La méthode du partage des poids

Cyril Favre-Martinoz & Emmanuel Gros

***Résumé* — L'objectif de cette note méthodologique est de décrire de façon rapide le partage des poids et les contextes d'application de la méthode. Le positionnement du partage des poids par rapport aux autres traitements post-collecte est également abordé.**

I. Cadre théorique et contextes d'application

En statistique d'enquête, on est parfois confronté à des situations dans lesquelles l'unité d'observation soit diffère de l'unité d'échantillonnage, soit peut être enquêtée par suite du tirage de différentes unités d'échantillonnage. Ce cadre est celui de toute enquête pour laquelle l'échantillon d'unités finales voulues résulte [1] de la sélection d'un ou plusieurs échantillons d'unités intermédiaires reliées aux unités finales.

Dans de telles situations, la méthode du partage des poids constitue la procédure de calcul de pondérations de référence et conduit à un estimateur sans biais, sous l'unique condition que toute unité finale soit reliée à au moins une unité intermédiaire.

En pratique, la méthode du partage des poids est essentiellement utilisée dans trois contextes spécifiques :

- lorsque l'échantillon d'unités finales a été sélectionné par ***sondage indirect (cf. II-A)*** : il s'agit du cadre d'application le plus naturel de la méthode, qui a précisément été développée dans ce contexte ;
- en cas de ***bases de sondage multiples (cf. II-B)***, *i.e.* lorsque l'échantillon d'unités finales résulte de la concaténation de plusieurs échantillons sélectionnés dans plusieurs bases de sondages non-disjointes ;
- lors de l'exploitation ***d'échantillons totalement ou partiellement panélisés (cf. II-C)*** : exploitation transversale d'un panel, exploitation transversale ou longitudinale d'un échantillon rotatif.

II. Description de la méthode

A. *Sondage indirect*

Afin d'illustrer la méthode de sondage indirect, on s'intéresse à l'exemple classique « parents-enfants ». On souhaite produire des estimations sur une population d'enfants (population d'intérêt), sachant qu'on ne dispose que d'une base de sondage constituée de parents. On sélectionne alors de façon probabiliste un échantillon de parents et on interroge tous les enfants des parents interrogés. Cette situation est illustrée par la figure 1. Les liens entre les deux bases correspondent aux liens de filiation.

1. Soit parce qu'on ne dispose pas d'une base de sondage permettant de sélectionner directement un échantillon d'unité finales, soit du fait d'un processus d'échantillonnage complexe (panélisation par exemple).

Plus généralement, le sondage indirect consiste à sélectionner un échantillon s^A dans une population U^A de taille N^A afin de produire une estimation pour une population cible U^B de taille N^B, en s'appuyant pour cela sur les liens qui existent entre les deux populations. On note Ω^B l'ensemble des unités échantillonnées indirectement dans la population U^B ayant au moins un lien avec une des unités échantillonnées. Pour estimer le total Y^B à partir des valeurs de y_i mesurées à partir de l'ensemble Ω^B, il est courant d'utiliser un estimateur de la forme :

$$\hat{Y}^B = \sum_{i \in \Omega^B} w_i y_i$$

où w_i est le poids d'estimation de l'unité i de Ω^B. Une façon classique de définir un jeu de poids produisant une estimation sans biais est de choisir le poids comme l'inverse de la probabilité d'inclusion. Malheureusement, dans le cas du sondage indirect, la détermination des probabilités d'inclusion des unités appartenant à l'échantillon Ω^B est souvent très complexe, voire impossible. En général, on ne dispose que du poids de sondage $d_j = 1/\pi_j$ de l'unité j appartenant à l'échantillon s^A, défini comme l'inverse de la probabilité d'inclusion π_j. Pour produire un estimateur sans biais, il faut alors avoir recours au partage des poids en définissant un système de liens L_{ij} entre deux unités i et j appartenant respectivement aux populations U^B et U^A. Ainsi s'il existe un lien entre l'unité i et l'unité j, L_{ij} sera égal à 1 et 0 sinon. L'estimateur résultant de cette méthode s'écrit alors :

$$\hat{Y}^B = \sum_{i \in \Omega^B} w_i y_i$$

où $w_i = \sum_{j \in s^A} d_j \frac{L_{ij}}{L_i}$ et $L_i = \sum_{j=1}^{N^A} L_{ij}$.

