N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

UN PLAN POUR LA COORDINATION DES ENQUÊTES AU NATIONAL AGRICULTURAL STATISTICS SERVICE

Phillip S. Kott[1]

RÉSUMÉ

Le National Agricultural Statistics Service (NASS) utilise de plus en plus fréquemment des variantes de l'échantillonnage de Poisson pour sélectionner les échantillons pour ses enquêtes sur les fermes. Pour une enquête particulière, les probabilités de sélection sont déterminées (idéalement) en utilisant la sélection de Brewer ou la sélection maximale de Brewer, selon le nombre de variables d'intérêt de l'enquête. Dans ces scénarios, les populations de fermes à l'intérieur des États ne sont pas stratifiées et une probabilité de sélection particulière est attribuée à chaque ferme. Combinée prudemment au calage, la nature aléatoire de l'échantillonnage de Poisson a un effet négligeable sur l'erreur quadratique moyenne d'un estimateur. Afin de limiter le nombre de fois qu'une ferme particulière est sélectionnée pour une enquête du NASS, l'auteur a proposé que l'organisme utilise un échantillonnage de Poisson à intervalles séquentiels (PIS) pour l'ensemble de ses nombreuses enquêtes. L'échantillonnage PIS permet d'éviter l'étape éventuellement difficile du calcul de la probabilité conditionnelle de sélection d'une ferme pour une nouvelle enquête sachant sa probabilité de sélection cible (inconditionnelle) pour cette enquête et sa situation de sélection dans les enquêtes précédentes.

MOTS CLÉS : Calage, échantillonnage de Poisson à intervalles séquentiels, sélection maximale de Brewer.

1. INTRODUCTION

Le National Agricultural Statistics Service (NASS) réalise une grande variété d'enquêtes couvrant plus ou moins la même population de fermes américaines. Le présent article décrit une méthode que le NASS est en train d'élaborer pour coordonner la sélection des échantillons d'enquête de façon à contrôler le nombre de fois que les unités de la population sont choisies.

Le NASS utilise de plus en plus fréquemment une variante de la *sélection maximale de Brewer (SMB)*, appelée à l'interne sélection « PPT multivariée » ou « PPTM », pour tirer l'échantillon d'enquêtes particulières. Pour expliquer la sélection maximale de Brewer, nous avons besoin d'une certaine notation. Supposons qu'une enquête est conçue pour estimer des statistiques sommaires pour K items. Soit y_{ik} le k^e item d'intérêt pour l'unité (ferme) i dans la population U et x_{ik} une mesure de taille connue (avant l'échantillonnage) de l'item k pour l'unité i. L'ensemble $\{x_{ik}\}$ représente les valeurs de contrôle.

Sous la SMB, la probabilité de sélection de l'unité i prend la forme :

$$\pi_i = \max_k \{n_k x_{ik}^{g(k)} / \Sigma_U x_{jk}^{g(k)}\}, \quad (1)$$

où les n_k sont des constantes prédéterminées. La théorie sur laquelle s'appuie l'équation (1) provient du modèle :

$$y_{ik} = x_{ik}\beta_k + \varepsilon_{ik}, \quad (2)$$

où $E(\varepsilon_{ik} \mid x_{ik}) = E(\varepsilon_{ik}, \varepsilon_{jk} \mid x_{ik}, x_{jk}) = 0$ pour $i \neq j$ et $E(\varepsilon_{ik}^2 \mid x_{ik}) \propto x_{ik}^{2g(k)}$ pour chaque item cible k. Voir Kott et Bailey (2000) pour des précisions.

1 Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030, pkott@nass.usda.gov

Il existe (au moins) deux moyens de tirer un échantillon SMB. Le premier est l'échantillonnage PPT systématique, où π_i est la mesure de taille, et 1 est l'intervalle d'échantillonnage. Sous ce plan, la taille de l'échantillon, n, est exactement $n^* = \Sigma_U \pi_i$ quand n^* est un nombre entier. Sinon, n est l'un des deux nombres entiers les plus proches de n^*. La liste à partir de laquelle l'échantillon PPT systématique est tiré peut être ordonnée de façon à assurer (ou à presque assurer) que les exigences de taille d'échantillon soient respectées pour certains types de fermes, telles que celles cultivant un produit rare, disons des oignons.

