est une loi hypergéométrique $\left(N - 1, N_j^{(1)}, n\right)$, parce qu'une personne dans U_1 ne peut pas être liée au lieu auquel elle appartient. Ainsi, la probabilité d'inclusion $\pi_j^{(1)}$ d'une personne $j \in U_1$ est donnée par

$$\pi_j^{(1)} = 1 - \Pr(j \notin S_0)\Pr(j \notin S_1 | j \notin S_0) = 1 - \left(1 - \frac{n}{N}\right)\Pr\left(X_j^{(1)} = 0 \middle| N_j^{(1)}, j \notin S_0\right) = 1 - \binom{N - N_j^{(1)}}{n} / \binom{N}{n}.$$

Veuillez noter que la probabilité d'inclusion $\pi_j^{(k)}$ dépend de $N_j^{(k)}$ et est par conséquent inconnue. C'est pourquoi nous l'estimerons en estimant $N_j^{(k)}$ au moyen d'une approche bayésienne.

Nous définirons d'abord une distribution initiale de $N_j^{(k)}$ à deux degrés. Dans le cas de $j \in U_2$, nous supposerons que $Pr\left(N_j^{(2)} \geq 1\right) = 1$, car sinon $j \in U_2$ ne peut pas être échantillonné. Ainsi, nous définirons comme distribution initiale de $N_j^{(2)}|\alpha_2, \beta_2$ la loi bêta-binomiale tronquée à zéro (N, α_2, β_2), c'est-à-dire

$$g_{N_j^{(2)}|N,\alpha_2,\beta_2}\left(n_j^{(2)}|N, \alpha_2, \beta_2\right) = \binom{N}{n_j^{(2)}} \frac{B(\alpha_2 + n_j^{(2)}, N + \beta_2 - n_j^{(2)})}{B(\alpha_2, \beta_2)} \frac{1}{1 - B(\alpha_2, N + \beta_2)/B(\alpha_2, \beta_2)}, \quad (1)$$

$n_j^{(2)} = 1, \ldots, N$. Dans le cas de $j \in U_1$, nous définirons comme distribution initiale de $N_j^{(1)}|\alpha_1, \beta_1$ la loi bêta-binomiale tronquée à zéro $(N-1, \alpha_1, \beta_1)$, c'est-à-dire

$$g_{N_j^{(1)}|N,\alpha_1,\beta_1}\left(n_j^{(1)}|N, \alpha_1, \beta_1\right) = \binom{N-1}{n_j^{(1)}} \frac{B(\alpha_1 + n_j^{(1)}, N - 1 + \beta_1 - n_j^{(1)})}{B(\alpha_1, \beta_1)} \frac{1}{1 - B(\alpha_1, N - 1 + \beta_1)/B(\alpha_1, \beta_1)}, \quad (2)$$

$n_j^{(1)} = 1, \ldots, N-1$. Notons que nous avons utilisé $N-1$ au lieu de N. Cela s'explique par le fait qu'une personne dans U_1 ne peut pas être liée au lieu auquel elle appartient et que par conséquent, la borne supérieure de son nombre de liens est $N-1$. Pour mettre fin à la spécification de la distribution initiale, nous devons définir la distribution conjointe de (α_k, β_k). Cependant, au lieu de spécifier une distribution pour ces paramètres, nous utiliserons la reparamétrisation proposée par Lee et Sabavala (1987), qui définissent les paramètres $\mu_k = \alpha_k/(\alpha_k + \beta_k)$ et $\rho_k = 1/(\alpha_k + \beta_k)$. Notons que la transformation inverse est $\alpha_k = \mu_k(1-\rho_k)/\rho_k$ et $\rho_k = (1-\mu_k)(1-\rho_k)/\rho_k$. Étant donné que les paramètres μ_k et ρ_k sont compris entre zéro et un, ils se définissent comme étant les distributions initiales de μ_k et ρ_k de distributions bêta avec les paramètres $\left(a^{(\mu_k)}, b^{(\mu_k)}\right)$ et $\left(a^{(\mu_k)}, b^{(\mu_k)}\right)$, respectivement. La fonction de densité de probabilité conjointe $g(\mu_k, \rho_k)$ de (μ_k, ρ_k) est définie comme le produit des fonctions de densité $g_{\mu_k}(\mu_k)$ et $g_{\rho_k}(\rho_k)$ de μ_k et ρ_k, respectivement, $k = 1, 2$.