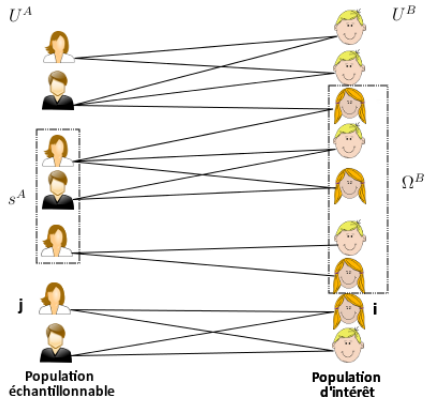

Fig. 1. Populations de parents et d'enfants avec des liens entre les deux

On notera que le nombre de liens L_i correspond au nombre total de liens de l'unité i avec la base de sondage initiale. Ce nombre total de liens L_i permet d'ajuster le poids associé

à l'unité i : plus le nombre de liens L_i est élevé, plus l'unité a une probabilité élevée d'être sélectionnée, il est par conséquent logique que son poids issu de la méthode de partage des poids w_i diminue. Il est important de noter que ce nombre de liens L_i est comptabilisé sur l'ensemble de la population U^A (et non pas seulement l'échantillon s^A). Ainsi dans le cas particulier de la sélection indirecte des enfants *via* leurs parents, il est nécessaire que chaque enfant appartenant à l'échantillon Ω^B soit capable d'indiquer qui sont ses parents dans l'ensemble de la population U^A.

Enfin, même si les poids issus de la méthode du partage des poids garantissent une estimation sans biais, ceux-ci ne sont pas nécessairement optimaux en termes de variance. Des résultats sur l'optimalité des estimateurs issus de la méthode du partage des poids sont donnés dans l'article de Deville et Lavallée (2006)

B. Bases multiples

Afin de pallier un éventuel défaut de couverture, il est courant d'effectuer plusieurs tirages d'échantillons dans plusieurs bases de sondage, dont l'intersection n'est pas nécessairement vide. Certaines unités du champ de l'enquête peuvent alors être sélectionnées avec une probabilité non nulle dans chacune des bases. Les poids de sondage utilisés pour l'estimation doivent tenir compte de cette spécificité. Les unités présentes dans les différentes bases de sondages constituent alors une population intermédiaire U^A de taille N^A permettant de couvrir l'ensemble de la population d'intérêt U^B. On se place dans le cas simple résumé par la figure 2. où l'échantillon est le résultat de deux tirages indépendants d'échantillon dans les bases 1 et 2. Un échantillon probabiliste s_1 de taille n_1 est tiré parmi les N_1 unités dans la base de sondage 1, et un échantillon probabiliste indépendant s_2 de taille n_2 est tiré parmi les N_2 unités de la base de sondage 2. L'unité j dans l'échantillon s_1 possède un poids $w_{j,1}$ correspondant à l'inverse de la probabilité d'inclusion de l'unité j dans la base 1. De façon similaire, l'unité k dans l'échantillon s_2 possède un poids $w_{k,2}$ égal à l'inverse de la probabilité d'inclusion dans la base 2. Si l'on prenait l'estimateur naïf $\hat{Y}^{HT} = \sum_{j\in s_1} w_{j,1}y_j + \sum_{k\in s_2} w_{k,2}y_k$ du total Y alors celui-ci sur-estimerait le total Y, à cause des « doubles comptes » résultant des unités appartenant à l'intersection de ces deux bases. De manière analogue au cas (II-A), en désignant par Ω^B l'échantillon obtenu en fusionnant les deux échantillons et en supprimant les doublons, on peut construire à l'aide du partage des poids un estimateur sans biais de la forme :

$$\hat{Y}^B = \sum_{i\in\Omega^B} \left(\sum_{j\in s_1\cup s_2} d_j \frac{L_{ij}}{L_i} \right) y_i \qquad (1)$$

où $d_j = w_{j,1}I_{j\in s_1} + w_{j,2}I_{j\in s_2}$ et $L_i = \sum_{j\in U_1\cup U_2} L_{ij}$.

On peut se ramener au cas précédent « parents-enfants », en considérant que l'unité dans la base 1 est l'équivalent du père et l'unité dans la base 2 est l'équivalent de la mère. Dans ce cas, L_i correspond aux nombres de bases dans lesquelles l'unité i aurait pu être échantillonnée. Si on se réfère à la figure 2, le poids de sondage final des unités échantillonnées dans les deux bases de sondage est égal à la somme des poids de l'unité considérée dans chaque base divisée par deux. Le poids de sondage des unités n'appartenant qu'à une seule des deux bases reste inchangé.

