JMS Journées de méthodologie statistique de l'Insee 2018

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
(**) Université de Neuchâtel

lionel.qualite@bfs.admin.ch

Mots-clés. Taux de réponse, Calage, Biais.

Résumé

Depuis 2010, l'Office Fédéral de la Statistique (OFS, Suisse) a la chance de disposer pour ses enquêtes auprès des ménages et de la population d'une base de sondage très riche en informations. Ces informations peuvent être mobilisées pour gagner en efficacité dans les enquêtes répétées.

Une caractéristique de nos enquêtes téléphoniques est le taux de réponse très différent entre les personnes pour lesquelles on dispose d'un numéro de téléphone avant l'enquête et celles auxquelles on doit demander un numéro de téléphone par écrit pour pouvoir les contacter. On propose ici d'adapter les plans de sondage pour anticiper cette non-réponse, éventuellement en utilisant également d'autres informations de la base de sondage. Plus précisément, on cherche des probabilités de sélection, fonction des probabilités de réponse, qui minimisent une variance pour un budget donné.

Une première étude démontre la possibilité de réaliser des gains en efficacité substantiels pour des enquêtes avec des probabilités de réponse hétérogènes si celles-ci peuvent être bien expliquées au moyen de quelques variables contenues dans la base de sondage.

Toutefois, la réalisation de ces gains n'est pas sans risques car elle nécessite d'utiliser les résultats d'enquêtes passées. De plus, ce que l'on gagne en précision pour une variable d'intérêt pourrait être perdu pour une autre variable.

Il est possible de limiter ces risques en restreignant la recherche des probabilités de sélection optimales à des fonctions choisies des principales variables qui expliquent la non-réponse. Des hypothèses ou des données d'enquêtes passées concernant les variables d'intérêt peuvent aussi être utilisées dans cette démarche.

Les résultats obtenus montrent assez logiquement que, sans information sur les variables d'intérêt, il faudrait sur-échantillonner les populations qui répondent le moins bien. Cela a pour conséquence de faire baisser le taux de réponse apparent des enquêtes. Toutefois, tant que les procédures de collecte et les probabilités individuelles de réponse ne sont pas diminuées, cela ne devrait pas avoir de conséquence sur de possibles biais d'estimation.

Abstract

Starting in 2010, the Swiss Federal Statistical Office (SFSO) has enjoyed using an information rich population and household sampling frame. It is possible to take advantage of these data in order to achieve higher efficiency with our surveys.

A characteristic of SFSO's telephone surveys is the wide discrepancy between response rates of units for whom a telephone number is readily available in our sampling frame and those of units to whom we have to ask for a contact number by means of a written request. We intend to adapt our sampling designs to this non-response configuration, and also consider using other data from our sampling frame. Specifically, we are looking for selection probabilities that are functions of response probabilities and that minimize a variance function while meeting a given budget.

A first investigation shows that subsantial efficiency improvements can be achieved in surveys with heterogeneous response probabilities that are well related to selected variables present in the sampling frame.

However, fulfilling these improvements does not go without risks as it entails using data from past survey occasions. Moreover, a gain on precision for given interest variable can turn out to be a loss for another interest variable.

These risks can be limited by restricting the search for optimal selection probabilities to selected functions of the key variables that explain non-response. Assumptions or data from past surveys regarding variables of interest may also be used in this approach.

The results obtained show quite logically that, without further information on variables of interest, the populations with lower response probabilities should be oversampled. This has the effect of diminishing the apparent response rate of surveys. However, as long as the collection procedures and individual response probabilities are not reduced, this should not have an impact on possible estimation biases.

Introduction

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. On se concentre ici sur l'utilisation des données disponibles dans la base de sondage pour anticiper la non-réponse.

La disparité des probabilités de réponse estimées entraine en effet une dispersion des poids d'extrapolation si les probabilités de tirages sont homogènes. Cette dispersion peut avoir un effet négatif sur la précision des estimateurs.

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.

1 Enquêtes auprès de la population suisse

Depuis 2010, l'OFS collecte les registres communaux et cantonaux de population, ainsi que des registres fédéraux. Ces données sont ensuite utilisées pour créer les bases de sondage trimestrielles des enquêtes auprès de la population et des ménages. Ces fichiers contiennent des informations sur

les personnes vivant en Suisse : date de naissance, sexe, état-civil, nationalité, permis de séjour, adresse, et sur la composition des ménages. Les numéros de téléphone fixe des personnes sont récupérés depuis la "base de données suisse des appels d'urgence", alimentée par les opérateurs de téléphonie. Par une procédure d'appariement, on arrive à associer un numéro de téléphone à environ 75% de la population des bases de sondage.