Le deuxième moyen de tirer un échantillon SMB est l'échantillonnage de Poisson. Sous cet échantillonnage, n est une variable aléatoire pour laquelle $E(n) = \Sigma_U \pi_i$ et $Var(n) = \Sigma_U \pi_i(1-\pi_i)$.

Dans les enquêtes où le NASS utilise la SMB, les poids d'échantillonnage, $\{w_i, i \in S\}$, sont calés sur toutes les variables de contrôle, autrement dit $\Sigma_S w_i x_{ik} = \Sigma_U x_{ik}$ pour tout k, où S dénote l'échantillon. Sous le modèle,

$$y_{ik} = \Sigma_g x_{ig}\beta_g + \varepsilon_i , \qquad (3)$$

où $E(\varepsilon_i | \{x_{jg}\}) = E(\varepsilon_i \varepsilon_{i'} | \{x_{jg}\}) = 0$ pour $i \neq i'$, les erreurs quadratiques moyennes prévues pour $t_k = \Sigma_S w_i x_{ik}$ sont les mêmes sous l'un et l'autre plan de sondage (de nouveau, voir Kott et Bailey, 2000). Le modèle de l'équation (3) est un peu plus général que celui décrit par l'équation (2).

À la section 2, nous décrivons une méthode pour minimiser le chevauchement de la sélection des unités entre deux échantillons SMB. À la section 3, nous présentons l'échantillonnage de Poisson à intervalles séquentiels (PIS). Bien qu'il soit plus limité que la méthode présentée à la section 2, il généralise simplement les situations où plus de deux échantillons doivent être coordonnés, comme cela est souvent le cas en pratique. À la section 4, nous examinons la coordination d'un nouvel échantillon à des échantillons antérieurs sélectionnés indépendamment. À la section 5, nous présentons une discussion.

2. UN CADRE DE TRAVAIL À DEUX ÉCHANTILLONS

Supposons que nous souhaitions tirer des échantillons pour deux enquêtes, A et B, à partir de la même population U. Nous avons les probabilités de sélection inconditionnelles (déterminées au moyen d'une formule comme l'équation (1)) pour chaque unité comprise dans U, π_A et π_B, pour les enquêtes A et B, respectivement. Nous supprimons l'indicateur de l'unité pour simplifier. Il est permis que π_A ou π_B soit nulle. Effectivement, les populations à partir desquelles les échantillons sont tirés pour A et B ne doivent pas coïncider strictement, autrement dit l'une peut contenir des unités qui ne figurent pas dans l'autre. Néanmoins, les unités doivent être *définies* de façon compatible.

Étant donné un échantillon tiré pour l'enquête A (appelé « échantillon A ») selon les probabilités de sélection prescrites, comment peut-on tirer pour l'enquête B un échantillon qui minimise le chevauchement des unités entre les deux échantillons? En nous concentrant sur une unité particulière de la population, nous pouvons faire ce qui suit :

Si $\pi_A + \pi_B \leq 1$ et que l'unité est dans l'échantillon A,
alors ne pas la choisir pour l'échantillon B.
Si $\pi_A + \pi_B \leq 1$ et que l'unité n'est pas dans l'échantillon A,
alors la choisir pour l'échantillon B avec la probabilité conditionnelle $p_{B|\bar{A}} = \pi_B/(1-\pi_A)$.
Si $\pi_A + \pi_B > 1$ et que l'unité est dans l'échantillon A,
alors la choisir pour l'échantillon B avec la probabilité conditionnelle $p_{B|A} = (\pi_A + \pi_B - 1)/\pi_A$.
Si $\pi_A + \pi_B > 1$ et que l'unité n'est pas dans l'échantillon A,
alors la choisir pour l'échantillon B avec certitude.

