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

APPLICATIONS DES MÉTHODES D'ÉCHANTILLONNAGE ADAPTÉ À L'EXAMEN DE PROBLÈMES DE SANTÉ PUBLIQUE

Myron J. Katzoff[1]

RÉSUMÉ

Les méthodes d'échantillonnage adapté se prêtent nettement à des applications à l'examen de nouveaux problèmes de santé publique. Tel serait précisément le cas pour l'examen des conséquences des attentats bioterroristes. Comme de tels attentats devraient initialement faire des victimes dans une petite partie seulement d'une population humaine, les quelques personnes infectées pourraient être difficiles à repérer. Les méthodes d'échantillonnage adapté sont conçues pour tirer parti des liens mutuels des éléments de l'échantillon. On peut donc s'attendre à obtenir des échantillons plus riches en unités d'intérêt que par les méthodes classiques d'échantillonnage. Dans ce document, nous examinons comment, par l'échantillonnage adapté, il est possible d'étendre les enquêtes nationales sur la santé de manière à pouvoir suivre et observer efficacement de nouvelles menaces pour la santé et repérer les gens exposés.

MOTS CLÉS : Échantillonnage adapté, échantillonnage en réseau, plans d'échantillonnage à marche aléatoire.

1. INTRODUCTION

Les méthodes d'échantillonnage adapté forment une catégorie importante de plans de sondage offrant les avantages suivants :

(1) on peut réunir un nombre suffisant d'unités d'intérêt dans un échantillon pour une estimation sûre des caractéristiques de sous-domaines pour des populations « fuyantes » ou difficilement repérables;
(2) on peut enrichir les échantillons (pour ce qui est, par exemple, des jeux de variables explicatives) pour des analyses secondaires de données d'enquête ou des études détaillées de problèmes;
(3) on peut contrôler le contenu en « cas » (il s'agira, par exemple, du nombre d'unités de l'échantillon correspondant à diverses combinaisons de variables démographiques) là où on peut douter de la qualité et de l'intérêt d'un échantillon constitué pour des estimations générales de population.

Comme les éclosions de morbidité et le bioterrorisme ne devraient initialement frapper qu'une faible partie d'une population humaine, il sera peut-être difficile de repérer les quelques personnes infectées ou les gens qui auront été en contact avec ces maladies ou ces agents biologiques. Ainsi, les méthodes d'échantillonnage adapté sont de nature à accroître la qualité de l'information statistique en vue de l'examen des conséquences de ces problèmes de santé publique.

Si on assure un contrôle dynamique des tailles d'échantillon de sous-domaines par les techniques d'échantillonnage adapté, les instruments de collecte de données pourront nous livrer assez de renseignements pour que nous puissions suivre l'efficacité des interventions dans le temps et constater comment et où se propage une maladie, ce qu'on peut accomplir plus efficacement en rééchantillonnant par adaptation et à intervalles réguliers les groupes touchés dans la population. Entre autres utilisations des données issues de l'application de méthodes d'échantillonnage adapté, mentionnons l'observation et la prévision du nombre de personnes touchées aux divers stades de la propagation d'une maladie (par une modélisation d'espace d'états, par exemple), l'actualisation et la validation des modèles mathématiques d'analyse de la dynamique des populations en cas d'épidémie.

[1]Myron J. Katzoff, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311, chemin Toledo, Hyattsville, Maryland, États-Unis, 20782 (mkatzoff@cdc.gov).

Ce qui distingue l'échantillonnage adapté des plans de sondage classiques est que l'échantillonnage des unités dépend des valeurs des variables observées pour les unités sélectionnées, ce qui a d'importantes conséquences sur l'inférence statistique reposant sur des données obtenues par échantillonnage adapté. Avec des unités qu'on sélectionne par adaptation en suivant les liens, on obtient des échantillons dont la structure probabiliste n'est pas si différente de celle des échantillons classiques. On ne saurait employer les estimateurs des plans de sondage classiques sans une correction quelconque à l'égard des changements de structure probabiliste apportés par ce suivi des liens. Pour avoir une certaine idée de ce qui rend un échantillonnage « adapté » ainsi que l'incidence sur les probabilités d'inclusion des unités, on doit se rappeler que, dans certaines méthodes de base en échantillonnage adapté, l'échantillon initial est tiré par la méthode classique et que des unités sont ajoutées par une règle de poursuite ou de non-poursuite de la sélection pour chaque unité initialement sélectionnée et les autres unités de l'échantillon en place selon l'information observée pour chacune de ces unités. On trouvera d'autres détails sur ces plans de sondage dans Thompson (1992).

