Autour de la coordination d'échantillons poissoniens

Lionel QUALITÉ(*)

(*)Office Fédéral de la Statistique et Université de Neuchâtel, Suisse

Introduction

L'Office Fédéral de la Statistique (OFS) utilise un système d'échantillonnage coordonné développé dans (9). Ce système est une extension de la méthode proposée dans (1). Chaque échantillon transversal sélectionné dans ce système provient d'un plan de sondage de Poisson. L'usage de cette procédure, avec la variabilité des tailles d'échantillons qu'elle implique, nous conduit à réviser les méthodes de planifications employées. L'allocation optimale pour des échantillons stratifiés est remplacée par un calcul de tailles espérées d'échantillons dans des domaines. Un problème particulier qui apparait avec ces tirages poissoniens et qui n'existait pas avec les plans stratifiés utilisés avant l'intoduction du système de coordination est le risque de sélectionner une échantillon dont la taille dans certains domaines est bien inférieure à la taille espérée. Le risque d'observer des échantillons trop petits dans certains domaines était en fait déjà présent dans la plupart des enquêtes si l'on considère que la non-réponse est une phase supplémentaire de sondage : celle-ci est usuellement modélisée par un plan de Bernoulli dans des domaines et conduit donc à une taille aléatoire. Le deuxième problème que nous avons rencontré et pour lequel une solution partielle a été trouvée est celui de la sélection d'échantillons où une corrélation entre les sélections des unités est exigée, par exemple lorsque l'on n'autorise la sélection que d'une unité par ménage. De tels échantillons ne peuvent être obtenus simplement avec le système de coordination, mais nous proposons d'utiliser celui-ci pour une des phases de tirage d'un plan à plusieurs phases qui fournit des échantillons vérifiant les contraintes. Le calcul des probabilités d'inclusion à chaque phase de tirage reste cependant trop complexe pour proposer une solution complète.

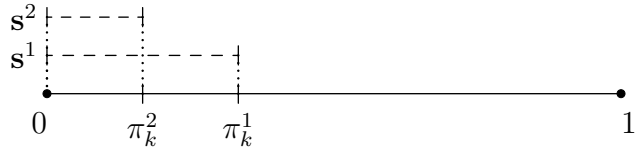
1 Échantillonnage poissonien coordonné

Dans (9), une méthode d'échantillonnage coordoné est proposée, qui permet d'atteindre une corrélation soit minimale soit maximale entre les indicatrices de sélection I_k^t d'une unité k dans différents échantillons s^t, pour tout k dans une population U. Cette méthode est une extension naturelle du plan de sondage de (1) pour deux échantillons. La méthode de (1) repose sur l'utilisation de nombres aléatoires permanents u_k dans $[0, 1]$, et sur la définition de zones de sélection, sous ensembles de $[0, 1]$, pour chaque unité k de telle façon que,

1. la longueur de la zone de sélection pour l'échantillon s^1 (resp. s^2) est égale à la probabilité d'inclusion voulue π_k^1 (resp. π_k^2),
2. le recouvrement des zones de sélection est minimale si l'on veut coordonner négativement, et maximale si l'on veut une coordination positive.

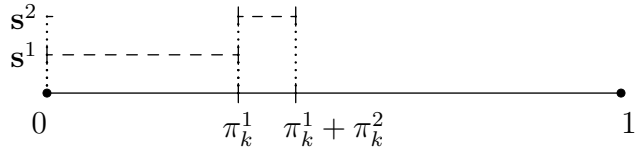
Pour une coordination positive, on définit par exemple la zone de sélection de k dans s^1 par $[0, \pi_k^1)$ et la zone de sélection de k dans s^2 par $[0, \pi_k^2)$. Ainsi, le segement $[0, 1]$ est divisé en trois intervalles représentés dans la Figure 1. Chacun de ces intervalles correspond à une valeur possible du vecteur aléatoire (I_k^1, I_k^2).