Observons que, si $\pi_A + \pi_B \leq 1$, alors l'unité peut être choisie uniquement pour l'échantillon A ou pour l'échantillon B, mais non pour les deux. En outre, la probabilité inconditionnelle que l'unité soit dans l'échantillon B est $(1 - \pi_A) \times p_{B|\bar{A}} = \pi_B$. Si $\pi_A + \pi_B > 1$, alors il est impossible d'assurer que l'unité ne figurera pas dans les deux échantillons. Notons que la probabilité inconditionnelle que l'unité soit dans l'échantillon B est $\pi_B = \pi_A \times p_{B|A} + (1 - \pi_A) \times p_{B|\bar{A}}$. Partant de cette

égalité, quand $\pi_A + \pi_B > 1$, nous minimisons $p_{B|A}$ en fixant $p_{B|\bar{A}}$ à la valeur la plus élevée à laquelle elle peut être fixée, c'est-à-dire 1.

L'utilisation de la technique de sélection de l'échantillon B avec les probabilités conditionnelles décrites plus haut est laissée à la discrétion de l'échantillonneur. A la section suivante, nous présentons une méthode qui s'appuie sur l'échantillonnage de Poisson pour les deux enquêtes et nous expliquons les avantages qu'il y a à accepter cette contrainte.

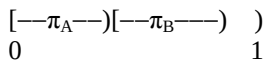
3. POISSON À INTERVALLES SÉQUENTIELS

Une méthode d'échantillonnage de Poisson très répandue consiste à commencer par attribuer à chaque unité de la population un numéro aléatoire permanent, r, issu de la distribution uniforme sur l'intervalle unitaire [0, 1). L'unité est ensuite sélectionnée dans l'échantillon A (disons) si, et uniquement si, $r < \pi_A$. En utilisant cette méthode pour l'échantillon A, nous tirons un échantillon de *Poisson à intervalles séquentiels* (PIS) pour l'enquête B qui concorde comme suit avec la méthode de minimisation du chevauchement décrite à la section précédente :

si $\pi_A + \pi_B \leq 1$, alors sélectionner l'unité pour l'échantillon B quand (et uniquement quand) $r \in [\pi_A,\ \pi_A + \pi_B)$;
si $\pi_A + \pi_B > 1$, alors sélectionner l'unité dans l'échantillon B quand $r \in [\pi_A, 1)$ ou $r < \pi_A + \pi_B - 1$.

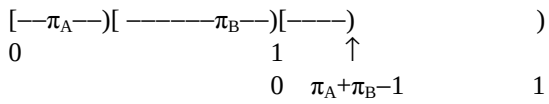
La figure 1 représente la première condition. Quand $\pi_A + \pi_B \leq 1$, si r est inférieur à π_A, alors l'unité est sélectionnée dans l'échantillon A. Si r est égal ou supérieur à π_A, mais inférieur à π_B, alors l'unité est sélectionnée dans l'échantillon B. Si r est plus grand que π_B, alors l'unité n'est sélectionnée dans aucun des deux échantillons. L'« intervalle d'échantillonnage » pour A est $[0, \pi_A)$, dont la longueur est π_A. L'intervalle pour l'échantillon B est $[\pi_A,\ \pi_A + \pi_B)$, dont la longueur est π_B.

Figure 1 : PIS quand $\pi_A + \pi_B \leq 1$



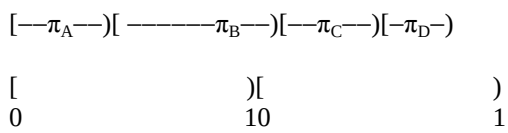
La figure 2 représente la deuxième condition. Si $\pi_A + \pi_B > 1$, alors l'intervalle pour l'échantillon A est le même qu'à la figure 1. L'intervalle pour l'échantillon B s'étend de π_A à 1, puis « enveloppe » [0, 1) afin d'inclure le secteur allant de 0 à $\pi_A + \pi_B - 1$ également.

