Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 207-216
Statistique Canada, N° 12-001-XPB au catalogue

Combinaison de l'échantillonnage par dépistage de liens et de l'échantillonnage en grappes pour estimer la taille de populations cachées : Une approche assistée par la méthode bayésienne

Martín H. Félix-Medina et Pedro E. Monjardin [1]

Résumé

Félix-Medina et Thompson (2004) ont proposé une variante de l'échantillonnage par dépistage de liens dans laquelle on suppose qu'une part de la population (qui n'est pas nécessairement la plus grande) est couverte par une liste d'emplacements disjoints où les membres de la population peuvent être trouvés avec une probabilité élevée. Après la sélection d'un échantillon d'emplacements, on demande aux personnes se trouvant à chacun de ces emplacements de nommer d'autres membres de la population. Les deux auteurs ont proposé des estimateurs du maximum de vraisemblance des tailles de population qui donnent des résultats acceptables à condition que, pour chaque emplacement, la probabilité qu'un membre de la population soit nommé par une personne se trouvant à cet emplacement, appelée probabilité de nomination, ne soit pas faible. Dans la présente étude, nous partons de la variante de Félix-Medina et Thompson, et nous proposons trois ensembles d'estimateurs des tailles de population dérivés sous une approche bayésienne. Deux des ensembles d'estimateurs sont obtenus en utilisant des lois a priori incorrectes des tailles de population, et l'autre en utilisant des lois a priori de Poisson. Cependant, nous n'utilisons la méthode bayésienne que pour faciliter la construction des estimateurs et adoptons l'approche fréquentiste pour faire les inférences au sujet des tailles de population. Nous proposons deux types d'estimateurs de variance et d'intervalles de confiance partiellement fondés sur le plan de sondage. L'un d'eux est obtenu en utilisant un bootstrap et l'autre, en suivant la méthode delta sous l'hypothèse de normalité asymptotique. Les résultats d'une étude par simulation indiquent que i) quand les probabilités de nomination ne sont pas faibles, chacun des ensembles d'estimateurs proposés donne de bon résultats et se comporte de façon fort semblable aux estimateurs du maximum de vraisemblance, ii) quand les probabilités de nomination sont faibles, l'ensemble d'estimateurs dérivés en utilisant des lois a priori de Poisson donne encore des résultats acceptables et ne présente pas les problèmes de biais qui caractérisent les estimateurs du maximum de vraisemblance et iii) les résultats précédents ne dépendent pas de la taille de la fraction de la population couverte par la base de sondage.

Mots clés : Approche bayésienne; capture-recapture; approche fondée sur le plan de sondage; population finie; population d'accès difficile; maximum de vraisemblance; approche fondée sur un modèle; base de sondage.

1. Introduction

L'échantillonnage par dépistage de liens (EDP) s'est avéré être une méthode convenant bien à l'échantillonnage de populations humaines cachées ou d'accès difficile, comme les drogués, les sans abri ou les travailleurs clandestins. Il consiste à sélectionner un échantillon initial de personnes parmi la population cible et de demander à ces personnes de nommer d'autres membres de la population. Les personnes nommées qui ne figurent pas déjà dans l'échantillon initial sont alors incluses dans l'échantillon et il peut leur être demandé de nommer d'autres personnes. Ce processus se poursuit jusqu'à ce qu'une règle d'arrêt préétablie soit satisfaite (pour une revue de l'échantillonnage par dépistage de liens, voir Spreen 1992, ainsi que Thompson et Frank 2000).

Bien que l'échantillonnage par dépistage de liens permette à l'échantillonneur de faire des inférences fondées sur un modèle valides au sujet d'un certain nombre de paramètres de population, en pratique, les hypothèses concernant l'échantillon initial sont difficiles à satisfaire. (Voir Snijders 1992, Frank et Snijders 1994, et Heckathorn 2002). Par exemple, Frank et Snijders (1994) ont élaboré une variante de l'échantillonnage par dépistage de liens dans laquelle l'échantillon initial est un échantillon de Bernoulli, c'est-à-dire que les éléments de l'échantillon initial sont sélectionnés indépendamment et avec probabilités égales; toutefois, dans les études réelles, le recrutement initial est généralement réalisé au moyen de dossiers de personnes fournis par des centres de soins de santé et des postes de police, ce qui introduit un biais de sélection appelé biais institutionnel.

La difficulté à satisfaire les hypothèses au sujet de l'échantillon initial dans les situations pratiques ont poussé Félix-Medina et Thompson (2004) à élaborer une variante de l'échantillonnage par dépistage de liens qui ne nécessite pas d'échantillon de Bernoulli initial. Ils supposent qu'une part, qui n'est pas nécessairement la plus grande, de la population cible est couverte par une base de sondage constituée d'emplacements accessibles où les membres de la population peuvent être trouvés avec une forte probabilité (par exemple, bars, hôpitaux, îlots d'habitations ou parcs).

1. Martin H. Félix-Medina et Pedro E. Monjardin, Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México.

Un échantillon aléatoire simple d'emplacements est sélectionné et les membres de la population appartenant à chaque emplacement sont identifiés. Enfin, comme dans l'échantillonnage par dépistage de liens ordinaire, il est demandé aux personnes se trouvant à chaque emplacement de nommer d'autres membres de la population.

Ces auteurs ont dérivé des estimateurs du maximum de vraisemblance (EMV) des tailles de population à partir de modèles probabilistes qui décrivent le nombre d'éléments découverts à chaque emplacement, ainsi que la probabilité qu'un membre de la population soit nommé à un emplacement, à laquelle on donne le nom de probabilité de nomination. Ils proposent aussi des estimateurs de variance fondés sur un modèle et partiellement fondés sur le plan de sondage, c'est-à-dire des estimateurs fondés à la fois sur le plan de sondage utilisé pour sélectionner l'échantillon initial et sur les modèles hypothétiques. Tout au long de l'article, nous parlerons d'estimateur « de type fondé sur le plan » pour faire référence à ce genre d'estimateur. Grâce à une étude par simulation, les auteurs ont montré que les EMV des tailles de population et leurs estimateurs de variance de type fondé sur le plan sont robustes aux écarts par rapport au modèle hypothétique, mais que les estimateurs de variance fondés sur un modèle ne sont pas robustes. En outre, ils ont constatés que les EMV ont tendance à surestimer gravement la taille de population si les probabilités de nomination sont faibles.

Comme l'ont indiqué ces auteurs, le problème de la surestimation qui se manifeste lorsque les probabilités de nomination sont faibles est dû à la petite quantité d'information que contient l'échantillon, quantité qui n'est pas suffisante pour obtenir des estimations stables des probabilités de nomination. Selon eux, un remède éventuel à ce problème consiste à suivre l'approche bayésienne pour construire des estimateurs dans lesquels sont intégrés des renseignements supplémentaires au sujet des paramètres de population.

Ici, nous utilisons la méthode bayésienne pour faciliter la construction des estimateurs des tailles de population, mais nous faisons les inférences selon une approche fréquentiste. Donc, en plus de calculer des estimateurs ponctuels, nous construisons des intervalles de confiance. Nous adoptons pour cela la stratégie proposée par Félix-Medina et Thompson (2004) en vue de construire des intervalles de confiance basés sur la loi normale et utilisons des estimateurs de variance de type fondé sur le plan obtenus par la méthode delta. En outre, nous construisons des intervalles de confiance bootstrap de type fondé sur le plan. Nous disons que cette approche inférentielle est « assistée par la méthode bayésienne ».