Dans le traitement que nous en faisons, l'« échantillonnage adapté » vise des plans de sondage par liens comme l'échantillonnage en réseau, boule de neige, par renvois, déterminé par l'enquêté, à marche aléatoire ou en grappes par adaptation. Il est aussi question de l'échantillonnage adapté par ensemble actif, cette nouvelle catégorie de plans d'échantillonnage (avec ceux qui font tout simplement figure de « cas d'espèce ») où des unités s'ajoutent à l'échantillon par un mélange de deux distributions dont on tire des unités soit avec de fortes probabilités en se reportant aux valeurs des variables mesurées pour un certain sous-ensemble ou « ensemble actif » d'unités de l'échantillon en place, soit avec de moindres probabilités par la structure de la base de sondage sans tenir compte des valeurs de toute variable observée.

Remerciements : Je me suis largement inspiré des concepts et des idées avancés par Monroe G. Sirken et Steven K. Thompson. J'ai aussi eu à mon profit de nombreux entretiens avec ces chercheurs de premier plan.

2. QUELQUES TERMES FRÉQUEMMENT EMPLOYÉS

Dans certaines études portant sur les plans d'échantillonnage adapté, les chercheurs ont jugé bon de recourir aux termes et aux concepts de la théorie des graphes pour décrire des populations avec des liens ou des réseaux. Si on utilise les diagrammes de la théorie des graphes, les nœuds de ces graphes représentent les éléments d'intérêt de la population et les arêtes (parfois appelées arcs), les traits qui unissent les unités de ces nœuds. Selon les applications, il importe également de distinguer les unités d'observation (unités de la population étudiée et donc les nœuds), des unités de sélection (unités de la base de sondage ou de l'univers qui se trouvent à la base d'un échantillon d'unités d'observation). Les unités d'observation et de sélection peuvent être identiques ou différentes.

3. TRAITS DISTINCTIFS DU PLAN D'ÉCHANTILLONNAGE ADAPTÉ DE LA NHIS

L'Enquête nationale par interviews sur la santé (National Health Interview Survey ou NHIS) a des caractéristiques qui en font un mécanisme attrayant sur lequel peuvent se greffer les méthodes d'échantillonnage adapté. Il en sera question dans les exemples présentés plus loin qui ont valeur d'illustration, bien que revêtant un caractère conjectural.

La NHIS est une enquête nationale à échantillonnage stratifié à plusieurs degrés où les ménages sont les unités de sélection au dernier degré. Les caractéristiques de conception que nous allons mentionner sont de nature à faciliter l'application locale, voire nationale des méthodes d'échantillonnage adapté le cas échéant :

(1) l'État est une variable de stratification;
(2) les unités primaires d'échantillonnage (UPE) sont les comtés, les groupes de comtés, les secteurs équivalents de comtés (paroisses et villes indépendantes, par exemple), les agglomérations, les divisions cantonales, les divisions civiles secondaires ou les régions statistiques métropolitaines (RSM);

(3) un certain nombre de grandes UPE sont sélectionnées avec certitude et constituent donc en soi des strates;
(4) s'il y a lieu et à condition que le plan de sondage général ne change pas entre périodes qui se suivent dans le temps, il est possible de grouper les échantillons « classiques » de ces périodes de manière à constituer des échantillons initiaux d'une taille suffisante pour un échantillonnage adapté d'unités d'observation pour les 80 régions et plus désignées comme régions statistiques métropolitaines regroupées (RSMR), régions statistiques métropolitaines primaires (RSMP) et RSM proprement dites.

4. EXEMPLE D'ÉCHANTILLONNAGE EN RÉSEAU PAR LIENS SYMÉTRIQUES

Supposons que l'objectif statistique est d'estimer dans chaque État le nombre de personnes atteintes d'une maladie d'origine génétique dans les trois plus grandes sous-populations asiatiques. On peut songer à cette fin à ajouter des gens à l'échantillon habituel de ménages de la NHIS en suivant les liens familiaux. Dans cet exemple, les adultes et les enfants qui appartiennent à la population d'un État seront les nœuds. Les arcs seront les liens familiaux : parents, frères et sœurs des parents, enfants et leurs frères et sœurs, enfants des frères et sœurs qui résident dans cet État. Supposons aussi que l'instrument d'enquête et le protocole d'interview sont conçus pour que l'inclusion de toute unité d'observation d'un groupe en lien comporte implicitement ou explicitement l'inclusion des données relatives à toutes les unités d'observation en lien dans ce groupe. Ces liens seront alors qualifiés de symétriques. (Il convient tout particulièrement de noter que l'inclusion au hasard de tout nœud dans le domaine d'intérêt donne lieu à l'inclusion dans l'échantillon des données relatives à tous les membres liés de la population.) À noter que le protocole d'interview pourrait permettre une diversité de modes d'interview ou de répondants par procuration, ce qui permettrait d'alléger les coûts de collecte de données dans certaines applications.