Figure 2 : PIS quand $\pi_A + \pi_B > 1$



Observons que l'échantillonnage PIS ne nécessite pas le calcul des probabilités conditionnelles de sélection, ce qui simplifie beaucoup les choses quand on doit coordonner plus de deux enquêtes. Considérons, par exemple, les quatre enquêtes de la figure 3. Tel qu'illustré, $\pi_A + \pi_B > 1$, mais $\pi_A + \pi_B + \pi_C + \pi_D \leq 2$. Autrement dit, l'unité ne doit pas figurer dans plus de deux échantillons. Il est facile de dire que si $r < \pi_A + \pi_B - 1$, alors l'unité sera comprise dans les échantillons A et B. Si $r \geq \pi_A + \pi_B - 1$ et $r < \pi_A$, alors l'unité sera comprise dans les échantillons A et C. Si $r \geq \pi_A$ et $r < \pi_A + \pi_B + \pi_C - 1$, alors l'unité sera comprise dans les échantillons B et C. Si $r \geq \pi_A + \pi_B + \pi_C - 1$ et $r < \pi_A + \pi_B + \pi_C + \pi_D - 1$, alors l'unité sera comprise dans les échantillons B et D. Enfin, si $r \geq \pi_A + \pi_B + \pi_C + \pi_D - 1$, l'unité sera comprise dans l'échantillon B uniquement.

Figure 3 : PIS pour quatre enquêtes



4. PSEUDO PIS

Une question qui s'est posée au NASS est celle de savoir s'il est possible de coordonner un nouvel échantillon SMB à des échantillons tirés antérieurement. Supposons que nous ayons Q échantillons tirés indépendamment (dénotés 1, ..., Q), que nous souhaitons coordonner à un nouvel échantillon SMB, B. Nous pouvons dénommer A l'union de tous les échantillons antérieurs. En nous concentrant de nouveau sur une unité particulière et en supprimant son identité de la notation, notons que $\pi_A = 1 - (1 - \pi_1) \times (1 - \pi_2) \times ... \times (1 - \pi_Q)$. Suit une procédure d'échantillonnage « pseudo PIS ».

Par souci de simplicité, nous supposons que $\pi_A \neq 1$. Observons que, si π_A est égal à 1, alors l'unité doit être sélectionnée avec certitude dans au moins un des échantillons sélectionnés antérieurement. Nous pouvons éliminer ces échantillons de A. Après cette étape (si elle est nécessaire), choisir ρ aléatoirement à partir de la distribution uniforme sur [0 , 1). Si l'unité est comprise dans l'échantillon A, lui attribuer le numéro aléatoire permanent $r = \rho\pi_A$. Sinon, lui attribuer $r = \rho(1 - \pi_A)$. Exécuter l'échantillonnage PIS pour l'échantillon B en commençant à π_A. La figure 1 s'applique de nouveau.

5. DISCUSSION

Ohlsson (1995) décrit plusieurs moyens de coordonner des échantillons en utilisant des numéros aléatoires permanents. La principale distinction entre ces méthodes et l'échantillonnage PIS est que les premières ont trait à des populations stratifiées. Bien que la définition des strates puissent varier d'un échantillon à l'autre, toutes les unités comprises dans une strate ont le même intervalle d'échantillonnage par rapport à l'échantillon sélectionné. Par contre, l'échantillonnage PIS a été conçu pour être utilisé avec des échantillons SMB n'ayant aucun identificateur de strate. De surcroît, les intervalles d'échantillonnage d'une unité de la population sont particuliers à la fois à l'échantillon et à l'unité.

Nous avons fait remarquer à la section 2 que π_A peut être nulle. Par conséquent, dans le cas de l'échantillonnage PIS, le traitement des « nouvelles unités », c'est-à-dire celles qui sont ajoutées à la population après qu'un ou plusieurs échantillons aient été sélectionnés, est un problème simple. Il suffit de prétendre que ces nouvelles unités ont toujours été comprises dans la population avec une probabilité de sélection nulle pour les échantillons tirés avant qu'elles ne figurent dans la population. Par exemple, dans le cas de deux échantillons, on attribue à une nouvelle unité survenant après le tirage de l'échantillon A un numéro aléatoire permanent, r, et on ne la sélectionne dans l'échantillon B que si, et uniquement si, $r < \pi_B$.