L'application du partage des poids permet d'obtenir dans le cas de bases de sondage multiples un estimateur sans biais du total, mais cet estimateur n'est pas nécessairement optimal en termes de précision. Plus précisément, l'estimateur (1) appartient à une classe plus large d'estimateurs sans biais de la forme :

$$\sum_{j\in s_1\cap\bar{U}_2} w_{j,1}y_j + \sum_{k\in s_2\cap\bar{U}_1} w_{k,2}y_k + \sum_{j\in U_1\cap U_2} \left[\Theta w_{j,1}I_{j\in s_1} + (1-\Theta)w_{j,2}I_{j\in s_2}\right] y_j$$

Le choix optimal du paramètre Θ a notamment été étudié par Hartley (1962, 1974). L'estimateur (1) issu de la méthode de partage des poids correspond au choix de $\Theta = 1/2$. En pratique, dans le cadre des enquêtes ménages de l'Insee, comme la variance des estimations est inversement proportionnelle à la taille d'échantillons, on choisit afin de limiter la dispersion des poids un paramètre[2] $\Theta = \frac{n_1}{n_1+n_2}$.

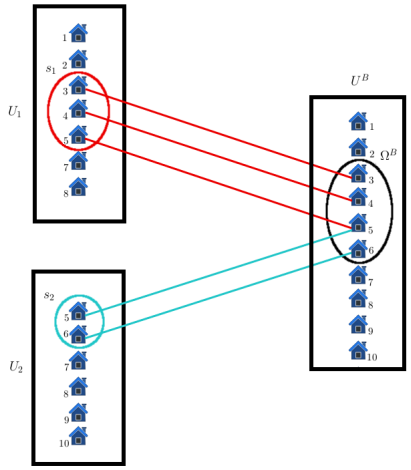

Fig. 2. Estimation en présence de bases de sondage multiples

C. Enquêtes répétées dans le temps et partage des poids

α. Exploitation transversale d'un panel

Un panel est un échantillon dont les unités sont interrogées au moins à deux reprises sur une période donnée : l'échantillon est sélectionné dans la population à la date initiale, puis les unités de cet échantillon sont suivies aussi longtemps que nécessaire pour les besoins de l'étude. Un panel est donc un échantillon qui représente la population à la date de son tirage. Il s'inscrit ainsi fondamentalement dans une approche *longitudinale* consistant à mesurer l'évolution d'un paramètre au cours du temps.

À première vue, l'utilisation d'un panel semble donc incompatible avec une approche *transversale* visant à estimer un paramètre à la date courante de l'enquête, puisque le panel ne représente pas la population courante mais la

2. Dans cette expression, les tailles d'échantillons n_1 et n_2 peuvent parfois être remplacées par le nombre de répondants dans chaque base. C'est notamment le cas lorsque les taux de réponse entre les deux bases utilisées diffèrent (par exemple, dans le cas où des modes de collecte différents sont utilisés sur les deux échantillons). Attention, l'extension de l'expression donnée pour le paramètre Θ ne s'étend pas de façon triviale pour davantage de bases de sondage.

population à la date de son tirage [3], et ne couvre donc pas de ce fait les unités entrées dans le champ visé entre la date de tirage du panel et la date courante. Cependant, lorsqu'il existe un concept « naturel » de regroupement des unités du panel, une utilisation astucieuse du sondage indirect va permettre d'obtenir un échantillon transversal à partir du panel.

Plaçons-nous dans le cas d'un panel d'individus : ces individus sont regroupés de façon naturelle au sein de logements. À partir de l'échantillon initial « d'individus panel », on va construire un échantillon transversal s_0 en enquêtant à la date courante tous les individus présents dans les logements contenant au moins un individu panel. Cette constitution de l'échantillon transversal par sondage indirect va permettre de prendre en compte les naissances, et donc de couvrir la population à la date courante [4]. Le poids des individus de cet échantillon transversal va ensuite être déterminé *via* la méthode du partage des poids. Chaque individu i d'un même logement ℓ va ainsi se voir attribuer le même poids $w_{i\ell}$, calculé comme la somme des poids de tirages $d_{k\ell}$ des individus panel k résidant dans le logement ℓ divisée par le nombre total L_ℓ d'individus du logement ℓ à la date courante qui étaient échantillonnables, à la date initiale, dans le panel s_0, soit :

$$w_{i\ell} = \frac{1}{L_\ell} \times \sum_{k \in s_0, k \in \ell} d_{k\ell}$$

β. Exploitations longitudinales et transversales d'un échantillon rotatif

Comme nous l'avons vu précédemment, un panel répond avant tout à une approche longitudinale visant à mesurer l'évolution d'un paramètre au cours du temps. Même si la méthode de sondage indirect évoquée précédemment associée à la sélection d'un échantillon complémentaire permet une exploitation transversale à partir d'un panel pur, cette méthode ne constitue qu'un pis-aller et peut s'avérer complexe à mettre en œuvre, en particulier pour ce qui est de la sélection de l'échantillon complémentaire.