La plupart des enquêtes de l'OFS auprès des ménages et de la population sont effectuées par téléphone, avec en outre la possibilité de plus en plus souvent offerte et mise en avant de remplir tout ou partie du questionnaire sur internet. Une exception notable à cela est le "relevé structurel", grande enquête annuelle qui reprend essentiellement le questionnaire des anciens recensements. Les personnes sélectionnées au relevé structurel peuvent répondre sur le questionnaire papier ou par internet. Une assistance téléphonique est néanmoins disponible.

Pour les enquêtes par téléphone, la connaissance d'un numéro fixe a une importance déterminante quant aux taux de réponse observés. En effet, un ménage ou une personne pour lequel on connaît (ou croit connaître) un numéro de téléphone va recevoir une lettre d'annonce pour l'informer qu'il a été sélectionné. Puis il sera activement contacté par l'institut de sondage mandaté, au numéro de téléphone connu. Un ménage ou une personne pour lequel on ne connaît pas de numéro de téléphone va, quant à lui, recevoir une lettre d'annonce avec une carte-réponse pour indiquer un numéro de téléphone auquel il peut être contacté.

De manière assez prévisible, la proportion de cartes-réponse retournées est relativement faible, de l'ordre de 20 à 30% selon les enquêtes. Le taux de réponse à l'entretien téléphonique est lui sensiblement plus élevé dans la population qui a renvoyé une carte réponse que dans la population directement contactée à un numéro connu (très grossièrement proche de 80% pour les premiers contre 50 à 60% pour les derniers).

Ainsi, la connaissance ou non d'un numéro de téléphone apparaît toujours comme le prédicteur le plus important de la non-réponse lors de la pondération des échantillons de répondants dans les enquêtes par téléphone. Les autres variables à disposition sont également considérées pour estimer des probabilités de réponse, à l'aide de régressions logistiques ou en créant des "groupes homogènes de réponse" par segmentation. Les caractères les plus importants sont habituellement la composition du ménage (personne seule, deux personnes majeures de sexe opposé, deux personnes majeures de sexe opposé et des mineurs, etc.), la classe d'âge et la nationalité.

Les principales enquêtes par échantillonnage de l'OFS sont répétées, certaines annuellement, d'autres tous les cinq ans (voir à ce sujet 2). Outre le relevé structurel, il s'agit des cinq enquêtes "thématiques" (sur la formation, la santé, les familles et générations, les langues religions et culture et sur la mobilité et les transports), de l'enquête sur la population active, de l'enquête sur le budget des ménages et de l'enquête sur les revenus et conditions de vie. Ainsi, l'on dispose d'informations sur les propensions à répondre l'année précédente ou avec un retard de cinq ans pour les enquêtes thématiques.

Les échantillons sont sélectionnés selon un plan de Poisson pour les enquêtes ménages et en deux phases pour les enquêtes "personnes" : par sélection de ménages selon un plan de Poisson dans une première phase, puis usuellement par sélection d'une personne cible dans chaque ménage sélectionné. Les taux de sondage de base sont uniformes parmi les personnes sélectionnables d'un même canton. Ces taux peuvent être augmentés par les cantons et communes qui souhaitent financer des extensions d'échantillon. En outre, les personnes de nationalité étrangère sont sur-sélectionnées pour l'enquête sur la population active.

2 Pondération et variance

Les pondérations ne sont pas calculées exactement de la même façon pour les différentes enquêtes, mais elles suivent toutes un schéma commun. La non-réponse est modélisée par une sélection aléatoire indépendante des répondants parmi les ménages et personnes enquêtées. Des "probabilités de réponse" sont estimées, soit à l'aide d'une régression logistique en utilisant des

caractéristiques connues dans la base de sondage, soit en constituant des "groupes de réponse homogène" par segmentation, toujours en utilisant les données de la base de sondage. Ainsi, un premier jeu de poids

$$d_i = 1/\pi_i \widehat{r}_i,\ i \in r,$$

est calculé pour toutes les unités i répondantes dans un échantillon observé r. Les π_i désignent les probabilités de sélection dans l'échantillon enquêté s, et les $\widehat{r}_i$ sont les probabilités de réponse estimées des répondants.