FIGURE 1 – coordination positive lorsque $\pi_k^2 \leq \pi_k^1$



Pour la coordination négative, les zones de sélection dans s^1 et s^2 sont par exemple définies par $[0, \pi_k^1)$ et $[\pi_k^1, \pi_k^1+\pi_k^2)$, si la somme des probabilités d'inclusion ne dépasse pas 1, comme dans la Figure 2. Dans le cas général, pour la coordination négative, le segment

FIGURE 2 – coordination négative lorsque $\pi_k^1 + \pi_k^2 \leq 1$

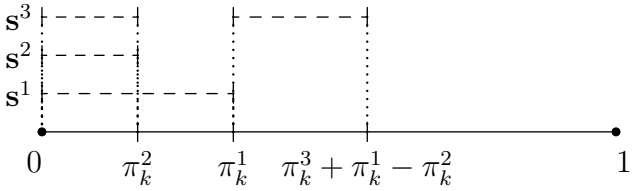


$[0, 1]$ est divisé en trois intervalles dont les limites sont données par 0, π_k^1, $(\pi_k^1+\pi_k^2)$modulo1 et 1.

Nous étendons cette idée à la sélection d'un nombre quelconque d'échantillons en définissant de manière récursive pour chaque nouvelle enquête une zone de sélection pour chaque unité. Le principe en est aisément compréhensible sur un example : disons que l'on a déjà sélectionné deux échantillons s^1 et s^2 avec coordination positive et que la situation est celle représentée sur la Figure 1. Supposons que l'on veuille sélectionner un troisième échantillon s^3 coordonné positivement avec s^2 et négativement avec s^1, la coordination avec s^2 ayant la priorité sur la coordination avec s^1, et que, par exemple, $\pi_k^3 > \pi_k^2$. Alors la zone de sélection pour s^3 contiendra la zone de sélection pour s^2 et un autre morceau de $[0, 1]$ de manière à respecter le mieux possible les règles de coordination. Dans ce cas, on choisira par exemple d'ajouter l'intervalle $[\pi_k^1, \pi_k^1 + \pi_k^3 - \pi_k^2)$ pour obtenir la zone de sélection dans s^3, de longueur totale π_k^3 représentée dans la Figure 3.

Plus formellement, au temps t et pour une chaque unité k, le segment $[0, 1]$ est divisé en une famille d'au plus $t + 1$ sous-intervalles. Chacun de ces sous-intervalles correspond à un historique possible de sélections de l'unité k (un échantillon longitudinal). L'ajout d'une nouvelle enquête s^{t+1} est obtenue en incluant dans la zone de sélection pour s^{t+1} des intervalles qui correspond aux échantillons longitudinaux les plus en adéquation avec les règles de coordination voulues, et usuellement en subdivisant l'un des intervalles en deux de manière à ce que la longueur totale de la zone de sélection soit exactement égale à la

FIGURE 3 – Coordination d'une troisième enquête



probabilité d'inclusion spécifiée π_k^{t+1}. Un ordre total sur les sous-intervalles est nécessaire pour réaliser cette opération. L'ordre que nous utilisons est obtenu par la fourniture d'un sens de coordination entre s^{t+1} et chacune des enquêtes passées et d'un ordre de priorité pour ces coordinations.

Les plans de sondages transversaux sont des plans de Poisson et donc sont de taille aléatoire. Si les coordinations sont toutes négatives et respectent l'ordre chronologique, le plan longitudinal pour chaque unité est systématique, ce qui est, pour l'espacement des sélections, le meilleur plan longitudinal (voir par exemple 8).

2 Developpements et utilisation dans la vie réelle

La méthode présentée dans la section 1 permet de sélectionner tous les types d'échantillons utilisés par l'OFS : les enquêtes ponctuelles, les panel mis à jour au bout de quelques années, et les panels rotatifs. Ces derniers sont en fait gérés comme des collections d'enquêtes. Par exemple, si le taux de rotation est de 20%, on commence par sélectionner cinq échantillons qui constituent le panel initial. Puis, chaque année, un nouveau cinquième est sélectionné, disjoint si possible de l'échantillon de l'année précédente, et quatre des sous-échantillons de l'année précédente sont mis à jour (i.e. on sélectionne quatre nouveaux sous-échantillons positivement coordonnés avec les sous-échantillons que l'on veut mettre à jour). De cette manière, le mouvement de la population (naissances, décès, et même fusions, reprises, dissolutions et scissions dans le cas des entreprises) est naturellement intégré à la mise à jour du panel.

