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

Techniques d'enquête

Combiner l'échantillonnage par dépistage de liens et l'échantillonnage en grappes pour estimer la taille d'une population cachée en présence de probabilités de lien hétérogènes

par Martín H. Félix-Medina, Pedro E. Monjardin
et Aida N. Aceves-Castro

Date de diffusion : le 17 décembre 2015

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2015

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2015
Vol. 41, nº 2, pp. 369-397
Statistique Canada, nº 12-001-X au catalogue

Combiner l'échantillonnage par dépistage de liens et l'échantillonnage en grappes pour estimer la taille d'une population cachée en présence de probabilités de lien hétérogènes

Martín H. Félix-Medina, Pedro E. Monjardin et Aida N. Aceves-Castro[1]

Résumé

Félix-Medina et Thompson (2004) ont proposé une variante de l'échantillonnage par dépistage de liens pour échantillonner des populations humaines cachées ou difficiles à joindre, comme les toxicomanes et les travailleurs de l'industrie du sexe. Dans cette variante, on commence par sélectionner un échantillon d'emplacements, puis on demande aux personnes trouvées dans les lieux échantillonnés de nommer d'autres membres de la population à inclure dans l'échantillon. Ces auteurs ont établi des estimateurs du maximum de vraisemblance de la taille de la population sous l'hypothèse que la probabilité qu'une personne soit nommée par une autre dans un lieu échantillonné (probabilité de lien) ne dépend pas de la personne nommée (hypothèse d'homogénéité). Dans le présent travail, nous étendons leur recherche au cas où les probabilités de lien sont hétérogènes et dérivons des estimateurs du maximum de vraisemblance inconditionnel et conditionnel de la taille de la population. Nous proposons aussi des intervalles de confiance par vraisemblance profilée et par bootstrap pour la taille de la population. Les résultats de nos études en simulation montrent qu'en présence de probabilités de lien hétérogènes, les estimateurs proposés donnent d'assez bons résultats à condition que les fractions d'échantillonnage soient relativement grandes, disons supérieures à 0,5, tandis que la performance des estimateurs calculés sous l'hypothèse d'homogénéité n'est pas bonne. Les résultats montrent aussi que les intervalles de confiance proposés ne sont pas très robustes aux écarts par rapport aux modèles supposés.

Mots-clés : Bootstrap; capture-recapture; échantillonnage par références enchaînées; estimateur du maximum de vraisemblance; intervalle de confiance par vraisemblance profilée; échantillonnage en boule de neige.

1 Introduction

Les méthodes d'échantillonnage classiques ne conviennent pas aux populations humaines cachées ou difficiles à joindre, comme les toxicomanes, les travailleurs de l'industrie du sexe et les sans-abri, en raison de l'absence de bases de sondage appropriées. Plusieurs méthodes d'échantillonnage particulières ont donc été proposées pour ces types de populations. (Voir Magnani, Sabin, Saidel et Heckathorn (2005) et Kalton (2009) pour des revues de certaines d'entre elles.) Selon Heckathorn (2002), deux méthodes d'échantillonnage pour les populations cachées sont couramment utilisées dans les études actuelles. L'une est l'échantillonnage d'emplacements (*location sampling*), également appelé échantillonnage espace-temps (*time-and-space sampling*) ou échantillonnage des lieux de sociabilité (*aggregation point sampling* ou *intercept point sampling*). L'autre est l'échantillonnage en boule de neige (*snowball sampling*), également appelé échantillonnage par dépistage de liens (*link-tracing sampling*) ou échantillonnage par références enchaînées (*chain referral sampling*).

L'échantillonnage d'emplacements comprend l'établissement d'une base de sondage d'unités primaires. Ces unités primaires sont des combinaisons d'emplacements et de segments horaires où et durant lesquels les individus de la population étudiée ont tendance à se réunir. Cette base de sondage n'est

1. Martín H. Félix-Medina, Pedro E. Monjardin et Aida N. Aceves-Castro, Facultad de Ciencias F sico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México. Courriel : mhfelix@uas.edu.mx.

pas considérée comme couvrant l'ensemble de la population. On sélectionne d'abord un échantillon probabiliste d'unités primaires, puis, de chaque unité échantillonnée, on tire un genre d'échantillon systématique d'individus. Bien que l'on puisse construire des estimateurs fondés sur le plan de sondage pour différents paramètres, le principal inconvénient de l'échantillonnage d'emplacements est que les inférences ne sont valides que pour la partie de la population couverte par la base de sondage. Pour des revues de cette méthode, consulter MacKellar, Valleroy, Karon, Lemp et Janssen (1996), Munhib, Lin, Stueve, Miller, Ford, Johnson et Smith (2001), McKenzie et Mistianen (2009), Semaan (2010) et Karon et Wejnert (2012). L'échantillonnage d'emplacements a été utilisé dans la *Young Men's Survey* pour estimer la séroprévalence du VIH chez les jeunes hommes qui ont des relations sexuelles avec des hommes. (Voir McKellar et coll. 1996.) Dans cette étude, les unités primaires étaient les lieux fréquentés par les jeunes hommes, tels que les clubs de danse, les bars et les lieux de rue.

Dans l'échantillonnage par dépistage de liens, on sélectionne un échantillon initial de membres de la population, puis on augmente la taille de cet échantillon en demandant aux personnes échantillonnées de nommer d'autres membres de la population ou d'aiguiller l'enquêteur vers d'autres membres de la population à inclure dans l'échantillon. On peut ensuite demander aux personnes nommées qui ne figurent pas dans l'échantillon initial d'aiguiller l'enquêteur vers d'autres personnes, et le processus peut se poursuivre de cette façon jusqu'à ce qu'une règle d'arrêt spécifiée soit satisfaite. Pour des revues de plusieurs variantes de l'échantillonnage par dépistage de liens, voir Spreen (1992), Thompson et Frank (2000) et Johnston et Sabin (2010). L'échantillonnage par dépistage de liens a été utilisé dans l'étude de la transmission hétérosexuelle du VIH/Sida de Colorado Springs. (Voir Potterat, Woodhouse, Rothenberg, Muth, Darrow, Muth et Reynolds 1993; Rothemberg, Woodhouse, Potterat, Muth, Darrow et Klovdahl 1995 et Potterat, Woodhouse, Muth, Rothenberg, Darrow, Klovdahl et Muth 2004.) Dans cette étude, on a obtenu un échantillon non probabiliste initial de personnes présumées être à haut risque d'acquérir ou de transmettre le VIH, et on leur a demandé d'énumérer tous leurs contacts personnels que l'on a également inclus dans l'échantillon.

Frank et Snijders (1994) ont proposé une variante de l'échantillonnage par dépistage de liens, qui permet à l'enquêteur d'estimer la taille de la population. Dans leur variante, ils supposent que l'échantillon initial est un échantillon de Bernoulli, c'est-à-dire que chaque élément de la population possède la même probabilité d'être inclus dans l'échantillon et que les inclusions sont indépendantes. En outre, ils supposent que la probabilité que la personne i dans l'échantillon initial aiguille l'enquêteur vers la personne j dans la population, probabilité que nous nommerons probabilité de lien, est une constante, et que les aiguillages sont indépendants. Nous nommerons la première prémisse additionnelle l'hypothèse d'homogénéité des probabilités de lien. En s'appuyant sur ces hypothèses, ces auteurs établissent plusieurs estimateurs de la taille de la population. Ils indiquent que leur méthode a donné des estimations raisonnables du nombre d'héroïnomanes à Groningen. Cependant, Dávid et Snijders (2002) ont signalé une sous-estimation du nombre de sans-abri à Budapest en utilisant cette méthode. Selon eux, la sous-estimation pourrait être causée par des divergences par rapport à l'hypothèse d'un échantillon initial de Bernoulli.

La difficulté à satisfaire, dans les applications pratiques de l'échantillonnage par dépistage de liens, l'hypothèse d'un échantillonnage initial bernoullien des membres de la population a poussé Félix-Medina et Thompson (2004) à élaborer une variante dans laquelle l'échantillon initial est tiré selon un plan probabiliste. Pour ce faire, ils supposent, comme dans l'échantillonnage d'emplacements, que l'enquêteur

peut construire une base de sondage des sites ou lieux où les membres de la population ont tendance à se réunir, comme les bars, les parcs et les quartiers ou îlots. La base de sondage est censée couvrir, non pas l'ensemble de la population, mais une partie seulement de celle-ci. Alors, un échantillon aléatoire simple sans remise (EASSR) des sites est sélectionné et les membres de la population qui appartiennent aux sites échantillonnés sont identifiés. Enfin, comme dans l'échantillonnage par dépistage de liens ordinaire, on demande aux personnes comprises dans l'échantillon initial de nommer d'autres membres de la population.

Ces auteurs proposent des estimateurs du maximum de vraisemblance (EMV) de la taille de la population établis sous un modèle probabiliste décrivant les nombres de personnes trouvées dans les lieux échantillonnés et un modèle considérant que les probabilités de lien entre les éléments de la population et les sites échantillonnés sont homogènes, c'est-à-dire qu'elles dépendent des sites échantillonnés, mais non des personnes éventuellement nommées. Plus tard, Félix-Medina et Monjardin (2006) ont considéré cette même variante de l'échantillonnage par dépistage de liens et proposé des estimateurs de la taille de la population également établis sous l'hypothèse d'homogénéité, mais en utilisant une méthode assistée par modèle bayésien, c'est-à-dire que les formes fonctionnelles des estimateurs sont obtenues en utilisant le modèle bayésien, mais que les inférences sont faites sous le modèle fréquentiste.

Même si la variante de l'échantillonnage par dépistage de liens proposée par Félix-Medina et Thompson (2004) n'a pas été utilisée dans une étude réelle, nous nous attendons à ce que les estimateurs de la taille de la population établis sous l'hypothèse d'homogénéité posent des problèmes de sous-estimation si cette hypothèse n'est pas satisfaite, comme cela se produit dans les études par capture-recapture. Nous pensons qu'il en est ainsi parce ces estimateurs ressemblent à ceux utilisés dans le cas de la méthode de capture-recapture.

Dans le présent article, nous étendons les travaux de Félix-Medina et Thompson (2004) au cas où les probabilités de lien dépendent des personnes nommées, autrement dit, nous supposons que les probabilités de lien sont hétérogènes. La présentation de l'article est la suivante. À la section 2, nous décrivons la variante de l'échantillonnage par dépistage de liens proposée par Félix-Medina et Thompson (2004). À la section 3, nous présentons un modèle pour les probabilités de lien qui tient compte de leur hétérogénéité et obtenons les EMV inconditionnel et conditionnel de la taille de la population. À la section 4, nous construisons des intervalles de confiance par vraisemblance profilée et par bootstrap pour la taille de la population. À la section 5, nous décrivons une procédure pour déterminer la taille de l'échantillon initial afin d'obtenir une valeur spécifiée de l'erreur relative de l'estimation. À la section 6, nous donnons les résultats de deux études en simulation et, enfin, à la section 7, nous présentons certaines conclusions et suggérons de futurs travaux de recherche.

2 Plan d'échantillonnage et notation

Puisque nous considérons dans l'étude la variante de l'échantillonnage par dépistage de liens proposée par Félix-Medina et Thompson (2004), nous la décrivons brièvement ici. Soit U une population finie constituée d'un nombre inconnu τ de personnes. Nous supposons qu'une partie U_1 de U est couverte par une base de sondage de N sites $A_1, \ldots, A_N$, où la probabilité de trouver les membres de la population est élevée. Nous supposons que nous disposons d'un critère qui nous permet d'affecter une personne présente

dans U_1 à un seul site dans la base de sondage. Il convient de souligner que nous supposons, non pas qu'une personne ne peut pas être trouvée dans différents sites, mais plutôt que, comme dans l'échantillonnage en grappes ordinaire, nous ne sommes capables de l'affecter qu'à un seul site, par exemple, le site où elle passe la plupart de son temps. Donc, nous pouvons considérer les sites qui figurent dans la base de sondage comme des grappes de personnes. Soit M_i le nombre de membres de la population qui appartiennent au site $A_i, i = 1, \ldots, N$. De l'hypothèse précédente, il découle que le nombre de personnes dans U_1 est $\tau_1 = \sum_1^N M_i$ et que le nombre de personnes dans la partie $U_2 = U - U_1$ de U qui n'est pas couverte par la base de sondage est $\tau_2 = \tau - \tau_1$.

Le plan d'échantillonnage est le suivant. Un EASSR S_A de n sites $A_1, \ldots, A_n$ est sélectionné dans la base de sondage et les M_i membres de la population qui appartiennent au site échantillonné A_i sont identifiés, $i = 1, \ldots, n$. Soit S_0 l'ensemble de personnes dans l'échantillon initial. Observons que la taille de S_0 est $M = \sum_1^n M_i$. On demande aux personnes dans chaque site échantillonné de nommer d'autres membres de la population. Nous dirons qu'une personne et un site sont liés si toute personne qui appartient à ce site nomme la personne en question. Enfin, soit S_1 et S_2 les ensembles de personnes dans $U_1 - S_0$ et U_2, respectivement, qui sont liées à un ou à plusieurs sites dans S_A.

3 Estimateurs du maximum de vraisemblance de τ_1, τ_2 et τ

3.1 Modèles probabilistes

Pour construire les EMV des τ, nous devons spécifier des modèles pour les variables observées. Donc, comme dans Félix-Medina et Thompson (2004), nous supposons que les nombres $M_1, \ldots, M_N$ de personnes qui appartiennent aux sites $A_1, \ldots, A_N$ sont des variables aléatoires de Poisson indépendantes de moyenne λ_1. Par conséquent, la loi conditionnelle conjointe de $(M_1, \ldots, M_n, \tau_1 - M)$ sachant que $\sum_1^N M_i = \tau_1$ est une loi multinomiale dont la fonction de masse de probabilité (fmp) est :

$$f(m_1, \ldots, m_n, \tau_1 - m) = \frac{\tau_1!}{\prod_1^n m_i!(\tau_1 - m)!} \left(\frac{1}{N}\right)^m \left(1 - \frac{n}{N}\right)^{\tau_1 - m}. \tag{3.1}$$

Pour modéliser les liens entre les membres de la population et les sites échantillonnés, nous définissons les variables aléatoires suivantes : $X_{ij}^{(k)} = 1$ si la personne j dans $U_k - A_i$ est liée au site A_i et $X_{ij}^{(k)} = 0$ si $j \in A_i$ ou que la personne n'est pas liée à $A_i, j = 1, \ldots, \tau_k, i = 1, \ldots, n$. Nous supposons que, étant donné l'échantillon S_A de sites, les $X_{ij}^{(k)}$ sont des variables aléatoires de Bernoulli de moyenne $p_{ij}^{(k)}$, où la probabilité de lien $p_{ij}^{(k)}$ satisfait le modèle de Rasch suivant :

$$p_{ij}^{(k)} = \Pr\left(X_{ij}^{(k)} = 1 \middle| \beta_j^{(k)}, S_A\right) = \frac{\exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}{1 + \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}, j \in U_k - A_i; i = 1, \ldots, n. \tag{3.2}$$

Il convient de souligner que ce modèle a été pris en considération par Coull et Agresti (1999) dans le contexte de l'échantillonnage par capture-recapture. Dans ce modèle, $\alpha_i^{(k)}$ est un effet fixe (non aléatoire) qui représente la possibilité qu'a la grappe A_i de former des liens avec les personnes comprises dans $U_k - A_i$, et $\beta_j^{(k)}$ est un effet aléatoire qui représente la propension de la personne $j \in U_k$ à être liée à une grappe. Nous supposons que $\beta_j^{(k)}$ suit une loi normale de moyenne 0 et de variance inconnue σ_k^2, et que ces variables sont indépendantes. Le paramètre σ_k^2 détermine le degré d'hétérogénéité des $p_{ij}^{(k)}$: de grandes valeurs de σ_k^2 impliquent un haut degré d'hétérogénéité.