2. Plan d'échantillonnage et notation

La structure de la population et le plan d'échantillonnage que nous considérons dans le présent article sont les mêmes que ceux proposés par Félix-Medina et Thompson (2004). En voici une brève description. Soit $U = \{u_1, \ldots, u_\tau\}$ une population humaine cachée de taille inconnue τ. Soit U_1 un sous-ensemble de U formé par un nombre inconnu τ_1 de personnes que l'on peut trouver dans différents emplacements accessibles, comme des bars, des parcs ou des îlots d'habitations. Ce plan d'échantillonnage s'appuie sur les hypothèses qu'il est possible de construire une base de sondage de N de ces emplacements et que le chercheur a établi une règle opérationnelle qui lui permet de déterminer si une personne appartient ou non à un emplacement figurant dans la base de sondage et, dans l'affirmative, de situer cet emplacement. Soulignons que l'on ne suppose pas que le sous-ensemble U_1 couvert par la base de sondage représente la partie principale de U et que, comme dans l'échantillonnage en grappes ordinaire, on suppose qu'une personne figurant dans la base de sondage n'appartient qu'à un seul emplacement. Soit A_i le i^e emplacement ou grappe dans la base de sondage et m_i le nombre de personnes qui appartiennent à $A_i, i = 1, \ldots, N$; alors $\tau_1 = \sum_1^N m_i$. Enfin, soit $U_2 = U - U_1$ la partie de U non couverte par la base de sondage et soit $\tau_2 = \tau - \tau_1$ sa taille.

Le plan d'échantillonnage est le suivant. Un échantillon $S_0 = \{A_1, \ldots, A_n\}$ de n grappes est tiré à partir de la base de sondage par échantillonnage aléatoire simple sans remise, et les m_i personnes qui appartiennent à chaque $A_i \in S_0$ sont identifiées. Notons que nous avons utilisé les indices $1, \ldots, n$ pour dénoter les grappes dans S_0; toutefois, cela ne signifie pas que les n premières grappes dans la base de sondage sont nécessairement les grappes contenues dans l'échantillon. Puis, on demande aux personnes comprises dans la grappe échantillonnée A_i de nommer des membres de U, mais seules les personnes nommées comprises dans $U - A_i$ sont prises en considération. Cette procédure est répétée pour chaque grappe $A_i \in S_0$. Par convention, nous dirons qu'une personne est nommée par une grappe si elle est nommée par au moins un membre de cette grappe. Les nominations à partir des diverses grappes sont faites indépendamment les unes des autres, et diverses stratégies de nomination peuvent être utilisées à différents emplacements. Par exemple, à l'emplacement A_i, les nominations pourraient être faites par les m_i membres, en tant que groupe, tandis que dans un autre emplacement, A_j, chacun des m_j membres pourrait faire des nominations séparément. Enfin, pour chaque personne nommée, le chercheur doit enregistrer le ou les emplacements qui l'ont nommé, ainsi que le segment U_1 ou U_2 de la population auquel elle appartient. Il convient de souligner que ce dernier élément

d'information peut être obtenu auprès de la personne qui a fait la nomination ou, si cela est impossible, durant une interview de la personne nommée.

La nomination de personnes par les grappes sera indiquée au moyen des matrices $\mathbf{X}_1 = [x_{ij}^{(1)}]_{n\times\tau_1}$ et $\mathbf{X}_2 = [x_{ij}^{(2)}]_{n\times\tau_2}$, où $x_{ij}^{(1)} = 1$ si la personne $u_j \in U_1 - A_i$ est nommée par la grappe A_i, et $x_{ij}^{(1)} = 0$ si $u_j \in A_i$ ou u_j n'est pas nommée par A_i. De même, $x_{ij}^{(2)} = 1$ si la personne $u_j \in U_2$ est nommée par la grappe A_i, et $x_{ij}^{(2)} = 0$ autrement. Comme l'on souligné Félix-Medina et Thompson (2004), $\mathbf{X}_1$ et $\mathbf{X}_2$ ne sont connues que jusqu'aux permutations de leurs colonnes, car les personnes ne sont pas étiquetées. Par conséquent, les inférences au sujet de τ_1 et τ_2 sont basées sur l'ensemble des dénombrements $\mathbf{y} = \{y_\omega\}$, où y_ω, $\omega \subseteq \mathbf{\Omega} = \{1, \ldots, n\}, \omega \neq \varnothing$, indique le nombre de personnes dans U qui sont nommées par chaque grappe échantillonnée A_i pour laquelle i est compris dans l'ensemble ω, mais autrement non. Par exemple, si $\omega = \{4, 7, 8\}$, y_ω serait le nombre de personnes dans U qui sont nommées par A_4, A_7 et A_8 uniquement.

3. Estimateurs des tailles de population basés sur les modes a posteriori

Félix-Medina et Thompson ont constaté la ressemblance entre leur plan d'échantillonnage et celui de l'échantillonnage par capture-recapture multiple (ECRM). Cela nous permet d'appliquer à notre cas certains des modèles bayésiens qui ont été proposés pour l'analyse de l'ECRM. Voir Fienberg, Johnson et Junker (1999) pour une revue des analyses bayésienne de l'ECRM. Dans la présente étude, nous utilisons un modèle pris en considération par Castledine (1981) pour les lois a priori des logits des probabilités de nomination, ainsi que certains modèles pour les lois a priori des tailles de population.

À l'instar de Félix-Medina et Thompson (2004), nous supposons que les tailles $m_1, \ldots, m_N$ des grappes $A_1, \ldots, A_N$ sont des réalisations de variables aléatoires de Poisson indépendantes $M_1, \ldots, M_N$ de moyenne λ_1. Nous dénotons par $p_i^{(k)}$ la probabilité qu'une personne comprise dans $U_k - A_i$ soit nommée par l'emplacement $A_i \in S_0$. Les probabilités $p_i^{(k)}$ sont appelées probabilités de nomination. En outre, nous supposons que, conditionnellement aux tailles $m_1, \ldots, m_n$ des grappes dans S_0, à τ_1 et τ_2, ainsi qu'aux $p_i^{(k)}$, les variables $x_{ij}^{(k)}$ sont des réalisations de variables aléatoires de Bernoulli indépendantes $X_{ij}^{(k)}$ de moyenne $p_i^{(k)}$, $i = 1, \ldots, n$ et $k = 1, 2$.

Félix-Medina et Thompson (2004) ont utilisé le fait que la loi conditionnelle conjointe de $(M_1, \ldots, M_n, \tau_1 - \sum_1^n M_i)$, sachant que $\sum_1^N m_i = \tau_1$, est une loi multinomiale dont les paramètres sont τ_1 et $(1/N, \ldots, 1/N, 1 - n/N)$, et ont appliqué une méthode utilisée par Darroch (1958) pour montrer que la fonction de vraisemblance de $\tau_1, \tau_2, \mathbf{p}_1 = \{p_i^{(1)}\}_1^n$ et $\mathbf{p}_2 = \{p_i^{(2)}\}_1^n$ est le produit des facteurs suivants :

$$f(\mathbf{m}_s|\tau_1) = \frac{\tau_1!}{(\tau_1 - m)!\prod_1^n m_i!}(1/N)^m(1 - n/N)^{\tau_1 - m}$$

$$f(\mathbf{y}^{(1-0)}|\mathbf{m}_s, \tau_1, \mathbf{p}_1) = \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!\prod_{\omega\neq\varnothing} y_\omega^{(1-0)}!}\prod_{i=1}^n [p_i^{(1)}]^{z_i^{(1-0)}} \times [1 - p_i^{(1)}]^{\tau_1 - m - z_i^{(1-0)}}$$

$$f(\mathbf{y}^{(A_1)}, \ldots, \mathbf{y}^{(A_n)}|\mathbf{m}_s, \mathbf{p}_1) = \prod_{i=1}^n \frac{m_i!}{(m_i - w_i)!\prod_{\omega\neq\varnothing} y_\omega^{(A_i)}!}[p_i^{(1)}]^{z_i^{(0)}} \times [1 - p_i^{(1)}]^{m - m_i - z_i^{(0)}}$$

$$f(\mathbf{y}^{(2)}|\mathbf{m}_s, \tau_2, \mathbf{p}_2) = \frac{\tau_2!}{(\tau_2 - r_2)!\prod_{\omega\neq\varnothing} y_\omega^{(2)}!}\prod_{i=1}^n [p_i^{(2)}]^{z_i^{(2)}}[1 - p_i^{(2)}]^{\tau_2 - z_i^{(2)}},$$