Le système de coordination est utilisé à l'OFS depuis octobre 2009 pour les enquêtes auprès des entreprises et depuis novembre 2010 pour les enquêtes auprès de la population. Il a eu un effet modeste dans le cas des enquêtes auprès de entreprises. En effet, dans ce type d'enquêtes, la plupart des unités sont soit dans des "strates" recensées, soit ont de très faibles probabilités d'inclusion. Dans les deux cas, les tirages coordonnés sont proches de tirages indépendants. De plus, seules quelques enquêtes ont basculé vers ce nouveau système, ce qui fait que l'effet de la coordination sur le nombre de sélections répétées reste faible. Il a toutefois permis de mettre à jour simplement et correctement l'enquête sur la valeur ajoutée, qui est un panel rotatif dans une population dynamique, et dont le plan a été fortement modifié entre 2009 et 2010. On a de plus l'assurance d'avoir fait, de manière démontrable, tout ce que l'on pouvait pour éviter la sélection répétée et inutile de mêmes unités. Pour les enquêtes auprès de la population, le besoin de coordination est devenu impératif suite à l'introduction d'une enquête annuelle dite *structurelle*, qui remplace le recensement traditionnel et qui touche environ 7% de la population.

Jusqu'à présent, deux limitations de la méthode ont nécessité des adaptations. Elles sont toutes deux dues au fait que les plans transversaux sont poissoniens. La simpli-

cité de ce plan de sondage est la raison pour laquelle on est capable d'implémenter un système d'enquêtes coordonnées malléable, mais il ne répond pas parfaitement à toutes les attentes que l'on peut avoir pour la sélection ponctuelle d'échantillons. Un aspect qui est fréquemment évoqué est sa taille aléatoire et la perte de précision qui est censée l'accompagner. Comme on le voit dans la section 2.1, cela est en fait sans conséquence sur la précision espérée de la stratégie de sondage. Cependant, il existe un risque que la réalisation obtenue autorise une précision bien plus faible que celle espérée, et des adaptations doivent être faites pour limiter ce risque. Un autre aspect problématique est que dans le cas de l'échantillonnage de Poisson, par nature, la sélection d'une unité ne dépend pas de la sélection d'une autre. Cependant, pour les enquêtes auprès des entreprises comme pour les enquêtes auprès de la population, deux types d'unités sont à considérer : les entreprises et les établissements dans le premier cas, les ménages et les personnes dans le second. Dans le cas des enquêtes en population, la procédure usuelle à l'OFS avant la création du registre de population était de sélectionner des numéros de téléphone dans une base de numéros, puis de sélectionner une personne dans chaque ménage atteint par téléphone. Ce qui fait que l'on est habitués à ne sélectionner qu'une personne par ménage, et que les enquêtes et questionnaires sont prévues pour cet état de fait. On verra dans la section 2.2 comment nous avons sélectionné des échantillons avec une unité par ménage en utilisant le système de coordination.

2.1 Planification avec des échantillons poissoniens

Lorsque nous avons mis en place le système de coordination, avec ses échantillons poissoniens, l'inquiétude la plus fréquemment rencontrée concernait la perte de précision que l'on imaginait résulter de la taille aléatoire des échantillons. Et il est vrai que pour des variables bien corrélées avec les probabilités d'inclusion, un plan de taille fixe utilisé avec l'estimateur de Horvitz-Thompson (voir 6) a une meilleure précision qu'un plan de taille aléatoire également utilisé avec l'estimateur de Horvitz-Thompson. Cependant, dans la pratique, ce n'est jamais vraiment l'estimateur de Horvitz-Thompson qui est utilisé, mais plutôt l'estimateur de Hájek (5) ou un estimateur calé (voir 3). Le calage peut par exemple être employé pour "corriger" la non-réponse. Alors, si les probabilités d'inclusion sont une des variables de calage, la variabilité de la taille de l'échantillon a un effet négligeable sur la précision de la stratégie de sondage, comme on peut l'apprécier sur l'exemple suivant, largement applicable.