Avant de conclure la présente sous-section, nous formulerons certains commentaires au sujet des modèles supposés. Premièrement, la loi multinomiale des M_i observés (qui est celle utilisée dans la fonction de vraisemblance) implique que les personnes sont distribuées indépendamment et avec probabilités égales sur les sites de la base de sondage. Cette hypothèse est difficile à satisfaire en pratique; cependant, comme nous le montrerons plus loin, la fonction de vraisemblance dépend des M_i essentiellement par la voie de leur somme M et, puisque NM/n est un estimateur de τ_1 fondé sur le plan de sondage, c'est-à-dire qu'il s'agit d'un estimateur exempt d'une loi, il s'ensuit que l'EMV de τ_1 sera également robuste aux écarts par rapport à la loi multinomiale des M_i. Néanmoins, les écarts par rapport à ce modèle auront une incidence sur la performance des estimateurs de variance et des intervalles de confiance calculés sous cette hypothèse. Deuxièmement, le modèle de Rasch donné par (3.2) implique ce qui suit : i) la probabilité de lien $p_{ij}^{(k)}$ dépend uniquement de deux effets : la sociabilité des personnes dans la grappe A_i et celle de la personne $j \in U_k - A_i$; ii) les deux effets sont additifs, et iii) pour tout site A_i dans la base de sondage et toute personne $j \in U - A_i$, $p_{ij}^{(k)} > 0$. Le modèle (3.2) est un cas particulier d'un modèle mixte linéaire généralisé. (Voir Agresti 2002, section 2.1, pour un bref examen de ce type de modèle.) Par conséquent, nous pourrions incorporer les structures de réseau des personnes comprises dans la grappe A_i et de la personne $j \in U_k - A_i$ pour modéliser la probabilité de lien $p_{ij}^{(k)}$ en étendant le modèle (3.2) à un modèle qui comprend les covariables associées à la personne j et à la grappe A_i, ainsi que leurs termes d'interaction. Cependant, en utilisant un modèle plus général que (3.2), nous rendrons le problème d'inférence beaucoup plus difficile que celui que nous devons résoudre dans la présente étude. Donc, malgré la simplicité relative du modèle (3.2), nous nous attendons à ce qu'il rende compte de l'hétérogénéité des probabilités de lien et qu'il nous permette de faire des inférences au sujet des τ dont au moins l'ordre de grandeur est correct.

3.2 Fonction de vraisemblance

Le moyen le plus facile de construire la fonction de vraisemblance consiste à la factoriser en différentes composantes. L'une d'elles est associée à la probabilité de sélectionner l'échantillon initial S_0, qui est donnée par la loi multinomiale (3.1), c'est-à-dire

$$L_{\text{MULT}}(\tau_1) \propto \frac{\tau_1!}{(\tau_1 - m)!}(1 - n/N)^{\tau_1 - m}.$$

Deux autres composantes sont associées aux probabilités conditionnelles des configurations de liens entre les personnes dans $U_k - S_0, k = 1, 2,$ et les grappes $A_i \in S_A$, sachant S_A. Pour obtenir ces

facteurs, nous devons calculer les probabilités de certains événements. Soit $\mathbf{X}_j^{(k)} = \left(X_{1j}^{(k)}, \ldots, X_{nj}^{(k)}\right)$ le vecteur de dimension n de variables indicatrices de lien $X_{ij}^{(k)}$ associées à la j^{e} personne dans $U_k - S_0$. Notons que $\mathbf{X}_j^{(k)}$ indique quelles grappes $A_i \in S_A$ sont liées à cette personne. Soit $\mathbf{x} = (x_1, \ldots, x_n)$ un vecteur dont le i^{e} élément est 0 ou $1, i = 1, \ldots, n.$ En raison des hypothèses que nous avons faites au sujet des lois des variables $X_{ij}^{(k)}$, la probabilité conditionnelle, sachant $\beta_j^{(k)}$, que $\mathbf{X}_j^{(k)}$ soit égale à $\mathbf{x}$, c'est-à-dire la probabilité que la j^{e} personne soit liée uniquement aux grappes $A_i \in S_A$ de manière telle que le i^{e} élément x_i de $\mathbf{x}$ soit égal à 1, est

$$\Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x} \,\middle|\, \beta_j^{(k)}, S_A\right) = \prod_{i=1}^{n} \left[p_{ij}^{(k)}\right]^{x_i} \left[1 - p_{ij}^{(k)}\right]^{1-x_i} = \prod_{i=1}^{n} \frac{\exp\left[x_i\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}.$$

Donc, la probabilité que le vecteur de variables indicatrices de lien associé à une personne sélectionnée aléatoirement dans $U_k - S_0$ soit égal à $\mathbf{x}$ est

$$\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k) = \int \prod_{i=1}^{n} \frac{\exp\left[x_i\left(\alpha_i^{(k)} + \sigma_k z\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \sigma_k z\right)} \phi(z)\, dz,$$

où $\boldsymbol{\alpha}_k = \left(\alpha_1^{(k)}, \ldots, \alpha_n^{(k)}\right)$ et $\phi(\cdot)$ désignent la densité de probabilité de la loi normale centrée réduite $[\mathrm{N}(0,1)]$.

Comme dans Coull et Agresti (1999), au lieu d'utiliser $\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ dans la fonction de vraisemblance, nous utilisons son approximation par quadrature gaussienne $\tilde{\pi}_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ donnée par

$$\tilde{\pi}_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k) = \sum_{t=1}^{q} \prod_{i=1}^{n} \frac{\exp\left[x_i\left(\alpha_i^{(k)} + \sigma_k z_t\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \sigma_k z_t\right)} \nu_t, \tag{3.3}$$

où q est une constante fixée, et $\{z_t\}$ et les valeurs de $\{\nu_t\}$ sont tirées de tables.

Nous pouvons maintenant calculer les deux facteurs susmentionnés de la fonction de vraisemblance. Soit $\Omega = \{(x_1, \ldots, x_n) : x_i = 0, 1; i = 1, \ldots, n\}$, l'ensemble de tous les vecteurs de dimension n tels que chacun de leurs éléments est 0 ou 1. Pour $\mathbf{x} = (x_1, \ldots, x_n) \in \Omega$, soit $R_{\mathbf{x}}^{(k)}$ la variable aléatoire qui indique le nombre de personnes distinctes dans $U_k - S_0$ dont les vecteurs de variables indicatrices de lien sont égaux à $\mathbf{x}$. Enfin, soit R_k la variable aléatoire qui indique le nombre de personnes distinctes dans $U_k - S_0$ qui sont liées à au moins une grappe $A_i \in S_A$. Notons que $R_k = \sum_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} R_{\mathbf{x}}^{(k)}$, où $\mathbf{0}$ désigne le vecteur de dimension n de zéros.

En raison des hypothèses que nous avons émises au sujet des lois des variables $X_{ij}^{(k)}$, la loi de probabilité conjointe conditionnelle des variables $\left\{R_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ et $\tau_1 - m - R_1$, sachant que $\{M_i = m_i\}_{i=1}^{n}$, est une loi multinomiale de paramètre de taille $\tau_1 - m$ et de probabilités $\left\{\pi_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ et $\pi_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)$, et celle des variables $\left\{R_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ et $\tau_2 - R_2$ est une loi multinomiale de paramètre de taille τ_2 et de probabilités $\left\{\pi_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ et $\pi_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)$.

Par conséquent, les facteurs associés aux probabilités des configurations des liens entre les personnes dans $U_k - S_0, k = 1, 2,$ et les grappes $A_i \in S_A$ sont

$$L_1(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) \propto \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!} \prod_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{r_{\mathbf{x}}^{(1)}} \left[\tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{\tau_1 - m - r_1}$$

et

$$L_2(\tau_2, \boldsymbol{\alpha}_2, \sigma_2) \propto \frac{\tau_2!}{(\tau_2 - r_2)!} \prod_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{r_{\mathbf{x}}^{(2)}} \left[\tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{\tau_2 - r_2}.$$

La dernière composante de la fonction de vraisemblance est associée à la probabilité conditionnelle, sachant S_A, de la configuration des liens entre les personnes dans S_0 et les grappes $A_i \in S_A$. Pour calculer ce facteur, observons pour commencer que, par définition des variables indicatrices $X_{ij}^{(k)}$, le i^{e} élément du vecteur des variables indicatrices de lien associé à une personne dans $A_i \in S_A$ est égal à zéro. Donc, soit $\Omega_{-i} = \{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) : x_j = 0,1,\ j \neq i,\ j = 1, \ldots, n\}$, c'est-à-dire l'ensemble de tous les vecteurs de dimension $(n-1)$ obtenu à partir des vecteurs dans Ω en omettant leur i^{e} coordonnée. Pour $\mathbf{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \Omega_{-i}$, soit $R_{\mathbf{x}}^{(A_i)}$ la variable aléatoire qui indique le nombre de personnes distinctes dans $A_i \in S_A$ dont les vecteurs de variables indicatrices de lien, quand la i^{e} coordonnée est omise, sont égaux à $\mathbf{x}$. Enfin, soit $R^{(A_i)}$ la variable aléatoire qui indique le nombre de personnes distinctes dans $A_i \in S_A$ qui sont liées à au moins un site $A_j \in S_A, j \neq i$. Notons que $R^{(A_i)} = \sum_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}} R_{\mathbf{x}}^{(A_i)}$, où $\mathbf{0}$ désigne le vecteur de dimension $(n-1)$ de zéros. Alors, comme dans les cas précédents, la loi de probabilité conjointe conditionnelle des variables $\{R_{\mathbf{x}}^{(A_i)}\}_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}}$ et $m_i - R^{(A_i)}$, sachant que $\{M_i = m_i\}_{i=1}^n$, est une loi multinomiale de paramètre de taille m_i et de probabilités $\{\pi_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)\}_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}}$ et $\pi_{\mathbf{0}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)$, où $\boldsymbol{\alpha}_1^{(-i)} = (\alpha_1^{(1)}, \ldots, \alpha_{i-1}^{(1)}, \alpha_{i+1}^{(1)}, \ldots, \alpha_n^{(1)})$ et

$$\pi_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1) = \int \prod_{j \neq i}^{n} \frac{\exp\left[x_j(\alpha_j^{(1)} + \sigma_1 z)\right]}{1 + \exp(\alpha_j^{(1)} + \sigma_1 z)} \phi(z)\, dz.$$

Par conséquent, la probabilité de la configuration des liens entre les personnes dans S_0 et les grappes $A_j \in S_A$ est donnée par le produit des probabilités multinomiales précédentes (une pour chaque $A_i \in S_A$), et conséquemment le facteur de la vraisemblance qui est associé à cette probabilité est

$$L_0(\boldsymbol{\alpha}_1, \sigma_1) \propto \prod_{i=1}^{n} \prod_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)\right]^{r_{\mathbf{x}}^{(A_i)}} \left[\tilde{\pi}_{\mathbf{0}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)\right]^{m_i - r^{(A_i)}},$$

où

$$\tilde{\pi}_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1) = \sum_{t=1}^{q} \prod_{j \neq i}^{n} \frac{\exp\left[x_j(\alpha_j^{(1)} + \sigma_1 z_t)\right]}{1 + \exp(\alpha_j^{(1)} + \sigma_1 z_t)} \nu_t, \tag{3.4}$$

est l'approximation par quadrature gaussienne de la probabilité $\pi_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)$.

Des résultats qui précèdent, il découle que la fonction de vraisemblance est donnée par

$$L(\tau_1, \tau_2, \alpha_1, \alpha_2, \sigma_1, \sigma_2) = L_{(1)}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) L_{(2)}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2),$$

où

$$L_{(1)}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) = L_{\text{MULT}}(\tau_1) L_1(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) L_0(\boldsymbol{\alpha}_1, \sigma_1)$$

et

$$L_{(2)}\left(\tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) = L_2\left(\tau_2, \boldsymbol{\alpha}_2, \sigma_2\right).$$

Dans les commentaires à la fin de la sous-section 3.1, nous avons indiqué que la fonction de vraisemblance dépend des M_i essentiellement par la voie de leur somme M. On peut le constater en remarquant que seul le facteur L_0 dépend directement des M_i. Les facteurs L_{MULT} et L_1 dépendent des M_i par la voie de M, tandis que le facteur $L_{(2)}$ ne dépend pas des M_i.

3.3 Estimateurs du maximum de vraisemblance inconditionnel

La maximisation numérique de la fonction de vraisemblance $L\left(\tau_1, \tau_2, \alpha_1, \alpha_2, \sigma_1, \sigma_2\right)$ par rapport aux paramètres donne les estimateurs du maximum de vraisemblance ordinaire ou inconditionnel (EMVI) $\hat{\tau}_k^{(U)}, \hat{\alpha}_k^{(U)}$ et $\hat{\sigma}_k^{(U)}$ de τ_k, α_k et $\sigma_k, k = 1, 2$. Par conséquent, l'EMVI de $\tau = \tau_1 + \tau_2$ est $\hat{\tau}^{(U)} = \hat{\tau}_1^{(U)} + \hat{\tau}_2^{(U)}$. Il n'existe pas d'expressions analytiques pour les EMVI; cependant, en utilisant l'approximation asymptotique $\partial \ln\left(\tau_k!\right)/\partial\tau_k \approx \ln\left(\tau_k\right)$, nous obtenons les approximations suivantes de $\hat{\tau}_1^{(U)}$ et $\hat{\tau}_2^{(U)}$:

$$\hat{\tau}_1^{(U)} = \frac{M + R_1}{1 - (1 - n/N)\,\tilde{\pi}_{\mathbf{0}}^{(1)}\left(\hat{\boldsymbol{\alpha}}_1^{(U)}, \hat{\sigma}_1^{(U)}\right)} \quad \text{et} \quad \hat{\tau}_2^{(U)} = \frac{R_2}{1 - \tilde{\pi}_{\mathbf{0}}^{(2)}\left(\hat{\boldsymbol{\alpha}}_2^{(U)}, \hat{\sigma}_2^{(U)}\right)}. \tag{3.5}$$

Notons qu'il ne s'agit pas d'expressions analytiques, puisque $\hat{\alpha}_k^{(U)}$ et $\hat{\sigma}_k^{(U)}$ dépendent de $\hat{\tau}_k^{(U)}$, $k = 1, 2$. Néanmoins, ces expressions sont utiles pour obtenir les formules des variances asymptotiques de $\hat{\tau}_1^{(U)}$ et $\hat{\tau}_2^{(U)}$.

3.4 Estimateurs du maximum de vraisemblance conditionnel

Un autre moyen d'obtenir les EMV de τ_k, α_k et σ_k consiste à suivre l'approche de Sanathanan (1972), qui donne des estimateurs du maximum de vraisemblance conditionnel (EMVC). Ces estimateurs sont numériquement plus simples à calculer que les EMVI. En outre, si l'on utilise des covariables dans le modèle pour la probabilité de lien $p_{ij}^{(k)}$, cette approche pourrait encore être utilisée pour obtenir des estimateurs de τ_k, α_k et σ_k, tandis que l'approche de la vraisemblance inconditionnelle ne pourrait pas l'être, puisque les valeurs des covariables associées aux éléments non échantillonnés seraient inconnues.

