

PLANS DE SONDAGE AUTO-PONDÉRÉS À PLUSIEURS DEGRÉS

Peter Slock, Denis Luminet et Camille Vanderhoeft¹

RÉSUMÉ

À l'Institut National belge de Statistique, la plupart des enquêtes auprès de ménages ou d'individus appliquent un plan de sondage à deux degrés. Nous exposons la motivation du sondage à plusieurs degrés et indiquons quelques choix possibles pour les unités primaires (UP). Les avantages et désavantages de différents plans de sondage pour les UP sont décrits et des exemples donnés. Le tirage des UP peut se faire avec ou sans remise; ces deux méthodes sont comparées, et les implications sur les degrés suivants considérées.

Un plan de sondage est ensuite choisi pour les unités secondaires (US). Nous prouvons une propriété d'auto-pondération pour un plan à 2 degrés souvent utilisé. La généralisation à plus de 2 degrés est brièvement envisagée.

1. DES PLANS À PLUSIEURS DEGRÉS – POURQUOI ?

1.1 Interviews à domicile

L'Institut National belge de Statistique (INS) confie souvent l'exécution d'une enquête à des enquêteurs, qui vont interviewer des individus ou des ménages chez eux. Cette organisation décentralisée a des répercussions importantes sur le plan de sondage des enquêtes.

Nous nous placerons dans le cas où les unités sont des ménages; cependant, les idées restent valables s'il s'agit d'autres types d'unités, par exemple des individus. Il s'agit d'interroger un échantillon de ménages (supposés) représentatifs de la population d'intérêt. En pratique, l'échantillon sera tiré dans une base de sondage².

L'échantillon sélectionné est réparti par l'INS entre des enquêteurs, lesquels sont chargés de rendre visite aux (membres des) ménages sélectionnés et les interroger en face-à-face. Le principal critère de répartition vise à allouer à un même enquêteur des ménages géographiquement proches, par exemple domiciliés dans la même commune (NUTS5 dans la nomenclature Eurostat) ou partie de commune. Notons que l'INS procède toujours à une stratification, mais les strates correspondent en général à des zones plus étendues comme les régions (NUTS1) ou les provinces (NUTS2).

En Belgique existent trois niveaux de subdivision des 589 communes: chacune est formée de lettres statistiques, elles-mêmes divisées en sections, lesquelles se composent de secteurs³. La localisation de chaque ménage figure au RNP (Registre National de la Population)⁴. La taille (exprimée en nombre d'individus ou en nombre de ménages) de ces lettres, sections et secteurs varie énormément, notamment selon le caractère urbain ou rural de la zone. Par conséquent, un tirage à un seul degré (tel un sondage aléatoire simple ou un sondage systématique) produirait vraisemblablement un échantillon dont certaines zones géographiques sont absentes, et où d'autres seraient représentées par de nombreux ménages.

¹ Peter.slock@statbel.mineco.fgov.be, Denis.luminet@statbel.mineco.fgov.be et Camille.vanderhoeft@statbel.mineco.fgov.be; Bureau de Méthodologie et de Coordination, Institut National de Statistique, 44 Rue de Louvain, B-1000 Bruxelles, Belgique.

² Par la suite, nous supposons que base de sondage et population d'intérêt coïncident.

³ La Belgique compte 2644 lettres statistiques (habitées), 6217 sections, et 19024 secteurs.

⁴ "Registre National (de la Population)": le RNP est une base de données officielle qui reprend l'ensemble de la population (individus et composition des ménages) résidant en Belgique, quelle qu'en soit la nationalité, et fournit des renseignements utiles comme l'adresse ou la date de naissance. On consultera <http://www.registrenational.fgov.be>.

1.2 Répartition géographique des enquêteurs

Au sein d'une strate h^5 , on attribue à chaque enquêteur un *groupe* formé d'un même nombre G de ménages. Afin de répartir les ménages uniformément entre les enquêteurs tout en garantissant que chacun est chargé d'unités d'une même zone géographique, il importe que chaque zone représentée dans l'échantillon le soit par au moins G unités. Il nous faut tenir compte de cette exigence dans le plan de sondage, avant de tirer les unités, par exemple en tirant d'abord des zones géographiques (UP), puis en y sélectionnant des ménages (US): *plan à deux degrés*.

Les zones devraient donc compter au moins G unités. Cette contrainte ne sera pas toujours satisfaite si les zones sont petites (des sections, a fortiori des secteurs). Des remèdes à ce problème seront exposés en Section 2.2.