Les poids d_i sont ensuite ajustés par une technique de calage (voir 1) pour respecter certains totaux calculés sur la base de sondage. Les variables de calage sont habituellement construites à partir des mêmes caractères qui ont servi à estimer les probabilités de réponse.

Pour certaines enquêtes dont les échantillons sont sélectionnés dans différentes bases de sondage, il y a en plus une étape de combinaison des poids qui peut être réalisée selon les cas soit avant soit après le calage.

Pour simplifier, on va considérer ici le cas d'une enquête dont l'échantillon est sélectionné dans une seule base de sondage. Les estimateurs de totaux s'écrivent

$$\widehat{Y} = \sum_{i \in r} w_i y_i,$$

où w_i est le poids d'extrapolation de l'unité i et y_i est le paramètre dont on veut estimer le total. Le poids w_i peut à son tour s'écrire

$$w_i = g_i d_i = \frac{g_i}{\pi_i \widehat{r}_i},$$

où g_i est le facteur d'ajustement, "g-weight", résultat du calage.

Un calcul exact de la variance de l'estimateur $\widehat{Y}$ est difficilement praticable : il dépend de la manière dont $\widehat{r}_i$ et g_i sont calculés et les fonctions considérées ne sont pas linéaires. De plus, une estimation exactement sans biais est souvent impossible, ne serait-ce que parce qu'on ne sélectionne qu'une seule personne par ménage dans les enquêtes auprès des personnes. (3) ont étudié le cas, voisin et souvent confondu avec ce qui est effectivement utilisé à l'OFS, de l'estimateur par régression et par dilatation des poids lorsque les $\widehat{r}_i$ sont estimés par maximum de vraisemblance.

Notons M la population des ménages et U la population des personnes sélectionnables à l'enquête. Si les "vraies" probabilités de réponse r_i étaient connues, on pourrait considérer $\widehat{Y}_d = \sum_{i \in r} y_i/\pi_i r_i$. Selon (3), la variance de $\widehat{Y}_d$ est en général plus grande que la variance de $\widehat{Y}_e = \sum_{i \in r} d_i y_i$. Pour une enquête ménages, on a :

$$\mathrm{var}(\widehat{Y}_d) = \sum_{m \in M} \frac{1 - \pi_m r_m}{\pi_m r_m} y_m^2,$$

et pour une enquête auprès des personnes,

$$\mathrm{var}(\widehat{Y}_d) = \sum_{i \in U} \frac{1 - \pi_i r_i}{\pi_i r_i} y_i^2 - \sum_{m \in M} \sum_{i \neq j \in m} y_i y_j.$$

Avec les probabilités de sélection et de réponses utilisés pour les enquêtes de l'OFS, le terme croisé $\sum_{m \in M} \sum_{i \neq j \in m} y_i y_j$ est négligeable devant le terme diagonal $\sum_{i \in U} \frac{1-\pi_i r_i}{\pi_i r_i} y_i^2$. On retient donc la variance approchée :

$$\mathrm{var}(\widehat{Y}_d) \approx \sum_{i \in U} \frac{1 - \pi_i r_i}{\pi_i r_i} y_i^2.$$

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 après calage (voir 1). Pour une enquête auprès des personnes, on écrit

$$e_i = y_i - \mathbf{x}_i' \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k,$$

où $\mathbf{x}_k$ est le vecteur des variables de calage de l'unité k, et $\left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$ est une inverse généralisée de $\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k'$. Pour les enquêtes auprès des entreprises et pour les enquêtes auprès des ménages, il faut parfois adapter ces définitions et introduire des poids dans la régression selon la manière dont le calage a été réalisé.

La variance approchée de $\widehat{Y}$ est définie par :

$$\text{var}_a(\widehat{Y}) = \sum_{i \in U} \frac{1 - \pi_i r_i}{\pi_i r_i} e_i^2. \tag{1}$$

Cette approximation est utilisée dans la suite de ce travail. Pour être tout à fait transparent, les pratiques d'estimation de variance à l'OFS ne sont pas basées sur cette approximation. En général, les sections de production utilisent les procédures "surveyfreq" et "surveymeans" de SAS, avec un paramétrage qui correspond à un plan stratifié avec des "strates" définies soit en regard des taux de sondages uniformes, soit en regard des indicatrices de domaines utilisées dans les calages. L'effet complet du calage est rarement, sinon jamais, intégré dans ces estimations.

