N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2004 : Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

2004

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2004 de Statistique Canada
Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

ÉCHANTILLONNAGE PAR LIENS AVEC UN ÉCHANTILLON INITIAL DE LIEUX EN SÉLECTION SÉQUENTIELLE : ESTIMATION DE TAILLE DE POPULATION

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

RÉSUMÉ

Dans cet exposé, nous modifions la version de l'échantillonnage par liens proposée par Félix-Medina et Thompson (2004) en prenant un échantillon séquentiel de lieux plutôt qu'un simple échantillon aléatoire. La version que nous avançons permet à l'échantillonneur d'exercer un certain contrôle sur la taille finale d'échantillon ou la précision de l'estimateur de taille de la population. Nous proposons d'estimer la taille de la population par un estimateur de maximum de vraisemblance proposé par Félix-Medina et Thompson (2004) ou par un estimateur issu de la méthode bayésienne et proposé par Félix-Medina et Monjardin (2004). Nous proposons en outre d'élaborer des intervalles de confiance de l'estimation de taille de la population par les méthodes bootstrap. Les résultats d'une étude de simulation indiquent que le plan de sondage que nous recommandons donne des résultats acceptables.

MOTS CLÉS : Bootstrap, échantillonnage séquentiel, maximum de vraisemblance, méthode d'inférence bayésienne, population finie, principe de la règle d'arrêt, traitement par plan de sondage.

1. INTRODUCTION

On a jugé que l'échantillonnage par liens (EL) convenait à un sondage de population humaine cachée et peu accessible, qu'il s'agisse des consommateurs de drogue, des sans-abri ou des travailleurs clandestins. L'idée à la base de cette méthode est de prendre d'abord un échantillon de membres de la population visée et ensuite de majorer la taille d'échantillon par intégration de candidats proposés par les membres de l'échantillon initial. On peut encore ajouter à l'échantillon en intégrant les candidats proposés à leur tour par les premiers candidats choisis. On poursuit l'échantillonnage de la sorte jusqu'au déclenchement de l'application d'une règle d'arrêt spécifiée.

On a avancé plusieurs versions de l'EL, mais ces derniers temps Félix-Medina et Thompson (2004) en ont proposé une nouvelle où l'échantillon initial est un échantillon aléatoire simple (EAS) sans remise de grappes ou de lieux accessibles comme les bars, les parcs ou les îlots d'habitation après prélèvement sur une base de sondage qui vise une partie seulement de la population. Comme dans l'EL habituel, on demande aux gens de chaque lieu sélectionné de proposer d'autres membres de la population. Par ce plan de sondage initial, les auteurs atteignent un double but : d'abord, ils évitent la tâche usuelle et difficile consistant à vérifier l'hypothèse d'un échantillon initial de Bernoulli; en second lieu, ils se reportent à la distribution probabiliste du plan de sondage initial pour élaborer des estimateurs de variance de plan de sondage qui se révèlent robustes à l'égard de certaines hypothèses de modélisation.

Dans ce document, nous présentons une version modifiée du plan de sondage proposé par Félix-Medina et Thompson (2004). Cette nouvelle version donnera à l'échantillonneur la possibilité d'exercer un certain contrôle sur la taille de l'échantillon final ou celle d'un de ses sous-ensembles comme les membres de l'échantillon initial ou les candidats proposés dans la partie de la population que n'appréhende pas la base de sondage. L'échantillonneur pourrait aussi être maître de la précision des estimateurs de taille de population. Notre stratégie sera de procéder à un premier échantillonnage séquentiel de lieux et à intégrer des candidats à cet échantillon par la suite. En se servant

[1] Martín H. Félix-Medina et Pedro E. Monjardin, Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, Mexique, 80010. Adresses de courrier électronique : mhfelix@uas.uasnet.mx et pemo@uas.uasnet.mx.

de règles d'arrêt appropriées, l'échantillonneur pourrait réaliser la taille spécifiée de l'échantillon final ou d'un de ses sous-échantillons ou parvenir à la précision voulue des estimateurs.

Voici comment s'organise notre propos : à la section 2, nous présentons la notation à employer dans tout l'exposé et décrivons le plan de sondage proposé; à la section 3, nous mentionnons les différents estimateurs de taille de population à employer avec ce plan de sondage; à la section 4, nous exposons l'application des méthodes bootstrap à l'élaboration d'intervalles de confiance par plan de sondage pour les tailles de population; à la section 5, nous livrons les résultats d'une étude de simulation pour l'observation du rendement du plan de sondage avancé; à la section 6 enfin, nous tirons un certain nombre de conclusions.