On peut utiliser l'échantillonnage PIS pour coordonner des échantillons d'enquête au fil des ans, mais un problème se pose si l'on utilise l'information recueillie au moyen d'une enquête pour mettre à jour les valeurs de contrôle d'une autre à laquelle elle est coordonnée. Pour le montrer, considérons de nouveau le cas de deux échantillons en nous concentrant sur une unité particulière de la population. L'intervalle $[\pi_A, \pi_A + \pi_B)$ donné à la figure 1 représente uniquement la probabilité que l'unité soit sélectionnée dans l'échantillon B *avant* que nous sachions si elle a été tirée dans l'échantillon A. Donc, nous ne pouvons pas utiliser l'information provenant de l'enquête A pour mettre à jour les valeurs de contrôle sur lesquelles se fonde le calcul de π_B.

Il existe un moyen limité de contourner cette contrainte si l'échantillonneur utilise l'information au sujet d'une unité provenant d'une enquête antérieure uniquement pour mettre à jour la probabilité de sélection de *cette unité*. Cela peut se faire dans le cas de l'échantillonnage de Poisson, puisque les probabilités de sélection des unités sont indépendantes. Considérons la situation de la figure 3. L'information provenant de l'enquête A ne peut pas être utilisée pour mettre à jour π_B ni π_C sans que l'échantillonnage PIS n'affecte les probabilités de sélection inconditionnelle réelles. L'information provenant de A *peut* être utilisée pour mettre à jour π_D. Une fois que l'échantillonneur a enveloppé l'intervalle unitaire

pour une unité, l'information recueillie à partir de l'échantillon associé à ce passage par l'intervalle unitaire peut être utilisée pour mettre à jour la probabilité de sélection inconditionnelle de l'unité pour un nouvel échantillon.

En permettant que les probabilités de sélection des unités pour un échantillon particulier soient fondées sur un ensemble variable de valeurs de contrôle, certaines étant des valeurs mises à jour et d'autres non, nous nous écartons de l'échantillonnage SMB pur décrit par l'équation (1). L'objectif consiste alors à tirer des échantillons raisonnablement efficaces de façon statistiquement défendable qui limitent le fardeau imposé aux répondants potentiels. En se souvenant de cela, on pourrait considérer de rajuster la valeur g(k) de l'équation (1) (fixée par le NASS à ¾ selon la recommandation de Brewer) à la baisse. Plus cette valeur sera petite, plus la probabilité qu'une unité particulière se trouve dans plus d'un échantillon sera faible. Il faudra approfondir l'étude de la sensibilité des erreurs quadratiques moyennes à la taille de g(k).

Enfin, notons que les échantillons de Poisson préconisés ici simplifient l'estimation de l'erreur quadratique moyenne (eqm) fondée sur la randomisation. Toutefois, l'estimation de l'eqm purement fondée sur la randomisation peut donner des résultats trompeurs dans le cas de l'échantillonnage de Poisson. Voir Kott (2000).

RÉFÉRENCES

Kott, P. S. (2000), « Poisson Sampling, Regression Estimation, and the Delete-a-Group Jackknife », unpublished report, Fairfax, Virginia: National Agricultural Statistics Service, http://www.nass.usda.gov/research/reports/Poisppdag9_jos.pdf.

Kott, P.S. and Bailey, J. T. (2000), « The Theory and Practice of Maximal Brewer Selection ». *Proceedings of the Second International Conference on Establishment Surveys*, *Invited Papers*, pp. 269-278.

Ohlsson, E. (1995), « Coordination of Samples Using Permanent Random Numbers », In B.G. Cox et al. (eds.) *Business Survey Methods*, New York: Wiley, pp. 153-169.