Soit une population de taille N, une variable d'intérêt y avec une variance ajustée S_y^2. On considère l'exemple très simple d'un sondage de Bernoulli accompagné de l'estimateur de Hájek, noté $\widehat{Y}_{Hj}(s)$, et un sondage aléatoire simple avec les mêmes probabilités d'inclusion $p = n/N$, accompagné de l'estimateur de Horvitz-Thompson, noté $\widehat{Y}_{HT}(s)$ (voir par exemple 10). Conditionnellement à la taille $n(s)$ réalisée pour l'échantillon bernoullien, et si $n(s) \neq 0$, on obtient

$$\text{var}\left(\widehat{Y}_{Hj}|n(s)\right) = N^2\left(1 - \frac{n(s)}{N}\right)\frac{S_y^2}{n(s)}, \text{ and } \text{E}\left(\widehat{Y}_{Hj}|n(s)\right) = Y, \tag{1}$$

où Y est le total de y dans la population. Pour continuer les calculs, il faut étendre la définition de $\widehat{Y}_{Hj}(s)$ à l'échantillon nul en choisissant une valeur pour $\widehat{Y}_{Hj}(\emptyset)$. Le biais de l'estimateur $\widehat{Y}_{Hj}(s)$ est alors égal à

$$\text{B}(\widehat{Y}_{Hj}) = (1-p)^N\left(\widehat{Y}_{Hj}(\emptyset) - Y\right), \tag{2}$$

et est donc de l'ordre de $\exp(-n)$ si N est assez grand. Dans la plupart des situations, $\exp(-n) \ll 1/n$ et ce biais peut être négligé. Pour simplifier les calculs, posons donc $\widehat{Y}_{Hj}(\emptyset) = Y$. Alors,

$$\text{var}(\widehat{Y}_{Hj}) = \text{var}\left\{\text{E}\left[\widehat{Y}_{Hj}|n(s)\right]\right\} + \text{E}\left\{\text{var}\left[\widehat{Y}_{Hj}|n(s)\right]\right\}, \tag{3}$$

se simplifie comme suit :

$$\begin{aligned}
\text{var}(\widehat{Y}_{Hj}) &= \text{E}\left\{\text{var}\left[\widehat{Y}_{Hj}|n(s)\right]\right\}, \\
&= \sum_{m=1}^{N} N^2\left(1-\frac{m}{N}\right)\frac{S_y^2}{m}\binom{N}{m}p^m(1-p)^{N-m}, \\
&= N^2 S_y^2\left[1-(1-p)^N\right]\left[\frac{1}{1-(1-p)^N}\sum_{m=1}^{N}\frac{1}{m}\binom{N}{m}p^m(1-p)^{N-m} - \frac{1}{N}\right].
\end{aligned}$$

Des approximations pour la somme dans le dernier terme sont données dans (11; 7; 4; 2). Elles permettent toutes de conclure que

$$\text{var}(\widehat{Y}_{Hj}) = \text{var}(\widehat{Y}_{HT}) + \mathcal{O}\left(n^{-2}\right). \tag{4}$$

Le vrai problème lié à la taille aléatoire est que dans certains petits domaines, l'échantillon sélectionné peut avoir une taille inférieure à ce qui est jugé acceptable avant même la phase de non-réponse. La variance conditionnelle à la taille obtenue sera alors élevée dans ces domaines. Afin de limiter le risque que cela arrivem on peut choisir de modifier l'allocation initialement prévue de manière à augmenter la taille espérée dans les domaines où le risque est le plus grand. Quand le taux de sondage est constant dans des domaines prédéterminés, on peut facilement calculer une approximation de la probabilité d'obtenir un échantillon plus petit qu'une limite donnée $P(n(s) < n_{min})$ en fonction du taux de sondage. On peut alors inverser cette fonction et déterminer un taux de sondage tel que $P(n(s) < n_{min}) \leq \alpha$ où α est le risque accepté d'obtenir un échantillon trop petit. Ceci nous conduit à modifier notre algorithme d'allocation et à accepter un résultat sub-optimal quant à l'estimation du total sur la population entière. Lorsqu'il y a un grand nombre de petits domaines, il peut devenir très couteux en terme de précision ou de taille d'échantillon moyenne d'imposer un paramètre α assez petit pour que la probabilité de n'avoir aucun domaine avec un échantillon trop petit reste élevée. Bien que ce soit un réel problème, il est en fait inhérent à toutes les enquêtes avec non-réponse lorsque l'on considère que celle ci est une phase de sondage bernoullienne ou multinomiale supplémentaire. Le coût de contrôle du risque lié à la non-réponse est alors généralement bien plus grand que celui lié au tirage poissonien.