La NHIS fait intervenir un jeu complexe de corrections de pondération d'échantillon conçu pour produire des estimations nationales des totaux de population sans biais de plan de sondage. Une pondération d'estimation nationale de population fait intervenir quatre facteurs :

(1) pondération de base qui est l'inverse des probabilités de sélection des ménages;
(2) correction de non-réponse des ménages;
(3) correction de rapport au premier degré permettant de tenir compte de la pondération « personnes » des UPE non sélectionnées avec certitude par rapport aux totaux de population pour 24 catégories de résidence et de race-ethnicité;
(4) correction de rapport au second degré permettant de tenir compte de la pondération « personnes » par rapport aux totaux indépendamment établis par le Census Bureau des États-Unis pour 88 catégories d'âge-sexe-race-ethnicité de la population civile hors établissement.

On trouvera plus de détails sur la finalité et l'utilité des corrections de pondération dans Botman, Moore, Moriarity et Parsons (2000). En principe, il serait possible d'apporter certaines modifications à la suite mentionnée de ces corrections afin de dégager des estimations pour chaque État, mais on ne devrait pas s'attendre à ce que des corrections efficaces de cellules soient les mêmes pour les divers États, puisque le mode efficace de pondération dans chaque cas pourrait dépendre de l'objectif statistique visé et comporter une vaste expérimentation rendue nécessaire par une intégration « sur mesure » des cellules de correction pour chacun des États.

Pour traiter uniquement des caractéristiques essentielles de l'estimateur de base par plan de sondage pour notre mode d'échantillonnage adapté, nous ne considérerons plus les corrections de pondération que nous imposerait fort probablement le traitement des problèmes de non-réponse et de base de sondage. C'est ainsi que nous pouvons voir l'établissement des principales données d'enquête pour chaque État comme passant par au plus deux étapes d'échantillonnage (1) au niveau des comtés ou d'autres unités appropriées comme les UPE et (2) au niveau des ménages à l'intérieur des UPE en tant qu'unités d'échantillonnage secondaires (sous-unités d'échantillonnage ou SUE). Pour l'unité d'observation ℓ, soit

$$y_\ell = \begin{cases} 1, & \text{si l'unité fait partie d'une sous-population} \\ & \text{asiatique d'intérêt et atteinte de la maladie;} \\ 0, & \text{dans les autres cas.} \end{cases} \qquad (1)$$

Si A_{ik} désigne dans l'État l'ensemble d'unités d'observation qui sont liées à l'unité de sélection k de l'UPE i et m_ℓ, le nombre d'unités de sélection auxquelles est liée l'unité d'observation ℓ dans ce même État, nous définissons la moyenne d'unités de sélection par

$$w_{ik} = \sum_{\ell \, in \, A_{ik}} \frac{y_\ell}{m_\ell}. \qquad (2)$$

Si on tire sans remise t UPE de T avec des probabilités d'inclusion π_i et que, dans chaque UPE, on sélectionne n_i ménages dans N_i sans remise par échantillonnage aléatoire simple, l'estimateur sans biais à deux degrés du nombre total de résidents de l'État appartenant au groupe cible pour la maladie sera :

$$\hat{\tau} = \sum_{i=1}^{t} \pi_i^{-1} \frac{N_i}{n_i} \sum_{k=1}^{n_i} w_{ik}. \qquad (3)$$

Il y a un apport à chaque estimation qui correspond à $\hat{\tau}$ pour chaque strate sur laquelle les UPE ne sont pas prélevées avec certitude. Comme le taux d'incidence par État sera peut-être d'un plus grand intérêt pour le chercheur, on notera qu'il serait facile de définir à cette fin des variables indicatrices sous la forme y_ℓ qui séparent la présence ou l'absence de la maladie de l'appartenance à un groupe racial. Bien sûr, le taux est alors le rapport de deux estimations de totaux pour chaque État, auquel cas les estimateurs du numérateur et du dénominateur comprendraient chacun des termes sous la forme $\hat{\tau}$.