L'idée qui sous-tend l'approche de Sanathanan est de factoriser la fmp des lois multinomiales des fréquences $R_{\mathbf{x}}^{(k)}$ des différentes configurations des liens comme il suit :

$$\begin{aligned}
L_1\left(\tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) &\propto f\left(\left\{r_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}, \tau_1 - m - r_1 \middle| \{m_i\}, \tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) \\
&= f\left(\left\{r_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \middle| \{m_i\}, \tau_1, r_1, \boldsymbol{\alpha}_1, \sigma_1\right) f\left(r_1 \middle| \{m_i\}, \tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) \\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}\left[\frac{\tilde{\pi}_{\mathbf{x}}^{(1)}\left(\boldsymbol{\alpha}_1, \sigma_1\right)}{1 - \tilde{\pi}_{\mathbf{0}}^{(1)}\left(\boldsymbol{\alpha}_1, \sigma_1\right)}\right]^{r_{\mathbf{x}}^{(1)}} \times \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!}\left[1 - \tilde{\pi}_{\mathbf{0}}^{(1)}\left(\boldsymbol{\alpha}_1, \sigma_1\right)\right]^{r_1}\left[\tilde{\pi}_{\mathbf{0}}^{(1)}\left(\boldsymbol{\alpha}_1, \sigma_1\right)\right]^{\tau_1 - m - r_1} \\
&= L_{11}\left(\boldsymbol{\alpha}_1, \sigma_1\right) L_{12}\left(\tau_1, \boldsymbol{\alpha}_1, \sigma_1\right)
\end{aligned}$$

et

$$
\begin{aligned}
L_2(\tau_2, \boldsymbol{\alpha}_2, \sigma_2) &\propto f\left(\left\{r_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}, \tau_2 - r_2 \middle| \{m_i\}, \tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) \\
&= f\left(\left\{r_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \middle| \{m_i\}, \tau_2, r_2, \boldsymbol{\alpha}_2, \sigma_2\right) f\left(r_2 \middle| \{m_i\}, \tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) \\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\frac{\tilde{\pi}_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)}{1 - \tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)}\right]^{r_{\mathbf{x}}^{(2)}} \times \frac{\tau_2!}{(\tau_2 - r_2)!}\left[1 - \tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{r_2}\left[\tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{\tau_2 - r_2} \\
&= L_{21}(\boldsymbol{\alpha}_2, \sigma_2) L_{22}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2).
\end{aligned}
$$

Observons que, dans chaque cas, le premier facteur $L_{k1}(\boldsymbol{\alpha}_k, \sigma_k)$ est proportionnel à la fmp conjointe conditionnelle des $\left\{R_{\mathbf{x}}^{(k)}\right\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, sachant que $\{M_i = m_i\}_1^n$ et $R_k = r_k$, ce qui est la loi multinomiale de paramètre de taille r_k et de probabilités $\left\{\tilde{\pi}_{\mathbf{x}}^{(k)} \middle/ \left[1 - \tilde{\pi}_{\mathbf{0}}^{(k)}\right]\right\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, et que cette loi ne dépend pas de τ_k. Notons aussi que les seconds facteurs $L_{12}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1)$ et $L_{22}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2)$ sont proportionnels aux fmp conditionnelles de R_1 et R_2, sachant que $\{M_i = m_i\}_1^n$, qui sont les lois $\text{Bin}\left(\tau_1 - m, 1 - \tilde{\pi}_{\mathbf{0}}^{(1)}\right)$ et $\text{Bin}\left(\tau_2, 1 - \tilde{\pi}_{\mathbf{0}}^{(2)}\right)$, respectivement, où $\text{Bin}(\tau, \theta)$ désigne la loi binomiale de paramètre de taille τ et de probabilité θ.

Les EMVC $\hat{\boldsymbol{\alpha}}_k^{(C)}$ et $\hat{\sigma}_k^{(C)}$ de $\boldsymbol{\alpha}_k$ et $\sigma_k, k = 1, 2$ s'obtiennent en maximisant numériquement

$$
L_{11}(\boldsymbol{\alpha}_1, \sigma_1) L_0(\boldsymbol{\alpha}_1, \sigma_1) \quad \text{et} \quad L_{21}(\boldsymbol{\alpha}_2, \sigma_2) \tag{3.6}
$$

par rapport à $(\boldsymbol{\alpha}_1, \sigma_1)$ et $(\boldsymbol{\alpha}_2, \sigma_2)$, respectivement. Notons que, dans (3.6), les facteurs ne dépendent pas de $\tau_k, k = 1, 2$.

Enfin, en introduisant les estimations $\hat{\boldsymbol{\alpha}}_k^{(C)}$ et $\hat{\sigma}_k^{(C)}$ dans les facteurs de la fonction de vraisemblance qui dépendent de $\tau_k, k = 1, 2$, et en maximisant ces facteurs, c'est-à-dire en maximisant $L_{12}\left(\tau_1, \hat{\boldsymbol{\alpha}}_1^{(C)}, \hat{\sigma}_1^{(C)}\right) L_{\text{MULT}}(\tau_1)$ et $L_{22}\left(\tau_2, \hat{\boldsymbol{\alpha}}_2^{(C)}, \hat{\sigma}_2^{(C)}\right)$ par rapport à τ_1 et τ_2, respectivement, nous obtenons que les EMVC $\hat{\tau}_1^{(C)}$ et $\hat{\tau}_2^{(C)}$ de τ_1 et τ_2 sont donnés par (3.5) en remplaçant $\hat{\boldsymbol{\alpha}}_k^{(U)}$ et $\hat{\sigma}_k^{(U)}$ par $\hat{\boldsymbol{\alpha}}_k^{(C)}$ et $\hat{\sigma}_k^{(C)}, k = 1, 2$. Observons que ces expressions pour $\hat{\tau}_1^{(C)}$ et $\hat{\tau}_2^{(C)}$ sont des expressions analytiques. L'EMVC de τ est $\hat{\tau}^{(C)} = \hat{\tau}_1^{(C)} + \hat{\tau}_2^{(C)}$.

4 Intervalles de confiance

Nous considérons deux types d'intervalles de confiance (IC) pour les tailles de population : les IC par vraisemblance profilée et par bootstrap.

4.1 Intervalles de confiance par vraisemblance profilée

Plusieurs auteurs, dont Cormack (1992), Evans, Kim et O'Brien (1996), Coull et Agresti (1999) et Gimenes, Choquet, Lamor, Scofield, Fletcher, Lebreton et Pradel (2005), ont indiqué que, dans le contexte de l'échantillonnage par capture-recapture, les intervalles de confiance par vraisemblance profilée (ICVP) donnent de meilleurs résultats que les IC de Wald classiques quand la taille de l'échantillon n'est pas

grande. Certains facteurs qui influent sur les propriétés des IC de Wald sont le biais des estimateurs de la taille de la population, le biais des estimateurs de variance et l'asymétrie des lois de probabilité des estimateurs de la taille de population. En outre, un IC de Wald pour la taille de population pourrait présenter l'inconvénient d'avoir une borne inférieure dont la valeur est plus petite que le nombre d'éléments capturés. Notons que, mis à part le premier facteur énuméré, aucun autre n'a d'incidence sur les propriétés des ICVP. En outre, Evans et coll. (1996), en s'inspirant de Ratkowsky (1988), indiquent que la nature non linéaire des estimateurs par capture-recapture est mieux approximée par les IC basés sur la vraisemblance que par les IC de Wald.

Puisque les estimateurs proposés ressemblent à ceux utilisés dans l'échantillonnage par capture-recapture et compte tenu des commentaires qui précèdent, nous devrions nous attendre à ce que, dans notre cas aussi, les ICVP donnent de meilleurs résultats que les IC de Wald. Il convient de souligner que, si nous voulions utiliser les IC de Wald, nous devrions calculer les estimateurs des variances des estimateurs proposés. Une autre option consiste à construire des estimateurs de variance asymptotiques en utilisant les résultats de Sanathanan (1972); cependant, si n est grand, disons égal ou supérieur à 20, l'obtention de ce type d'estimateur nécessite de très importantes ressources informatiques, parce que chaque estimateur requiert la construction d'une matrice symétrique de dimensions $(n+1)\times(n+1)$ dont les éléments sont les sommes de 2^n termes.

Afin d'obtenir les ICVP pour τ_1, τ_2 et τ, nous suivrons l'approche de Coull et Agresti (1999). Donc, pour des valeurs fixées τ_1, τ_2 et τ des tailles de population, soit r_{10}, r_{20} et r_{00} des nombres réels non négatifs tels que $\tau_1 = m + r_1 + r_{10}, \tau_2 = r_2 + r_{20}$ et $\tau = m + r_1 + r_2 + r_{00}$, où m, r_1 et r_2 sont les valeurs observées des variables aléatoires M, R_1 et R_2. Alors, les ICVP à $100(1-\alpha)\%$ pour τ_1, τ_2 et τ sont définis comme étant les ensembles suivants : $\left\{\tau_1 = m + r_1 + r_{10} : -2\ln[\Lambda_1(r_{10})] \le \chi^2_{1,1-\alpha}\right\}$, $\left\{\tau_2 = r_2 + r_{20} : -2\ln[\Lambda_2(r_{20})] \le \chi^2_{1,1-\alpha}\right\}$ et $\left\{\tau = m + r_1 + r_2 + r_{00} : -2\ln[\Lambda(r_{00})] \le \chi^2_{1,1-\alpha}\right\}$, respectivement, où

$$\Lambda_1(r_{10}) = \max_{\alpha_1,\sigma_1} L_{(1)}(m + r_1 + r_{10}, \alpha_1, \sigma_1) \big/ L_{(1)}(\hat{\tau}_1, \hat{\alpha}_1, \hat{\sigma}_1),$$

$$\Lambda_2(r_{20}) = \max_{\alpha_2,\sigma_2} L_{(2)}(r_2 + r_{20}, \alpha_2, \sigma_2) \big/ L_{(2)}(\hat{\tau}_2, \hat{\alpha}_2, \hat{\sigma}_2)$$

et

$$\Lambda(r_{00}) = \max_{r_{10},\alpha_1,\alpha_2,\sigma_1,\sigma_2} L(m + r_1 + r_{10}, r_2 + r_{00} - r_{10}, \alpha_1, \alpha_2, \sigma_1, \sigma_2) \big/ L(\hat{\tau}_1, \hat{\tau}_2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}_1, \hat{\sigma}_2),$$

$\hat{\tau}_k, \hat{\alpha}_k$ et $\hat{\sigma}_k$ sont soit les EMVI ou les EMVC de τ_k, α_k et $\sigma_k, k = 1, 2$, et $\chi^2_{1,1-\alpha}$ est le $100(1-\alpha)^{\text{e}}$ quantile de la loi du khi-deux à 1 degré de liberté.

Bien que les ICVP pour τ_2 ne soient pas affectés par une variation extrapoissonnienne éventuelle des M_i [ou strictement parlant, une dispersion extramultinomiale de $(M_1, \ldots, M_n)$], comme ils sont obtenus à partir de la fonction de vraisemblance $L_{(2)}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2)$ qui ne dépend pas de ces variables, nous ne nous attendons pas à ce que les ICVP pour τ_1 et τ soient robustes à une variation extrapoissonnienne des M_i; par conséquent, nous considérons les ICVP ajustés pour τ_1 et τ qui tiennent compte de cette variation supplémentaire. Comme l'ont suggéré Gimenes et coll. (2005), nous construisons les ICVP ajustés comme les précédents, mais remplaçons la valeur de $\chi^2_{1,1-\alpha}$ par la valeur de $\left(s^2_M / \bar{m}\right) F_{1,n-1,1-\alpha}$, où $\bar{m} = m/n$ et $s^2_M = \sum_1^n (m_i - \bar{m})^2 \big/ (n-1)$ sont la moyenne et la variance d'échantillon des m_i, et $F_{1,n-1,1-\alpha}$ est le

$100(1-\alpha)^{e}$ quantile de la distribution F à 1 et à $n-1$ degrés de liberté. Observons que l'on obtient $s_M^2/\bar{m}$ en divisant par $n-1$ la valeur de la statistique de test du khi-deux de Pearson pour tester l'hypothèse que la loi conditionnelle des M_i, sachant que $\sum_1^n M_i = m$, est multinomiale de paramètre de taille m et de vecteur de probabilités $(1/n,\ldots,1/n)$. Les ICVP ajustés devraient être utilisés si l'hypothèse nulle de la loi multinomiale conditionnelle des M_i était rejetée au seuil de signification de 100α %, c'est-à-dire si $s_M^2/\bar{m} > \chi^2_{n-1,1-\alpha}/(n-1)$, sinon, les ICVP non ajustés devraient être utilisés.

Il convient de souligner que le calcul des ICVP est une tâche gourmande en ressources informatiques; par conséquent, il faut utiliser des algorithmes numériques efficients, tels que celui proposé par Venzon et Moolgavkar (1988).

4.2 Intervalles de confiance par bootstrap

Nous présentons une variante de la méthode du bootstrap pour construire les IC pour les tailles de population τ_1, τ_2 et τ basés sur les EMVI ou les EMVC. La variante proposée s'obtient en combinant la version bootstrap pour populations finies proposées par Booth, Butler et Hall (1994) et la variante du bootstrap paramétrique (voir Davison et Hinkley 1997, chapitre 2). Cette version du bootstrap est une extension de celle utilisée par Félix-Medina et Monjardin (2006) dans le cas de probabilités de lien homogènes.

Comme notre version proposée du bootstrap est une variante paramétrique, nous avons besoin des estimations de tous les paramètres associés aux modèles supposés. Les seuls paramètres qui n'ont pas encore été estimés jusqu'ici sont les effets aléatoires $\beta_j^{(k)}$. Nous allons maintenant obtenir un prédicteur de $\beta_j^{(k)}$. Donc, sachant le sous-ensemble $U_k - S_0$ ou $A_i \in S_A$ qui contient l'élément j, la fonction de densité conjointe conditionnelle de $\mathbf{X}_j^{(k)}$ et $\beta_j^{(k)}$ est

$$\begin{aligned} f\left(\mathbf{x}_j^{(k)}, \beta_j^{(k)} \middle| j \in U_k - S_0, S_A\right) &= \Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x}_j^{(k)} \middle| \beta_j^{(k)}, j \in U_k - S_0, S_A\right) f\left(\beta_j^{(k)}\right) \\ &\propto \prod_{i=1}^{n} \left[p_{ij}^{(k)}\right]^{x_{ij}^{(k)}} \left[1 - p_{ij}^{(k)}\right]^{1-x_{ij}^{(k)}} \exp\left[-\left(\beta_j^{(k)}\right)^2 \Big/ 2\sigma_k^2\right] \\ &\quad \text{si } j \in U_k - S_0,\ k = 1,2, \end{aligned}$$

ou

$$\begin{aligned} f\left(\mathbf{x}_j^{(1)}, \beta_j^{(1)} \middle| j \in A_{i'} \in S_A, S_A\right) &\propto \prod_{i \neq i'}^{n} \left[p_{ij}^{(1)}\right]^{x_{ij}^{(1)}} \left[1 - p_{ij}^{(1)}\right]^{1-x_{ij}^{(1)}} \exp\left[-\left(\beta_j^{(1)}\right)^2 \Big/ 2\sigma_1^2\right] \\ &\quad \text{si } j \in A_{i'} \in S_A,\ i' = 1,\ldots,n. \end{aligned}$$

Nous allons utiliser comme prédiction ou estimation de $\beta_j^{(k)}$ la valeur $\hat{\beta}_j^{(k)}$ qui maximise la fonction de densité conjointe conditionnelle de $\mathbf{X}_j^{(k)}$ et $\beta_j^{(k)}$ en fixant les paramètres α_k et σ_k à leur EMVI ou leur EMVC. Par cette procédure, $\hat{\beta}_j^{(k)}$ est donné par la solution de l'équation :

$$\sum_{i=1}^{n} x_{ij}^{(k)} - \sum_{i=1}^{n} \frac{\exp\left[\hat{\alpha}_i^{(k)} + \beta_j^{(k)}\right]}{1 + \exp\left[\hat{\alpha}_i^{(k)} + \beta_j^{(k)}\right]} - \frac{1}{\hat{\sigma}_k^2}\beta_j^{(k)} = 0 \quad \text{if } j \in U_k - S_0,\ k = 1,2,$$

ou

$$\sum_{i \neq i'}^{n} x_{ij}^{(1)} - \sum_{i \neq i'}^{n} \frac{\exp\left[\hat{\alpha}_i^{(1)} + \beta_j^{(1)}\right]}{1 + \exp\left[\hat{\alpha}_i^{(1)} + \beta_j^{(1)}\right]} - \frac{1}{\hat{\sigma}_1^2}\beta_j^{(1)} = 0 \quad \text{if} \quad j \in A_{i'} \in S_A,\ i' = 1, \ldots, n,$$

où $\hat{\alpha}_i^{(k)}$ et $\hat{\sigma}_k$ désignent les EMVI ou les EMVC de $\alpha_i^{(k)}$ et $\sigma_k, i = 1, \ldots, n; k = 1, 2$. Notons que cette équation implique que le prédicteur $\hat{\beta}_j^{(k)}$ de $\beta_j^{(k)}$ dépend du nombre de grappes qui sont liées à l'élément j, mais non des grappes particulières auxquelles cet élément est lié. Donc, si deux personnes j et j' dans $U_k - S_0$ sont liées au même nombre de grappes dans S_A, les prédicteurs $\hat{\beta}_j^{(k)}$ et $\hat{\beta}_{j'}^{(k)}$ sont égaux l'un à l'autre. Il en est de même pour deux personnes dans $A_i \in S_A$.

