OPTIMISATION DES PLANS DE SONDAGE POUR LES ENQUÊTES AUPRÈS DES PERSONNES ET DES MÉNAGES PAR L'UTILISATION DE PRÉDICTEURS DE NON-RÉPONSE

Lionel QUALITÉ (&**), Clément CHEVALIER (**)*

(*) Office Fédéral de la Statistique, Section Méthodes statistiques
(**) Université de Neuchâtel, Institut de Statistique

lionel.qualite@bfs.admin.ch

Mots-clés : taux de réponse, calage, biais

Résumé

Nous présentons un exercice simple de calcul d'allocation dans le cas des enquêtes conduites par l'Office Fédéral de la Statistique (OFS, Suisse) auprès des personnes et des ménages. Ce travail est particulièrement adapté au cas d'enquêtes qui présentent des probabilités de réponse individuelles hétérogènes, sélectionnées dans une base de sondage qui contient des bons prédicteurs de ces probabilités de réponse, et pour lesquelles on dispose de données permettant d'ajuster un modèle de réponse prévisionnel. Les estimations obtenues en utilisant l'enquête « Micro-recensement formation de base et formation continue (MZB) » de 2016 montrent la possibilité d'obtenir une réduction de variance d'environ 14% par rapport aux tirages actuels. Une optimisation complète n'étant pas sans risques, une solution simplifiée a été proposée aux sections d'enquête avec la possibilité de se prémunir contre ces risques au prix d'une efficacité moindre.

L'OFS utilise, pour ses enquêtes auprès des personnes et des ménages, une base de sondage construite à partir des registres cantonaux et communaux de population ainsi que de registres fédéraux. Cette base de sondage est mise à jour chaque trimestre et contient les informations démographiques de base sur la population (sexe, âge, nationalité, état-civil…) ainsi que la composition des ménages. On dispose, en outre, d'un numéro de téléphone fixe pour environ 75% de la population dans ces registres. Chaque année, plusieurs enquêtes sont réalisées par entretiens téléphoniques avec parfois une possibilité de télé-déclaration. Ces enquêtes sont répétées, certaines annuellement, d'autres tous les cinq ans. La disponibilité ou non d'un numéro de téléphone a une influence considérable sur les taux de réponse observés : environ 51% pour les personnes avec numéro de téléphone, et 25% pour les personnes dont on ne connaît pas le numéro. Ces dernières sont invitées par courrier à fournir un numéro de contact sur une carte réponse. Les échantillons actuels sont sélectionnés avec des probabilités égales pour les personnes ou les ménages en utilisant un plan de Poisson (exactement pour les ménages, approximativement pour les personnes). Les cantons et communes ont la possibilité de financer un doublement de leur échantillon.

Pour un plan de Poisson avec probabilités d'inclusion π_i , et probabilités de réponses indépendantes r_i connues, la variance d'estimation par expansion du total des y_i dans la population U peut s'écrire :

$$Var(\hat{Y}) = \sum_U \frac{1-\pi_i r_i}{\pi_i r_i} y_i^2 \quad (1).$$

Lorsqu'un calage est utilisé, ce qui est systématiquement le cas des enquêtes de l'OFS, les y_i peuvent être remplacés dans cette équation par les résidus de régression e_i des y_i sur les variables de calage pour obtenir une variance approchée (voir Deville et Särndal 1992).

Pour une enquête généraliste, sans hypothèse sur les variables d'intérêt, on considère que les y_i, ou plutôt les e_i, sont échangeables dans la population (ceci doit naturellement être adapté dans le cas d'enquêtes auprès des entreprises par exemple). On est alors amené à minimiser la somme des $\frac{1}{\pi_i r_i}$ sous contrainte de coût $C = \sum_U \pi_i C_i$, où C_i est l'espérance sous le mécanisme de réponse du coût associé à la sélection d'une unité i. Par exemple : $C_i = C_i^{nr}(1 - r_i) + C_i^r r_i$ où C_i^{nr} est le coût associé à une unité non répondante et C_i^r le coût associé à une unité répondante. La résolution symbolique conduit à choisir

$$\pi_i = C.S^{-1}.(r_i C_i)^{-\frac{1}{2}}, \text{ où } S = \sum_U \left(\frac{C_i}{r_i}\right)^{\frac{1}{2}}.$$

Certains π_i ainsi obtenus peuvent être supérieurs à 1 pour des unités avec très faibles probabilités de réponse et quelques itérations sont à envisager. Dans le cas de l'enquête MZB, en prédisant les r_i à l'aide d'un modèle logistique relativement complet, on arrive à un rapport de variance estimé de 0.86 entre cette allocation et un plan bernoullien de même coût, ou bien encore à une réduction du coût d'environ 14% pour une précision fixée. La variance (1) peut en outre aisément être estimée pour des variables d'intérêt dans le cas d'enquêtes répétées. On peut également dans ce cas chercher des π_i optimaux pour des variables d'intérêt si l'on se restreint à des π_i dans une classe donnée de fonctions des variables de la base de sondage (ce qui a été fait pour l'Enquête Familles et Générations de 2018).

Les risques encourus en choisissant ces probabilités de sélection optimales sont de deux natures : une erreur de prédiction sur les r_i pourrait conduire à des poids d'extrapolation élevés (si des r_i très petits sont largement surestimés). D'autre part, l'optimisation pourrait faire perdre de la précision pour une variable d'intérêt dans une configuration défavorable. En effet, la procédure conduit à surreprésenter les unités qui répondent le moins au détriment des autres. Une variable faiblement dispersée parmi les premières et fortement dispersée parmi les secondes pourrait donc pâtir de cette optimisation.

Prudemment, nous avons proposé dans un premier temps d'utiliser un modèle de réponse très simple, n'utilisant que la connaissance ou non d'un numéro de téléphone fixe, avec un gain d'efficacité potentiel de l'ordre de 12%. On peut alors choisir un compromis entre réduction du budget, gains d'efficacité et tolérance à des rapports de dispersion défavorables entre ces deux sous-populations. On peut ainsi réduire le budget en s'assurant que la précision sera meilleure pour toute variable d'intérêt dont la variance parmi les unités sans numéro de téléphone connu est au moins égale à une fraction choisie de la variance parmi les unités avec numéro de téléphone connu.

Enfin, le fait de surreprésenter les unités qui répondent le moins conduit à faire baisser le taux de réponse apparent à l'enquête. Il passerait par exemple de 45% à 39% pour l'enquête MZB. Se pose alors la question du risque de biais d'estimation. On peut en donner l'approximation au premier ordre suivante :

$$B \approx \sum_U y_i \left[\frac{r_i}{E(r_i)} - 1\right]'$$

où $E(r_i)$ désigne l'espérance sous le plan de sondage de l'estimateur de la probabilité de réponse. Ce biais ne dépend ainsi pas des π_i pour autant que le plan permette une bonne estimation des probabilités individuelles de réponse.

Bibliographie

[1] Deville J.-C., Särndal C.-E., « Calibration Estimators in Survey Sampling », *Journal of the American Statistical Association*, vol 87, n° 418, pp 376-382, juin 1992.