On trouvera dans Katzoff, Sirken et Thompson (2002) des expressions simples par base de sondage pour la variance de $\hat{\tau}$ et son estimateur. Toutefois, dans toute application qui comporte une utilisation supplémentaire de la NHIS, on se doit de tenir compte, dans l'estimation des variances, de nombreux effets d'échantillonnage dont nous n'avons pas parlé. À des fins pratiques d'analyse de données, on aura besoin de structures générales d'estimation de variance qui sont d'une application facile avec les logiciels dont nous disposons. Ces points sont examinés en détail dans Botman, Moore, Moriarity et Parsons (2000).

Quelques observations finales s'imposent au sujet de cet exemple. Il peut se trouver qu'une lecture attentive du premier exemple dans Katzoff, Sirken et Thompson (2002) permette de constater une différence subtile entre les deux exemples, bien que les formes algébriques des estimateurs soient les mêmes. Des liens de notre exemple, on peut dire qu'ils sont en forte connexité, terme qui, dans un échantillonnage par liens complets, désigne l'adjonction de tous les membres d'un groupe en lien pour toute unité d'observation. Dans la pratique, on peut ne pas définir de liens en forte connexité et le suivi des liens complets pourrait se révéler peu pratique à cause de restrictions de ressources.

5. EXEMPLE D'ÉCHANTILLONNAGE ADAPTÉ À MARCHE ALÉATOIRE

Dans cet exemple, l'objectif statistique est d'estimer la proportion d'une population locale, d'une RSMR, d'une RSMP ou d'une RSM qui est exposée à un agent biologique contagieux et transmissible. Là encore, l'échantillon de la NHIS est notre point de départ. Les nœuds sont les unités de la population et les axes sont déterminés par les

mécanismes de contact et de transmission au sein de la population. Dans l'application de la méthode à marche aléatoire que nous décrivons, l'incidence des caractéristiques du plan d'échantillonnage NHIS diminue à mesure que progresse l'échantillonnage. En fait, on peut quelque peu perdre de vue l'échantillon initial de personnes dans la population des ménages, parce qu'on suit et contrôle l'échantillonnage par liens de manière à obtenir asymptotiquement les probabilités voulues de sélection, ainsi que nous allons le décrire. L'idée de base est de s'attacher aux tendances naturelles des populations et à se laisser guider juste assez par l'échantillonnage par liens pour que puissent être dégagées des estimations simples qui soient représentatives de toute la population. La stratégie que nous retenons vient de Thompson (2003).

Pour un tel échantillonnage à marche aléatoire, on a besoin d'une structure de chaîne de Markov qui est apériodique et irréductible. Définissons un ensemble fondamental de probabilités de passage par

$$q_{ij} = \begin{cases} \dfrac{(1-d)}{N} + \dfrac{d\, a_{ij}}{a_{i.}}, & \text{si } a_{i.} > 0 \\ \dfrac{1}{N}, & \text{si } a_{i.} = 0, \end{cases} \qquad (4)$$

où $a_{ii} = 0$, pour $i \neq j$, a_{ij} étant le nombre de liens entre les unités i et j et où $0 < d < 1$ et $a_{i.} \overset{\text{def}}{=} \sum_j a_{ij}$, soit le degré de sortie pour le nœud i. Posons ensuite que les membres de la population font l'objet d'une sélection approximative dans une distribution stationnaire donnée $\{\pi_1, \pi_2, \pi_3, \ldots\}$. On peut déterminer les probabilités π_i comme fonction de variables démographiques. Par notre connaissance de la vulnérabilité de chacun à l'agent biologique par exemple, on peut scinder la population en deux et sélectionner des gens dans le premier groupe à un taux d'échantillonnage qui sera double de celui de l'autre groupe. Posons aussi que, d'après les données d'un recensement récent, on sait que le premier groupe compte N_1 personnes et le second, N_2. Dans ce cas, on pourrait songer à sélectionner dans le premier avec des probabilités $\dfrac{2}{2N_1 + N_2}$ et dans le second avec des probabilités $\dfrac{1}{2N_1 + N_2}$, de sorte que π_i prenne une et une seulement de ces valeurs.