2. NOTATION ET PLAN DE SONDAGE

Nous posons la même structure de population que dans Félix-Medina et Thompson (2004). Ainsi, nous supposons qu'une population humaine finie et cachée $U = \{u_1, \ldots u_\tau\}$ d'une taille inconnue τ peut se scinder en deux parties U_1 et U_2 de tailles inconnues τ_1 et $\tau_2 = \tau - \tau_1$. Nous supposons en outre que la partie U_1 est appréhendée par une base de sondage comprenant N lieux accessibles $A_1, \ldots, A_N$ comme des bars, des parcs ou des îlots d'habitation de taille (nombre de personnes dans ces lieux) $m_1, \ldots, m_N$. Enfin, nous posons qu'il est possible de déterminer si quelqu'un appartient ou non à la région visée par la base de sondage, et dans ce cas, le lieu auquel appartient cette personne. À noter que $\tau_1 = \sum_1^N m_i$, une personne étant rattachable à un seul lieu.

Voici notre plan de sondage. On sélectionne un lieu dans la base de sondage par échantillonnage aléatoire simple sans remise. Soit A_1 le lieu sélectionné. On reconnaît les m_1 membres de A_1 à qui on demande de proposer des membres de la population en dehors de A_1, c'est-à-dire dans $U_1 - A_1$. Par convention, nous dirons qu'un candidat est proposé par un lieu si au moins un de ses membres fait cette proposition. Il convient de noter que diverses stratégies de proposition de noms sont applicables. Ainsi, les m_1 membres pourraient proposer des noms comme groupe ou il pourrait s'agir de chacun des m_1 membres. Enfin, pour chaque nom proposé, la région U_1 ou U_2 à laquelle appartient le candidat est inscrite.

Une fois que l'activité de proposition de noms cesse dans ce lieu, on vérifie si une règle d'arrêt est respectée ou non. Si elle l'est, la démarche d'échantillonnage prend fin et, dans les autres cas, on continue jusqu'à satisfaire à cette règle d'arrêt. À noter que les propositions émanant des différents lieux doivent être indépendantes et que les stratégies de proposition de noms peuvent varier selon les lieux. À noter également que, au terme de la démarche d'échantillonnage, nous aurons un échantillon initialement ordonné $S_0 = (A_1, \ldots, A_n)$ des n lieux en sélection séquentielle. Nous appellerons le plan de sondage qui donne S_0 plan d'échantillonnage aléatoire simple (EAS) séquentiel ou EASS.

Il est possible d'employer toute règle d'arrêt qui est uniquement fonction des données d'observation. Si le but est, par exemple, d'exercer un contrôle sur la taille de l'échantillon final ou d'un de ses sous-ensembles comme le groupe de candidats proposés en U_1 ou en U_2, nous pourrions mettre fin à cette démarche lorsque la taille d'échantillon atteint une borne supérieure spécifiée ou la dépasse pour la première fois. De même, si l'objectif est de contrôler les coûts d'échantillonnage, la démarche prendrait fin lorsque ce coût atteint une valeur spécifiée ou la dépasse pour la première fois. Un dernier exemple : si le but est de contrôler la précision d'un des estimateurs de τ_1, τ_2 ou τ, nous cesserons d'échantillonner lorsque l'estimateur de la variance de cet estimateur ou la longueur de l'intervalle de confiance reposant sur cet estimateur atteint une valeur spécifiée ou tombe au-dessous de cette valeur pour la première fois. Il importe de noter que, avec le plan de sondage avancé, il est difficile de mettre fin à la démarche d'échantillonnage lorsqu'on atteint précisément la valeur spécifiée, et ce, parce que les membres de

l'échantillon initial sont sélectionnés en grappes – et non pas un à un – et que le processus de proposition de noms cesse dans un lieu lorsque plus personne ne propose de noms.

Nous terminerons cette partie de notre exposé en présentant en majeure partie la notation employée dans tout le document. Soit $m = \sum_1^n m_i$ le nombre de membres de S_0, r_1 et r_2 les nombres respectifs de candidats distincts proposés dans $U_1 - S_0$ et dans U_2, et $z_1^{(1)}$ et $z_1^{(2)}$ les nombres respectifs de membres proposés par le lieu A_i dans $U_1 - A_i$ et U_2; $i = 1,\ldots,n$; $x_{ij}^{(1)}$ correspond à l'unité si le candidat u_j dans $U_1 - A_i$ est proposé par le lieu A_i et devient nul dans les autres cas; $x_{ij}^{(2)}$ est égal à 1 si le candidat u_j dans U_2 est proposé par le lieu A_i et à 0 dans les autres cas.