Dès lors que l'on cherche à concilier des objectifs d'exploitations transversales et longitudinales, on va donc privilégier l'utilisation d'un échantillon rotatif. Un échantillon rotatif est un échantillon réunissant des panels tirés à des dates différentes et dont la durée de vie est constante et limitée, le système étant conçu de manière à ce qu'à chaque campagne d'enquête, un panel entre dans l'échantillon et un panel en sorte. Le schéma en figure 3 (inspiré de ceux présents dans [4] au chapitre IV.3.3) récapitule la situation pour un échantillon rotatif renouvelé au quart.

Ce système d'échantillon rotatif permet conjointement des exploitations longitudinales et transversales :

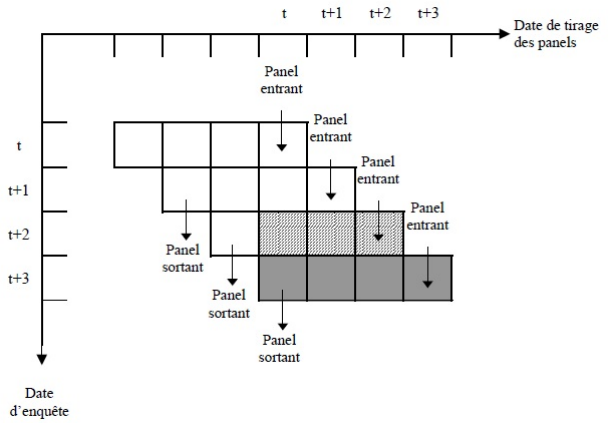

Fig. 3. Échantillon rotatif renouvelé au quart

- pour estimer l'évolution d'un paramètre entre deux dates – par exemple ici entre t+2 et t+3 –, on s'appuiera sur l'échantillon longitudinal – constitué ici de la réunion des trois panels hachurés en figure 3 qui sont les seuls à être enquêtés à la fois en t+2 et t+3 ;
- pour estimer un paramètre à une date d'enquête donnée – par exemple ici en t+3 –, on s'appuiera sur l'échantillon transversal – constitué ici de la réunion des quatre panels grisés en figure 3 enquêtés tous les quatre en t+3 .

Ainsi, dans cette configuration, les échantillons longitudinaux et transversaux sont tous deux constitués par union de différents panels, chaque panel étant représentatif de la population à la date de son tirage. Les pondérations associées à ces échantillons longitudinaux et transversaux vont là encore être déterminées par partage des poids [5] :

- pour l'échantillon longitudinal, le poids d'un individu sera égal à son poids de tirage dans le panel ***via* lequel il a été sélectionné** divisé par le nombre de panels dans lesquels l'individu aurait pu être échantillonné ;
- pour l'échantillon transversal, un premier partage des poids selon la procédure détaillée en II-C-α est à effectuer panel par panel. On opère ensuite un second partage des poids consistant à diviser, pour chaque individu, son poids issu du premier partage des poids par le nombre de panels au travers desquels le ménage dans lequel il réside aurait pu être échantillonné.

III. Déterminer les liens

La détermination des liens dans la méthode du partage des poids est un point crucial. En effet, la qualité de la méthode, et en particulier le caractère sans biais de celle-ci, est conditionnée par une évaluation correcte des liens entre les unités d'échantillonnage et les unités d'observation. De plus, il est nécessaire que chacune des unités de la base d'intérêt ait au moins un lien avec les unités échantillonnables pour garantir le caractère sans biais des estimateurs issus du

3. En général moins les sorties de champ observées entre la date de tirage du panel et la date courante de l'enquête.

4. En pratique, un défaut de couverture subsiste avec cette approche, puisque les individus occupant des logements dans lesquels il ne peut y avoir d'individu panel – immigrants habitant un logement ne comprenant que des immigrants par exemple – échappent à l'enquête. Ce défaut de couverture résiduel peut être traité par la sélection d'un échantillon complémentaire tiré directement dans la population courante.