Dans la suite de l'exposé, nous désignerons par $[\hat{\tau}_k]$ l'entier le plus proche de $\hat{\tau}_k$, où $\hat{\tau}_k$ désigne soit l'EMVI ou l'EMVC de $\tau_k, k = 1, 2$. Les étapes de la procédure bootstrap proposée sont les suivantes. i) Construire un vecteur de population $\mathbf{m}_{\text{Boot}}$ de N valeurs des m_i en répétant N/n fois, sous l'hypothèse que N/n est un entier, l'échantillon observé de n tailles de grappe $\mathbf{m}_s = \{m_1, \ldots, m_n\}$. Si N/n n'est pas un entier, c'est-à-dire si $N = an + b$, où a et $b, b < n$ sont des entiers positifs, alors répéter a fois $\mathbf{m}_s$ et ajouter à cet ensemble un EASSR de b valeurs des m_i sélectionné à partir de $\mathbf{m}_s$. ii) Pour chaque $k = 1, 2$, construire un vecteur de population $\hat{\alpha}_{\text{Boot}}^{(k)}$ de dimension N dont les éléments sont les estimations $\hat{\alpha}_i^{(k)}$ des $\alpha_i^{(k)}$ associés aux grappes dont les tailles m_i sont dans $\mathbf{m}_{\text{Boot}}$. iii) Construire un vecteur de population $\hat{\beta}_{\text{Boot}}^{(0)}$ dont les éléments sont les estimations $\hat{\beta}_j^{(1)}$ des $\beta_j^{(1)}$ associés aux personnes qui appartiennent aux grappes dont les tailles m_i sont dans $\mathbf{m}_{\text{Boot}}$. Observons que la dimension de ce vecteur n'est pas nécessairement $[\hat{\tau}_1]$, mais est égale à la somme des m_i dans $\mathbf{m}_{\text{Boot}}$. iv) Construire un vecteur de population $\hat{\beta}_{\text{Boot}}^{(1)}$ de dimension $[\hat{\tau}_1]$ dont les m premiers éléments sont les estimations $\hat{\beta}_j^{(1)}$ des $\beta_j^{(1)}$ associés aux personnes dans S_0; les $[\hat{\tau}_1] - m$ éléments restants sont les r_1 estimations $\hat{\beta}_j^{(1)}$ des $\beta_j^{(1)}$ associés aux personnes dans S_1 et les $[\hat{\tau}_1] - m - r_1$ estimations $\hat{\beta}_j^{(1)}$ des $\beta_j^{(1)}$ associés aux personnes non échantillonnées dans U_1. Ces $[\hat{\tau}_1] - m$ éléments $\hat{\beta}_j^{(1)}$ sont placés aléatoirement après les m premiers éléments $\hat{\beta}_j^{(1)}$ de $\hat{\beta}_{\text{Boot}}^{(1)}$. v) Construire un vecteur de population $\hat{\beta}_{\text{Boot}}^{(2)}$ de dimension $[\hat{\tau}_2]$ dont les r_2 premiers éléments sont les estimations $\hat{\beta}_j^{(2)}$ des $\beta_j^{(2)}$ associés aux personnes dans S_2 et les $[\hat{\tau}_2] - r_2$ éléments restants sont les estimations $\hat{\beta}_j^{(2)}$ des $\beta_j^{(2)}$ associés aux personnes non échantillonnées dans U_2. vi) Sélectionner un EASSR de n valeurs m_i à partir de $\mathbf{m}_{\text{Boot}}$. Soit $S_A^{\text{Boot}} = \{i_1, \ldots, i_n\}$ l'ensemble d'indices des m_i dans l'échantillon. En outre, soit $A_i^{\text{Boot}} = \left(\sum_{t=1}^{i-1} m_t, \sum_{t=1}^{i} m_t\right) \cap \mathbb{Z}$ l'ensemble d'indices j associé aux éléments dans la grappe dont l'indice est $i \in S_A^{\text{Boot}}$, où m_t est le t^{e} élément de $\mathbf{m}_{\text{Boot}}$ et $\mathbb{Z}$ est l'ensemble de nombres entiers. Enfin, soit $S_0^{\text{Boot}} = \bigcup_{i \in S_A^{\text{Boot}}} A_i^{\text{Boot}}$. vii) Pour chaque $i \in S_A^{\text{Boot}}$ et $j \in \{1, \ldots, [\hat{\tau}_2]\}$, générer une valeur $x_{ij}^{(2)}$ par échantillonnage à partir de la loi de Bernoulli de moyenne $\hat{p}_{ij}^{(2)}$ donnée par (3.2), mais en remplaçant $\alpha_i^{(2)}$ et $\beta_j^{(2)}$ par leurs estimations $\hat{\alpha}_i^{(2)}$ et $\hat{\beta}_j^{(2)}$. De même, pour chaque $i \in S_A^{\text{Boot}}$ et $j \in \{1, \ldots, [\hat{\tau}_1]\} - A_i^{\text{Boot}}$, générer une valeur $x_{ij}^{(1)}$ par échantillonnage à partir de la loi de Bernoulli de moyenne $\hat{p}_{ij}^{(1)}$, où la valeur de $\hat{\beta}_j^{(1)}$ qui est utilisée pour calculer $\hat{p}_{ij}^{(1)}$ est obtenue à partir de $\hat{\beta}_{\text{Boot}}^{(0)}$ si $j \in S_0^{\text{Boot}}$, et à partir de $\hat{\beta}_{\text{Boot}}^{(1)}$ autrement. viii) Calculer les estimations de τ_1, τ_2 et τ en suivant la même procédure que celle utilisée pour calculer les estimations originales $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$.

ix) Répéter les étapes vi) à viii) un nombre suffisamment grand B de fois. Soit $\hat{\tau}_{1,b}^{\text{Boot}}, \hat{\tau}_{2,b}^{\text{Boot}}$ et $\hat{\tau}_b^{\text{Boot}}$ les estimations obtenues dans le b^{e} échantillon bootstrap, $b = 1, \ldots, B$.

L'étape finale de la variante bootstrap que nous proposons est la construction des IC pour les tailles de population. Il existe plusieurs moyens de le faire. L'un consiste à construire les intervalles sans émettre aucune hypothèse distributionnelle pour les estimateurs $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. La méthode de base et la méthode des centiles sont des exemples de cette option. (Voir Davison et Hinkley 1997, chapitre 5, pour la description de ces méthodes.) Dans la méthode de base, un IC à $100(1-\alpha)\,\%$ pour τ est donné par $\left[2\hat{\tau} - \hat{\tau}_{1-\alpha/2}^{\text{Boot}}, 2\hat{\tau} - \hat{\tau}_{\alpha/2}^{\text{Boot}}\right]$, et dans le cas de la méthode des centiles, l'IC est donné par $\left[\hat{\tau}_{\alpha/2}^{\text{Boot}}, \hat{\tau}_{1-\alpha/2}^{\text{Boot}}\right]$, où $\hat{\tau}_{\alpha/2}^{\text{Boot}}$ et $\hat{\tau}_{1-\alpha/2}^{\text{Boot}}$ sont les points $\alpha/2$ inférieur et supérieur de la loi empirique obtenue à partir de $\hat{\tau}_b^{\text{Boot}}, b = 1, \ldots, B$. Bien que ce type d'option ait de bonnes propriétés de robustesse, elle requiert un grand nombre B d'échantillons bootstrap, disons $B = 1\,000$, ce qui pourrait être un sérieux problème si le calcul de $\hat{\tau}$ est onéreux.

Une autre option pour construire les IC consiste à émettre l'hypothèse d'une loi pour $\hat{\tau}$ et à utiliser l'échantillon bootstrap pour estimer les paramètres de cette loi. Dans ce cas, le nombre B d'échantillons bootstrap requis n'est pas aussi grand, disons que $50 \leq B \leq 200$ est généralement suffisant. Des exemples de cette option sont l'hypothèse que $\hat{\tau}$ suit une loi normale et l'hypothèse que $\hat{\tau} - \nu$ suit une loi log-normale, où ν est le nombre d'éléments échantillonnés. Dans le premier cas, un IC à $100(1-\alpha)\,\%$ de τ est l'IC bien connu de Wald donné par $\hat{\tau} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})}$, tandis que dans le deuxième cas, l'IC est $[\nu + (\hat{\tau} - \nu)/c,\ \nu + (\hat{\tau} - \nu) \times c]$, où $c = \exp\left\{z_{\alpha/2}\sqrt{\ln\left[1 + \hat{V}(\hat{\tau})/(\hat{\tau} - \nu)^2\right]}\right\}$, $z_{\alpha/2}$ est le point $\alpha/2$ supérieur de la loi normale centrée réduite et $\hat{V}(\hat{\tau})$ est une estimation de la variance de $\hat{\tau}$. (Voir Williams, Nichols et Conroy 2002, section 14.2, pour une description de ce type d'IC.) Il faut souligner que, dans le cas des IC fondés sur la loi log-normale pour τ_1, τ_2 et τ, les valeurs de ν sont $m + r_1, r_2$ et $m + r_1 + r_2$, respectivement.

Un estimateur $\hat{V}(\hat{\tau})$ de la variance de $\hat{\tau}$ pourrait être calculé en utilisant la variance d'échantillon de l'échantillon bootstrap $\hat{\tau}_b^{\text{Boot}}, b = 1, \ldots, B$. Toutefois, cet estimateur n'est pas robuste aux valeurs extrêmes de $\hat{\tau}_b^{\text{Boot}}$ qui seront vraisemblablement obtenues avec les estimateurs proposés quand les taux d'échantillonnage ne sont pas suffisamment grands. Par conséquent, l'utilisation d'un estimateur robuste de $V(\hat{\tau})$ est une meilleure stratégie. Une possibilité consiste à utiliser la proposition 2 de Huber pour estimer conjointement les paramètres de position et d'échelle à partir de l'échantillon bootstrap. (Voir Staudte et Sheather 1990, section 4.5, pour une description de cette méthode.) En particulier, l'estimation du paramètre d'échelle est une estimation de l'écart-type $\sqrt{\hat{V}(\hat{\tau})}$ de $\hat{\tau}$.

5 Détermination de la taille de l'échantillon

Nous présentons une procédure pour déterminer la taille de l'échantillon initial n. Cette procédure est fondée sur des hypothèses rigoureuses, mais comme l'a mentionné l'un des examinateurs de l'article, elle pourrait néanmoins être très utile aux chercheurs qui souhaitent appliquer ce plan d'échantillonnage.

La première étape consiste à calculer les variances asymptotiques des estimateurs proposés. Bien que les variances dépendent de plusieurs paramètres inconnus, nous pouvons les simplifier en supposant que les effets $\alpha_i^{(k)}$ des sites échantillonnés sont homogènes, c'est-à-dire que $\alpha_i^{(k)} = \alpha^{(k)}, i = 1, \ldots, n$;

$k = 1, 2$. Sous cette prémisse, les probabilités $\pi_{\mathbf{x}}^{(k)}$ et $\pi_{\mathbf{x}}^{(A_i)}$ que les vecteurs de variables indicatrices de lien associés aux personnes sélectionnées aléatoirement dans $U_k - S_0$ et S_0, respectivement, soient égaux à $\mathbf{x}$ dépend uniquement du nombre de 1 qui figurent dans le vecteur $\mathbf{x}$. Donc, leurs approximations par quadrature gaussienne, données par (3.3) et (3.4), se simplifient en

$$\tilde{\pi}_x^{(k)}\left(\boldsymbol{\theta}^{(k)}\right) = \tilde{\pi}_x^{(k)}\left(\alpha^{(k)}, \sigma_k\right) = \sum_{t=1}^{q} \frac{\exp\left[x\left(\alpha^{(k)} + \sigma_k z_t\right)\right]}{\left[1 + \exp\left(\alpha^{(k)} + \sigma_k z_t\right)\right]^n} \nu_t, \; x = 0, 1, \ldots, n,$$

et

$$\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(k)}\right) = \tilde{\pi}_x^{(A)}\left(\alpha^{(1)}, \sigma_1\right) = \sum_{t=1}^{q} \frac{\exp\left[x\left(\alpha^{(1)} + \sigma_1 z_t\right)\right]}{\left[1 + \exp\left(\alpha^{(1)} + \sigma_1 z_t\right)\right]^n} \nu_t, \; x = 0, 1, \ldots, n-1,$$

où $\boldsymbol{\theta}^{(k)} = \left(\theta_1^{(k)}, \theta_2^{(k)}\right) = \left(\alpha^{(k)}, \sigma_k\right)$.

En suivant la procédure de Sanathanan (1972), nous obtenons que les variances asymptotiques des estimateurs proposés sont données par

$$V(\hat{\tau}_k) = \tau_k \big/ \left(D_k - \mathbf{B}_k' \mathbf{A}_k^{-1} \mathbf{B}_k\right), \; k = 1, 2, \text{ et } V(\hat{\tau}) = V(\hat{\tau}_1) + V(\hat{\tau}_2),$$

où $\mathbf{A}_k = \left[a_{ij}^{(k)}\right]$ est une matrice de dimensions 2×2 dont l'élément $a_{ij}^{(k)}$ est

$$a_{ij}^{(k)} = \begin{cases} \left(1 - \dfrac{n}{N}\right) \displaystyle\sum_{x=0}^{n} \binom{n}{x} \frac{1}{\tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)} \left[\frac{\partial \tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_i^{(1)}}\right] \left[\frac{\partial \tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_j^{(1)}}\right] \\ + \dfrac{n}{N} \displaystyle\sum_{x=0}^{n-1} \binom{n-1}{x} \frac{1}{\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)} \left[\frac{\partial \tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_i^{(1)}}\right] \left[\frac{\partial \tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_j^{(1)}}\right] & \text{si } k = 1 \\ \displaystyle\sum_{x=0}^{n} \binom{n}{x} \frac{1}{\tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)} \left[\frac{\partial \tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}{\partial \theta_i^{(2)}}\right] \left[\frac{\partial \tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}{\partial \theta_j^{(2)}}\right] & \text{si } k = 2, \end{cases}$$

$\mathbf{B}_k$ est un vecteur bidimensionnel dont les éléments sont

$$b_i^{(k)} = -\left[\partial \tilde{\pi}_0^{(k)}\left(\boldsymbol{\theta}^{(k)}\right) \big/ \partial \theta_i^{(k)}\right] \big/ \tilde{\pi}_0^{(k)}(\boldsymbol{\theta}^{(k)}), \; i = 1, 2$$

et D_k est un nombre réel donné par

$$D_k = \begin{cases} \left[1 - (1 - n/N)\, \tilde{\pi}_0^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)\right] \big/ \left[(1 - n/N)\, \tilde{\pi}_0^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)\right] & \text{si } k = 1 \\ \left[1 - \tilde{\pi}_0^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)\right] \big/ \tilde{\pi}_0^{(2)}\left(\boldsymbol{\theta}^{(2)}\right) & \text{si } k = 2. \end{cases}$$

Il convient de souligner que, dans le calcul des variances asymptotiques, nous avons émis les hypothèses que $\tau_k \to \infty, k = 1, 2, m_i \to \infty, i = 1, \ldots, N$ et que N et n sont des nombres fixes.

Pour obtenir les valeurs numériques des variances, nous devons spécifier les valeurs pour $\tau_k, \alpha^{(k)}, \sigma_k$ et n. Un moyen de le faire consiste à attribuer des valeurs à τ_k, à σ_k et à la proportion $\tilde{\pi}_{1+}^{(k)}$ de personnes dans U_k qui sont liées au moins à un site particulier A_i, qui est commune à tous les sites et est plus facile à spécifier que $\alpha^{(k)}$. Alors, pour un n donné, la valeur de $\alpha^{(k)}$ est la solution de l'équation

$$\tilde{\pi}_{1+}^{(k)} - \sum_{x=1}^{n} \binom{n-1}{x-1} \tilde{\pi}_{x}^{(k)} \left(\alpha^{(k)}, \sigma_k\right) = 0.$$

Une fois que $\alpha^{(k)}, k = 1, 2,$ est obtenu pour un n donné, nous pouvons calculer les valeurs numériques des variances $V(\hat{\tau}_k)$ et $V(\hat{\tau})$; les racines carrées des variances relatives $\sqrt{V(\hat{\tau}_k)/\tau_k^2}$ et $\sqrt{V(\hat{\tau})/\tau^2}$, et les fractions d'échantillonnage $f_1 = 1 - (1 - n/N)\tilde{\pi}_0^{(1)}$, $f_2 = 1 - \tilde{\pi}_0^{(2)}$ et $f = (f_1 \times \tau_1 + f_2 \times \tau_2)/\tau$. Si les valeurs des racines carrées des variances relatives n'étaient pas satisfaisantes, nous pourrions essayer différentes valeurs de n jusqu'à ce que nous obtenions des valeurs satisfaisantes.