2.2 Ajout de contraintes sur les sélections simultannées d'unités

Afin de limiter la charge d'enquête au sein des ménages, et dans l'espoir d'obtenir des meilleurs taux de réponse, ou bien pour éviter de collecter inutilement plusieurs fois la même information, il est courant de ne pas sélectionner plus d'une unité par ménage. La plupart des enquêtes de l'OFS étaient jusqu'à maintenant conçues de cette façon. Avant 2010, la seule base de sondage exploitable pour ces enquêtes était un répertoire de numéros de téléphones fixes livré par les opérateurs de télécommunication. Après avoir sélectionné

un échantillon du numéros et listé les unités du ménage contacté, un individu par ménage était alors retenu. La nouvelle base de sondage est un registre de population obtenu en assemblant les registres tenus par les communes et l'état fédéral (toutes les personnes résidant en Suisse ont obligation de s'enregistrer auprès des autorités de leur commune de résidence, de l'office des migrations ou d'un office pour les travailleurs étrangers). Ces registres ont pour but premier d'enregistrer les individus et non les ménages. Il a semblé adequat de construire une base de sondage à partir de ces registres au niveau des personnes en utilisant le numéro d'assurance sociale comme identifiant. Toutefois, un identifiant de ménage est également disponible pour quasiment toute la population, et deviendra obligatoire en 2013.

Bien que des échantillons ne comportant qu'une unité par ménage ne puissent être obtenus directement avec notre système de tirages coordonnés, une procédure en deux phases permet de sélectionner de tels échantillons. Les calculs pour réaliser ces tirages correctement sont relativement simples quand les probabilités d'inclusion voulues dans chaque ménage sont nulles pour les individus hors population cible et égales entre elles pour toutes les autres personnes. Elles peuvent devenir complexes autrement. Le problème devient également complexe si l'on veut sélectionner un échantillon à plusieurs dates. C'est le cas d'une enquête menée à l'OFS : le registre de population est mis à jour tous les trois mois, et une partie de l'échantillon doit être sélectionnée dans le registre de décembre, une autre dans le registre de mars et une dernière dans le registre de juin, sans que l'on ne resélectionne un membre d'un ménage déjà atteint. L'évolution de la structure des ménages va alors parfois rendre impossible un calcul explicite des probabilités de tirage de première phase correctes dans les ménages dont la composition a changé.

Nous avons employé la procédure suivante :

1. dans un premier temps, un échantillon est sélectionné via le système de coordination avec des probabilités d'inclusion que l'on doit calculer de telle manière que les probabilités d'inclusion finales soient correctes,
2. puis, dans les ménages où plusieurs unités ont été sélectionnées, l'une de ces unités, choisie au hasard, est conservée, et les autres sont éliminées de l'échantillon,
3. enfin, dans le cas des enquêtes tirées en plusieurs vagues, les sélections dans les ménages qui avaient été déjà précédemment enquêtés sont filtrées.

Avec cette procédure, quelques sélections sont enregistrées à tort dans le système de coordination, puisque quelques unités sont sélectionnées dans la première phase, mais ne seront pas réellement interrogées. Cependant, les probabilités d'inclusion enregistrées sont proches des probabilités d'inclusion réelles, exception faite des très grands ménages (pour lesquels il arrive même que la probabilité de sélection en première phase vaille 1). La procédure permet au final d'avoir une bonne coordination entre les enquêtes, sauf pour ces très grands ménages.

Lorsque la troisième étape n'est pas nécessaire, la probabilité d'inclusion p_k pour la première phase de sélection d'une unité k dans un ménage $\mathcal{M}$ de taille m_k peut être calculée en fonction de la probabilité d'inclusion finale π_k en notant que

$$\pi_k = \frac{1}{m_k}\left[1-(1-p_k)^{m_k}\right]. \tag{5}$$

On obtient alors $p_k = 1-(1-m_k\pi_k)^{\frac{1}{m_k}}$. Dans le cas général de probabilités d'inclusion

inégales, il faudrait résoudre le système d'équations (6) en les p_i, $i \in \mathcal{M}$:

$$\pi_k = p_k \sum_{n=0}^{m_k-1} \sum_{i_1 \neq \cdots \neq i_n \neq k \in \mathcal{M}} \frac{1}{n+1} \prod_{j \in 1,\ldots,n} p_{i_j} \prod_{\ell \notin i_1,\ldots,i_n,k} (1-p_\ell)\,,\ k \in \mathcal{M}. \tag{6}$$

Malheureusement il n'y a généralement pas de solution explicite à ces équations. Des procédures de résolution numérique semblent fonctionner sur quelques exemples choisis, mais il n'est pas certain qu'elles puissent être utilisées raisonnablement sur une population de près de huit millions d'unités.