3. ESTIMATEURS DE TAILLE DE POPULATION

Dans le plan de sondage de Félix-Medina et Thompson (2004), ceux-ci proposent des estimateurs de maximum de vraisemblance (EMV) des tailles de population τ_1, τ_2 et τ. Pour obtenir ces estimateurs, ils posent que les tailles $m_1,\ldots,m_N$ des lieux $A_1,\ldots,A_N$ sont des réalisations de variables aléatoires indépendantes de Poisson $M_1,\ldots,M_N$ dont la moyenne est λ_1 et que, compte tenu de l'échantillon initial, les valeurs de $x_{ij}^{(1)}$ et $x_{ij}^{(2)}$ sont des réalisations de variables aléatoires indépendantes de Bernoulli $X_{ij}^{(1)}$ et $X_{ij}^{(2)}$ aux moyennes respectives $p_i^{(1)}$ et $p_i^{(2)}$, $i = 1,\ldots,n$. À partir de ces hypothèses et suivant la méthode de Darroch (1958), ils dégagent les EMV $\tilde{\tau}_1$ et $\tilde{\tau}_2$ de τ_1 et τ_2 comme solution des équations non linéaires suivantes :

$$\tilde{\tau}_1 = \frac{M + R_1}{1 - (1 - n/N)\prod_1^n (1 - \tilde{p}_i^{(1)})} \text{ et}$$

$$\tilde{\tau}_2 = \frac{R_2}{1 - \prod_1^n (1 - \tilde{p}_i^{(2)})},$$

où $\tilde{p}_i^{(1)} = Z_i^{(1)} / (\tilde{\tau}_1 - M_i)$ et $\tilde{p}_i^{(2)} = Z_i^{(2)} / \tilde{\tau}_2$ sont les EMV de $p_i^{(1)}$ et $p_i^{(2)}$, $i = 1,\ldots,n$, et où M, R_1, R_2, $Z_i^{(1)}$ et $Z_i^{(2)}$ sont les variables aléatoires qui donnent respectivement les valeurs m, r_1, r_2, $z_i^{(1)}$ et $z_i^{(2)}$. L'EMV $\tilde{\tau}$ de τ est alors donné par $\tilde{\tau} = \tilde{\tau}_1 + \tilde{\tau}_2$.

Les auteurs ont procédé à une étude de simulation pour découvrir que, lorsque les probabilités de proposition de noms $p_i^{(2)}$, $i = 1,\ldots,n$, sont petites, l'estimateur $\tilde{\tau}_2$ est très instable et surestime grandement τ_2. C'est ce problème qui a amené Félix-Medina et Monjardin (2004) à employer la méthode bayésienne pour élaborer des estimateurs exempts de biais. Ces auteurs proposent trois jeux d'estimateurs, mais nous ne considérerons ici que le jeu tiré des distributions initiales suivantes de τ_1 et τ_2 :

Distributions gamma de Poisson :

$\pi(\tau_1 \mid \lambda_1) \propto (N\lambda_1)^{\tau_1} / \tau_1!$ et $\pi(\lambda_1) \propto \lambda_1^{a_1 - 1} e^{-b_1 \lambda_1}$,

$\pi(\tau_2 \mid \lambda_2) \propto \lambda_2^{\tau_2} / \tau_2!$ et $\pi(\lambda_2) \propto \lambda_2^{a_2 - 1} e^{-b_2 \lambda_2}$,

où a_1, b_1, a_2 et b_2 sont des constantes connues, où τ_1 et τ_2 sont conditionnellement indépendants étant donné λ_1 et λ_2 et où λ_1 et λ_2 sont aussi des variables aléatoires indépendantes.

Dans le cas des probabilités de proposition de noms $p_i^{(k)}$, $i = 1, \ldots, n$, $k = 1,2$, les auteurs ne prennent pas les distributions initiales de ces variables, mais les logits $\alpha_i^{(k)} = \log\left[p_i^{(k)} / (1 - p_i^{(k)})\right]$ des $p_i^{(k)}$ et ils posent les distributions initiales suivantes des $\alpha_i^{(k)}$:

$$\alpha_i^{(k)} \mid \theta_k \sim N(\theta_k, \sigma_k^2) \text{ et } \theta_k \sim N(\mu_k, \gamma_k{}^2),$$

$i = 1, \ldots, n$, $k = 1,2$, où $N(\theta_k, \sigma_k^2)$ remplace la distribution normale à moyenne θ_k et à variance σ_k^2, où σ_k^2, μ_k et γ_k^2 sont des constantes connues et où les $\alpha_i^{(k)}$ sont conditionnellement indépendants étant donné θ_k.