Nous avons programmé cette procédure en langage R et les lecteurs intéressés peuvent obtenir le programme sur demande auprès des auteurs. Pour illustrer la procédure, supposons que nous ayons une base de sondage de $N = 150$ sites et que nous attribuions les valeurs $\tau_1 = 1\,200$, $\tau_2 = 400$, $\sigma_1 = \sigma_2 = 1$, $\tilde{\pi}_{1+}^{(1)} = 0{,}05$ et $\tilde{\pi}_{1+}^{(2)} = 0{,}04$; alors, pour $n = 15$, nous obtenons que $V(\hat{\tau}_1) = 4\,780{,}8$, $V(\hat{\tau}_2) = 11\,525{,}3$, $V(\hat{\tau}) = 16\,306{,}1$, $\sqrt{V(\hat{\tau}_1)}/\tau_1 = 0{,}06$, $\sqrt{V(\hat{\tau}_2)}/\tau_2 = 0{,}27$, $\sqrt{V(\hat{\tau})}/\tau = 0{,}08$, $f_1 = 0{,}50$, $f_2 = 0{,}38$ et $f = 0{,}47$.

6 Études Monte Carlo

6.1 Populations construites en utilisant des données artificielles

Nous avons construit quatre populations artificielles; une description de chacune est présentée au tableau 6.1. Soulignons que, pour les populations I, III et IV, les $N = 150$ valeurs des m_i ont été obtenues par échantillonnage d'une loi de Poisson, tandis que pour la population II, elles ont été obtenues par tirage à partir d'une loi binomiale négative tronquée à zéro. En outre, pour les populations I et II, les probabilités de lien $p_{ij}^{(k)}$ ont été générées au moyen du modèle de Rasch (3.2). Pour la population III, elles ont été générées au moyen de ce modèle, mais les effets aléatoires $\beta_j^{(k)}$ ont été obtenus par échantillonnage d'une loi t de Student normée à six degrés de liberté et de variance égale à l'unité plutôt que par échantillonnage d'une loi normale centrée réduite. Enfin, pour la population IV, les $p_{ij}^{(k)}$ ont été générées au moyen du modèle à classes latentes suivant proposé par Pledger (2000) dans le contexte des études par capture-recapture : $p_{ij}^{(k)} = \exp\left[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}\right] \Big/ \left\{1 + \exp\left[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}\right]\right\}$, $i = 1, \ldots, n$; $j = 1, 2,$ et $k = 1, 2$. Dans ce modèle, les personnes présentes dans U_k sont réparties en deux classes latentes $(j = 1, 2)$ en fonction de leur propension à être liées aux grappes échantillonnées. La probabilité qu'une personne aléatoire dans U_k soit dans la classe j est $p_j^{(k)}$ et les $p_{ij}^{(k)}$ sont les mêmes pour toutes les personnes dans la classe j.

L'expérience en simulation a été exécutée en procédant à la sélection répétée de r échantillons à partir de chaque population en utilisant le plan d'échantillonnage décrit à la section 2 avec une taille d'échantillon initiale de $n = 15$. Donc, chaque fois que la valeur m_i a été incluse dans un échantillon initial, la valeur $x_{ij}^{(k)}$ a été obtenue par échantillonnage à partir d'une loi de Bernoulli de moyenne $p_{ij}^{(k)}$. En raison des valeurs attribuées à n et aux paramètres qui figurent dans l'expression de $p_{ij}^{(k)}$, les taux d'échantillonnage résultants étaient de $f_1 \approx 0{,}5$ et $f_2 \approx 0{,}4$. Soulignons que les caractéristiques des populations et des échantillons utilisés dans la présente étude n'étaient pas motivées par celles d'une étude

réelle, puisque ce plan d'échantillonnage n'a pas encore été appliqué. Donc, les populations et les échantillons ont été construits uniquement dans le but d'analyser les propriétés des estimateurs ponctuels et des estimateurs d'intervalle proposés.

Tableau 6.1
Paramètres des populations simulées

Population I	Population II	Population III	Population IV
$N = 150$	$N = 150$	$N = 150$	$N = 150$
$M_i \sim$ Poisson	$M_i \sim$ binomiale négative tronquée à zéro	$M_i \sim$ Poisson	$M_i \sim$ Poisson
$E(M_i) = 8$	$E(M_i) = 8$	$E(M_i) = 8$	$E(M_i) = 8$
$V(M_i) = 8$	$V(M_i) = 24$	$V(M_i) = 8$	$V(M_i) = 8$
$\tau_1 = 1\,209$	$\tau_1 = 1\,208$	$\tau_1 = 1\,209$	$\tau_1 = 1\,209$
$\tau_2 = 400$	$\tau_2 = 400$	$\tau_2 = 400$	$\tau_2 = 400$
$\tau = 1\,609$	$\tau = 1\,608$	$\tau = 1\,609$	$\tau = 1\,609$
$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0{,}001}$	$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0{,}001}$	$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0{,}001}$	$\alpha_i^{(k)} = \dfrac{-12}{M_i^{1/2} + 0{,}001}$
$c_1 = -5{,}45$	$c_1 = -5{,}45$	$c_1 = -5{,}45$	$\mu^{(1)} = -1{,}1;\ \mu^{(2)} = -1{,}2$
$c_2 = -5{,}85$	$c_2 = -5{,}85$	$c_2 = -5{,}85$	$\beta_1^{(k)} = 1{,}5;\ \beta_2^{(k)} = 0$
$\beta_j^{(k)} \sim N(0,1)$	$\beta_j^{(k)} \sim N(0,1)$	$\beta_j^{(k)} \sim \sqrt{2/3}\,T_6$	$(\alpha\beta)_{ij}^{(k)} \sim N\left(0;1{,}25^2\right)$
			$p_1^{(k)} = 0{,}3 = 1 - p_2^{(k)}$

Pour chaque échantillon, nous avons considéré les estimateurs de τ_1, τ_2 et τ suivants : les EMVI proposés $\hat{\tau}_1^{(U)}, \hat{\tau}_2^{(U)}$ et $\hat{\tau}^{(U)}$; les EMVC proposés $\hat{\tau}_1^{(C)}, \hat{\tau}_2^{(C)}$ et $\hat{\tau}^{(C)}$; les EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ proposés par Félix-Medina et Thompson (2004) et obtenus sous l'hypothèse des probabilités de lien homogènes, et les estimateurs assistés par un modèle bayésien $\bar{\tau}_1, \bar{\tau}_2$ et $\bar{\tau}$ proposés par Félix-Medina et Monjardin (2006), également obtenus sous l'hypothèse d'homogénéité et en utilisant les lois initiales ci-après pour τ_1, τ_2 et $\alpha_i^{(k)} = \ln\left[p_i^{(k)} / \left(1 - p_i^{(k)}\right)\right]$, où $p_i^{(k)}$ est donnée par (3.2), mais en posant que $\beta_j^{(k)} = 0$: $\xi(\tau_1 \mid \lambda_1) \propto (N\lambda_1)^{\tau_1} / \tau_1!$ et $\xi(\lambda_1) \propto \lambda_1^{a_1 - 1} \exp(-b_1\lambda_1)$; $\xi(\tau_2 \mid \lambda_2) \propto \lambda_2^{\tau_2} / \tau_2!$ et $\xi(\lambda_2) \propto \lambda_2^{a_2 - 1} \exp(-b_2\lambda_2)$, et $\xi\left(\alpha_i^{(k)} \middle| \theta_k\right) \propto \exp\left[-\left(\alpha_i^{(k)} - \theta_k\right)^2 / 2\sigma_k^2\right]$ et $\xi(\theta_k) \propto \exp\left[-(\theta_k - \mu_k)^2 / 2\gamma_k^2\right]$, où $a_1 = 1{,}0,\ b_1 = 0{,}1, a_2 = 6{,}0, b_2 = 0{,}01, \mu_k = -3{,}5$ et $\sigma_k^2 = \gamma_k^2 = 9{,}0$. Ces valeurs attribuées aux paramètres des lois initiales ont rendu ces dernières pratiquement non informatives. Les approximations par quadrature gaussienne (3.3) et (3.4) des probabilités $\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ et $\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{-i}, \sigma_1\right)$ ont été calculées en utilisant $q = 40$ termes.

Pour évaluer la performance d'un estimateur $\hat{\tau}$ de τ, disons, nous avons examiné son biais relatif (br), la racine carrée de son erreur quadratique moyenne relative $\left(\sqrt{\mathrm{eqmr}}\right)$, et les médianes des erreurs

relatives d'estimation (mder) et des erreurs relatives absolues d'estimation (mdera), définis respectivement par $\text{br} = \sum_1^r (\hat{\tau}_i - \tau)/(r\tau)$, $\sqrt{\text{eqmr}} = \sqrt{\sum_1^r (\hat{\tau}_i - \tau)^2/(r\tau^2)}$, mder = médiane $\{(\hat{\tau}_i - \tau)/\tau\}$ et mdera = médiane $\{|(\hat{\tau}_i - \tau)/\tau|\}$, où $\hat{\tau}_i$ est la valeur de $\hat{\tau}$ obtenue dans le i^{e} échantillon, qui, dans le cas des estimateurs ponctuels, est de 10 000.

Nous avons également considéré les IC à 95 % suivants pour les τ : les ICVP proposés et les ICVP ajustés pour la variation extrapoissonnienne; les IC par bootstrap proposés basés sur $B = 100$ échantillons bootstrap et construits en supposant une loi log-normale pour $\hat{\tau} - \nu$ et en estimant $\sqrt{\hat{V}(\hat{\tau})}$ au moyen de l'estimateur du paramètre d'échelle de la proposition 2 de Huber avec valeur d'ajustement de $d = 1{,}5$; les IC de Wald fondés sur le plan de sondage obtenus à partir des EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ et proposés par Félix-Medina et Thompson (2004), et les IC de Wald fondés sur le plan de sondage obtenus à partir des estimateurs bayésiens $\bar{\tau}_1, \bar{\tau}_2$ et $\bar{\tau}$ et proposés par Félix-Medina et Monjardin (2006). Il convient de souligner que les ICVP et les ICVP ajustés ont été calculés en utilisant la méthode de Venzon et Moolgavkar (1988), ou un algorithme basé sur la définition d'un ICVP quand la première méthode n'a pas permis de trouver les bornes des intervalles.

Pour évaluer la performance d'un IC, nous avons examiné la probabilité de couverture (pc), la moyenne des longueurs relatives (mlr) et la médiane des longueurs relatives (mdlr), définies comme étant la proportion d'échantillons pour lesquels le paramètre était contenu dans l'intervalle, et la moyenne et la médiane des longueurs des intervalles divisées par la valeur du paramètre, respectivement. Comme l'exécution d'une étude en simulation sur les IC en utilisant un grand nombre d'échantillons répétés est une tâche qui prend beaucoup de temps, nous avons évalué la performance des ICVP pour τ_1 et τ_2 en utilisant $r = 1\,000$ échantillons; celle des ICVP pour τ en utilisant $r = 500$ échantillons, et celle des IC bootstrap en utilisant $r = 250$ échantillons. La performance des IC basés sur les estimateurs calculés sous l'hypothèse d'homogénéité a été évaluée en utilisant $r = 10\,000$ échantillons. L'étude numérique a été exécutée en utilisant le langage de programmation et le logiciel R [R Development Core Team (2013)].

Les résultats de l'étude sur les estimateurs des tailles de population sont présentés au tableau 6.2 et aux figures 6.1 et 6.2. Les principaux résultats sont les suivants. Les distributions des estimateurs EMVI $\hat{\tau}_1^{(U)}$ et EMVC $\hat{\tau}_1^{(C)}$ étaient plus ou moins symétriques autour de τ_1; donc, les deux mesures du biais (br et mder) avaient des valeurs similaires, ainsi que les deux mesures de variabilité ($\sqrt{\text{eqmr}}$ et mdera). Ces estimateurs ont donné tous deux des résultats acceptables, sauf pour la population III pour laquelle l'estimateur $\hat{\tau}_1^{(U)}$ présentait des problèmes modérés de biais et $\hat{\tau}_1^{(C)}$ présentait des problèmes de biais un peu plus graves. Les distributions des estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ étaient étalées vers la droite et possédaient de très longues queues. Par conséquent, les valeurs de br et eqmr avaient tendance à être grandes. Cependant, en ce qui concerne les médianes de leurs erreurs relatives (mder), ces estimateurs présentaient des problèmes modérés de biais pour les populations I et III et des problèmes sérieux pour la population IV. Pour ce qui est des médianes de leurs erreurs relatives absolues (mdera), ces estimateurs présentaient des problèmes modérés d'instabilité dans les trois premières populations et des problèmes sérieux dans la quatrième population. Les distributions des estimateurs $\hat{\tau}^{(U)}$ et $\hat{\tau}^{(C)}$ étaient similaires à celles des estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$; donc, les quantités br et $\sqrt{\text{eqmr}}$ étaient plus sensibles aux grandes valeurs que les quantités mder et mdera. Ces estimateurs donnaient tous deux des résultats acceptables

pour les populations I, II et IV; quoique, dans cette dernière population, les valeurs de $\sqrt{\text{eqmr}}$ étaient grandes pour les raisons mentionnées antérieurement. Dans la population III, les deux estimateurs posaient des problèmes de biais.

L'écart par rapport à la loi de Poisson supposée des M_i accroissait la variabilité de tous les estimateurs proposés, mais les augmentations n'étaient pas importantes, de sorte que nous considérons que ces estimateurs possèdent certaines propriétés de robustesse aux écarts par rapport à cette hypothèse. Les estimateurs proposés de τ_1 et τ étaient en outre robustes à l'écart par rapport au modèle de Rasch supposé pour les $p_{ij}^{(1)}$ (quoique les valeurs de $\sqrt{\text{eqmr}}$ des estimateurs de τ étaient grandes, celles de la médiane de leurs erreurs relatives absolues ne l'étaient pas). L'écart par rapport à la loi normale supposée des effets $\beta_j^{(k)}$ faisait que tous les estimateurs proposés présentaient des problèmes de surestimation. Aucun des deux types d'estimateurs proposés, EMVI et EMVC, ne donnait uniformément des résultats meilleurs que l'autre, mais les EMVI donnaient des résultats un peu meilleurs que les EMVC dans un plus grand nombre de cas que l'inverse.