Nous nous guidons dans cette marche aléatoire vers la distribution stationnaire $\{\pi_1, \pi_2, \pi_3, \ldots\}$ par la méthode de Hastings (1970), qui nous impose une matrice de probabilités de passage aux éléments diagonaux

$$P_{ij} = q_{ij}\alpha_{ij}, \qquad (5)$$

où $\alpha_{ij} = \min\{1, \pi_j q_{ji} / \pi_i q_{ij}\}$ et où les éléments de la diagonale sont $P_{ii} = 1 - \sum_{i \neq j} P_{ij}$. Cette technique livre des données permettant de quantifier l'estimateur quotient généralisé de la moyenne

$$\hat{\mu} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \Big/ \sum_{i=1}^{n} \frac{1}{\pi_i}, \qquad (6)$$

où n est la taille d'échantillon avec ou sans répétition. À noter que, si $\pi_i = \pi$ pour tout i, $\hat{\mu}$ est simplement la moyenne arithmétique. Les études de simulation à éléments finis réalisées à ce jour nous indiquent que la distribution d'échantillonnage de $\hat{\mu}$ comporte une moyenne très proche de la moyenne connue pour des valeurs modérées de n si les π_i de la distribution stationnaire n'ont pas de valeurs extrêmes. Ce serait là une direction importante pour la vérification de résultats asymptotiques.

Pour l'estimateur de la variance de $\hat{\mu}$, Thompson (2003) a envisagé d'utiliser la distribution de permutation d'échantillon en conditionnalité par les estimateurs exhaustifs. C'est une méthode qui peut toutefois donner des estimations négatives de variance. Pour éviter la difficulté, Thompson (2003) propose un estimateur fondé sur quelques itérations indépendantes de la marche aléatoire.

6. OBSERVATIONS EN CONCLUSION

Dans le premier exemple, il fallait que chaque élément en lien soit entièrement échantillonné en application du protocole d'échantillonnage pour des estimations sans biais. Dans le second, il est difficile de voir combien longtemps la marche aléatoire devrait se poursuivre pour que les effets asymptotiques se manifestent. Il reste que la grande famille des techniques d'échantillonnage adapté par ensemble actif (EAEA) comprend bien des plans de sondage qui visent à résoudre de tels problèmes et dont nous n'avons pas parlé. Elle permet l'application d'un critère d'adjonction d'unités par variables de lien et mode, tout en donnant de la souplesse dans le contrôle de la taille d'échantillon et la détermination des unités d'échantillonnage qui dictent l'intégration à l'échantillon de nouveaux membres de la population. Dans les méthodes d'estimation applicables à des plans de sondage EAEA, on emploie le détail des structures probabilistes à chaque stade de la sélection lorsqu'on ajoute des unités et on exploite des idées comme celle des estimateurs exhaustifs et des méthodes de Monte Carlo afin de dégager des estimations sans biais.

Dans les futures études consacrées aux plans de sondage EAEA, on examinera d'autres mécanismes d'échantillonnage et d'adjonction d'unités. Il sera notamment question d'estimations asymptotiques par base de sondage en sus des estimateurs habituels de variance pour les moyennes. Dans le cas des plans de sondage où on recourt à un échantillon supplémentaire, il faut mieux comprendre l'incidence des paramètres dans le tirage de l'échantillon initial et, en particulier, les problèmes de répartition (quelle devrait être la taille de l'échantillon initial dans divers plans de sondage?). Il y a enfin le besoin considérable de pleinement développer les méthodes d'estimation par modèle pour des plans d'échantillonnage EAEA.

RÉFÉRENCES

Botman, S. L., Moore, T. F., Moriarity, C. L. et Parsons, V. L. (2000), "Design and Estimation for the National Health Interview Survey, 1995-2004", *Vital Health Statistics*, Series 2, 130, Hyattsville, Maryland: National Center for Health Statistics.

Hastings, W. K. (1970), "Monte Carlo Sampling Using Markov Chains and Their Applications", *Biometrika*, 57, pp. 97-109.

Katzoff, M. J., Sirken, M. G. et Thompson, S. K. (2002), "Proposals for Adaptive and Link-Tracing Sampling designs in Health Surveys", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 1772-1775.

Thompson, S. K. (1992), *Sampling*, New York: Wiley.

Thompson, S. K. (2003), "Simulation Program for Link-Tracing Designs: Designs that Take Advantage of inherent Link-Tracing Tendencies and that Avoid Truncation Problems", unpublished report, Hyattsville, Maryland: National Center for Health Statistics.