Ils complètent les hypothèses au sujet des distributions initiales en posant que les vecteurs aléatoires (τ_k, λ_k) et $(\alpha_i^{(k)}, \theta_k)$, $i = 1, \ldots, n$, $k = 1,2$, sont réciproquement indépendants.

Ils proposent d'estimer $(\tau_1, \tau_2, \alpha_1, \alpha_2)$, où $\alpha_k = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, $k = 1,2$, par la moyenne du mode de leur codistribution postérieure. Par cette stratégie, ils dégagent un estimateur $(\hat{\tau}_1, \hat{\tau}_2, \hat{\alpha}_1, \hat{\alpha}_2)$ de $(\tau_1, \tau_2, \alpha_1, \alpha_2)$ comme solution du système suivant d'équations non linéaires :

$$\hat{\tau}_1 = \frac{M + R_1 + (1 - n/N)[N(a_1 - 1)/(N + b_1)] \prod_1^n (1 - \hat{p}_i^{(1)})}{1 - (1 - n/N)[N/(N + b_1)] \prod_1^n (1 - \hat{p}_i^{(1)})};$$

$$\hat{p}_i^{(1)} = \frac{Z_i^{(1)}}{\hat{\tau}_1 - M_i} - \frac{\hat{\alpha}_i^{(1)} - \hat{\bar{\alpha}}_1}{(\hat{\tau}_1 - M_i)\sigma_1^2} - \frac{\hat{\bar{\alpha}}_1 - \mu_1}{n(\hat{\tau}_1 - M_i)\upsilon_1};\ i = 1, \ldots n;$$

$$\hat{\tau}_2 = \frac{R_2 + [(a_2 - 1)/(1 + b_2)] \prod_1^n (1 - \hat{p}_i^{(2)})}{1 - [1/(1 + b_2)] \prod_1^n (1 - \hat{p}_i^{(2)})};$$

$$\hat{p}_i^{(2)} = \frac{Z_i^{(2)}}{\hat{\tau}_2} - \frac{\hat{\alpha}_i^{(2)} - \hat{\bar{\alpha}}_2}{\hat{\tau}_2 \sigma_2^2} - \frac{\hat{\bar{\alpha}}_2 - \mu_2}{n \hat{\tau}_2 \upsilon_2};\ i = 1, \ldots, n;,$$

où $\hat{\bar{\alpha}}_k = \sum_1^n \hat{\alpha}_i^{(k)} / n$ et $\upsilon_k = \gamma_k^2 + \sigma_k^2 / n,\ k = 1,2$. Ils proposent d'estimer τ par $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

Dans le cas de notre plan de sondage, la seule différence d'avec le plan avancé par Félix-Medina et Thompson (2004) est que la taille de l'échantillon initial est aléatoire dans le premier de ces plans et fixe dans le second. Si nous suivons le principe de la règle d'arrêt, les EMV à employer dans notre cas sont ceux que proposent Félix-Medina et Thompson (2004) et, pour la même raison, les estimateurs tirés de la méthode bayésienne par Félix-Medina et Monjardin (2004) seraient aussi applicables en pareil cas.

Pour conclure cette partie de notre exposé, nous ferons les deux observations suivantes. D'abord, bien que Félix-Medina et Thompson (2004) jugent que l'EMV $\tilde{\tau}_2$ est instable lorsque les probabilités de proposition de noms sont petites, il est possible d'éviter ce problème dans notre plan de sondage en appliquant une règle d'arrêt qui garantit un minimum de candidats proposés dans U_2, ou encore une précision spécifiée de l'estimateur $\tilde{\tau}_2$. En second lieu, Félix-Medina et Monjardin (2004) se sont servis de cette méthode uniquement pour élaborer des estimateurs de taille de population, mais leurs inférences relèvent de la méthode fréquentiste, en ce sens que les tailles de population sont traitées comme des paramètres fixes et que les distributions probabilistes des estimateurs sont à la base des inférences au sujet des paramètres. Les auteurs qualifient cette technique de méthode bayésienne assistée, puisqu'elle ressemble à l'échantillonnage assisté par modèle que proposent Särndal et coll. (1992). Nous employons également la méthode bayésienne assistée pour nos inférences au sujet des tailles de population lorsque nous utilisons les estimateurs antérieurement décrits qui sont issus de la méthode de Bayes.