Tableau 6.2
Biais relatifs, racines carrées des erreurs quadratiques moyennes relatives et médianes des erreurs relatives ainsi que des erreurs relatives absolues des estimateurs des tailles de population

Population		I				II				III				IV			
Taux d'échantillonnage		f_1 0,51		f_2 0,40		f_1 0,50		f_2 0,40		f_1 0,51		f_2 0,40		f_1 0,51		f_2 0,40	
Estimateur		br	$\sqrt{\text{eqmr}}$	mder	mdera	br	$\sqrt{\text{eqmr}}$	mder	mdera	br	$\sqrt{\text{eqmr}}$	mder	mdera	br	$\sqrt{\text{eqmr}}$	mder	mdera
EMV	$\hat{\tau}_1^{(U)}$	-0,01	0,06	-0,01	0,04	-0,00	0,08	-0,00	0,05	0,10	0,11	0,10	0,10	$-0,04_6^{0,20}$	0,08	-0,03	0,05
incond.	$\hat{\tau}_2^{(U)}$	-0,06	0,26	-0,11	0,17	$0,06^{0,02}$	0,35	-0,01	0,16	0,16	0,43	0,07	0,18	$0,04_2^{15}$	2,2	-0,19	0,25
hétér.	$\hat{\tau}^{(U)}$	-0,02	0,08	-0,03	0,05	$0,01^{0,02}$	0,10	0,01	0,06	0,11	0,15	0,10	0,10	$-0,02_7^{15}$	0,55	-0,6	0,08
EMV.	$\hat{\tau}_1^{(C)}$	-0,00	0,07	-0,01	0,05	0,01	0,07	0,00	0,05	0,18	0,19	0,17	0,17	$-0,05^{1,6}$	0,09	-0,05	0,07
cond.	$\hat{\tau}_2^{(C)}$	-0,04	0,26	-0,09	0,17	0,09	0,38	0,01	0,16	0,18	0,46	0,10	0,18	$0,12_2^{21}$	2,4	-0,14	0,23
hétér.	$\hat{\tau}^{(C)}$	-0,01	0,08	-0,02	0,05	0,03	0,11	0,01	0,06	0,18	0,22	0,16	0,16	$-0,00_2^{23}$	0,61	-0,06	0,08
EMV	$\tilde{\tau}_1$	-0,28	0,28	-0,28	0,28	-0,31	0,31	-0,31	0,31	-0,30	0,30	-0,30	0,30	-0,18	0,19	-0,18	0,18
homo-	$\tilde{\tau}_2$	-0,40	0,40	-0,40	0,40	-0,40	0,40	-0,40	0,40	-0,40	0,40	-0,40	0,40	-0,30	0,32	-0,32	0,32
gènes	$\tilde{\tau}$	-0,31	0,31	-0,31	0,31	-0,33	0,33	-0,33	0,33	-0,32	0,33	-0,32	0,32	-0,21	0,22	-0,21	0,21
EB	$\bar{\tau}_1$	-0,28	0,28	-0,28	0,28	-0,31	0,31	-0,31	0,31	-0,30	0,30	-0,30	0,30	-0,18	0,19	-0,18	0,18
homo-	$\bar{\tau}_2$	-0,39	0,39	-0,39	0,39	-0,39	0,40	-0,39	0,39	-0,39	0,40	-0,39	0,39	-0,27	0,30	-0,29	0,29
gènes	$\bar{\tau}$	-0,31	0,31	-0,31	0,31	-0,33	0,33	-0,33	0,33	-0,32	0,32	-0,32	0,32	-0,20	0,21	-0,20	0,20

Note : Les résultats sont fondés sur 10^4 échantillons. Un chiffre en exposant indique le pourcentage d'échantillons pour lesquels l'estimateur n'a pas été calculé à cause de problèmes de convergence numérique et un chiffre en indice inférieur indique le nombre de valeurs de l'estimateur qui étaient supérieures à 10^5 et qui n'ont pas été utilisées pour calculer son br et son eqmr. Les bornes supérieures pour les erreurs Monte Carlo des estimations de br et de $\sqrt{\text{eqmr}}$ des estimateurs des τ étaient les suivantes : τ_1 : 0,001 et 0,001; τ_2 : 0,004 et 0,011 pour les populations I à III, et 0,027 et 0,39 pour la population IV; τ : 0,001 et 0,002 pour les populations I à III, et 0,007 et 0,095 pour la population IV.

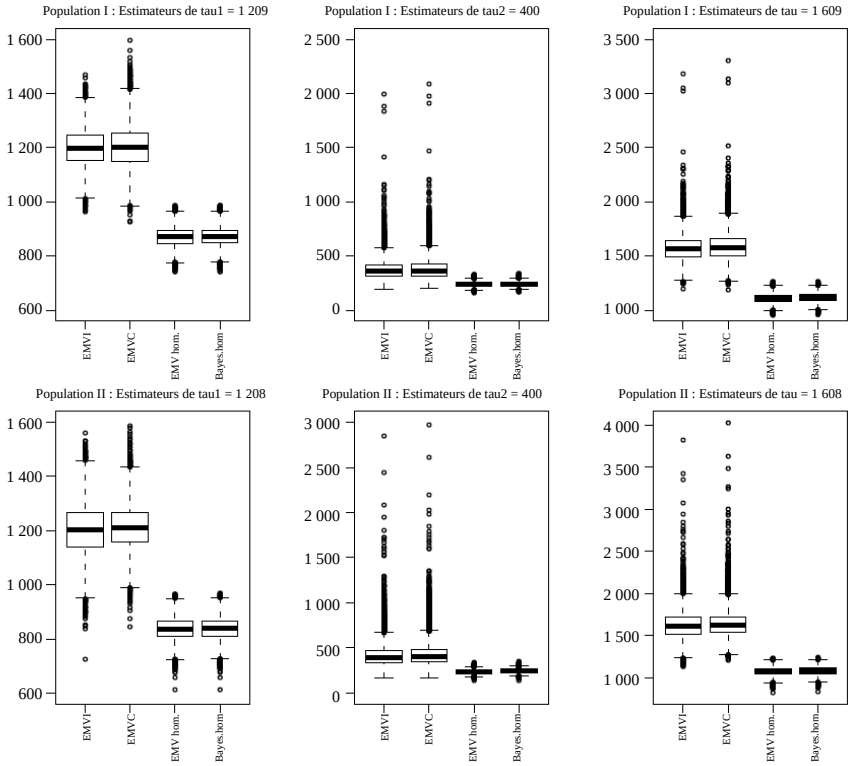

Figure 6.1 Boîtes à moustache pour les valeurs des estimateurs de τ_1, τ_2 et τ dans les populations I et II.

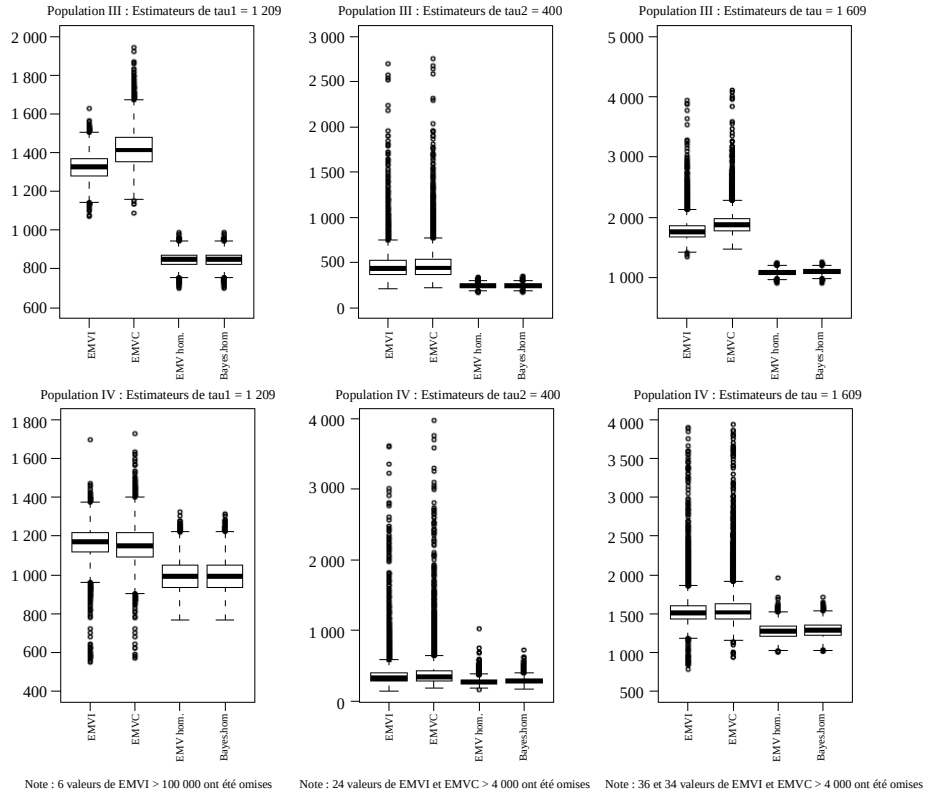

Figure 6.2 Boîtes à moustache pour les valeurs des estimateurs de τ_1, τ_2 et τ dans les populations III et IV.

En ce qui concerne les estimateurs établis sous l'hypothèse d'homogénéité des $p_{ij}^{(k)}$, les EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ ainsi que les estimateurs assistés par un modèle bayésien $\breve{\tau}_1, \breve{\tau}_2$ et $\breve{\tau}$ présentaient un comportement très semblable, caractérisé par de graves problèmes de biais qui détérioraient leur performance.

Notons que les pourcentages d'échantillons pour lesquels les estimateurs proposés n'ont pas été calculés en raison de problèmes de convergence numérique, ainsi que les nombres d'échantillons pour lesquels les valeurs des estimateurs étaient supérieures à 10^5 et n'ont pas été utilisées pour calculer les résultats présentés de br et de $\sqrt{\text{eqmr}}$ parce qu'ils auraient été sérieusement affectés, n'étaient pas grands, sauf pour la population IV. Comme l'a mentionné un examinateur, calculer le br et l'eqmr d'un estimateur en utilisant uniquement les valeurs disponibles inférieures à 10^5 favorise les estimateurs proposés. Nous sommes d'accord avec cette observation et, pour cette raison, nous avons également présenté les mesures de performance des estimateurs basées sur les médianes des erreurs relatives et des erreurs relatives absolues, qui sont robustes aux grandes valeurs des estimateurs. Donc, si nous supposions que, chaque fois qu'un estimateur n'a pas été calculé, sa valeur était très grande et que nous calculions les valeurs des mesures de performance des estimateurs fondées sur les médianes en utilisant l'ensemble complet d'observations, les résultats ne seraient pas différents de ceux présentés au tableau 6.2 et nos conclusions fondées sur ces mesures ne seraient pas modifiées.

Les résultats de l'étude en simulation sur les IC à 95 % sont présentés au tableau 6.3. Les principaux résultats sont les suivants. Tous les ICVP et les ICVP ajustés pour τ_1, à savoir ceux fondés sur l'EMVI $\hat{\tau}_1^{(U)}$ et ceux fondés sur l'EMVC $\hat{\tau}_1^{(C)}$, présentaient de bonnes valeurs de la probabilité de couverture (pc) pour la population I. Les ICVP ajustés présentaient aussi de bonnes valeurs de la pc pour la population II, mais pas les ICVP ajustés dont les valeurs de la pc étaient relativement faibles. Pour la population III, les valeurs de la pc de tous les ICVP et ICVP ajustés pour τ_1 étaient faibles, tandis que pour la population IV, les valeurs n'étaient que légèrement faibles. Une bonne caractéristique de ces IC était qu'ils présentaient des valeurs assez acceptables de la mlr et de la mdlr dans chacune des situations examinées. L'ICVP pour τ_2 basé sur $\hat{\tau}_2^{(U)}$ et celui basé sur $\hat{\tau}_2^{(C)}$ présentaient des valeurs acceptables de la pc pour toutes les populations, sauf la population IV, où les valeurs étaient un peu faibles. Cependant, dans tous les cas, la mlr et la mdlr de ces IC étaient si grandes qu'elles n'étaient pas utiles pour faire des inférences raisonnables. Les deux type d'ICVP pour τ, ceux fondés sur l'EMVI $\hat{\tau}^{(U)}$ et ceux fondés sur l'EMVC $\hat{\tau}^{(C)}$, ont donné des résultats acceptables pour les populations I et II, quoique les moyennes de leurs longueurs relatives étaient grandes pour la population II, parce que cette mesure n'est pas robuste aux grandes valeurs des longueurs des intervalles. Pour les autres populations, ces IC posaient des problèmes de faible couverture et/ou de grandes longueurs relatives; donc, leur performance n'était pas bonne. Les deux types d'ICVP ajustés ne donnaient des bons résultats que pour la population I, et présentaient de grandes valeurs de leurs longueurs relatives pour les autres populations. Ni l'un ni l'autre type d'IC, c'est-à-dire ceux fondés sur les EMVI et ceux fondés sur les EMVC, ne donnait systématiquement de meilleurs résultats que l'autre, mais ceux basés sur les EMVI donnaient d'un peu meilleurs résultats que ceux fondés sur les EMVC dans un plus grand nombre de cas qu'inversement.

En ce qui concerne les IC bootstrap, nous constatons que les deux types d'IC pour τ_1, celui fondé sur $\hat{\tau}_1^{(U)}$ et celui fondé sur $\hat{\tau}_1^{(C)}$, donnent de bons résultats pour les populations I, II et IV, quoique pour cette dernière, les valeurs de la pc soient un peu faibles. Pour la population III, la valeur de la pc est très faible à

cause des biais des estimateurs ponctuels de τ_1. Les deux types d'IC bootstrap pour τ_2 ont donné de mauvais résultats pour toutes les populations, parce que les valeurs de leurs longueurs relatives étaient grandes. Enfin, les deux types d'IC pour τ ont en général donné de bons résultats. Notons que les valeurs de leur mlr avaient tendance à être grandes, parce que cette mesure n'est pas robuste aux grandes valeurs des longueurs, tandis que les valeurs de leur mdlr étaient acceptables. Ni l'un ni l'autre des deux types d'IC bootstrap ne donnait systématiquement de meilleurs résultats que l'autre, mais les IC basés sur les EMVI donnaient dans la plupart des cas de meilleurs résultats que ceux basés sur les EMVC.

Tableau 6.3
Probabilités de couverture, et moyennes et médianes des longueurs relatives des intervalles de confiance à 95 % pour les tailles de population

Population	I			II			III			IV		
Taux d'échantillonnage	f_1 0,51		f_2 0,40	f_1 0,50		f_2 0,40	f_1 0,51		f_2 0,40	f_1 0,51		f_2 0,40
Intervalle de conf.	pc	mlr	mdlr	pc	mlr	mdlr	pc	mlr	mdlr	pc	mlr	mdlr
ICVP- $\hat{\tau}_1^{(U)}$	0,95	0,22	0,22	0,85	0,22	0,22	0,61	0,24	0,24	$0,89^{1,6}$	0,23	0,24
ICVP adj.- $\hat{\tau}_1^{(U)}$	0,94	0,24	0,24	0,98	0,42	0,41	0,69	0,27	0,26	$0,90^{1,6}$	0,25	0,26
ICVP- $\hat{\tau}_2^{(U)}$	0,94	1,4	0,98	0,95	$2,8_2$	1,3	0,95	2,6	1,6	$0,77^{19}$	$7,1_6$	1,4
ICVP- $\hat{\tau}^{(U)}$	0,95	0,66	0,59	0,97	0,91	0,65	1,0	1,0	0,79	$0,86^{21}$	$2,1_1$	0,65
ICVP adj.- $\hat{\tau}^{(U)}$	0,92	0,75	0,62	$1,0^{7,0}$	$5,8_{33}$	2,1	$1,0^{0,20}$	$1,7_1$	0,87	$0,90^{22}$	$2,7_4$	0,78
IC-bootstrap- $\hat{\tau}_1^{(U)}$	0,94	0,23	0,23	0,94	0,33	0,31	0,59	0,25	0,25	$0,89^{0,40}$	0,24	0,25
IC-bootstrap- $\hat{\tau}_2^{(U)}$	0,87	1,4	0,86	0,97	3,9	1,6	0,97	4,7	2,1	$0,83^{13}$	$4,6_3$	0,90
IC-bootstrap- $\hat{\tau}^{(U)}$	0,93	0,38	0,28	0,99	0,97	0,49	0,98	1,1	0,52	$0,89^{13}$	$1,2_3$	0,31
ICVP- $\hat{\tau}_1^{(C)}$	0,96	0,24	0,23	0,91	0,25	0,24	0,76	0,31	0,29	$0,90^{2,1}$	0,25	0,25
ICVP adj.- $\hat{\tau}_1^{(C)}$	0,95	0,26	0,25	0,99	0,44	0,42	0,81	0,33	0,32	$0,90^{2,1}$	0,27	0,27
ICVP- $\hat{\tau}_2^{(C)}$	0,94	1,4	0,98	0,95	$2,7_2$	1,3	0,95	2,5	1,6	$0,85^{25}$	$7,5_4$	1,7
ICVP- $\hat{\tau}^{(C)}$	0,95	0,64	0,54	$0,94^{1,6}$	1,2	0,62	$0,86^{2,0}$	$3,1_1$	0,74	$0,93^{30}$	$2,9_1$	0,84
ICVP adj.- $\hat{\tau}^{(C)}$	0,96	0,82	0,59	$1,0^{7,6}$	$7,2_{33}$	2,6	$0,90^{3,8}$	$3,6_6$	1,2	$0,94^{32}$	$2,9_5$	0,90
IC-bootstrap- $\hat{\tau}_1^{(C)}$	0,96	0,29	0,29	0,98	0,31	0,31	0,48	0,40	0,39	$0,89^{1,6}$	0,31	0,30
IC-bootstrap- $\hat{\tau}_2^{(C)}$	0,90	1,5	1,0	0,98	4,7	1,7	0,97	5,5	2,4	$0,94^{24}$	$3,4_6$	1,3
IC-bootstrap- $\hat{\tau}^{(C)}$	0,95	0,44	0,34	1,0	1,1	0,49	0,96	1,3	0,62	$0,94^{25}$	$0,88_6$	0,43
IC de Wald- $\tilde{\tau}_1$	0,00	0,09	0,09	0,00	0,08	0,08	0,00	0,08	0,08	0,08	0,13	0,13
IC de Wald- $\tilde{\tau}_2$	0,00	0,17	0,16	0,00	0,17	0,16	0,00	0,16	0,16	0,18	0,34	0,31
IC de Wald- $\tilde{\tau}$	0,00	0,08	0,08	0,00	0,07	0,07	0,00	0,07	0,07	0,03	0,13	0,13
IC de Wald- $\bar{\tau}_1$	0,00	0,09	0,09	0,00	0,09	0,09	0,00	0,08	0,08	0,11	0,15	0,15
IC de Wald- $\bar{\tau}_2$	0,00	0,17	0,17	0,00	0,17	0,17	0,00	0,17	0,17	0,25	0,36	0,33
IC de Wald- $\bar{\tau}$	0,00	0,08	0,08	0,00	0,08	0,08	0,00	0,08	0,07	0,04	0,15	0,14