où $\mathbf{m}_s = \{m_i\}_1^n; m = \sum_1^n m_i$ est la valeur observée de la variable aléatoire M qui indique le nombre de personnes dans S_0; $\mathbf{y}^{(1-0)} = \{y_\omega^{(1-0)}\}_{\omega\neq\varnothing}, \mathbf{y}^{(2)} = \{y_\omega^{(2)}\}_{\omega\neq\varnothing}$, et $\mathbf{y}^{(A_i)} = \{y_\omega^{(A_i)}\}_{\omega\neq\varnothing}, A_i \in S_0$, sont les ensembles de dénombrements obtenus à partir de $\mathbf{y}$, qui correspondent aux dénombrements de personnes nommées dans $U_1 - S_0, U_2$ et $A_i \in S_0$, respectivement; $z_i^{(0)} = \sum_{j\neq i}^n \sum_{\omega\supset i} y_\omega^{(A_j)}$, $z_i^{(1-0)} = \sum_{\omega\supset i} y_\omega^{(1-0)}$ et $z_i^{(2)} = \sum_{\omega\supset i} y_\omega^{(2)}$ sont les valeurs observées des variables aléatoires $Z_i^{(0)}, Z_i^{(1-0)}$ et $Z_i^{(2)}$ qui indiquent le nombre de personnes distinctes dans $S_0, U_1 - S_0$ et U_2, respectivement, qui sont nommées par A_i; et $r_1 = \sum_{\omega\neq\varnothing} y_\omega^{(1-0)}, r_2 = \sum_{\omega\neq\varnothing} y_\omega^{(2)}$ et $w_i = \sum_{\omega\neq\varnothing} y_\omega^{(A_i)}$ sont les valeurs observées des variables aléatoires R_1, R_2 et W_i qui indiquent le nombre de personnes distinctes dans $U_1 - S_0, U_2$ et A_i, respectivement, qui sont nommées par au moins une des grappes comprises dans S_0.

Nous allons maintenant nous pencher sur le problème de la définition des lois a priori de $\tau_1, \tau_2, \mathbf{p}_1$ et $\mathbf{p}_2$. Dans le cas de τ_1 et τ_2, nous considérerons les trois modèles qui suivent pour les lois a priori :

Lois de Poisson-Gamma

$\pi(\tau_1|\lambda_1) \propto (N\lambda_1)^{\tau_1}/\tau_1!$ et $\pi(\lambda_1) \propto \lambda_1^{a_1 - 1}e^{-b_1\lambda_1}$,
$\pi(\tau_2|\lambda_2) \propto \lambda_2^{\tau_2}/\tau_2!$ et $\pi(\lambda_2) \propto \lambda_2^{a_2 - 1}e^{-b_2\lambda_2}$,
où a_1, b_1, a_2, b_2 sont des constantes connues, et (τ_1, λ_1) et (τ_2, λ_2) sont indépendants.

Lois de Jeffreys

$\pi(\tau_k) \propto 1/\tau_k$, où $k = 1, 2$, et τ_1 et τ_2 sont des variables aléatoires indépendantes.

Lois uniformes

$\pi(\tau_k) \propto 1$, où $k = 1, 2$, et τ_1 et τ_2 sont des variables aléatoires indépendantes.

La loi a priori de Poisson de τ_1 définie dans le premier cas a pour motivation le fait que $\tau_1 = \sum_1^N M_i$, et que M_i est une variable de Poisson de moyenne λ_1. Soulignons que ce cas permet aux chercheurs d'utiliser l'information au sujet de τ_1 et τ_2 qui est connue avant l'observation de l'échantillon. Par ailleurs, les lois définies dans les deux autres cas ne sont pas informatives.

Dans le cas des probabilités de nomination $p_i^{(k)}$, à l'instar de Castledine (1981), nous supposons qu'elles sont échangeables et nous utiliserons le modèle normal à deux degrés pour les logits $\alpha_i^{(k)} = \log[p_i^{(k)}/(1-p_i^{(k)})]$ des $p_i^{(k)}$:

$$\alpha_i^{(k)} | \theta_k \sim N(\theta_k, \sigma_k^2),$$

et

$$\theta_k \sim N(\mu_k, \gamma_k^2); i = 1, \ldots, n, \ k = 1, 2, \qquad (1)$$

où $N(\theta_k, \sigma_k^2)$ représente la loi normale de moyenne θ_k et de variance σ_k^2; σ_k^2, μ_k et γ_k^2 sont des constantes connues; et les $\alpha_i^{(k)}$ sont conditionnellement indépendants sachant θ_k. Sous l'hypothèse d'échangeabilité, les $\alpha_i^{(k)}$ ne sont pas indépendants, mais l'information au sujet de n'importe lequel d'entre eux est utilisée pour obtenir des renseignements sur n'importe lequel des $\alpha_i^{(k)}$. Naturellement, si nous voulions des lois a priori indépendantes pour les $\alpha_i^{(k)}$, nous pourrions obtenir un modèle normal à un degré à partir de (1) en fixant $\theta_k = \mu_k$ et $\gamma_k^2 = 0, k = 1, 2$.

Enfin, nous supposerons que tous les vecteurs aléatoires (τ_k, λ_k) et (α_k, θ_k), où $\alpha_k = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, $k = 1, 2$, sont mutuellement indépendants.

Bien que nous ayons défini trois types de loi a priori pour τ_1 et τ_2, elles peuvent être traitées de manière uniforme, parce que les lois marginales a priori de τ_1 et τ_2, obtenues à partir des lois de Poisson-Gamma, sont les lois binomiales négatives :

$$\pi(\tau_1) \propto \frac{\Gamma(\tau_1 + a_1)}{\tau_1!}\left(\frac{N}{N + b_1}\right)^{\tau_1}$$

et

$$\pi(\tau_2) \propto \frac{\Gamma(\tau_2 + a_2)}{\tau_2!}\left(\frac{1}{1 + b_2}\right)^{\tau_2}, \qquad (2)$$

où $\Gamma(\cdot)$ dénote la fonction Gamma. Les lois de Jeffreys et les lois uniformes sont des cas limites de (2) obtenus en supposant que $a_k = b_k = 0, k = 1, 2,$ et $a_k = 1, b_k = 0, k = 1, 2,$ respectivement. Notons que la loi Gamma n'est pas définie pour ces valeurs de a_k et b_k; toutefois, pour la dérivation des estimateurs, nous pouvons utiliser ces valeurs dans (2).

La loi conjointe a posteriori de τ_1, τ_2, α_1 et α_2 peut être exprimée sous la forme

$$\pi(\tau_1, \tau_2, \alpha_1, \alpha_2 \,|\, \text{données})$$

$$\propto \frac{(N-n)^{\tau_1}\Gamma(\tau_1 + a_1)}{(\tau_1 - m - r_1)!(N + b_1)^{\tau_1}} \prod_{i=1}^{n} \frac{\exp[\alpha_i^{(1)} z_i^{(1)}]}{\left[1 + \exp[\alpha_i^{(1)}]\right]^{\tau_1 - m_i}}$$

$$\times \exp\left[\begin{array}{l} -\dfrac{\sum_{i=1}^{n}(\alpha_i^{(1)} - \bar{\alpha}^{(1)})^2}{2\sigma_1^2} \\ -\dfrac{(\bar{\alpha}^{(1)} - \mu_1)^2}{2v_1} \end{array}\right] \frac{\Gamma(\tau_2 + a_2)}{(\tau_2 - r_2)!(b_2 + 1)^{\tau_2}}$$

$$\times \prod_{i=1}^{n} \frac{\exp[\alpha_i^{(2)} z_i^{(2)}]}{\left[1 + \exp[\alpha_i^{(2)}]\right]^{\tau_2}} \exp\left[\begin{array}{l} -\dfrac{\sum_{i=1}^{n}(\alpha_i^{(2)} - \bar{\alpha}^{(2)})^2}{2\sigma_2^2} \\ -\dfrac{(\bar{\alpha}^{(2)} - \mu_2)^2}{2v_2} \end{array}\right] \qquad (3)$$

où $z_i^{(1)} = z_i^{(0)} + z_i^{(1-0)}$ est la valeur observée de la variable aléatoire $Z_i^{(1)} = Z_i^{(0)} + Z_i^{(1-0)}$ qui indique le nombre de personnes distinctes dans U_1, soit dans S_0 ou dans $U_1 - S_0$, qui sont nommées par A_i; $\bar{\alpha}^{(k)}$ est la moyenne arithmétique des $\alpha_i^{(k)}$; et $v_k = \gamma_k^2 + \sigma_k^2/n, k = 1, 2$.