4. INTERVALLES DE CONFIANCE BOOTSTRAP

Nous emploierons la version des méthodes bootstrap avancée par Félix-Medina et Monjardin (2004) pour élaborer des intervalles de confiance par plan de sondage. Dans cette version, nous tenons compte tant du plan de sondage pour l'échantillonnage initial des lieux que de la procédure de proposition de noms. C'est pourquoi nous échantillonnons les lieux à partir d'une population finie artificielle à l'aide de la version bootstrap proposée par Gross (1980). Nous passons à la procédure de proposition de noms en échantillonnant les variables indicatrices $X_{ij}^{(k)}$ à partir de leurs distributions estimées à l'aide de la version paramétrique bootstrap. (On trouvera une description de cette version bootstrap dans Efron et Tibshirani, 1993.) Nous dégageons des intervalles en points ou en percentiles.

Voici les étapes de l'application de cette procédure que nous proposons. (i) On élabore une population finie artificielle de N valeurs de m_i en itérant N/n fois – en supposant que N/n est un nombre entier – l'échantillon prélevé de n valeurs de taille de lieux $m_1, \ldots, m_n$. Si N/n n'est pas un entier, on applique la procédure de Booth et coll. (1994) pour élaborer la population finie. En d'autres termes, si $N = kn + r$, où k et r sont des entiers positifs, on itère k fois l'échantillon prélevé de n valeurs de taille de lieux et ajoute à cet ensemble de m_i un échantillon de r valeurs de m_i constitué par échantillonnage aléatoire simple sans remise à partir de l'échantillon observé de n valeurs de taille de lieux. (ii) On prélève un échantillon aléatoire simple sans remise et de taille un sur la population artificielle des m_i. Soit m_j l'élément sélectionné. (iii) On tire des échantillons de valeurs de taille $\hat{\tau}_1 - m_j$ et $\hat{\tau}_2$ de distributions de Bernoulli à moyennes respectives $\hat{p}_j^{(1)}$ et $\hat{p}_j^{(2)}$, où $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{p}_j^{(1)}$ et $\hat{p}_j^{(2)}$ sont les estimations de τ_1, τ_2, $p_j^{(1)}$ et $p_j^{(2)}$ établies à partir de l'échantillon initial. Ces échantillons simulent les propositions de noms faites par le lieu A_j. (iv) On détermine si l'échantillon prélevé aux étapes (ii) et (iii) satisfait ou non à la règle d'arrêt appliquée à l'échantillon initial. Si tel est le cas, on met fin à la procédure d'échantillonnage et établit les estimations $\hat{\tau}_1$, $\hat{\tau}_2$ et $\hat{\tau}$ de τ_1, τ_2 et τ par les mêmes expressions que pour les estimations initiales. Dans les autres cas, on reprend les étapes (i) à (iv). (v) On reprend les étapes (i) à (iv) un grand

nombre B de fois. Les distributions bootstrap de τ_1, τ_2 et τ sont les distributions empiriques respectivement tirées des jeux de valeurs B de $\hat{\tau}_1$, $\hat{\tau}_2$ et $\hat{\tau}$. (vi) On élabore les $100(1-\alpha)\%$ intervalles de confiance bootstrap pour τ_1, τ_2 et τ en points ou en percentiles (voir Davison et Hinkley, 1997, chapitre 5, pour une description de ces méthodes). Dans le premier cas, l'intervalle pour τ est $[2\hat{\tau} - \tau^{(1-\alpha/2)}, 2\hat{\tau} - \tau^{(\alpha/2)}]$ et, dans le second, $[\tau^{(\alpha/2)}, \tau^{(1-\alpha/2)}]$, où $\tau^{(\alpha/2)}$ et $\tau^{(1-\alpha/2)}$ sont les points $\alpha/2$ et $1-\alpha/2$ de la distribution bootstrap de τ et où $\hat{\tau}$ est l'estimation de τ tirée de l'échantillon initial.

Il convient de noter que, si la règle d'arrêt dépendait de la précision d'un des estimateurs $\hat{\tau}_1$, $\hat{\tau}_2$ ou $\hat{\tau}$, nous devrions établir cet estimateur à l'étape (iv) pour juger si on a satisfait à la règle d'arrêt.

En dehors de l'élaboration d'intervalles de confiance, on pourrait obtenir des estimateurs simples par plan de sondage des variances de τ_1, τ_2 et τ en calculant les variances d'échantillon à partir des jeux correspondants de valeurs B des estimateurs.