Note : Un chiffre en exposant indique le pourcentage d'échantillons pour lesquels l'intervalle de confiance (IC) n'a pas été calculé à cause de problèmes de convergence numérique et un chiffre en indice inférieur indique le nombre de valeurs de la longueur relative de l'IC qui dépassaient 10^5 et qui n'ont pas été utilisées pour calculer la mlr. Une borne supérieure (BS) des erreurs Monte Carlo (EMC) des pc était de 0,03. Les BS pour les EMC des estimations de la mlr étaient les suivantes : ICVP et ICVP adj. pour τ_1 : 0,003; ICVP pour τ_2 : 0,78; ICVP et ICVP adj. pour τ : 0,07 pour la population I et 0,56 pour les populations II à IV; IC bootstrap pour τ_1 : 0,005; IC bootstrap pour τ_2 : 0,1 pour la population I et 1,5 pour les populations II à IV; IC bootstrap pour τ : 0,02 pour la population I et 0,37 pour les populations II à IV.

En ce qui concerne les IC basés sur les estimateurs ponctuels calculés sous l'hypothèse d'homogénéité, tous présentaient une valeur nulle de la pc, sauf pour la population IV, où les valeurs différaient de zéro,

mais étaient quand même trop faibles. La mauvaise performance de ces IC en ce qui concerne la pc découlait des grands biais des estimateurs ponctuels. Donc, malgré les très petites valeurs de la mlr de ces intervalles, les très faibles valeurs de leur pc ne permettaient pas de faire des inférences raisonnables.

Observons que, pour les trois premières populations, les pourcentages d'échantillons pour lesquels les IC proposés n'ont pas été calculés en raison de problèmes de convergence numérique, ainsi que le nombre d'échantillons pour lesquels les valeurs de la longueur relative des IC étaient supérieures à 10^5 et qui n'ont pas été utilisées pour calculer les résultats présentés de la mlr parce que ceux-ci auraient été gravement affectés, n'étaient pas grands (moins de 4 %), sauf dans le cas des ICVP ajustés fondés sur $\hat{\tau}^{(U)}$ et sur $\hat{\tau}^{(C)}$ qui, pour la population II, n'ont pas été calculés dans environ 7 % des échantillons et dont les moyennes de leurs longueurs relatives ont été calculées sans utiliser les 33 valeurs supérieures à 10^5. Cependant, pour la population IV, certains pourcentages étaient proches de 20 % et d'autres proches de 30 %. Ces pourcentages élevés étaient attribuables, en partie, aux valeurs relativement grandes des pourcentages d'échantillons pour lesquels les estimateurs ponctuels correspondants n'avaient pas été calculés. Il est clair que calculer les mesures de performance d'un IC en utilisant uniquement les cas pour lesquels l'intervalle a été obtenu ou calculer la mlr en utilisant uniquement les échantillons pour lesquels les longueurs relatives étaient inférieures à 10^5 favorise les IC proposés. Cependant, notons que dans presque tous les cas où ces pourcentages étaient élevés, disons supérieurs à 5 %, la mlr ainsi que la mdlr des intervalles étaient suffisamment grandes pour que ces intervalles ne soient pas utiles pour faire des inférences. Par conséquent, si la performance de ces IC n'était pas bonne sous cette évaluation favorable, elle ne sera pas bonne sous une évaluation plus équitable. Les exceptions à cette tendance étaient l'ICVP basé sur $\hat{\tau}_1^{(U)}$, et les deux types d'IC bootstrap pour τ, qui donnaient des résultats acceptables et qui n'étaient pas calculés pour un pourcentage élevé d'échantillons. Donc, les résultats de ces IC doivent être interprétés avec prudence.

6.2 Population construite en utilisant les données de l'étude de Colorado Springs sur la transmission du VIH/Sida

Dans cette étude en simulation, nous avons construit une population en utilisant des données provenant de l'étude de Colorado Springs sur la transmission hétérosexuelle du VIH/Sida. Comme il est indiqué dans l'introduction du présent article, cette étude épidémiologique était axée sur une population de personnes qui vivaient dans la région métropolitaine de Colorado Springs de 1982 à 1992 et qui couraient un risque élevé d'acquérir et de transmettre le VIH. Cette population comprenait des toxicomanes, des travailleurs de l'industrie du sexe et leurs contacts personnels, définis comme étant les personnes avec lesquelles ils entretenaient d'étroites relations sociales, sexuelles ou toxicomaniaques. Dans cette étude, 595 répondants initiaux ont été sélectionnés de façon non aléatoire dans une clinique de traitement des maladies transmises sexuellement, dans une clinique pour toxicomanes, par accès direct et par l'entremise de travailleurs de rue. On a demandé aux répondants de donner la liste complète de leurs contacts personnels, à la suite de quoi un total de 7 379 contacts qui ne figuraient pas dans l'ensemble de répondants initiaux ont été nommés et inclus dans l'étude. Dans notre étude en simulation, l'ensemble U_1 a été défini comme étant l'ensemble des 595 répondants initiaux et, à l'instar de Félix-Medina et Monjardin (2010), ils ont été

groupés en $N = 105$ grappes de tailles m_i générées en effectuant un tirage à partir d'une loi binomiale négative tronquée à zéro de paramètre de taille 2,5 et de probabilité $2/3$. La moyenne et la variance d'échantillon des 105 valeurs de m_i étaient 5,67 et 15,03, respectivement. Il convient de souligner que la plupart des personnes qui étaient affectées à une même grappe provenaient de la même source originale de recrutement. Une personne a été définie comme étant liée à une grappe s'il s'agissait d'un contact personnel d'au moins un élément dans la grappe en question. Puisqu'environ 95 % des 7 379 contacts des répondants initiaux n'étaient liés qu'à une seule grappe et que cela pouvait avoir une incidence sur la performance des estimateurs proposés, dans notre étude, nous avons défini l'ensemble U_2 comme étant le sous-ensemble des 7 379 contacts formé par les 415 personnes qui étaient liées à au moins deux grappes auxquelles nous avons ajouté les 379 travailleurs de l'industrie du sexe qui étaient liés à une grappe seulement. Donc, $\tau_1 = 595, \tau_2 = 794$ et $\tau = 1\,389$. Soulignons que cette population est la même que celle appelée « population réduite » par Félix-Medina et Monjardin (2010).

Nous avons fixé les tailles des échantillons initiaux sélectionnés à partir de la population à $n = 25$. Cette valeur de n a donné les taux d'échantillonnage $f_1 = 0{,}46$ et $f_2 = 0{,}37$. L'expérience en simulation a été exécutée comme la précédente, sauf que chaque fois que la valeur m_i était contenue dans un échantillon initial, toutes les personnes liées à la grappe i ont été incluses dans l'échantillon. Nous avons utilisé le même nombre r de répliques et le même nombre B d'échantillons bootstrap que dans l'étude précédente. En outre, les valeurs des paramètres des lois initiales qui ont été utilisées pour construire les estimateurs assistés par un modèle bayésien $\bar{\tau}_k$ et la valeur de q utilisée pour calculer les formules de quadrature gaussienne (3.3) et (3.4) étaient les mêmes que celles utilisées dans l'étude précédente.

Les résultats de l'étude en simulation sont présentés au tableau 6.4. Nous constatons que, parmi les estimateurs proposés des tailles de population, seuls les estimateurs de τ_1 ne présentent pas de problèmes de biais ni de problèmes d'instabilité. Les estimateurs de τ_2 et τ, particulièrement les estimateurs de τ_2, présentent de graves problèmes de biais qui ont une incidence sur leur performance. En raison de la performance des estimateurs ponctuels, seuls les ICVP ajustés et les IC par bootstrap pour τ_1 donnent des résultats acceptables, quoique les valeurs de la pc des IC bootstrap étaient un peu faibles. Les ICVP non ajustés pour τ_1 possèdent des valeurs faibles de la pc à cause de la divergence par rapport à la loi de Poisson supposée des M_i. Les ICVP et les IC bootstrap pour τ_2 et τ présentaient des très grandes valeurs de la mlr et de la mdlr rendant ces intervalles inutiles. Observons que les pourcentages d'échantillons pour lesquels les estimateurs ponctuels et d'intervalle proposés de τ_1 n'ont pas été calculés en raison de problèmes de convergence numérique étaient faibles (inférieurs à 1,2 %). Par conséquent, ces estimateurs n'étaient pratiquement pas favorisés par la procédure d'évaluation. Dans le cas des estimateurs ponctuels et d'intervalle proposés de τ_2 et τ, ces pourcentages étaient grands. Cependant, si la performance de ces estimateurs n'était pas bonne sous cette évaluation favorable, elle ne sera pas bonne dans des conditions d'évaluation plus équitables.

En ce qui concerne les estimateurs ponctuels calculés sous l'hypothèse d'homogénéité, nous constatons que les EMV $\tilde{\tau}_1$ et $\tilde{\tau}_2$ présentaient des problèmes de biais qui ont eu une incidence sur leur performance; par contre, l'estimateur $\tilde{\tau}$ ne présentait pas de problème de biais et sa performance était acceptable. Le petit biais de cet estimateur pourrait tenir au fait que le biais négatif de $\tilde{\tau}_1$ a été annulé par

le biais positif de $\tilde{\tau}_2$. Les estimateurs assistés par un modèle bayésien ont donné des résultats comparables aux précédents, quoique dans ce cas, l'estimateur $\bar{\tau}_2$ de τ_2 ne présentait que de légers problèmes de biais. Les valeurs de la pc des IC de Wald basés sur les EMV et sur les estimateurs assistés par un modèle bayésien étaient faibles. Cependant, puisque les valeurs du br de ces intervalles étaient acceptables, les intervalles pour τ_2 et τ pourraient fournir certains renseignements au sujet de ces paramètres.

Tableau 6.4
Résultats des simulations obtenus pour les estimateurs et les intervalles de confiance pour une population construite en utilisant des données de l'étude de Colorado Springs

		Estimateurs ponctuels				Intervalles de confiance			
Estimateur		**br**	$\sqrt{\text{eqmr}}$	**mder**	**mdera**	**Intervalle de conf.**	**pc**	**mder**	**mdera**
EMV hétér. incond.	$\hat{\tau}_1^{(U)}$	$-0{,}00_1^{0{,}03}$	0,10	-0,01	0,07	ICVP- $\hat{\tau}_1^{(U)}$	0,75	0,24	0,24
						ICVP adj.- $\hat{\tau}_1^{(U)}$	0,95	0,41	0,41
	$\hat{\tau}_2^{(U)}$	$1{,}7_{16}^{3{,}5}$	4,5	0,79	0,79	ICVP- $\hat{\tau}_2^{(U)}$	$0{,}39^{8{,}3}$	10_{18}	3,7
						ICVP- $\hat{\tau}^{(U)}$	$0{,}83^{8{,}6}$	$5{,}7_6$	2,1
						ICVP adj.- $\hat{\tau}^{(U)}$	$0{,}99^{21}$	11_{47}	7,5
	$\hat{\tau}^{(U)}$	$0{,}95_{17}^{3{,}5}$	2,6	0,46	0,46	IC-bootstrap- $\hat{\tau}_1^{(U)}$	0,91	0,37	0,37
						IC-bootstrap- $\hat{\tau}_2^{(U)}$	$0{,}86^{3{,}2}$	11_{29}	3,6
						IC-bootstrap- $\hat{\tau}^{(U)}$	$0{,}88^{3{,}2}$	$6{,}3_{28}$	2,0
EMV hétér. cond.	$\hat{\tau}_1^{(C)}$	$0{,}01^{0{,}57}$	0,12	0,01	0,08	ICVP- $\hat{\tau}_1^{(C)}$	$0{,}81^{1{,}2}$	0,30	0,27
						ICVP adj.- $\hat{\tau}_1^{(C)}$	$0{,}97^{1{,}2}$	0,45	0,44
	$\hat{\tau}_2^{(C)}$	$1{,}7_{10}^{4{,}7}$	4,5	0,80	0,80	ICVP- $\hat{\tau}_2^{(C)}$	$0{,}39^{10}$	$9{,}6_{17}$	3,6
						ICVP- $\hat{\tau}^{(C)}$	$0{,}89^{20}$	$6{,}2_9$	2,6
						ICVP adj.- $\hat{\tau}^{(C)}$	$1{,}0^{27}$	14_{32}	9,3
	$\hat{\tau}^{(C)}$	$0{,}96_{10}^{5{,}2}$	2,6	0,46	0,46	IC-bootstrap- $\hat{\tau}_1^{(C)}$	$0{,}86^{1{,}2}$	0,35	0,35
						IC-bootstrap- $\hat{\tau}_2^{(C)}$	$0{,}90^{8{,}0}$	$9{,}7_{35}$	3,9
						IC-bootstrap- $\hat{\tau}^{(C)}$	$0{,}91^{9{,}2}$	$5{,}9_{34}$	2,2
EMV homo-gènes	$\tilde{\tau}_1$	-0,22	0,23	-0,22	0,22	IC de Wald- $\tilde{\tau}_1$	0,06	0,16	0,16
	$\tilde{\tau}_2$	0,21	0,34	0,16	0,18	IC de Wald- $\tilde{\tau}_2$	0,71	0,60	0,53
	$\tilde{\tau}$	0,02	0,17	-0,00	0,10	IC de Wald- $\tilde{\tau}$	0,73	0,35	0,31
EB homo-gènes	$\bar{\tau}_1$	-0,22	0,23	-0,22	0,22	IC de Wald- $\bar{\tau}_1$	0,13	0,22	0,22
	$\bar{\tau}_2$	0,12	0,22	0,10	0,13	IC de Wald- $\bar{\tau}_2$	0,72	0,45	0,43
	$\bar{\tau}$	-0,02	0,13	-0,04	0,09	IC de Wald- $\bar{\tau}$	0,70	0,27	0,26

Note : Un chiffre en exposant indique le pourcentage d'échantillons pour lesquels l'estimateur ou l'intervalle de confiance (IC) n'a pas été calculé en raison de problèmes de convergence numérique. Un chiffre en indice inférieur indique le nombre de valeurs de l'estimateur ou de la longueur relative d'un IC qui étaient supérieures à 10^5 et qui n'ont pas été utilisées pour calculer le br et la $\sqrt{\text{eqmr}}$ de l'estimateur de la mlr ou de l'IC. Les bornes supérieures (BS) des erreurs Monte Carlo (EMC) des estimations du br et de la $\sqrt{\text{eqmr}}$ des estimateurs des τ étaient les suivantes : τ_1 : 0,001 et 0,001; τ_2 : 0,05 et 0,31, et τ : 0,03 et 0,18. Une BS pour l'EMC des estimations de la pc était de 0,02. Les BS pour les EMC des estimations de la mlr étaient les suivantes : ICVP et ICVP adj. pour τ_1 : 0,003; ICVP pour τ_2 et τ : 0,59; ICVP adj. pour τ : 0,93; IC bootstrap pour τ_1, τ_2 et τ : 0,005,1,5 et 0,88, respectivement.