Puisque nous ne pouvons pas calculer l'intégrale analytique de (3) par rapport à $\alpha_i^{(1)}$ et à $\alpha_i^{(2)}$, nous n'essayerons pas d'obtenir les expressions pour les lois a posteriori de τ_1 et τ_2, mais, comme dans Castledine (1981), nous utiliserons le mode de $\pi(\tau_1, \tau_2, \alpha_1, \alpha_2|\text{données})$ comme estimateur de $(\tau_1, \tau_2, \alpha_1, \alpha_2)$. En adoptant cette stratégie, nous avons que l'estimateur proposé est la solution du système d'équations :

$$\hat{\tau}_1 = \frac{M + R_1 + (1 - n/N)[N(a_1 - 1)/(N + b_1)]\prod_{i=1}^{n}(1 - \hat{p}_i^{(1)})}{1 - (1 - n/N)[N/(N + b_1)]\prod_{i=1}^{n}(1 - \hat{p}_i^{(1)})};$$

$$\hat{p}_i^{(1)} = \frac{\exp\{\hat{\alpha}_i^{(1)}\}}{1 + \exp\{\hat{\alpha}_i^{(1)}\}} = \frac{Z_i^{(1)}}{\hat{\tau}_1 - M_i} - \frac{\hat{\alpha}_i^{(1)} - \hat{\bar{\alpha}}^{(1)}}{(\hat{\tau}_1 - M_i)\sigma_1^2}$$

$$- \frac{\hat{\bar{\alpha}}^{(1)} - \mu_1}{n(\hat{\tau}_1 - M_i)v_1}; i = 1, \ldots, n; \qquad (4)$$

$$\hat{\tau}_2 = \frac{R_2 + [(a_2 - 1)/(1 + b_2)]\prod_{i=1}^{n}(1 - \hat{p}_i^{(2)})}{1 - [1/(1 + b_2)]\prod_{i=1}^{n}(1 - \hat{p}_i^{(2)})};$$

$$\hat{p}_i^{(2)} = \frac{\exp\{\hat{\alpha}_i^{(2)}\}}{1 + \exp\{\hat{\alpha}_i^{(2)}\}} = \frac{Z_i^{(2)}}{\hat{\tau}_2} - \frac{\hat{\alpha}_i^{(2)} - \hat{\bar{\alpha}}^{(2)}}{\hat{\tau}_2\sigma_2^2}$$

$$- \frac{\hat{\bar{\alpha}}^{(2)} - \mu_2}{n\hat{\tau}_2 v_2}; i = 1, \ldots, n; \qquad (5)$$

où $\hat{\bar{\alpha}}^{(k)} = \sum_1^n \hat{\alpha}_i^{(k)}/n, k = 1, 2$. Il s'ensuit qu'un estimateur de τ est $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

Les formes de ces estimateurs sont fondamentalement des ajustements des formes de l'EMV proposées par Félix-Medina et Thompson (2004), de sorte qu'est intégrée dans les estimateurs proposés l'information initiale au sujet de τ_k et $\alpha_i^{(k)}, i = 1, \ldots, n;\ k = 1, 2$. En outre, comme la fait remarquer un examinateur, l'estimateur $\hat{p}_i^{(k)}$ a la forme de l'EMV de $p_i^{(k)}$ suivi par des termes de rétrécissement, l'un étant celui de $\alpha_i^{(k)}$ vers la moyenne arithmétique $\hat{\bar{\alpha}}^{(k)}$ et l'autre, celui de $\hat{\bar{\alpha}}^{(k)}$ vers la moyenne a priori μ_k.

4. Intervalles de confiance pour les tailles de population

Comme nous l'avons indiqué plus haut, nous utiliserons l'approche fréquentiste pour obtenir des intervalles de confiance de type fondé sur le plan de sondage qui sont robustes aux écarts par rapport à la loi de Poisson hypothétique des M_i. Nous examinerons des intervalles bootstrap et des intervalles de Wald basés sur une approximation normale (voir Agresti 2002, page 13 et Evans, Kim et O'Brien 1996 pour la terminologie la plus récente).

4.1 Intervalles de confiance bootstrap

Nous utiliserons une version du bootstrap obtenue en combinant la variante du bootstrap pour les populations finies proposée par Booth, Butler et Hall (1994) et la variante paramétrique du bootstrap (voir Davison et Hinkley 1997, chapitre 2).

Les étapes de la méthode que nous proposons sont les suivantes. (i) Construire une population artificielle de N valeurs des m_i en répétant N/n fois, en supposant que N/n est un nombre entier, l'échantillon sélectionné de n tailles de grappe $m_1, \ldots, m_n$. Si $N = kn + r$, où k et r sont des entiers positifs, construire la population en répétant k fois l'échantillon sélectionné de n tailles de grappe et ajouter à cet ensemble de tailles m_i un échantillon aléatoire simple sans remise (EASSR) de r valeurs des m_i sélectionnées à partir de l'échantillon observé de n tailles de grappe. ii) Sélectionner un EASSR de n tailles à partir de la population des m_i. Soit $i_1, \ldots, i_n$ les indices des m_i dans l'échantillon. (iii) Pour chaque $i = i_1, \ldots, i_n$, tirer des échantillons de tailles $\hat{\tau}_1 - m_i$ et $\hat{\tau}_2$ à partir des lois de Bernoulli de moyennes $\hat{p}_i^{(1)}$ et $\hat{p}_i^{(2)}$, respectivement, où $\hat{\tau}_1, \hat{\tau}_2, \hat{p}_i^{(1)}$ et $\hat{p}_i^{(2)}$ sont les estimations de $\tau_1, \tau_2, p_i^{(1)}$ et $p_i^{(2)}$ calculées d'après l'échantillon observé original. Ces échantillons simulent les valeurs des ensembles $\{x_{ij}^{(1)}\}$ et $\{x_{ij}^{(2)}\}$ de variables indicatrices. iv) Calculer les estimations de τ_1, τ_2 et τ à partir des échantillons tirés aux étapes (ii) et (iii) en suivant la même méthode que celle utilisée pour calculer les estimations originales $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. (v) Obtenir les lois bootstrap de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$ en répétant les étapes (i) à (iv) un grand nombre B de fois et en calculant les lois empiriques à partir des ensembles de B valeurs de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. (vi) Construire les intervalles de confiance bootstrap $100(1-\alpha)\,\%$ pour τ_1, τ_2 et τ en utilisant la méthode de base ou celle des centiles (voir Davison et Hinkley 1997, chapitre 5 pour la description de ces méthodes). Dans la méthode de base, l'intervalle pour τ est $[2\hat{\tau} - \hat{\tau}^{(1-\alpha/2)}, 2\hat{\tau} - \hat{\tau}^{(\alpha/2)}]$, et dans la méthode des centiles, il est $[\hat{\tau}^{(\alpha/2)}, \hat{\tau}^{(1-\alpha/2)}]$, où $\hat{\tau}^{(\alpha/2)}$ et $\hat{\tau}^{(1-\alpha/2)}$ sont les points $\alpha/2$ inférieur et supérieur de la distribution bootstrap de l'estimation originale $\hat{\tau}$ de τ.

Notons que cette variante du bootstrap ne repose pas sur l'utilisation de la loi de Poisson hypothétique des M_i, mais sur le plan d'échantillonnage utilisé pour sélectionner l'échantillon initial de grappes. Donc, nous pouvons considérer que les intervalles de confiance résultants sont robustes aux écarts par rapport à la loi hypothétique des M_i.