La fonction de vraisemblance est fondée sur les variables $X_j^{(k)}$ associées aux personnes de l'échantillon. Ainsi, si $j \in S_2$, alors $\Pr\left(X_j^{(2)} \geq 1 \middle| j \in S_2\right) = 1$, et par conséquent, la distribution conditionnelle de $X_j^{(2)}$, étant donné que $j \in S_2$ et $N_j^{(2)}$, est la loi hypergéométrique tronquée à zéro $\left(N, N_j^{(2)}, n\right)$, c'est-à-dire

$$f_{X_j^{(2)}|j\in S_2,N_j^{(2)}}\left(x_j^{(2)}|j \in S_2, n_j^{(2)}\right) = \left[\binom{n_j^{(2)}}{x_j^{(2)}}\binom{N - n_j^{(2)}}{n - x_j^{(2)}} / \binom{N}{n}\right]\left[1 - \binom{N - n_j^{(2)}}{n} / \binom{N}{n}\right]^{-1}$$

De même, si $j \in S_1$, la distribution conditionnelle de $X_j^{(1)}$, étant donné que $j \in S_1$ et $N_j^{(1)}$, est la loi hypergéométrique tronquée à zéro $\left(N-1, N_j^{(1)}, n\right)$, c'est-à-dire

$$f_{X_j^{(1)}|j\in S_1,N_j^{(1)}}\left(x_j^{(1)}|j \in S_1, n_j^{(1)}\right) = \left[\binom{n_j^{(1)}}{x_j^{(1)}}\binom{N - 1 - n_j^{(1)}}{n - x_j^{(1)}} / \binom{N-1}{n}\right]\left[1 - \binom{N - 1 - n_j^{(1)}}{n} / \binom{N-1}{n}\right]^{-1},$$

tandis que si $j \in A_i \in S_A$, la distribution conditionnelle de $X_j^{(A_i)}$, étant donné que $j \in A_i \in S_A$ et $N_j^{(A_i)}$, est la loi hypergéométrique tronquée à zéro $\left(N-1, N_j^{(A_i)}, n-1\right)$, c'est-à-dire

$$f_{X_j^{(A_i)}|j\in A_i\in S_A,N_j^{(A_i)}}\left(x_j^{(A_i)}|j \in A_i \in S_A, n_j^{(A_i)}\right) = \binom{n_j^{(A_i)}}{x_j^{(A_i)}}\binom{N - 1 - n_j^{(A_i)}}{n - 1 - x_j^{(A_i)}} / \binom{N-1}{n-1},$$

où nous avons noté par $N_j^{(A_i)}$ et $X_j^{(A_i)}$ les variables qui comptent le nombre de lieux dans la base de sondage et dans l'échantillon S_A, respectivement, qui sont liés à la personne $j \in A_i \in S_A$.