Notons que G , représentant le nombre d'unités pouvant être interrogées par un enquêteur au cours d'une période limitée (mettons, 1 mois) est généralement assez bas (entre 20 et 40). Bien entendu, un enquêteur capable d'interviewer davantage d'unités peut se voir attribuer plusieurs groupes.

Dans chaque strate, nous tirerons un certain nombre d'UP (premier degré), et ensuite un nombre constant G d'unités secondaires (et finales, dans ce cas) pour chaque tirage d'une UP (second degré). L'intérêt de choisir une petite valeur pour G est que, avec la partition de la base de sondage par zone géographique, la plupart des «candidates UP» contiendront au moins G unités, permettant (le cas échéant) le tirage de G US.

D'autres motifs que la proximité géographique peuvent justifier un plan à plusieurs degrés. En 2004, l'INS a tiré (dans le RNP) un échantillon d'individus en vue d'une étude sur la prévalence de l'hépatite. Les UP se composaient alors d'individus aux mêmes caractéristiques selon des critères géographiques (lettre statistique), mais aussi démographiques ((catégorie d') âge et sexe). Au second degré, des groupes d'effectif constant ont été tirés, et les individus contactés en plusieurs vagues espacées dans le temps. La construction de ces UP ainsi que la procédure de tirage et de contact est détaillée dans [Slock & Vanderhoeft, 2004]; cette enquête était organisée par voie postale, sans recourir à des enquêteurs.

Observons enfin que, dans d'autres situations, un plan à plus de deux degrés se justifie (voir section 3.2).

2. DES PLANS À PLUSIEURS DEGRÉS – COMMENT ?

2.1 Taille des groupes

Le plan de sondage postule un nombre constant G d'US par groupe. En désignant par n l'effectif désiré (en termes d'unités secondaires), $n/G = n_i$ ⁶ tirages d'UP seront requis, en supposant que les ensembles d'US tirés pour les n_i tirages d'UP sont disjoints et contiennent G unités distinctes chacun. Nous ferons l'hypothèse que n_i est entier, autrement dit que n est multiple de la taille de groupe G ; ceci peut être assuré par l'allocation de l'échantillon entre les strates. En vue d'obtenir n unités secondaires au total, un groupe de $n_i = G$ US sera tiré (à probabilités égales, par exemple par Sondage Aléatoire Simple «SAS») pour chaque tirage d'UP; par exemple, si une UP i a été tirée 3 fois au premier degré ($m_i = 3$), nous tirerons au second degré 3 groupes de G US. Les détails seront donnés en section 3.

2.2 Base de sondage pour les UP

Afin de constituer une base de sondage pour les UP, le plus simple est de partir d'un niveau géographique (commune, lettre, section ou secteur). Néanmoins, si certaines UP comptent moins de G US, il faudra les fusionner dans la base de sondage (avant leur tirage).

Un exemple issu de l'EFT (Enquête sur les Forces de Travail), où $G = 23$ et les UP sont en principe des sections:

Au sein d'une lettre statistique,

- Si pas de «petite» section (moins de 25⁷ ménages): OK
- Si une seule «petite» section, la combiner avec la plus petite des grandes sections de la lettre: OK
- Si plusieurs «petites» sections, les combiner...

⁵ Nous ne considérerons dorénavant qu'une strate, et omettrons l'indice h . Tout le raisonnement pour une strate reste valable lorsqu'il y en a plusieurs, car l'échantillonnage se fait indépendamment.

⁶ L'indice i indique qu'il s'agit de l'effectif du premier degré de tirage.

⁷ 25 = $G + 2$, où les 2 US (ménages) supplémentaires constituent une marge nous préservant contre une éventuelle diminution du nombre d'US dans cette section entre le tirage des UP et des US.

- c1) si cette combinaison compte au moins 25 ménages, OK
 c2) si cette combinaison compte moins de 25 ménages, combiner (la combinaison) avec la plus petite des grandes sections de la lettre, OK⁸

On peut se référer à [Luminet, 2004].

2.3 Tirage des UP

Les UP peuvent être tirées de diverses manières:

- Avec remplacement (AR) ou sans remplacement (SR)
- Probabilités égales ou inégales (p.ex. probabilité proportionnelle à la taille)
- Nombre de tirages d'UP par strate: fixe ou non
- Tirage systématique ou autre (p. ex. Lahiri).