Si l'on souhaite également calculer les estimations bootstrap des variances de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$, il est possible d'obtenir des estimations simples en calculant les variances d'échantillon des ensembles de B valeurs de ces estimateurs.

4.2 Intervalles de confiance de Wald

Bien que nous ne démontrions pas théoriquement ici que les estimateurs proposés des tailles de population suivent asymptotiquement une loi normale, nous supposerons que la loi normale est une approximation raisonnable des lois des estimateurs. Donc, nous construisons pour les tailles de population des intervalles de confiance de Wald $100(1-\alpha)\,\%$ de type fondé sur le plan de sondage de la forme $\hat{\tau}_k \pm z_{1-\alpha/2}\sqrt{\hat{\mathbf{V}}(\hat{\tau}_k)}$, où $z_{1-\alpha/2}$ est le point $\alpha/2$ supérieur de la loi normale standard, et $\hat{\mathbf{V}}(\hat{\tau}_k)$ est un estimateur de type fondé sur le plan de sondage de la variance de $\hat{\tau}_k$.

Pour construire ce genre d'intervalle, nous commençons par dériver des estimateurs de la variance de type fondé sur le plan de sondage en suivant la même stratégie que celle utilisée par Félix-Medina et Thompson (2004). Celle-ci consiste à remplacer la distribution des tailles de grappes par celle du plan d'échantillonnage utilisé pour sélectionner l'échantillon initial S_0. Nous utilisons pour cela la formule :

$$\mathbf{V}(\hat{\tau}_k) = \mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_k|\mathbf{m}_s)] + \mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_k|\mathbf{m}_s)], \qquad (6)$$

où $\mathbf{E}_\xi(\hat{\tau}_k|\mathbf{m}_s)$ et $\mathbf{V}_\xi(\hat{\tau}_k|\mathbf{m}_s)$ dénotent les opérateurs d'espérance et de variance conditionnelles fondées sur le modèle, sachant que $\mathbf{M}_s = \mathbf{m}_s$; et $\mathbf{E}_p(\cdot)$ et $\mathbf{V}_p(\cdot)$ dénotent les opérateurs d'espérance et de variance fondées sur le plan de sondage. Donc, nous obtenons les estimateurs de variance en appliquant (6) aux approximation de Taylor de

premier ordre $\hat{\tau}_1^*$ et $\hat{\tau}_2^*$ de $\hat{\tau}_1$ et $\hat{\tau}_2$, respectivement, autour des espérances fondées sur le modèle de $c_s^{(1)} = (\mathbf{M}_s, \mathbf{Z}_s^{(1)}, R_1)$ et $c_s^{(2)} = (\mathbf{Z}_s^{(2)}, R_2)$, où $\mathbf{Z}_s^{(k)} = (Z_1^{(k)}, \ldots Z_n^{(k)})$, $k = 1, 2$.

En utilisant la stratégie décrite antérieurement et le fait que $Z_i^{(1)}|\mathbf{m}_s \sim \text{bin}(\tau_1 - m_i, p_i^{(1)})$ et $R_1|\mathbf{m}_s \sim \text{bin}(\tau_1 - m, 1 - Q_1)$, où $Q_1 = \prod_{i=1}^n (1 - p_i^{(1)})$, nous avons qu'un estimateur de $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_1^*|\mathbf{m}_s)]$ est

$$\hat{\mathbf{V}}_{11} = n(1 - n/N)\hat{K}^2 \frac{1}{n-1}\sum_{i=1}^n (m_i - \bar{m})^2, \tag{7}$$

où $\bar{m} = n^{-1}\sum_1^n m_i$; $\hat{K} = -\hat{Q}_1/[\hat{A}_1(\hat{\tau}_1 - m - r_1)]$; $\hat{Q}_1 = \prod_{i=1}^n (1 - \hat{p}_i^{(1)})$;

$$\hat{A}_1 = \sum_{i=1}^n \frac{(\hat{p}_i^{(1)})^2}{\hat{B}_i^{(1)}} - \hat{C}_1 + \frac{1}{\hat{\tau}_1 + a_1 - 1} - \frac{1}{\hat{\tau}_1 - m - r_1};$$

$$\hat{B}_i^{(1)} = (\hat{\tau}_1 - m_i)\hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)}) + \sigma_1^{-2}, \; i = 1, \ldots, n;$$

et

$$\hat{C}_1 = \frac{(v_1^{-1} - n\sigma_1^{-2})\left[n^{-1}\sum_{i=1}^n \hat{p}_i^{(1)}/\hat{B}_i^{(1)}\right]^2}{1 + n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(1)}}. \tag{8}$$

En outre, puisque $\mathbf{Cov}(Z_i^{(1)}, R_1|\mathbf{m}_s) = (\tau_1 - m)Q_1 p_i^{(1)}$, un estimateur de $\mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_1^*|\mathbf{m}_s)]$ a la forme

$$\hat{\mathbf{V}}_{12} = \hat{A}_1^{-2}\left\{\begin{array}{l} \displaystyle\sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}}\right)^2 (\hat{\tau}_1 - m_i)\hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)}) \\ \displaystyle + \frac{(\hat{\tau}_1 - m)\hat{Q}_1(1 - \hat{Q}_1)}{(\hat{\tau}_1 - m - r_1)^2} \\ \displaystyle - \frac{2(\hat{\tau}_1 - m)\hat{Q}_1}{\hat{\tau}_1 - m - r_1}\sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}}\right)\hat{p}_i^{(1)} \end{array}\right\}, \tag{9}$$

où

$$\hat{D}_1 = \frac{n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n \hat{p}_i^{(1)}/\hat{B}_i^{(1)}}{1 + n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(1)}}.$$

Par conséquent, un estimateur de type fondé sur le plan de sondage de $\mathbf{V}(\hat{\tau}_1)$ est $\hat{\mathbf{V}}(\hat{\tau}_1) = \hat{\mathbf{V}}_{11} + \hat{\mathbf{V}}_{12}$.

Dans le cas de $\hat{\tau}_2^*$, puisque $Z_i^{(2)}|\mathbf{m}_s \sim \text{bin}(\tau_2, p_i^{(2)})$ et $R_2|\mathbf{m}_s \sim \text{bin}(\tau_2, 1 - Q_2)$, où $Q_2 = \prod_{i=1}^n (1 - p_i^{(2)})$, il s'ensuit que $\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)$ ne dépend pas de $\mathbf{m}_s$, et conséquemment que $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)] \approx 0$. Donc, puisque $\mathbf{Cov}(Z_i^{(2)}, R_2|\mathbf{m}_s) = \tau_2 Q_2 p_i^{(2)}$, un estimateur de $\mathbf{V}(\hat{\tau}_2)$ est

$$\hat{\mathbf{V}}(\hat{\tau}_2) = \hat{A}_2^{-2}\left\{\begin{array}{l} \displaystyle\sum_{i=1}^n \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}}\right)^2 \hat{\tau}_2 \hat{p}_i^{(2)}(1 - \hat{p}_i^{(2)}) \\ \displaystyle + \frac{\hat{\tau}_2\hat{Q}_2(1 - \hat{Q}_2)}{(\hat{\tau}_2 - r_2)^2} \\ \displaystyle - \frac{2\hat{\tau}_2\hat{Q}_2}{\hat{\tau}_2 - r_2}\sum_{i=1}^n \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}}\right)\hat{p}_i^{(2)} \end{array}\right\} \tag{10}$$

où $\hat{Q}_2 = \prod_{i=1}^n (1 - \hat{p}_i^{(2)})$,

$$\hat{A}_2 = \sum_{i=1}^n \frac{(\hat{p}_i^{(2)})^2}{\hat{B}_i^{(2)}} - \hat{C}_2 + \frac{1}{\hat{\tau}_2 + a_2 - 1} - \frac{1}{\hat{\tau}_2 - r_2},$$

$$\hat{B}_i^{(2)} = \hat{\tau}_2 \hat{p}_i^{(2)}(1 - \hat{p}_i^{(2)}) + \sigma_2^{-2}, \quad i = 1, \ldots, n,$$

$$\hat{C}_2 = \frac{(v_2^{-1} - n\sigma_2^{-2})\left[n^{-1}\sum_{i=1}^n \hat{p}_i^{(2)}/\hat{B}_i^{(2)}\right]^2}{1 + n^{-1}(v_2^{-1} - n\sigma_2^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(2)}},$$

et

$$\hat{D}_2 = \frac{n^{-1}(v_2^{-1} - n\sigma_2^{-2})n^{-1}\sum_{i=1}^n \hat{p}_i^{(2)}/\hat{B}_i^{(2)}}{1 + n^{-1}(v_2^{-1} - n\sigma_2^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(2)}}.$$