3 Possibilité d'optimisation

3.1 Calcul de probabilités de sélection optimales

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. On fait ici appel à un modèle de super-population où chaque permutation des valeurs e_i dans la population U est équiprobable. Ceci doit naturellement être adapté dans le cas d'enquêtes auprès des entreprises par exemple, pour lesquelles les e_i sont usuellement dépendants d'une variable de taille.

Si les e_i sont échangeables, la variance approchée de $\widehat{Y}$ (moyenne sous le modèle de super-population) est déterminée par

$$V = \sum_{i \in U} (1 - \pi_i r_i)/\pi_i r_i.$$

Les probabilités de sélection optimales sous une contrainte de coût sont alors obtenues lorsque les π_i minimisent la somme des $1/\pi_i r_i$.

Considérons par exemple une fonction de coût définie par

$$C = \sum_{i \in U} \pi_i C_i,$$

où C_i est l'espérance sous le mécanisme de réponse du coût entrainé par la sélection de l'unité i dans l'échantillon s. Pour les besoins de l'OFS, on peut se contenter d'une fonction de cette forme, avec $C_i = C_i^{nr}(1 - r_i) + C_i^r r_i$, 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 en π_i, sans autre contrainte, conduit à choisir

$$\pi_i = \frac{C}{S}(r_i C_i)^{-\frac{1}{2}}, \text{ où } S = \sum_{k \in U} (C_k/r_k)^{\frac{1}{2}}. \tag{2}$$

Certains π_i ainsi calculés peuvent être supérieurs à 1 pour des unités avec des faibles probabilités de réponse. La solution du problème de minimisation avec les contraintes supplémentaires $\pi \leq$

1 pour tout i peut facilement être obtenue directement ou bien de manière itérative (les π qui dépassent 1 sont bloqués à cette valeur et on recommence l'optimisation sur le reste de la population...).

Si aucun des π_i n'est égal à 1, la valeur minimale de V est

$$V_{opt} = \frac{S^2}{C} - N,$$

où N est la taille de U. On peut comparer cela à la valeur de V lorsque les π_i sont uniformes et égaux à $C/\sum_{i\in U} C_i$. On a alors, pour le plan de Bernoulli correspondant :

$$V_{bern} = \frac{\sum_{i\in U} C_i \sum_{i\in U} 1/r_i}{C} - N.$$

3.2 Estimation des gains d'efficacité potentiels

Les seules inconnues dans le calcul de V_{opt}, V_{bern} et des π_i optimaux sont les probabilités de réponse r_i. Dans les enquêtes de l'OFS, les r_i sont en principe estimés en utilisant des prédicteurs z_i connus dans toute la base de sondage. Les $\widehat{r}_i$ peuvent donc être calculés non seulement pour l'échantillon sélectionné s et l'échantillon de répondants r, mais aussi pour l'ensemble de la population du cadre. Ce dernier point est en outre nécessaire si l'on souhaite effectivement adapter les probabilités de sélection à la non-réponse anticipée.

Supposons que l'on dispose des résultats d'une enquête précédente, aux caractéristiques suffisamment proches de la nouvelle enquête. On peut alors estimer sans trop de difficulté le potentiel de gain de précision V_{opt}/V_{bern}. Le terme S, par exemple, peut être estimé par

$$\widehat{S}_r = \sum_{i\in r} d_i \left(\widehat{C}_i/\widehat{r}_i\right)^{\frac{1}{2}},$$

où $\widehat{C}_i = C_i^{nr}(1-\widehat{r}_i) + C_i^r \widehat{r}_i$, ou bien

$$\widehat{S}_w = \sum_{i\in r} w_i \left(\widehat{C}_i/\widehat{r}_i\right)^{\frac{1}{2}},$$

mais aussi par

$$\widehat{S}_s = \sum_{i\in s} d_i\widehat{r}_i \left(\widehat{C}_i/\widehat{r}_i\right)^{\frac{1}{2}},$$

et par

$$\widehat{S}_U = \sum_{i\in U} \left(\widehat{C}_i/\widehat{r}_i\right)^{\frac{1}{2}}.$$

Les d_i et w_i désignent ici les poids d'extrapolation de l'enquête passée, tandis que les C_i^r et C_i^{nr} sont les coûts de la nouvelle enquête.