7 Conclusions et futurs travaux de recherche suggérés

Les résultats des études en simulation réalisées dans le cadre de la présente recherche indiquent que les deux estimateurs proposés de τ_1, à savoir $\hat{\tau}_1^{(U)}$ et $\hat{\tau}_1^{(C)}$, donnent des résultats raisonnablement bons, ce qui est une preuve de leur robustesse à plusieurs types d'écarts par rapport au modèle supposé. (Quoique $\hat{\tau}_1^{(C)}$ semble être sensible aux écarts par rapport à la loi normale supposée des $\beta_j^{(k)}$.) Par ailleurs, les deux estimateurs proposés de τ_2, à savoir $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$, posent des problèmes de biais et surtout des problèmes d'instabilité si la fraction d'échantillonnage dans U_2 n'est pas suffisamment grande, disons si elle n'est pas supérieure à 50 %. En outre, de petites fractions d'échantillonnage conjuguées à des écarts par rapport au modèle supposé des probabilités de lien augmentent le risque de problèmes de convergence numérique. Les deux estimateurs proposés de τ, à savoir $\hat{\tau}^{(U)}$ et $\hat{\tau}^{(C)}$, donnent des résultats comparables aux estimateurs $\hat{\tau}_1^{(U)}$ et $\hat{\tau}_1^{(C)}$ si τ_1 est beaucoup plus grand que τ_2 (comme dans le cas des populations artificielles), donnent des résultats comparables aux estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ si τ_2 est beaucoup plus grand que τ_1 (comme dans le cas de la population de Colorado Springs), et produisent une combinaison des performances des estimateurs de τ_1 et τ_2 si les valeurs de ces paramètres ne sont pas très différentes l'une de l'autre. Enfin, les estimateurs calculés sous l'hypothèse de probabilités de lien homogènes présentent de graves problèmes de biais si cette hypothèse n'est pas satisfaite.

Il convient de mentionner que notre conclusion au sujet des estimateurs proposés de τ_2 est basée sur les résultats de plusieurs petites études en simulation qui ont été exécutées avec des fractions d'échantillonnage supérieures à celles utilisées dans les études Monte Carlo présentées dans le présent article. Dans une étude effectuée avec les populations artificielles, nous avons augmenté les valeurs des probabilités de lien $p_{ij}^{(k)}$ de manière que leurs valeurs moyennes soient $\bar{p}_{ij}^{(1)} \approx 0{,}088$ et $\bar{p}_{ij}^{(2)} \approx 0{,}071$, et nous avons maintenu les tailles des échantillons initiaux à $n = 15$. Ces changements ont donné les fractions d'échantillonnage $f_1 \approx 0{,}65$ et $f_2 \approx 0{,}55$. Dans une autre étude toujours avec les populations artificielles, nous avons réduit les valeurs de $p_{ij}^{(k)}$ de sorte que $\bar{p}_{ij}^{(1)} = \bar{p}_{ij}^{(2)} \approx 0{,}016$ et augmenté les tailles des échantillons initiaux à $n = 78$ pour les populations I à III et à $n = 67$ pour la population IV. Ces changements ont donné les fractions d'échantillonnage $f_1 \approx 0{,}78$ et $f_2 \approx 0{,}55$. Dans les deux études, les estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ ont donné des résultats acceptables. (Les résultats ne sont pas présentés.) Ces résultats indiquent que ces estimateurs semblent aussi avoir des propriétés de robustesse aux écarts par rapport aux modèles supposés, à condition que les fractions d'échantillonnage utilisées soient grandes.

Cependant, dans une étude portant sur la population de Colorado Springs en utilisant des échantillons initiaux de taille $n = 42$ qui ont donné les fractions d'échantillonnage $f_1 \approx 0{,}64$ et $f_2 \approx 0{,}56$, les estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ présentaient de graves problèmes de biais ($\text{br} \approx 1{,}0$ et $\text{mder} \approx 0{,}85$) qui ont affecté les valeurs de leur eqmr $\left(\sqrt{\text{eqmr}} \approx 1{,}0\right)$ et de la $\text{mdera}\left(\text{mdera} \approx 0{,}85\right)$. Pourquoi ces estimateurs n'ont-ils pas donné de bons résultats même pour de grandes fractions d'échantillonnage ? Nous pensons que la mauvaise performance de ces estimateurs est une conséquence de la très petite valeur moyenne des $p_{ij}^{(2)}$ $\left(\bar{p}_{ij}^{(2)} \approx 0{,}018\right)$ et de la façon dont les études Monte Carlo ont été exécutées. Pour clarifier cette affirmation, notons ce qui suit. Quand $\bar{p}_{ij}^{(2)}$ est très petite, disons inférieure à 0,02, le nombre prévu d'éléments dans U_2 qui sont liés à au moins un site dans la base de sondage est nettement inférieur à τ_2. Par exemple, dans la population I, quand $\bar{p}_{ij}^{(2)} = 0{,}015$, ce nombre prévu était environ

$300 < 400 = \tau_2$. Par conséquent, si la fraction d'échantillonnage f_2 est suffisamment grande, les estimations $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ seront proches de τ_2 et seront beaucoup plus grandes que le nombre prévu d'éléments liés à au moins un site dans la base de sondage. Donc, si nous supposions que la population de Colorado Springs était générée selon un processus aléatoire, les 794 contacts qui sont liés à au moins un site dans la base de sondage et que nous avons utilisés comme valeur de τ_2 représenteraient une valeur beaucoup plus petite que la taille réelle de U_2. Par conséquent, la performance de $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ en tant qu'estimateurs de la valeur supposée 794 de τ_2 serait très mauvaise. Cette explication a été proposée et confirmée par les résultats d'une petite étude en simulation que nous avons réalisée en utilisant la population I (dans laquelle chacune des hypothèses est satisfaite), mais au lieu de suivre la procédure décrite à la sous-section 5.1, nous avons généré l'ensemble complet de valeurs $x_{ij}^{(2)}$ de $X_{ij}^{(2)}$ en échantillonnant des lois de Bernoulli de moyenne $p_{ij}^{(2)}, i = 1,\ldots,N; j = 1,\ldots,400$ et nous les avons maintenues fixes. Ensuite, nous avons défini la valeur de τ_2 comme étant le nombre d'éléments de U_2 liés à au moins un site dans la base de sondage. Nous avons considéré deux cas : de grandes valeurs de $p_{ij}^{(2)}$ $\left(\bar{p}_{ij}^{(2)} = 0{,}071\right)$ et de petites valeurs $\left(\bar{p}_{ij}^{(2)} = 0{,}015\right)$. Dans le premier cas, $\tau_2 = 388$, tandis que dans le second, $\tau_2 = 300$. Pour obtenir des résultats comparables, les tailles des échantillons initiaux ont été fixées à $n = 15$ dans le premier cas et à $n = 78$ dans le second, de sorte que dans les deux cas le nombre d'éléments échantillonnés était environ de 220. Les résultats de l'étude numérique ont montré que, pour de grandes valeurs de $p_{ij}^{(2)}$, les estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ donnaient de bons résultats (parce que $388 \approx 400$), tandis que pour de petites valeurs de $p_{ij}^{(2)}$, ces estimateurs donnaient de mauvais résultats (parce que $300 << 400$). Nous pensons que ces seconds résultats illustrent et expliquent ceux obtenus pour la population de Colorado Springs.

En ce qui concerne les deux types d'IC proposés, à savoir les IC par vraisemblance profilée et les IC par bootstrap, nous pouvons conclure qu'ils nécessitent de plus grandes tailles d'échantillon que les estimateurs ponctuels pour donner des résultats raisonnablement bons. Ils sont plus sensibles aux écarts par rapport aux modèles supposés que les estimateurs ponctuels. En outre, si de petites fractions d'échantillonnage sont utilisées et qu'il existe des écarts par rapport au modèle supposé pour les $p_{ij}^{(k)}$, les problèmes de convergence numérique seront plus fréquents que dans le cas des estimateurs ponctuels.

Compte tenu des observations qui précèdent, nous concluons que, dans les applications réelles de cette méthodologie d'échantillonnage, une bonne stratégie consiste à construire une base de sondage couvrant la partie la plus grande possible de la population cible. De cette façon, τ_1 sera proche de τ, et les estimations de τ_1 pourront être utilisées comme estimations de τ. L'avantage de cette stratégie est que les estimateurs $\hat{\tau}_1^{(U)}$ et $\hat{\tau}_1^{(C)}$ donnent de meilleurs résultats que les estimateurs $\hat{\tau}_2^{(U)}$ et $\hat{\tau}_2^{(C)}$ parce que les premiers incorporent l'information au sujet des tailles de grappe m_i. En outre, cette stratégie permet d'utiliser l'estimateur fondé sur le plan de sondage $N\sum_1^n m_i/n$ comme estimateur de τ. L'autre facteur dont il faut tenir compte pour obtenir de bonnes estimations est que les fractions d'échantillonnage doivent être grandes, disons supérieures à 0,5. Cette suggestion est en harmonie avec le résultat présenté par Xi, Watson et Yip (2008), qui, dans le contexte d'études par capture-recapture, indiquent qu'en présence de probabilités de capture hétérogènes, une taille de population comprise entre 300 et 500, et un nombre d'occasions d'échantillonnage compris entre 10 et 20, la fraction d'échantillonnage minimale pour obtenir des estimations fiables est au moins de 60 %. Puisque l'estimation de τ_2 équivaut fondamentalement au

même problème que celui de l'estimation de la taille de population dans une étude par capture-recapture, nous pensons que cette conclusion s'applique aussi à notre situation. Dans cet ordre d'idées, nous avons élaboré une méthode pour déterminer la taille de l'échantillon initial afin d'obtenir les valeurs souhaitées de $\sqrt{V(\hat{\tau}_k)/\tau_k^2}, k = 1, 2$ et $\sqrt{V(\hat{\tau})/\tau^2}$. Bien que cette procédure soit fondée sur des hypothèses sévères, telles que l'homogénéité des effets $\alpha_i^{(k)}$ associés aux sites et la nécessité que les valeurs de m_i soient grandes, les résultats semblent être satisfaisants. Ainsi, la situation illustrée à la fin de la section 5 correspond à celle des populations artificielles considérées dans l'étude Monte Carlo, et nous pouvons voir que les résultats obtenus par notre procédure sont très proches de ceux signalés pour les populations I et II (tableau 6.2), où les estimateurs de τ_2 et τ donnent des résultats acceptables.

Enfin, malgré leurs inconvénients, les estimateurs ponctuels et d'intervalle proposés représentent une meilleure option pour faire des inférences au sujet de la taille de population que ceux fondés sur l'hypothèse de probabilités de lien homogènes. Manifestement, notre proposition devra être améliorée. Les deux principaux problèmes qui doivent être examinés dans le cadre de futures études sont l'instabilité des estimateurs de τ_2 quand la fraction d'échantillonnage n'est pas suffisamment grande et la performance insatisfaisante des intervalles de confiance. Une solution possible aux deux problèmes consisterait à utiliser l'approche bayésienne pour construire des estimateurs dans lesquels est intégrée l'information au sujet des paramètres dont le chercheur dispose avant l'échantillonnage. Les estimateurs ponctuels et d'intervalle obtenus par cette approche pourraient être plus stables que ceux proposés dans le présent article en raison de l'information supplémentaire utilisée pour les construire. Une autre solution possible du problème de manque de robustesse des intervalles de confiance consisterait à remplacer l'hypothèse que les M_i suivent une loi de Poisson par l'hypothèse d'une loi plus souple, comme la binomiale négative, et l'hypothèse d'une loi normale pour les effets $\beta_j^{(k)}$ par l'hypothèse d'une des lois ordinairement utilisées pour augmenter la robustesse des estimateurs, telles qu'un mélange de lois normales ou une loi t de Student.

Remerciements

La présente étude a été financée par la subvention PROFAPI 2008/054 de la *Universidad Autónoma de Sinaloa* et la subvention APOY-COMPL-2008/89777 du *Consejo Nacional de Ciencia y Tecnología*, au Mexique. Nous remercions John Potterat et Steve Muth de nous avoir permis d'utiliser les données de l'étude de Colorado Springs. Nous remercions aussi les examinateurs de leurs suggestions et commentaires utiles qui nous ont permis d'améliorer le manuscrit.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis, Second edition*. 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.

Cormack, R.M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics*, 48, 567-576.

Coull, B.A., et Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55, 294-301.

Dávid, B., et Snijders, T.A.B. (2002). Estimating the size of the homeless population in Budapest, Hungary. *Quality & Quantity*, 36, 291-303.

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 Monjardin, P.E. (2006). 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. *Techniques d'enquête*, 32, 2, 207-216.

Félix-Medina, M.H., et Monjardin, P.E. (2010). Combining link-tracing sampling and cluster sampling to estimate totals and means of hidden human populations. *Journal of Official Statistics*, 26, 603-631.

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.

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

Gimenes, O., Choquet, R., Lamor, R., Scofield, P., Fletcher, D., Lebreton, J.-D. et Pradel, R. (2005). Efficient profile-likelihood confidence intervals for capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics*, 10, 1-13.

Heckathorn, D.D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Johnston, L.G., et Sabin, K. (2010). Sampling hard-to-reach populations with respondent driven sampling. *Methodological Innovations Online*, 5, 2, 38-48.

Kalton, G. (2009). Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales. *Techniques d'enquête*, 35, 2, 133-152.

Karon, J.M., et Wejnert, C. (2012). Statistical methods for the analysis of time-location sampling data. *Journal of Urban Health*, 89, 565-586.

MacKellar, D., Valleroy, L., Karon, J., Lemp, G. et Janssen, R. (1996). The young men's survey: Methods for estimating HIV sero-prevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111, supplément 1, 138-144.

Magnani, R., Sabin, K., Saidel, T. et Heckathorn, D. (2005). Review of sampling hard-to-reach populations for HIV surveillance. *AIDS*, 19, S67-S72.

McKenzie, D.J., et Mistiaen, J. (2009). Surveying migrant households: a comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society, Series A*, 172, 339-360.

Munhib, F.B., Lin, L.S., Stueve, A., Miller, R.L., Ford, W.L., Johnson, W.D. et Smith, P. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports*, 116, supplement 1, 216-222.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56, 434-442.

Potterat, J.J., Woodhouse, D.E., Muth, S.Q., Rothenberg, R.B., Darrow, W.W., Klovdahl, A.S. et Muth, J.B. (2004). Network dynamism: History and lessons of the Colorado Springs study. Dans *Network Epidemiology: A Handbook for Survey Design and Data Collection*, (Éd., M. Morris), New York : Oxford University Press, 87-114.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. et Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratkowsky, D.A. (1988). *Handbook of Nonlinear Regression Models*. New York : Marcel Dekker.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. et Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. Dans *Social Networks, Drug Abuse, and HIV Transmission*, (Éds., R.H. Needle, S.G. Genser et R.T. II Trotter) NIDA Research Monograph 151. Rockville, MD : National Institute of Drug Abuse, 3-19.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

Semaan, S. (2010). Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodological Innovations Online*, 5, 2, 60-75.

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

Staudte, R.G., et Sheather, S.J. (1990). *Robust Estimation and Testing*. New York : John Wiley & Sons, Inc.

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, 1, 99-112.

Venzon, D.J., et Moolgavkar, S.H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37, 87-94.

Williams, B.K., Nichols, J.D. et Conroy, M.J. (2002). *Analysis and Management of Animal Populations*. San Diego, California : Academic Press.

Xi, L., Watson, R. et Yip, P.S.F. (2008). The minimum capture proportion for reliable estimation in capture-recapture models. *Biometrics*, 64, 242-249.