Enfin, puisque la non-dépendance de $\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)$ par rapport à $\mathbf{m}_s$ implique que $\mathbf{Cov}(\hat{\tau}_1^*, \hat{\tau}_2^*) \approx 0$, il s'ensuit qu'un estimateur de la variance de $\hat{\tau}$ est $\hat{\mathbf{V}}(\hat{\tau}) = \hat{\mathbf{V}}(\hat{\tau}_1) + \hat{\mathbf{V}}(\hat{\tau}_2)$.

5. Étude de Monte Carlo

Nous avons considéré quatre populations, décrites chacune au tableau 1. Dans la paire formée par les populations I et II, la base de sondage couvrait environ 45 % de la population, tandis que dans la paire formée par les populations III et IV, elle couvrait environ 70 % de la population. Les populations de chaque paire étaient fort semblables, sauf le fait que, dans l'une des populations de chaque paire, la loi des M_i était une loi de Poisson, tandis que dans l'autre il s'agissait d'une loi binomiale négative. Les probabilités de nomination $p_i^{(k)}, i = 1, \ldots, N$, $k = 1, 2$, ont été générées en utilisant le modèle $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, où les valeurs de β_k étaient fixées de façon que les valeurs suivantes de $\bar{p}^{(k)} = \sum_1^N p_i^{(k)}/N$ soient obtenues. Pour les populations I et II : $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,05, 0,01)$ et $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,01, 0,002)$. Pour les populations III et IV : $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,05, 0,03)$ et $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,01, 0,006)$. Le modèle employé pour générer les $p_i^{(k)}$ est un modèle utilisé dans les méthodes prises-effort (voir Seber 1982, chapitre 7 pour une description de ces méthodes). Comme l'a souligné un

rédacteur associé, ce modèle implique que le nombre de personnes nommées par la grappe A_i est l'espérance $(\tau_1 - m_i)\ (1-\exp(-\beta_1 m_i)) + \tau_2(1-\exp(-\beta_2 m_i))$ et que, par conséquent, le nombre de personnes nommées est approximativement proportionnel à m_i. Notons que le modèle échangeable hypothétique pour $p_i^{(k)}$ ne postule pas ce genre de relation avec m_i. Puisque l'estimation de $p_i^{(k)}$ dépend principalement de $z_i^{(k)}$, le nombre de personnes dans U_k nommées par la grappe A_i, nous nous attendons à ce que l'omission de cette relation n'ait pas d'incidence sur l'efficacité de l'estimateur de $p_i^{(k)}$. Darroch (1958) a montré, dans le cas de l'estimation du maximum de vraisemblance, que l'on ne réalise aucun gain significatif en émettant l'hypothèse d'un modèle prises-effort.

Tableau 1
Paramètres des populations simulées

Population I	Population II	Population III	Population IV
$N = 250$	$N = 250$	$N = 250$	$N = 250$
M_i Poisson	M_i Binomiale nég.	M_i Poisson	M_i Binomiale nég.
$\mathbf{E}(M_i) = 7,2$	$\mathbf{E}(M_i) = 7,2$	$\mathbf{E}(M_i) = 7,2$	$\mathbf{E}(M_i) = 7,2$
$\mathbf{V}(M_i) = 7,2$	$\mathbf{V}(M_i) = 24,48$	$\mathbf{V}(M_i) = 7,2$	$\mathbf{V}(M_i) = 24,48$
$\tau_1 = 1\,811$	$\tau_1 = 1\,872$	$\tau_1 = 1\,811$	$\tau_1 = 1\,872$
$\tau_2 = 2\,200$	$\tau_2 = 2\,200$	$\tau_2 = 700$	$\tau_2 = 700$
$\tau = 4\,011$	$\tau = 4\,072$	$\tau = 2\,511$	$\tau = 2\,572$
$\tau_1/\tau = 0,45$	$\tau_1/\tau = 0,46$	$\tau_1/\tau = 0,72$	$\tau_1/\tau = 0,73$

Pour les populations I et II, les valeurs des paramètres des lois a priori étaient $\sigma_k^2 = 25, \mu_k = -3,5, \gamma_k^2 = 25, k = 1, 2, a_1 = 1, b_1 = 0,1, a_2 = 7,84, b_2 = 0,0028,$ de sorte que $\mathbf{E}(\lambda_1) = 10, \mathbf{V}(\lambda_1) = 100, \mathbf{E}(\lambda_2) = 2\,800,$ et $\mathbf{V}(\lambda_2) = 10^6$. Pour les populations III et IV les valeurs des paramètres étaient $\sigma_k^2 = 9, \mu_k = -3,5, \gamma_k^2 = 9, k = 1,2, a_1 = 1, b_1 = 0,1, a_2 = 8, b_2 = 0,01,$ de sorte que $\mathbf{E}(\lambda_1) = 10, \mathbf{V}(\lambda_1) = 100, \mathbf{E}(\lambda_2) = 800$ et $\mathbf{V}(\lambda_2) = 80\,000$. Ces valeurs impliquent que les lois a priori sont bien dispersées sur les intervalles relativement grands qui contiennent les paramètres d'intérêt.

Nous avons réalisé l'expérience par simulation comme il suit. À partir de chaque population de $N = 250$ valeurs de m_i, nous avons sélectionné un EASSR de $n = 25$ valeurs. À partir de la grappe A_i dans l'échantillon, nous avons généré les valeurs de $X_{ij}^{(1)}$ et $X_{ij}^{(2)}$ en tirant des échantillons de taille $\tau_1 - m_i$ et τ_2 à partir de lois de Bernoulli de moyenne $p_i^{(1)}$ et $p_i^{(2)}$, respectivement. Ces données ont été utilisées pour calculer les estimateurs suivants des tailles de population : l'ensemble d'EMV $\tilde{\tau}_1, \tilde{\tau}_2,$ et $\tilde{\tau} = \tilde{\tau}_1 + \tilde{\tau}_2$ proposé par Félix-Medina et Thompson (2004); ainsi que les trois ensembles d'estimateurs bayésiens $\hat{\tau}_1^a, \hat{\tau}_2^a$ et $\hat{\tau}^a = \hat{\tau}_1^a + \hat{\tau}_2^a,\ a = U, J, P,$ obtenus en utilisant comme lois a priori les lois uniforme (U), de Jeffreys (J) et de Poisson (P) respectivement. En outre, nous avons calculé les estimateurs de variance et les intervalles de confiance. Nous avons calculé les intervalles bootstrap selon la méthode de base, sauf les intervalles fondés sur les estimateurs $\hat{\tau}_1^P, \hat{\tau}_2^P$ et $\hat{\tau}^P$, qui ont été calculés par la méthode des centiles. Tous les estimateurs bootstrap ont été obtenus en utilisant 2 000 échantillons bootstrap. Enfin, les propriétés des estimateurs ponctuels et d'intervalle ont été évaluées en utilisant $r = 10\,000$ essais de la méthode qui précède.