Pour estimer $N_j^{(k)}$, $k = 1, 2$ et $N_j^{(A_i)}$, $i = 1, \dots, n$, nous utiliserons une approche par classes latentes. Ainsi, dans le cas d'une personne $j \in S_2$, nous définirons le vecteur des variables d'indicateur latent associé à cette personne comme étant $C_j^{(2)} = \left(C_{1j}^{(2)}, \dots, C_{Nj}^{(2)}\right)$, où $C_{lj}^{(2)} = 1$ si $N_j^{(2)} = l$ et $C_{lj}^{(2)} = 0$ sinon, $l = 1, \dots, N$. Dans le cas d'une personne $j \in S_1$, son vecteur associé de variables d'indicateur latent est $C_j^{(1)} = \left(C_{0j}^{(1)}, \dots, C_{N-1j}^{(1)}\right)$, où $C_{lj}^{(1)} = 1$ si $N_j^{(1)} = l$ et $C_{lj}^{(1)} = 0$ sinon, $l = 0, \dots, N-1$. Enfin, dans le cas d'une personne $j \in A_i \in S_A$, le vecteur est $C_j^{(A_i)} = \left(C_{0j}^{(A_i)}, \dots, C_{N-1j}^{(A_i)}\right)$, où $C_{lj}^{(A_i)} = 1$ si $N_j^{(A_i)} = l$ et $C_{lj}^{(A_i)} = 0$ sinon, $l = 0, \dots, N-1$. Notons que $N_j^{(k)}$ détermine $C_j^{(k)}$ et inversement; par conséquent, le problème de l'estimation de $N_j^{(k)}$ équivaut à celui de l'estimation de $C_j^{(k)}$. Il en va de même pour $N_j^{(A_i)}$ et $C_j^{(A_i)}$, $A_i \in S_A$. Nous nous concentrerons sur l'estimation de $C_j^{(k)}$, $k = 1, 2$, et $C_j^{(A_i)}$, $A_i \in S_A$. La fonction de masse de la probabilité conditionnelle (fmp), $g_{C_j^{(2)}|N,\mu_2,\rho_2}\left(c_j^{(2)}|N, \mu_2, \rho_2\right)$, de $C_j^{(2)}$ étant donné que μ_2 et ρ_2 est le un (1) de la distribution multinomiale $\left(1, \left\{g_{N_j^{(2)}|N,\alpha_2,\beta_2}(l|N, \alpha_2^*, \beta_2^*)\right\}_{l=1}^{N}\right)$, tandis que la fmp conditionnelle, $g_{C_j^{(1)}|N,\mu_1,\rho_1}\left(c_j^{(1)}|N, \mu_1, \rho_1\right)$, de $C_j^{(1)}$, ainsi que la fmp conditionnelle, $g_{C_j^{(A_i)}|N,\mu_1,\rho_1}\left(c_j^{(A_i)}|N, \mu_1, \rho_1\right)$, de $C_j^{(A_i)}$, étant donné μ_1 et ρ_1, sont toutes deux égales au un (1) de la même distribution multinomiale $\left(1, \left\{g_{N_j^{(1)}|N,\alpha_1,\beta_1}(l|N, \alpha_1^*, \beta_1^*)\right\}_{l=0}^{N-1}\right)$, où $g_{N_j^{(2)}|N,\alpha_2,\beta_2}$ et $g_{N_j^{(1)}|N,\alpha_1,\beta_1}$ sont donnés respectivement par (1) et (2), et α_k^* et β_k^* sont donnés par la fonction inverse de (μ_k, ρ_k), $k = 1, 2$.

En supposant que les vecteurs $\left(X_j^{(k)}, C_j^{(k)}\right)$ associés aux r_k personnes dans S_k sont mutuellement indépendants, nous obtiendrons, étant donnés μ_k et ρ_k, la distribution conditionnelle conjointe suivante :

$$f_{X^{(k)},C^{(k)}|\mu_k,\rho_k}\left(x^{(k)}, c^{(k)}|\mu_k, \rho_k\right) = \prod_{j=1}^{r_k} f_{X_j^{(k)}|j\in S_k,C_j^{(k)}}\left(x_j^{(k)}|j \in S_k, c_j^{(k)}\right) g_{C_j^{(k)}|N,\mu_k,\rho_k}\left(c_j^{(k)}|N, \mu_k, \rho_k\right), \quad \text{où}$$