Vu leurs tailles x_i très différentes, nous tirerons les UP i avec probabilité proportionnelle à la taille (AR ou SR; nous verrons que l'échantillonnage SR n'est pas toujours possible), comme expliqué dans [SSW, 1992] (p88-99). Ceci implique que les zones les plus peuplées seront en général sélectionnées (et parfois plusieurs fois dans le cas AR), tandis que les plus petites (en particulier, celles de moins de G ménages) ne le seront que rarement. De nombreux algorithmes existent pour tirer un échantillon avec probabilité proportionnelle à la taille (AR ou, ce qui est en général plus compliqué, SR); nous présentons ici les propriétés générales, indépendantes de l'algorithme. Notons que n'importe quelle variable quantitative (nombre de rues, nombre de logements...) pourrait être prise comme «taille» des UP, mais en règle générale c'est le nombre d'US (ici, de ménages) qui détermine la taille. Sous certaines hypothèses, le plan de sondage à deux degrés sera alors auto-pondéré (par strate). Nous supposons un nombre fixe n_i de tirages d'UP, ce qui facilite l'échantillonnage et l'estimation, et donne un meilleur contrôle sur l'effectif des US (donc sur le budget!). Soit n'_i ($\leq n_i$) le nombre d'UP distinctes sélectionnées, ce nombre n'est pas fixe en cas de tirage AR (voir 2.3.2).

2.3.1 π PT (SR) ou PPT (AR) au premier degré?

Il peut être avantageux, voire indispensable, de tirer PPT (probabilité proportionnelle à la taille, **avec** remplacement).

Un exemple: pour l'Enquête sur le Budget des Ménages (EBM), les 3 régions (NUTS1) constituent les strates, les zones géographiques (candidates UP) sont les lettres statistiques. On tire annuellement 84 groupes dans la région de Bruxelles (NUTS1), laquelle se compose de 27 lettres: le tirage AR est par conséquent impératif.

Un autre atout de PPT est de permettre qu'une UP de grande taille soit sélectionnée plus d'une fois avec une probabilité élevée; on tirera G US par *tirage* de l'UP. Par contre, avec π PT (probabilité proportionnelle à la taille, **sans** remplacement), les grandes UP i (taille $x_i \geq X/n_i$, où $X \equiv \sum_{i=1}^{N_i} x_i$ avec N_i le nombre d'UP dans la base de sondage) feront nécessairement partie de l'échantillon au premier degré ($\pi_{1,i} = 1$), mais on n'y tirera que G US au second degré (avec un algorithme de SAS), diminuant la probabilité de sélection $\pi_{(ij)}$ d'une US j appartenant à l'UP i , comme illustré par l'exemple⁹:

Degré 1: $n_i = 50$; $x_1 = 10^5$, $x_2 = 20, \dots, x_{N_i} = 100$; $X = 10^6$;

Degré 2: $G = 20$

$$\pi_{1,1} = \min\left\{\frac{n_i \cdot x_1}{X}, 1\right\} = \min\{5, 1\} = 1; \pi_{j1} = \frac{20}{10^5} = 2 \cdot 10^{-4}; \pi_{(1j)} = \pi_{1,1} \cdot \pi_{j1} = 1 \cdot (2 \cdot 10^{-4}) = 2 \cdot 10^{-4}$$

$$\pi_{1,2} = \min\left\{\frac{(n_i - 1) \cdot x_2}{X - x_1}, 1\right\} = 1,089 \cdot 10^{-3}; \pi_{j2} = \frac{20}{20} = 1; \pi_{(2j)} = \pi_{1,2} \cdot \pi_{j2} = (1,089 \cdot 10^{-3}) \cdot 1 = 1,089 \cdot 10^{-3} \neq \pi_{(1j)}$$

⁸ OK...sauf si la lettre elle-même compte moins de 25 ménages, cas exceptionnel.

⁹ Nous supposons, pour la simplicité, que l'UP 1 est la seule UP i avec $x_i \geq X/n_i$ et donc la seule UP à traiter à part (ramener $\frac{n_i \cdot x_i}{X}$ à 1) pour calculer les $\pi_{1,i}$ (la formule peut être généralisée facilement dans le cas où il y a

plusieurs UP dans ce cas), et que la réduction de $\pi_{1,i}$ à 1 n'entraîne pas que la valeur de $\frac{(n_i - 1) \cdot x_i}{X - x_i}$ excède 1 pour d'autres UP i . Dans un tel cas l'application itérative des formules de l'exemple résoudra ce problème.