Si l'on doit également utiliser la troisième étape, une équation similaire à 6 est relativement facile à trouver. Mais elle dépend des probabilités de sélection des unités de ménages aux vagues précédentes de l'enquête. Si la composition du ménage à changé, il est possible que ces probabilités ne soient pas les mêmes pour toutes les personnes du ménage. Dans ce cas, nous n'avons pas de solution explicite pour calculer les probabilités de sélection correctes. Heureusement, la structure des ménages n'évolue pas très rapidement. En effet, si l'on compare le registre de décembre 2010 et celui de mars 2010, on constate que seuls 2% des ménages ont changé. Pour les 98% restants, nous avons pu réaliser un tirage qui respectait exactement les probabilités d'inclusion planifiées. Pour les 2% problèmatiques, une procédure simplifiée a été utilisée avec des probabilités d'inclusion très légèrement différentes de celles prévues, et calculables a posteriori.

3 Conclusion

Le système d'échantillonnage coordonné que nous utilisons a permis jusqu'à maintenant de sélectionner plus de 930'000 unités pour participer à 21 enquêtes auprès de la population. La simplicité du plan transversal poissonien, malgrè ses défauts, nous permet de conserver un volume de données et de calculs raisonnables. Il nous a aussi permis de faire face à des besoins qui n'avaient pas été exprimés avant la mise en place du système : la nécessité de ne sélectionner qu'une unité par ménage, y compris dans des enquêtes sélectionnées en plusieurs vagues. La solution proposée à ce problèeme conduit à une coordination dégradée pour les grands ménages, ce qui fait que 10 unités ont déjà été sélectionnées deux fois dans des très grands ménages. Ce chiffre est à rapprocher des 42'200 sélections doubles de mêmes unités et 800 sélections triples que l'on aurait eues si les enquêtes avaient été tirées de manière indépendante.

Références

[1] Brewer, K. R. W., Early, L. J., and Joyce, S. F. Selecting several samples from a single population. *Australian Journal of Statistics 3* (1972), 231–239.

[2] David, F., and Johnson, N. Reciprocal bernoulli and poisson variables. *Metron 18* (1956), 77–81.

[3] Deville, J.-C., and Särndal, C.-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association 87* (1992), 376–382.

[4] Grab, E., and Savage, I. Tables of the expected value of 1/x for positive bernoulli and poisson variables. *Journal of the American Statistical Association 49* (1954), 169–177.

[5] Hájek, J. Discussion of an essay on the logical foundations of survey sampling, part on by d. basu. In *Foundations of Statistical Inference* (Toronto, Canada, 1971), V. P. Godambe and D. A. Sprott, Eds., Holt, Rinehart, Winston, p. 326.

[6] Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47* (1952), 663–685.

[7] Marciniak, E., and Wesolowski, J. Asymptotic eulerian expansions for binomial and negative binomial reciprocals. *Proceedings of the American Mathematical Society 127*, 11 (1999), 3329–3338.

[8] Nedyalkova, D., Qualité, L., and Tillé, Y. Tirages coordonnés d'échantillons à entropie maximale. Technical report, University of Neuchâtel, 2009.

[9] Qualité, L. *Unequal probability sampling and repeated surveys.* Thèse de doctorat, Université de Neuchâtel, Neuchâtel, Suisse, 2009.

[10] Särndal, C.-E., Swensson, B., and Wretman, J. H. *Model Assisted Survey Sampling.* Springer, New York, 1992.

[11] Thionet, P. Sur le moment d'ordre (-1) de la distribution tronquée. application à l'échantillonnage de hájek. *Publ. Inst. Statist. Univ. Paris 12 31 :827* (1963), 93–102.