Nous avons évalué les propriétés d'un estimateur, disons $\hat{\tau}$, par son biais relatif et la racine carrée de son erreur quadratique moyenne relative, définis comme étant $\text{r} - \text{biais} = \sum_1^r (\hat{\tau}_i - \tau)/(r\tau)$ et $\sqrt{\text{r} - \text{eqm}} = \sqrt{\sum_1^r (\hat{\tau}_i - \tau)^2/(r\tau^2)}$, où $\hat{\tau}_i$ est la valeur de $\hat{\tau}$ obtenue lors du i^{e} essai. Nous avons également évalué les propriétés d'un estimateur de variance par son biais relatif et la racine carrée de son erreur quadratique moyenne relative, qui ont été définis de la même façon que ceux d'un estimateur de la taille de population, mais en utilisant la variance déterminée empiriquement plutôt que la variance réelle. Enfin, nous avons évalué les propriétés des intervalles de confiance à 95 % par leur probabilité de couverture et leur longueur moyenne.

6. Résultats et discussion

Faute d'espace, aux tableaux 2 à 4, nous présentons que certains résultats de l'étude numérique. Toutefois, les commentaires qui suivent ont trait à l'ensemble complet de résultats.

Malgré les limites de l'étude par simulation, nous pouvons conclure que le principal facteur qui influe sur les propriétés des estimateurs et des intervalles de confiance et la grandeur des $p_i^{(k)}$. Lorsque celles-ci sont grandes et indépendamment de la loi des M_i et de la taille de la fraction τ_1/τ couverte par la base de sondage, chacun des estimateurs des τ et des intervalles de confiance de type fondé sur le plan de sondage (Wald ou bootstrap) donne de bons résultats. Toutefois, lorsque les $p_i^{(k)}$ sont faibles et malgré tous les autres facteurs, seuls les estimateurs bayésiens $\hat{\tau}_k^P$ donnent des résultats acceptables. Il mérite d'être souligné que, si les $p_i^{(k)}$ sont faibles, les estimateurs bayésiens $\hat{\tau}_k^U$ et $\hat{\tau}_k^J$ donnent de meilleurs résultats que l'EMV $\tilde{\tau}_k$; toutefois, les propriétés de $\hat{\tau}_k^U$ et $\hat{\tau}_k^J$ ne sont pas suffisamment bonnes pour rendre les inférences fiables.

Les intervalles de confiance bootstrap pour τ_1 basés sur $\hat{\tau}_1^P$ ne sont pas aussi bons que les intervalles de Wald lorsque les $p_i^{(k)}$ sont faibles ou que les M_i ne suivent pas une loi de Poisson. L'explication de ce résultat et l'élaboration de meilleurs intervalles bootstrap sont des sujets qui devront être étudiés de façon plus approfondie.

Tableau 2
Biais relatif et racine carrée de l'erreur quadratique moyenne relative des estimateurs des tailles de population

	Population I				Population II				Population III				Population IV			
$\overline{p}_1$	0,05		0,01		0,05		0,01		0,05		0,01		0,05		0,01	
$\overline{p}_2$	0,01		0,002		0,01		0,002		0,03		0,006		0,03		0,006	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\tilde{\tau}_1$	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09
$\tilde{\tau}_2$	0,01	0,12	0,24[a]	0,78[a]	0,01	0,13	0,21[a]	0,76[a]	0,00	0,06	0,17[b]	0,67[b]	0,00	0,06	0,16[c]	0,63[c]
$\tilde{\tau}$	0,01	0,07	0,13[a]	0,43[a]	0,01	0,07	0,12[a]	0,42[a]	0,00	0,02	0,05[b]	0,19[b]	−0,00	0,02	0,04[c]	0,18[c]
$\hat{\tau}_1^U$	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09
$\hat{\tau}_2^U$	0,02	0,13	0,14[a]	0,65[a]	0,01	0,12	0,14[a]	0,65[a]	0,00	0,06	0,13	0,65	0,00	0,06	0,13	0,71
$\hat{\tau}^U$	0,01	0,07	0,08[a]	0,36[a]	0,01	0,07	0,08[a]	0,36[a]	0,00	0,02	0,03	0,19	−0,00	0,02	0,03	0,20
$\hat{\tau}_1^J$	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09
$\hat{\tau}_2^J$	−0,00	0,12	−0,14	0,48	−0,00	0,12	−0,14	0,48	−0,00	0,06	−0,04	0,37	−0,00	0,06	−0,04	0,35
$\hat{\tau}^J$	−0,00	0,07	−0,08	0,27	−0,00	0,07	−0,08	0,27	−0,00	0,02	−0,02	0,11	−0,00	0,02	−0,02	0,12
$\hat{\tau}_1^P$	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09
$\hat{\tau}_2^P$	0,02	0,12	0,07	0,20	0,02	0,11	0,07	0,20	0,00	0,06	0,00	0,18	0,00	0,06	0,01	0,18
$\hat{\tau}^P$	0,01	0,06	0,04	0,11	0,01	0,06	0,03	0,11	0,00	0,02	−0,00	0,07	−0,00	0,02	−0,00	0,08

Nota : $r\beta$, biais relatif; $r\varepsilon^2$, erreur quadratique moyenne relative; $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$, EMV. Les indices supérieurs *U*, *J* et *P* des estimateurs $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$ indiquent des estimateurs bayésiens basés sur une loi uniforme, de Jeffreys et de Poisson-Gamma à deux degrés, respectivement. Les résultats sont fondés sur 10^4 essais. Les indices supérieurs *a*, *b* et *c* indiquent des résultats obtenus en ne tenant pas compte de 8 %, de 15 % et de 21 % des essais. Les essais omis étaient ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10^4.

Tableau 3
Probabilité de couverture et longueur moyenne des intervalles de confiance à 95 %

	Population I								Population II							
	$\overline{p}_1 \approx 0{,}05$, $\overline{p}_2 \approx 0{,}01$				$\overline{p}_1 \approx 0{,}01$, $\overline{p}_2 \approx 0{,}002$				$\overline{p}_1 \approx 0{,}05$, $\overline{p}_2 \approx 0{,}01$				$\overline{p}_1 \approx 0{,}01$, $\overline{p}_2 \approx 0{,}002$			
	Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald	
	cp	$\overline{l}$	cp	$\overline{l}$	cp	$\overline{l}$	cp	$\overline{l}$	cp	$\overline{l}$	cp	$\overline{l}$	cp	$\overline{l}$	cp	$\overline{l}$
$\tilde{\tau}_1^M$	NC	NC	0,95	129	NC	NC	0,94	398	NC	NC	0,93	127	NC	NC	0,76	400
$\tilde{\tau}_2^M$	NC	NC	0,95	1 044	NC	NC	0,90[a]	8 181[a]	NC	NC	0,95	1 029	NC	NC	0,90[a]	7 764[a]
$\tilde{\tau}^M$	NC	NC	0,95	1 052	NC	NC	0,90[a]	8 200[a]	NC	NC	0,95	1 037	NC	NC	0,90[a]	7 784[a]
$\tilde{\tau}_1^D$	0,95	130	0,95	129	0,92	399	0,94	404	0,97	147	0,95	137	0,96	642	0,92	657
$\tilde{\tau}_2^D$	0,94	1 110	0,95	1 044	0,74	L_1	0,90[a]	8 181[a]	0,94	1 129	0,95	1 029	0,74	L_1	0,90[a]	7 764[a]
$\tilde{\tau}^D$	0,94	1 118	0,95	1 052	0,75	L_1	0,90[a]	8 201[a]	0,95	1 139	0,95	1 038	0,78	L_1	0,90[a]	7 819[a]
$\hat{\tau}_1^U$	0,94	131	0,95	129	0,92	412	0,94	403	0,97	150	0,94	137	0,97	668	0,93	657
$\hat{\tau}_2^U$	0,94	1 116	0,95	1 049	0,72	L_2	0,89[a]	6 887[a]	0,94	1 128	0,95	1 028	0,73	L_2	0,89[a]	6 738[a]
$\hat{\tau}^U$	0,94	1 124	0,95	1 057	0,73	L_2	0,90[a]	6 908[a]	0,95	1 139	0,95	1 038	0,77	L_2	0,90[a]	6 796[a]
$\hat{\tau}_1^J$	0,95	131	0,95	128	0,93	412	0,94	402	0,96	151	0,95	137	0,96	666	0,92	652
$\hat{\tau}_2^J$	0,93	1 043	0,94	998	0,58	3 122	0,71	3 142	0,93	1 057	0,93	985	0,60	3 074	0,72	3 095
$\hat{\tau}^J$	0,93	1 052	0,94	1 007	0,60	3 199	0,72	3 178	0,94	1 072	0,93	995	0,68	3 276	0,73	3 188
$\hat{\tau}_1^P$	0,94	131	0,95	129	0,91	411	0,94	402	0,89	151	0,95	137	0,86	666	0,93	654
$\hat{\tau}_2^P$	0,97	997	0,95	957	1,00	1 506	0,92	1 573	0,97	1 000	0,95	943	1,00	1 510	0,92	1 577
$\hat{\tau}^P$	0,97	1 006	0,95	966	1,00	1 575	0,94	1 624	0,97	1 011	0,95	953	1,00	1 679	0,95	1 710