Ainsi, le π PT (SR) accorde à un ménage de la grande UP 1 une probabilité 5,4 fois moindre qu'à un ménage de la petite UP 2: comparons les probabilités finales $\pi_{(1,j)}$ et $\pi_{(2,j)}$. Le plan n'est pas auto-pondéré!

Le PPT peut être implémenté via un algorithme systématique, en prenant $a = X/n_i$ comme pas (fractionnaire) et x_i comme largeur de l'UP (candidate) i dans la base de sondage. Dès que $x_i \geq 2 \cdot a$ pour une UP i , cette unité sera prise au moins deux fois et un plan π PT (SR) est impossible à réaliser par cet algorithme.

2.3.2 Un effectif fixe

En général, un nombre fixe de tirages d'UP n_i est souhaité au premier degré. Dans le cas SR, n_i égale le nombre n'_i d'UP distinctes sélectionnées (pour autant qu'il y ait n_i UP distinctes!), et $\sum_{i=1}^{N_i} \pi_i = n_i$.

En revanche, dans le cas AR, n'_i peut être strictement inférieur à n_i . On a cependant $\sum_{i=1}^{N_i} M_i = n_i$, où pour chaque i , la variable aléatoire discrète M_i prend pour valeur $m_i (\leq n_i)$ le nombre de tirages de l'UP i (n'_i termes de la somme sont différents de zéro), autrement dit la multiplicité de sélection de i .

Au deuxième degré, on tire G US par tirage de l'UP i . Si l'UP i a été tirée m_i fois ($m_i \geq 1$), on y prendra $G m_i$ US lors du second degré de tirage.

L'effectif total (en termes d'unités secondaires = unités finales) est donc $\sum_{i=1}^{N_i} m_i \cdot G$, traditionnellement noté n .

3. L'AUTO-PONDERATION

3.1 Plans à deux degrés

Considérons à présent les US (c'est-à-dire les unités finales, ici les ménages). On désire souvent leur donner des probabilités d'inclusion égales, à la fois pour simplifier l'extrapolation et pour des raisons d'équité¹⁰. On suppose un plan SAS au deuxième degré, les $\pi_{j|M_i=m_i}$ ci-dessous ne dépendent donc pas de l'US j appartenant à l'UP i .

Pour vérifier cette propriété, penchons-nous sur les probabilités finales¹¹ d'inclusion (du premier ordre) $\pi_{(ij)}$. Notons $N_{i,i}$ et $s_{i,i}$ respectivement le nombre d'US (candidates) et l'échantillon sélectionné dans l'UP i .

$$\begin{aligned} \pi_{j|M_i=m_i} &= P(j \in s_{i,i} \mid M_i = m_i) = \frac{m_i \cdot G}{N_{i,i}} \text{ (si } \frac{m_i \cdot G}{N_{i,i}} \leq 1, \text{ voir condition (2) ci-dessous; sinon } \pi_{j|M_i=m_i} = 1) \\ \pi_{(ij)} &= \sum_{m_i} \pi_{j|M_i=m_i} \cdot P(M_i = m_i) = \sum_{m_i=0}^{n_i} \frac{G}{N_{i,i}} m_i \cdot P(M_i = m_i) = \frac{G}{N_{i,i}} \cdot \sum_{m_i=0}^{n_i} m_i \cdot P(M_i = m_i) = \frac{G}{N_{i,i}} \cdot E(M_i) \end{aligned} \quad (1.1)$$

Ces sommations parcourent les valeurs m_i de M_i ($m_i \leq n_i$), et la dernière n'est autre que l'espérance mathématique de la variable aléatoire discrète M_i . Nous ne ferons pas d'hypothèse particulière sur la distribution des M_i , ainsi la description qui suit ne dépend pas de l'algorithme de tirage utilisé pour le plan PPT.

Si la probabilité de sélection de l'UP i vaut p_i (avec $\sum_{i=1}^{N_i} p_i = n_i$) à chacun des n_i tirages, alors

$$E(M_i) = n_i \cdot p_i. \quad (1.2)$$

Vu que $p_i = x_i/X$ (PPT), nous obtenons

$$E(M_i) = n_i p_i = \frac{n_i}{X} x_i. \quad (1.3)$$

¹⁰ Conformément à un article de la loi statistique belge: *Les personnes appelées à répondre sont désignées [...] suivant une méthode impliquant, pour toutes les personnes comprises dans une même catégorie, la même probabilité d'être astreintes à déclarer.*

¹¹ 'finale' en considérant les deux degrés.