5. ÉTUDE DE MONTE CARLO

Nous avons considéré deux populations finies artificielles dégagées à l'aide des valeurs paramétriques employées par Félix-Medina et Monjardin (2004). On trouvera au tableau 1 une description de chaque population. Nous avons tiré les probabilités de proposition de noms $p_i^{(k)}$, $k = 1,2$, du modèle $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, où les valeurs de β_k sont fixées de manière que nous puissions obtenir les valeurs approchées suivantes de $E(p_i^{(k)})$: $(E(p_i^{(1)}), E(p_i^{(2)})) \cong (0.05, 0.03)$ et $(E(p_i^{(1)}), E(p_i^{(2)})) \cong (0.01, 0.006)$. Les valeurs des paramètres des distributions initiales ont été les suivantes : $\sigma^2 = 9$, $\mu_k = -3.5$, $\gamma_k^2 = 9$, $k = 1,2$, $a_1 = 1$, $b_1 = 0.1$, $a_2 = 8$ et $b_2 = 0.01$, si bien que $E(\lambda_1) = 10$, $V(\lambda_1) = 100$, $E(\lambda_2) = 800$ et $V(\lambda_2) = 80000$.

Tableau 1. Paramètres des populations simulées

Population	N	Distribution de M_i	$E(M_i)$	$V(M_i)$	τ_1	τ_2	τ	τ_1/τ
I	250	Poisson	7,2	7,2	1 828	700	2 528	0,72
II	250	Distribution binomiale négative	7,2	24,4	1 861	700	2 561	0,73

Voici comment s'est faite l'étude de simulation. Sur chaque population de $N = 250$ valeurs de m_i, nous avons prélevé un échantillon séquentiel de valeurs par un plan de sondage EASS (échantillonnage aléatoire simple séquentiel). Nous avons mis fin à la procédure d'échantillonnage lorsque le nombre de candidats proposés dans U_2 a atteint les 250 ou dépassé cette valeur une première fois. Nous avons établi les nombres respectifs de candidats proposés dans U_1 et U_2 par le lieu A_i de l'échantillon par $X_{i1}^{(1)} + \cdots + X_{i\tau_1 - m_i}^{(1)}$ et $X_{i1}^{(2)} + \cdots + X_{i\tau_2}^{(2)}$, où les $X_{ij}^{(1)}$ et les $X_{ij}^{(2)}$ étaient des échantillons sur distribution de Bernoulli à moyennes $p_i^{(1)}$ et $p_i^{(2)}$. Cette procédure a été itérée $r = 2000$ fois. Nous avons aussi considéré la version de l'échantillonnage par liens avancée par Félix-Medina et Thompson (2004), c'est-à-dire la version où l'échantillon initial de lieux est prélevé par échantillonnage aléatoire simple. Nous avons fait correspondre les tailles respectives des échantillons initiaux en EAS à la taille moyenne des 2 000 échantillons initiaux en EASS.

Nous avons observé le rendement des EMV $\tilde{\tau}_1$, $\tilde{\tau}_2$ et $\tilde{\tau}$ et des estimateurs $\hat{\tau}_1$, $\hat{\tau}_2$ et $\hat{\tau}$ issus de la méthode bayésienne. Nous avons évalué le rendement d'un estimateur $\hat{\tau}$, disons, par son biais relatif et la racine carrée de son erreur quadratique moyenne, ce que nous avons défini comme biais r = $\sum_1^r(\hat{\tau}_i - \tau)/(r\tau)$ et $\sqrt{\text{r} - \text{eqm}} = \sqrt{\sum_1^r(\hat{\tau}_i - \tau)^2/(r\tau^2)}$ respectivement, où $\hat{\tau}_i$ est la valeur de $\hat{\tau}$ obtenue dans le i^{e} échantillon simulé.

Nous avons en outre observé le rendement des intervalles de confiance à 95 % pour les tailles de population respectivement dégagées par les méthodes bootstrap en points et en percentiles. Nous avons apprécié le rendement de l'intervalle par sa longueur moyenne et ses probabilités de couverture.

Voici les grands résultats de notre étude de simulation. En ce qui concerne le rendement des estimateurs de taille de population (voir le tableau 2), nous savons que, dans le cas de la version EL en EASS initial, les biais r des estimateurs de τ_2 et τ étaient supérieurs aux biais r correspondants dans le cas de la version en EAS initial. Les biais r n'étaient cependant pas importants; ils étaient de moins de 0,1 et n'avaient aucune incidence marquée sur le rendement de l'estimateur. En fait, dans le cas de la version en EASS initial, les $\sqrt{\text{r} - \text{eqm}}$ des estimateurs de τ_2 et τ n'étaient que légèrement supérieurs à ceux de la version en EAS initial. Par ailleurs, le rendement de chacun des estimateurs de τ_1 était pour ainsi dire le même dans les deux versions. Enfin, les estimateurs tirés de la méthode bayésienne étaient un peu plus efficients que les EMV et chacun des estimateurs observés était robuste quant à l'écart par rapport à la distribution posée de Poisson des valeurs de taille de lieux.