Nota : cp, probabilité de couverture; $\overline{l}$, longueur moyenne. Les indices supérieurs *M* et *D* des EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ indiquent des intervalles de confiance fondés sur le modèle et fondés sur le plan, respectivement. Intervalles de confiance bootstrap calculés sur 2 000 échantillons bootstrap. NC, non calculé. Résultats fondés sur 10^4 essais. L'indice supérieur *a* indique des résultats obtenus en ne tenant pas compte de 8 % des essais. Les essais omis sont ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10^4. L_1 et L_2 indiquent des longueurs supérieures à 10^9 et 10^4, respectivement.

Tableau 4
Biais relatif et racine carrée de l'erreur quadratique moyenne relative des estimateurs de variance

	Population I								Population II							
	$\bar{p}_1 \approx 0,05$, $\bar{p}_2 \approx 0,01$				$\bar{p}_1 \approx 0,01$, $\bar{p}_2 \approx 0,002$				$\bar{p}_1 \approx 0,05$, $\bar{p}_2 \approx 0,01$				$\bar{p}_1 \approx 0,01$, $\bar{p}_2 \approx 0,002$			
	Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\tilde{\tau}_1^M$	NC	NC	0,01	0,17	NC	NC	−0,04	0,08	NC	NC	−0,20	0,31	NC	NC	−0,64	0,65
$\tilde{\tau}_2^M$	NC	NC	0,01	0,49	NC	NC	1,9[a]	5,3[a]	NC	NC	−0,02	0,64	NC	NC	1,8[a]	5,4[a]
$\tilde{\tau}^M$	NC	NC	0,01	0,48	NC	NC	1,9[a]	5,3[a]	NC	NC	−0,02	0,64	NC	NC	1,7[a]	5,3[a]
$\tilde{\tau}_1^D$	0,03	0,19	0,01	0,17	−0,02	0,17	−0,00	0,17	0,08	0,46	−0,07	0,28	−0,05	0,40	−0,01	0,37
$\tilde{\tau}_2^D$	0,16	0,62	0,01	0,49	L_1	L_2	1,9[a]	5,3[a]	0,20	1,10	−0,02	0,64	L_2	L_2	1,7[a]	5,3[a]
$\tilde{\tau}^D$	0,15	0,61	0,01	0,48	L_1	L_2	1,9[a]	5,3[a]	0,20	1,10	−0,02	0,64	L_2	L_2	1,7[a]	5,3[a]
$\hat{\tau}_1^U$	0,02	0,20	−0,01	0,17	0,03	0,19	−0,01	0,17	0,14	0,51	−0,06	0,28	0,05	0,37	0,01	0,37
$\hat{\tau}_2^U$	0,13	0,62	−0,01	0,49	0,24	1,20	1,7[a]	4,6[a]	0,22	0,92	−0,00	0,62	0,30	1,40	1,6[a]	6,4[a]
$\hat{\tau}^U$	0,13	0,61	−0,01	0,48	0,24	1,20	1,6[a]	4,5[a]	0,23	0,91	0,01	0,61	0,30	1,40	1,6[a]	6,2[a]
$\hat{\tau}_1^J$	0,06	0,21	0,02	0,17	0,05	0,19	−0,01	0,17	0,12	0,50	−0,08	0,28	0,00	0,35	−0,04	0,36
$\hat{\tau}_2^J$	0,07	0,51	−0,03	0,44	−0,25	0,66	−0,11	1,40	0,13	0,69	−0,03	0,55	−0,25	0,74	−0,13	1,50
$\hat{\tau}^J$	0,06	0,50	−0,03	0,43	−0,25	0,66	−0,12	1,40	0,12	0,68	−0,03	0,53	−0,24	0,72	−0,15	1,40
$\hat{\tau}_1^P$	0,03	0,20	−0,01	0,17	0,03	0,18	−0,02	0,17	0,16	0,52	−0,05	0,28	0,05	0,37	0,01	0,37
$\hat{\tau}_2^P$	0,07	0,34	−0,02	0,35	−0,07	0,16	−0,03	0,12	0,10	0,42	−0,01	0,41	−0,06	0,17	−0,01	0,16
$\hat{\tau}^P$	0,06	0,34	−0,02	0,34	−0,05	0,14	−0,02	0,11	0,10	0,42	−0,01	0,41	−0,03	0,15	0,01	0,16

Nota : $r\beta$, biais relatif; $r\varepsilon^2$, erreur quadratique moyenne relative. Les indices supérieurs M et D des EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ indiquent des estimateurs de variance fondés sur le modèle et fondés sur le plan, respectivement. Intervalles de confiance bootstrap calculés sur 2 000 échantillons bootstrap. NC, non calculé. Résultats basés sur 10^4 essais. L'indice supérieur *a* indique des résultats obtenus en ne tenant pas compte de 8 % des essais. Les essais omis étaient ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10^4. L_1 et L_2 indiquent des valeurs supérieures à 10^2 et 10^4, respectivement.

Enfin, les meilleures propriétés de l'ensemble d'estimateurs $\hat{\tau}_k^P$ sont une conséquence de la plus grande quantité d'information qu'ils utilisent. Bien que nous nous soyons servis de lois a priori relativement uniformes pour les τ_k, l'information qu'ils fournissent est suffisante pour éviter les problèmes de biais et de forte variabilité observés pour les autres estimateurs. Nous avons réalisé certains essais par simulations supplémentaires et les résultats (qui ne sont pas présentés dans les tableaux) indiquent qu'à condition que les lois a priori soient maintenues relativement uniformes, les estimations ne sont pas affectées par les valeurs de leurs paramètres. Manifestement, une information initiale erronée combinée à de faibles valeurs des $p_i^{(k)}$ aura une incidence sur les estimations. À titre d'exemple, mentionnons une loi a priori de τ_2 dont la densité de probabilité est fortement concentrée autour d'une valeur très éloignée de la valeur réelle de τ_2. Cependant, nous pensons que si le chercheur dispose d'information correcte, même si elle est vague, il vaut la peine d'utiliser l'ensemble d'estimateurs $\hat{\tau}_k^P$.

Remerciements

Cette étude a été financée par la subvention UASIN-EXB-01-01 du PROMEP et par la subvention PAFI-UAS-2002-I-MHFM-0 de l'UAS. Nous remercions Eduardo Gutierrez, le rédacteur associé et les examinateurs de leurs suggestions et commentaires constructifs.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis.* 2ième édition. New York : John Wiley & Sons, Inc.

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

Castledine, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.

Darroch, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.

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

Evans, M.A., Kim, H.-M. et O'Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 131-140.

Félix-Medina, M.H., et Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.

Fienberg, S.E., Johnson, M.S. et Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society*, A, 162, 383-405.

Frank, O., et Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Heckathorn, D.D. (1994). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Seber, G.A.F. (1982). *The Estimation of Animal Abundance*. 2ième édition. London : Griffin.

Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Méthodologie Sociologique*, 36, 59-70.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.

Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.