$f_{X_j^{(k)}|j\in S_k,C_j^{(k)}}$ est la loi hypergéométrique tronquée à zéro de $X_j^{(k)}$, $X^{(k)}$ est le vecteur des variables $X_j^{(k)}$, et $C^{(k)}$ est la matrice dont les colonnes sont les vecteurs $C_j^{(k)}$, $j \in S_k$, $k = 1, 2$. De plus, en supposant que les vecteurs $\left(X_j^{(A_i)}, C_j^{(A_i)}\right)$ associés aux m personnes dans S_0 sont mutuellement indépendants, nous obtiendrons, étant donnés μ_1 et ρ_1, la distribution conditionnelle conjointe suivante :

$$f_{X^{(0)},C^{(0)}|,\mu_1,\rho_1}\left(x^{(0)}, c^{(0)}|\mu_1, \rho_1\right) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} f_{X_j^{(A_i)}|j\in A_i\in S_A,C_j^{(A_i)}}\left(x_j^{(A_i)}|j \in S_0, c_j^{(A_i)}\right) g_{C_j^{(A_i)}|N,\mu_1,\rho_1}\left(c_j^{(A_i)}|N, \mu_1, \rho_1\right),$$

où $f_{X_j^{(A_i)}|j\in A_i\in S_A,C_j^{(A_i)}}$ est la loi hypergéométrique tronquée à zéro de $X_j^{(A_i)}$, $X^{(0)}$ est le vecteur des variables $X_j^{(A_i)}$, et $C^{(0)}$ est la matrice dont les colonnes sont les vecteurs $C_j^{(A_i)}$, $j \in A_i$, $i = 1, \dots, n$.

La fonction de densité de la probabilité conjointe de la distribution finale de $\left(C^{(0)}, C^{(1)}, C^{(2)}, \mu_1, \mu_2, \rho_1, \rho_2\right)$ est

$$g_{C^{(0)},C^{(1)},C^{(2)},\mu_1,\mu_2,\rho_1,\rho_2|data}\left(c^{(0)}, c^{(1)}, c^{(2)}, \mu_1, \mu_2, \rho_1, \rho_2|data\right) = g_{C^{(0)},C^{(1)},\mu_1,\rho_1|data}\left(c^{(0)}, c^{(1)}, \mu_1, \rho_1|data\right) g_{C^{(2)},\mu_2,\rho_2|data}\left(c^{(2)}, \mu_2, \rho_2|data\right),$$

où

$$g_{C^{(0)},C^{(1)},\mu_1,\rho_1|data}\left(c^{(0)}, c^{(1)}, \mu_1, \rho_1|data\right) \propto f_{X^{(1)},C^{(1)}|\mu_1,\rho_1}\left(x^{(1)}, c^{(1)}|\mu_1, \rho_1\right) f_{X^{(0)},C^{(0)}|,\mu_1,\rho_1}\left(x^{(0)}, c^{(0)}|\mu_1, \rho_1\right) g(\mu_1, \rho_1)$$

et

$$g_{C^{(2)},\mu_2,\rho_2|data}\left(c^{(2)}, \mu_2, \rho_2|data\right) \propto f_{X^{(2)},C^{(2)}|\mu_2,\rho_2}\left(x^{(2)}, c^{(2)}|\mu_2, \rho_2\right) g(\mu_2, \rho_2).$$

Par conséquent, pour ce qui est de la distribution finale, $(C^{(0)}, C^{(1)}, \mu_1, \rho_1)$ et $(C^{(2)}, \mu_2, \rho_2)$ sont indépendants.

Les inférences au sujet des paramètres d'intérêt sont effectuées au moyen d'un échantillonnage de Gibbs. Ainsi, on obtient la distribution conditionnelle finale de chacun des paramètres $C_j^{(A_i)}$, $j \in A_i \in S_A$, $C_j^{(k)}$, $j \in S_k$, μ_k et ρ_k, $k = 1, 2$, compte tenu du reste des paramètres. On met en œuvre la procédure en spécifiant le nombre de chaînes et la longueur T de chaque chaîne, ainsi que les valeurs initiales des paramètres μ_k et ρ_k. Ensuite, à chaque itération de