Pour ce qui est du rendement des intervalles de confiance (voir le tableau 3), les intervalles obtenus dans la version en EASS initial étaient moindres que ceux de la version en EAS initial. Dans la première de ces versions, les probabilités de couverture des intervalles bootstrap en points étaient relativement proches de la valeur nominale, à savoir 0,95, alors que, dans la seconde, les probabilités correspondantes de l'intervalle relatif à l'estimateur $\hat{\tau}_1$ par la méthode bayésienne n'étaient pas aussi proches de cette valeur nominale que les probabilités des autres intervalles. Les probabilités de couverture des intervalles bootstrap en percentiles n'étaient pas non plus aussi proches de 0,95 que celles des intervalles correspondants en points. Parfois, elles étaient de moins de 0,9 et, parfois aussi, même de moins de 0,8. Les intervalles de confiance fondés sur les estimateurs tirés de la méthode bayésienne étaient d'un rendement supérieur à ceux qui reposaient sur les EMV. En s'écartant de la distribution posée de Poisson des valeurs de taille de lieux, on accroissait la longueur des intervalles de τ_1 et τ, et les probabilités de couverture s'éloignaient un peu de la valeur nominale. Les longueurs et les probabilités de couverture des intervalles de τ_2 n'étaient pas touchées si on s'écartait de la distribution de Poisson.

Tableau 2. Biais relatifs et racine carrée des erreurs quadratiques moyennes relatives des estimateurs de taille de population (résultats après 2 000 itérations)

	Population I							Population II						
EASS initial		$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$		$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$
E($p_i^{(1)}$)≈0,05	E(n)=14,56	0,001	0,051	0,015	0,001	0,044	0,013	E(n)=14,47	0,002	0,039	0,012	0,001	0,033	0,010
E($p_i^{(2)}$)≈0,03	E(R_2)=263,3	0,035	0,129	0,044	0,034	0,114	0,040	E(R_2)=264,2	0,038	0,126	0,045	0,038	0,112	0,041
E($p_i^{(1)}$)≈0,01	E(n)=68,96	-0,000	0,088	0,024	-0,001	0,074	0,020	E(n)=67,78	0,001	0,079	0,023	0,001	0,067	0,019
E($p_i^{(2)}$)≈0,006	E(R_2)=257,2	0,022	0,153	0,045	0,022	0,131	0,039	E(R_2)=257,3	0,030	0,146	0,046	0,030	0,126	0,041
Initial SRS														
E($p_i^{(1)}$)≈0,05	E(n)=15	-0,001	0,010	0,002	-0,001	0,007	0,001	E(n)=15	-0,004	0,012	0,000	-0,005	0,008	-0,001
E($p_i^{(2)}$)≈0,03	E(R_2)=259,7	0,033	0,115	0,040	0,033	0,103	0,038	E(R_2)=263,0	0,039	0,118	0,042	0,039	0,104	0,040
E($p_i^{(1)}$)≈0,01	E(n)=69	-0,000	0,016	0,004	-0,001	0,009	0,002	E(n)=68	-0,001	0,013	0,003	-0,002	0,006	0,000
E($p_i^{(2)}$)≈0,006	E(R_2)=241,2	0,023	0,131	0,039	0,023	0,114	0,035	E(R_2)=241,9	0,030	0,128	0,041	0,030	0,112	0,037

La première ligne de chaque cellule présente les biais relatifs des estimateurs et la seconde, la racine carrée des erreurs quadratiques moyennes relatives des estimateurs;. $\tilde{\tau}_k$ est l'estimateur de maximum de vraisemblance et $\hat{\tau}_k$, l'estimateur tiré de la méthode bayésienne et de la distribution antérieure de Poisson. Les espérances de n et R_2 ont été dégagées par simulation.

Tableau 3. Longueur moyenne et probabilités de couverture des intervalles de confiance bootstrap à 95 % pour les tailles de population (après 2 000 itérations)