Ce résultat apparaît dans [Cochran, 1977] (formule (9A.9), p253), avec d'autres notations et sous conditions plus sévères. La propriété (1.3) reste valable pour des algorithmes séquentiels de liste comme le tirage systématique généralisé (voir [Vanderhoeft, 2004]): $E(M_i) = \frac{x_i}{a} = \frac{n_i}{X} x_i$ où $a = \frac{X}{n_i}$ est le pas dans la liste dans laquelle chaque UP est représentée par un intervalle de longueur x_i .

Introduisant (1.3) dans $\pi_{(ij)} = \frac{G}{N_{H,i}} \cdot E(M_i)$ et supposant que $x_i = N_{H,i}$ (donc $X = \sum_{i=1}^{N_H} N_{H,i} \equiv N$), on obtient

$$\pi_{(ij)} = \frac{G}{x_i} \cdot n_i \frac{x_i}{X} = \frac{n_i G}{X} \equiv \frac{n_i n_H}{X} = \frac{n}{X} = \frac{n}{N}, \quad (1.4)$$

montrant que le plan à deux degrés est auto-pondéré (par strate, rappelons-le), pour autant que:

- (1) la taille x_i de chaque UP i soit exactement le nombre d'unités secondaires $N_{H,i}$ au moment du tirage de ces US; ceci n'est pas toujours le cas en pratique, puisque les UP sont souvent sélectionnées longtemps à l'avance pour permettre la désignation d'enquêteurs, alors que les ménages ne le seront que juste avant le travail de terrain, afin de minimiser les erreurs de (sous- ou sur-)couverture.
- (2) la condition $N_{H,i} \geq m_i \cdot G$ soit remplie. En cas de PPT totalement indépendant (n_i tirages indépendants, $M_i \sim \text{Bin}(n_i, p_i)$), il peut arriver (même si la probabilité est faible) que $m_i = n_i$. Ceci est impossible¹² avec un algorithme séquentiel systématique, pour lequel m_i ne peut prendre que les valeurs $\left\lfloor \frac{x_i}{a} \right\rfloor$ ou $\left\lceil \frac{x_i}{a} \right\rceil$: $\left\lceil \frac{x_i}{a} \right\rceil$ est alors une borne supérieure pour m_i , et en pratique (taux de sondage $n/N \leq 0,5$) il suffira de vérifier que $x_i \geq G$ si la condition (1) ($x_i = N_{H,i}$) est satisfaite. Notre logiciel d'échantillonnage (pour PPT et π PPT) vérifie si cette condition de taille est remplie pour l'UP (candidate).

3.2 Généralisation à 3 degrés ou plus

Dans certaines circonstances, un plan à 3 degrés (ou même davantage) est indiqué. Reprenons le cas de l'EFT: une fois tirées les UP (essentiellement des sections), on pourrait tirer comme unités secondaires dans l'UP des catégories de ménages (p.ex. selon le nombre de membres du ménage), les ménages n'étant sélectionnés que comme unités tertiaires et finales, et on peut montrer que (moyennant certaines conditions!) les unités finales seront tirées à probabilités égales. Ce procédé avait été envisagé, mais nous avons finalement donné la préférence à un plan à deux degrés, avec tirage des ménages (= US) systématiquement selon la taille (donc par catégorie de ménages), ce qui n'impose pas de taille minimum aux unités primaires.

RÉFÉRENCES

- [Cochran, 1977]: Cochran, W.G., *Sampling Techniques* (3rd edition, 1977), New York: Wiley.
- [Luminet, 2004]: Luminet, D., «Echantillonnage de l'Enquête sur les Forces de Travail 2005», rapport interne, Statistics Belgium, Bruxelles, Belgique.
- [Slock & Vanderhoeft, 2004]: Slock, P., Vanderhoeft, C., «An incremental 2-stage sampling plan for a Flemish hepatitis prevalence study: accumulation of respondents over successive waves», *Proceedings of the "European Conference on Quality and Methodology in Official Statistics (Q2004)"*, October 2004; disponible sur http://statbel.fgov.be/studies/home_en.asp.
- [SSW, 1992]: Särndal, C.-E., Swensson, B., Wretman, J., *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- [Vanderhoeft, 2004]: Vanderhoeft, C., «Some notes on Random Number Generators, Sampling and Algorithms», rapport interne, Statistics Belgium, Bruxelles, Belgique.

¹² Sauf si $\left\lceil \frac{x_i}{a} \right\rceil = n_i$.