l'algorithme, une valeur de chaque paramètre est échantillonnée à partir de sa distribution conditionnelle étant données les valeurs les plus récentes du reste des paramètres. Cela permet de simuler les valeurs des paramètres $N_j^{(k)}$ et $N_j^{(A_i)}$, les valeurs des probabilités d'inclusion $\pi_j^{(k)},\ j \in S_k^*,\ k = 1,2$ et $\pi_j^{(A_i)}, j \in A_i, i = 1, \ldots, n$, et par conséquent les valeurs des estimateurs de Horvitz-Thompson des tailles, des totaux et des moyennes. Nous simulons également les valeurs des estimateurs de multiplicité (Birnbaum et Sirken, 1965) des tailles, des totaux et des moyennes, ainsi que les valeurs des estimateurs de leurs variances conditionnelles étant données les valeurs de $N_j^{(k)}$ et $N_j^{(A_i)}$. L'idée du calcul des estimations des variances des estimateurs de multiplicité est qu'elles sont plus faciles à calculer que celles des estimateurs de Horvitz-Thompson, et elles ont été utilisées dans le calcul des estimations des variances des estimateurs de Horvitz-Thompson.

Une fois que les itérations de l'algorithme d'échantillonnage de Gibbs ont été effectuées, les estimations des variances inconditionnelles des estimateurs de Horvitz-Thompson et de multiplicité ont été calculées au moyen de la formule à deux degrés $\hat{V}\left(\hat{\theta}^{(k)}\right) = V\left[E\left(\hat{\theta}^{(k)}|\widehat{\{N_j^{(k)}\}}_{J\in S_k^*}\right)\right] + E\left[V\left(\hat{\theta}_M^{(k)}|\widehat{\{N_j^{(k)}\}}_{J\in S_k^*}\right)\right]$, où $\hat{\theta}^{(k)}$ désigne soit l'estimateur de Horvitz-Thompson soit l'estimateur de multiplicité du paramètre de population θ_k ($that\ is, \tau_k$, Y_k ou $\bar{Y}_k$), $V\left[E\left(\hat{\theta}^{(k)}|\widehat{\{N_j^{(k)}\}}_{J\in S_k^*}\right)\right]$ est la variance de l'échantillon des valeurs simulées de l'estimateur $\hat{\theta}^{(k)}$ (après suppression des valeurs correspondant à la période de rodage), et $E\left[V\left(\hat{\theta}_M^{(k)}|\widehat{\{N_j^{(k)}\}}_{J\in S_k^*}\right)\right]$ est la moyenne de l'échantillon des valeurs simulées de l'estimateur de la variance conditionnelle de l'estimateur de multiplicité $\hat{\theta}_M^{(k)}$ de θ_k. L'estimation de la variance de l'estimateur $\hat{\theta}$ du paramètre θ de la population entière U est obtenue par l'addition des variances inconditionnelles des estimateurs $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$.

4. Étude par la méthode Monte Carlo

Pour observer les performances des estimateurs proposés, des estimateurs de leurs variances et des intervalles de confiance de Wald fondés sur ces estimateurs, nous avons réalisé une petite étude par simulations. Nous avons donc utilisé les données de l'Étude longitudinale nationale sur la santé des adolescents (Add Health) recueillies au cours de l'année scolaire 1994-1995 pour constituer une population. Voir dans Harris (2013) la description de cette étude. Plus précisément, les données de l'école secondaire et de l'école intermédiaire qui l'alimente dans la collectivité 50 ont servi à constituer une population U de $\tau = 2487$ éléments, divisée en sous-populations U_1 et U_2 de tailles $\tau_1 = 1800$ et $\tau_2 = 687$, respectivement. Les éléments dans U_1 ont été groupés en $N = 150$ grappes de tailles $m_i,\ i = 1, \ldots, N$, dont les valeurs ont été générées à partir d'une distribution binomiale négative avec une moyenne et une variance respectivement égales à 12 et 24. Un élève de U était lié à une grappe si l'un des élèves de la grappe indiquait qu'il était son ami. La variable d'intérêt associée à chaque élève était le nombre d'amis qu'il ou elle nommait. Les totaux de population étaient $Y_1 = 10162$, $Y_2 = 2631$ et $Y = 12793$; et les moyennes de population étaient $\bar{Y}_1 = 5{,}65$, $\bar{Y}_2 = 3{,}83$ et $\bar{Y} = 5{,}14$. L'étude a été réalisée par sélection répétée de 1 000 échantillons dans la population au moyen du plan de sondage décrit dans la section 2. La taille de l'échantillon initial S_A de grappes était $n = 20.$. Pour chaque échantillon, des inférences à propos de chaque paramètre ont été obtenues au moyen de l'algorithme d'échantillonnage de Gibbs à deux chaînes, chacune d'une longueur de 4 000, et une période de rodage de 2 000. L'étude a été réalisée au moyen de l'environnement logiciel R aux fins de calcul statistique (R Core Team, 2022).