	Population I						Population II					
EASS initial	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$
$E(p_i^{(1)})\approx 0{,}05$	$E(n)=14{,}56$						$E(n)=14{,}47$					
$E(p_i^{(2)})\approx 0{,}03$	$E(R_2)=263{,}3$						$E(R_2)=264{,}2$					
Longueur	254,9	360,8	446,1	254,9	304,8	398,1	294,7	356,7	466,1	294,7	302,0	421,2
moyenne	254,9	360,8	446,1	254,9	304,8	398,1	294,7	356,7	466,1	294,7	302,0	421,2
Probabilités	0,947	0,946	0,961	0,945	0,932	0,946	0,968	0,929	0,960	0,969	0,915	0,948
de couverture	0,939	0,905	0,902	0,939	0,937	0,928	0,880	0,920	0,895	0,880	0,941	0,912
$E(p_i^{(1)})\approx 0{,}01$	$E(n)=68{,}96$						$E(n)=67{,}78$					
$E(p_i^{(2)})\approx 0{,}006$	$E(R_2)=257{,}2$						$E(R_2)=257{,}3$					
Longueur	178,4	402,2	442,5	177,2	322,1	368,3	250,6	397,4	471,8	248,0	319,4	404,2
moyenne	178,4	402,2	442,5	177,2	322,1	368,3	250,6	397,4	471,8	248,0	319,4	404,2
Probabilités	0,949	0,963	0,975	0,948	0,948	0,965	0,985	0,947	0,971	0,985	0,935	0,962
de couverture	0,946	0,762	0,802	0,945	0,814	0,864	0,869	0,790	0,807	0,867	0,841	0,837
EAS initial												
$E(p_i^{(1)})\approx 0{,}05$	$E(n)=15$						$E(n)=15$					
$E(p_i^{(2)})\approx 0{,}03$	$E(R_2)=259{,}7$						$E(R_2)=263{,}0$					
Longueur	218,2	336,7	414,5	218,2	290,7	364,9	254,9	360,8	446,1	254,9	304,8	398,1
Moyenne	218,2	336,7	414,5	218,2	290,7	364,9	254,9	360,8	446,1	254,9	304,8	398,1
Probabilités	0,942	0,930	0,959	0,896	0,913	0,924	0,968	0,943	0,981	0,846	0,916	0,908
de couverture	0,924	0,953	0,949	0,889	0,964	0,938	0,859	0,966	0,934	0,734	0,963	0,871
$E(p_i^{(1)})\approx 0{,}01$	$E(n)=69$						$E(n)=68$					
$E(p_i^{(2)})\approx 0{,}006$	$E(R_2)=241{,}2$						$E(R_2)=241{,}9$					
Longueur	162,4	359,5	396,7	135,6	306,0	335,6	222,7	360,9	426,2	151,3	304,6	341,4
Moyenne	162,4	359,5	396,7	135,6	306,0	335,6	222,7	360,9	426,2	151,3	304,6	341,4
Probabilités	0,925	0,929	0,949	0,831	0,905	0,914	0,981	0,931	0,963	0,771	0,910	0,897
de couverture	0,912	0,943	0,946	0,814	0,966	0,943	0,835	0,954	0,921	0,598	0,967	0,868

La taille des échantillons bootstrap est de 1 000 unités. Dans chaque cellule, la première ligne livre les résultats des intervalles bootstrap en points et la seconde, ceux des intervalles en percentiles.

6. CONCLUSIONS

Dans ce document, nous avons proposé une version EL (échantillonnage par liens) où l'échantillon initial vient d'un EAS séquentiel de lieux. Avec cette version, l'échantillonneur exerce un certain contrôle sur la taille de l'échantillon final, le nombre de candidats proposés, le coût de l'échantillonnage ou la précision des estimateurs. On peut estimer les tailles de population soit par les EMV proposés par Félix-Medina et Thompson (2004) ou par les estimateurs bayésiens de Félix-Medina et Monjardin (2004). Par les méthodes bootstrap, on peut élaborer des intervalles de confiance pour les tailles de population.

Les résultats de l'étude de simulation indiquent que les estimateurs de taille de population utilisés avec la version EL proposée sont d'un rendement acceptable. Il reste que, à cause du caractère aléatoire de la taille de l'échantillon initial, les estimateurs sont un peu moins efficients dans cette version que dans la version en EAS initial. Les estimateurs issus de la méthode bayésienne sont d'une efficience légèrement supérieure à celle des EMV. Disons

enfin que les intervalles de confiance dégagés par la méthode bootstrap en points offrent des propriétés de couverture supérieures à celles de la méthode en percentiles.

REMERCIEMENTS

Ces travaux de recherche ont fait l'objet des subventions UASIN-EXB-01-01 et PIFI-2003-25-28 du Secretaría de Educación Pública, ainsi que d'une subvention de la Coordinación General de Investigación y Postgrado de la Universidad Autónoma de Sinaloa.

RÉFÉRENCES

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

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

Efron, B. et Tibshirani, R. J. (1993), *Introduction to the Bootstrap*, New York: Chapman & Hall.

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

Félix-Medina, M. H. et Monjardin, P. E. (2004), "Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations: a Bayesian Assisted Approach", document propose à Techniques d'enquête.

Gross, S. (1980), "Median Estimation in Sample Surveys", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 181-184.

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