5. On présente ici les résultats dans un contexte légèrement simplifié où l'on néglige la probabilité (très faible en pratique) qu'un même individu soit sélectionné dans plus d'un panel. On se référera au chapitre IV.3.3 de [4] pour le détail des calculs et les formules générales.

partage des poids.

Cette détermination des liens peut s'effectuer de différentes façons suivant le contexte d'utilisation de la méthode de partage des poids. Dans le cas classique du sondage indirect (cf. II-A) et dans le cas d'échantillons panélisés (cf. II-C), on ajoute dans le questionnaire une question spécifique. Par exemple, dans le cas des panels, on demande à l'individu sélectionné s'il appartenait au champ de l'enquête aux dates correspondants aux tirages des échantillons panélisés. Dans le cas de bases de sondages multiples, il est parfois possible d'effectuer un appariement entre les bases. Cela permet de déterminer à quelle(s) base(s) de sondage appartiennent les unités sélectionnées indirectement.

IV. Traitements post-collecte

On se contente ici de donner les grandes lignes relatives aux traitements post-collecte appliqués à des échantillons sur lesquels un partage des poids a été effectué.

A. Partage des poids et correction de la non-réponse

On distingue deux types de non-réponse fondamentalement différents en cas de partage de poids (en sus de la non-réponse partielle classique) :

- la non-réponse totale d'unité : elle est en général traitée en amont du partage des poids. Une procédure de repondération est d'abord mise en œuvre sur le ou les échantillons de tirage, avant d'effectuer le partage des poids en s'appuyant sur les répondants du ou des échantillons de tirage avec leurs poids corrigés de la non-réponse. Le traitement de la non-réponse totale d'unité pour un panel ou un échantillon rotatif est plus complexe, et décrit dans [4] au chapitre IV.3 ;
- la non-réponse de lien : il s'agit d'une non-réponse partielle portant sur la ou les variables du questionnaire permettant de déterminer les liens associés à une unité répondante. Plusieurs méthodes, détaillées dans [5], permettent de traiter cet épineux problème, spécifique aux échantillons impliquant un partage de poids. On peut par exemple modéliser, sur l'échantillon des répondants, chaque variable de lien en fonction de variables auxiliaires à l'aide d'une régression logistique, et ensuite appliquer ce modèle de régression logistique aux non-répondants pour imputer les liens manquants.

B. Partage des poids et calage sur marges

L'interaction entre partage des poids et calage sur marges, et en particulier l'ordre des opérations, va dépendre des informations auxiliaires dont on dispose :

- si on dispose uniquement ou très majoritairement de marges relatives à la population cible des unités finales, on procédera au calage sur marges après partage des poids, en s'appuyant sur l'échantillon d'unités finales répondantes après correction de la non-réponse et partage des poids ;
- si on dispose uniquement ou très majoritairement de marges relatives à la / aux population(s) d'unités intermédiaires, on procédera au calage sur marges avant partage des poids, en s'appuyant sur le / les échantillon(s) d'unités intermédiaires répondantes après correction de la non-réponse ;
- si on dispose simultanément de marges relatives aux populations d'unités intermédiaires et finales, il est possible, *via* une modification *ad hoc* des variables de calage relatives aux unités finales, de se ramener à un calage portant uniquement sur l'échantillon d'unités intermédiaires. Dans ce cas, on procédera donc à ce calage spécifique avant partage des poids, en s'appuyant sur le / les échantillon(s) d'unités intermédiaires répondantes après correction de la non-réponse. Cette procédure de calage, plus générale mais plus complexe que les précédentes, est détaillée dans [6] au paragraphe 7.2.

Là encore, la problématique du calage sur marges dans le cas d'un panel ou d'un échantillon rotatif est spécifique et plus complexe, et toujours décrite dans [4] au chapitre IV.3.

References

[1] Deville, J.-C., Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, Vol. 32, No 2, p. 185.

[2] Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association. pp. 203-206.

[3] Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhya*, Vol. 36, No 997, p. 118.

[4] Ardilly, P. (2006). *Les techniques de sondages*. Éditions Technip, Paris.

[5] Xiaojian X., Lavallée, P. (2009). Traitements de la non-réponse de lien dans l'échantillonnage indirect. *Techniques d'enquête*, Vol. 35, No 2, pp. 165-177.

[6] Lavallée, P. (2007). *Indirect sampling*. Springer, 2007, New York.

Département des méthodes statistiques
Version n° 1, diffusée le 10 octobre 2017