Tableau 4-1
Résultats de l'étude par la méthode Monte Carlo fondés sur 1 000 échantillons répétés sélectionnés à partir d'une population artificielle établie au moyen des données de l'Étude longitudinale nationale sur la santé des adolescents.

Estimateurs de Horvitz-Thompson		Tailles			Totaux			Moyennes		
		$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}$	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}$
Estimateurs des paramètres de population	Paramètres de pop.	1 800	687	2 487	10 162	2 631	12 793	5,65	3,83	5,14
	Moyenne	1 602,8	730,3	2 333,1	9 556,4	3 089,3	12 645,7	6,0	4,2	5,4
	Biais relatif	-0,11	0,06	-0,06	-0,06	0,17	-0,01	0,06	0,11	0,06
	$\sqrt{EQM}$ relative	0,12	0,16	0,09	0,08	0,24	0,06	0,06	0,12	0,06

Estimateurs des écarts-types	Écart-type	85,2	99,1	144,3	504,6	414,8	711,0	0,08	0,17	0,09
	Moyenne	144,2	117,8	222,7	916,2	516,9	1 215,0	0,11	0,25	0,13
	Biais relatif	0,69	0,19	0,54	0,81	0,25	0,71	0,45	0,49	0,41
	$\sqrt{EQM}$ relative	0,75	0,29	0,60	0,87	0,33	0,76	0,48	0,51	0,44
Intervalles conf. 95 %	Prob. couverture	0,78	0,98	0,93	0,96	0,96	0,99	0,08	0,70	0,40
	Longueur relative	0,31	0,67	0,35	0,35	0,77	0,37	0,08	0,26	0,10

Estimateurs de multiplicité		Tailles			Totaux			Moyennes		
		$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\tilde{Y}_1$	$\tilde{Y}_2$	$\tilde{Y}$	$\tilde{\bar{Y}}_1$	$\tilde{\bar{Y}}_2$	$\tilde{\bar{Y}}$
Estimateurs des paramètres de population	Paramètres de pop.	1 800	687	2 487	10 162	2 631	12 793	5,65	3,83	5,14
	Moyenne	1 579,5	731,0	2 310,4	9 581,8	3 157,1	12 738,8	6,1	4,3	5,5
	Biais relatif	-0,12	0,06	-0,07	-0,06	0,20	-0,00	0,08	0,13	0,07
	$\sqrt{\text{EQM}}$ relative	0,13	0,16	0,09	0,08	0,25	0,06	0,08	0,14	0,08
Estimateurs des écarts-types	Écart-type	83,1	99,0	142,5	502,3	411,6	704,2	0,09	0,16	0,10
	Moyenne	149,1	124,0	229,1	945,9	543,4	1 248,9	0,11	0,26	0,14
	Biais relatif	0,80	0,25	0,61	0,88	0,32	0,77	0,27	0,63	0,37
	$\sqrt{\text{EQM}}$ relative	0,84	0,33	0,66	0,93	0,39	0,82	0,32	0,64	0,40
Intervalles conf. 95 %	Prob. couverture	0,75	0,98	0,93	0,98	0,96	0,99	0,02	0,53	0,19
	Longueur relative	0,33	0,71	0,36	0,35	0,81	0,38	0,08	0,27	0,11

Les résultats de l'étude sont indiqués dans le tableau 4-1. Nous constatons que les estimateurs de Horvitz-Thompson des tailles, des totaux et des moyennes présentaient certains problèmes de biais, bien que la plupart des valeurs des biais relatifs étaient, en valeur absolue, inférieures ou proches de 0,1, sauf celle de l'estimateur $\hat{Y}_2$, qui était une valeur relativement grande. En général, les valeurs des racines carrées des erreurs quadratiques moyennes relatives des estimateurs étaient acceptables, c'est-à-dire qu'elles étaient inférieures ou proches de 0,1, sauf celles des estimateurs de τ_2 et Y_2, qui étaient assez grandes. Ainsi, de façon générale, les performances des estimateurs de Horvitz-Thompson étaient acceptables. Pour ce qui est des estimateurs des écarts-types des estimateurs de Horvitz-Thompson, ils présentaient de graves problèmes de surestimation. Dans le cas des intervalles de confiance de 95 % des tailles et des totaux de population, ils présentaient des valeurs acceptables des probabilités de couverture, sauf l'intervalle de τ_1, qui avait une valeur relativement petite de la probabilité de couverture. Les longueurs relatives de ces intervalles étaient aussi acceptables, sauf celles des intervalles de τ_2 et Y_2, qui étaient relativement grandes. Toutefois, les valeurs des probabilités de couverture étaient très faibles pour les intervalles des moyennes de population. Le problème était que leurs longueurs étaient très petites, ce qui, conjugué aux petits biais des estimateurs ponctuels des moyennes, a produit les valeurs très faibles des probabilités de couverture. Néanmoins, même si l'on tient compte des très petites valeurs des probabilités de couverture de ces intervalles, nous pensons qu'elles fournissent une bonne information sur les moyennes parce que les intervalles sont très courts et assez proches des valeurs réelles des moyennes. Enfin, en ce qui concerne les estimateurs de multiplicité des tailles, des totaux et des moyennes, les estimateurs de leurs écarts-types et leurs intervalles de confiance correspondants de 95 %, nous pouvons dire que leurs performances étaient similaires, mais légèrement inférieures à celles des estimateurs de Horvitz-Thompson.

Remerciements

Cette recherche a été soutenue par la subvention PROFAPI-2022, PRO_A1_014 de l'Universidad Autonoma de Sinaloa.

Bibliographie

Birnbaum, Z.W. et M.G. Sirken (1965), « Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates », *Vital and Health Statistics*, Ser. 2, nº 11. Washington: Government Printing Office.

Félix-Medina, M.H. et S.K. Thompson (2004), « Combining Cluster Sampling and Link-Tracing Sampling to Estimate the Size of Hidden Populations », *Journal of Official Statistics*, 20, p. 19-38.

Félix-Medina, M.H. et P.E. Monjardin (2006), « Combining Link-Tracing Sampling and Cluster Sampling and to Estimate the Size of Hidden Populations: a Bayesian-Assisted Approach », 32, p. 187-195.

Félix-Medina, M.H. et P.E. Monjardin (2010), « Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations », *Journal of Official Statistics*, 26, p. 603-631.

Félix-Medina, M.H., P.E. Monjardin, et A.N. Aceves Castro (2015), « Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities », *Survey Methodology*, 41, p. 349-376.

Félix-Medina, M.H. (2021), « Combining Cluster Sampling and Link-Tracing Sampling to Estimate Totals and Means of Hidden Populations in Presence of Heterogeneous Probabilities of Links », *Journal of Official Statistics*, 37, p. 865-905.

Harris, K.M. (2013), « The Add Health Study: Design and Accomplishments », rapport non publié. Disponible à l'adresse :
htpps://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf.

Lee, J.C. et D.J. Sabavala (1987), « Bayesian Estimation and Prediction for the Beta-Binomial Model », *Journal of Business and Economic Statistics*, 5, p. 357-367.

R Core Team (2022), « R: A Language and Environment for Statistical Computing ». Vienne, Autriche : R Foundation for Statistical Computing. Disponible à : https://www.R-project.org.

Tourangeau, R. (2014), « Defining hard-to-survey populations », dans R. Tourangeauet et coll. (eds.*) Hard-to-Survey Populations*, Cambridge : Cambridge University Press, p. 3-20.