Les estimateurs S_w et S_r ont l'avantage de n'utiliser que des données déjà disponibles : les $\widehat{r}_i$ ont été calculés lors de la pondération de l'enquête.

La somme $\widehat{S}_U$ permet quant à elle d'utiliser une base de sondage U différente de celle de l'enquête passée, et donc de tenir compte de lévolution de la population.

Dans le cas de l'enquête MZB de 2016, en prédisant les r_i pour toute la base de sondage à l'aide d'un modèle logistique relativement complet, on arrive à un rapport de variance estimé V_{opt}/V_{bern} de 0.86. En prenant le problème dans l'autre sens, on arrive à un potentiel de réduction de coût d'environ 14% pour une valeur de V donnée. Le potentiel de gain était meilleur encore, s'approchant de 20%, pour la précédente enquête MZB en 2011. En effet, l'écart entre les taux de réponse des personnes avec un numéro de téléphone connu dans la base de sondage et celles sans

numéro de téléphone connu était plus important. Le développement des possibilités de réponse par voie électronique, avec des accès fournis par envoi postal, semble faire diminuer l'impact de la connaissance d'un numéro de téléphone sur les taux de réponse.

La grande majorité des réponses du relevé structurel est reçue par retour du questionnaire papier. Le taux de réponse global à l'enquête est proche de 90%. Le potentiel de gain identifié en anticipant la non-réponse est très faible, de l'ordre de 3%.

3.3 Limites

Un calcul des probabilités de sélection optimales comme dans la Section 3.1 suivi de leur estimation n'est pas nécessairement le choix que l'on retiendra à l'OFS.

Le calcul des probabilités de sélection optimales dans la Section 3.1 suppose que l'on connaît des probabilités de réponse r_i pour toute la base de sondage. Substituer des prédictions $\widehat{r}_i$ à ces r_i nécessite de pouvoir faire ces prédictions. Considérons que l'on souhaite utiliser pour ce faire un modèle ajusté sur une enquête précédente, par exemple si l'on peut écrire $\widehat{r}_i = r(\mathbf{z}_i, \widehat{\alpha})$, ou $\mathbf{z}_i$ sont les prédicteurs utilisés dans le modèle et $\widehat{\alpha}$ est un paramètre estimé avec les données d'enquête. Il faut alors que les prédicteurs z_i soient disponibles pour toutes les unités de la (nouvelle) base de sondage. C'est le cas de la majorité des enquêtes auprès des personnes et des ménages à l'OFS. Mais il arrive pour certaines enquêtes que l'on utilise également de l'information relevée lors de la collecte pour estimer les probabilités de réponse. Cette information n'est disponible que pour des unités qui étaient sélectionnées dans l'échantillon s.

En outre, pour peu que le modèle de prédiction soit un peu élaboré, on est souvent confronté à des valeurs très basses de $\widehat{r}_i$ lorsque l'on calcule des pour toutes les unités de la base de sondage. Il s'agit de prédictions pour des unités qui cumulent des caractéristiques associées avec un faible taux de réponse. De plus une mauvaise prédiction des r_i peut conduire à une allocation sous-optimale et des mauvaises surprises lors de la pondération de l'enquête une fois terminée avec en particulier des poids élevés attribués à des unités qui ont finalement moins bien répondu que ce qui était prévu.

D'autre part, l'optimisation conduit à échantillonner avec une probabilité de sélection élevée les unités qui répondent le moins bien. Si cela a du sens pour une enquête, on peut craindre que les sollicitations répétées de ces unités diminue encore leur propension à répondre à de nouvelles enquêtes.

4 Une solution moins aggressive

Pour pallier les problèmes soulevés, et avoir des résultats moins dépendants des conditions de l'enquête précédente, on propose de résoudre un problème d'optimisation un peu moins ambitieux. Il s'agit de chercher les probabilités de sélection optimales en se restreignant à une famille paramétrée de fonctions de certaines variables $\mathbf{u}_i$ de la base de sondage : $\pi_i = \pi(\mathbf{u}_i, \beta)$, oú β est un paramêtre à estimer. Ainsi, les donnés nécessaires au calcul des probabilités dínclusion sont disponibles pour toute les unités, et un choix prudent de π permet d'avoir des probabilités d'inclusion moins sensibles aux probabilités de réponse et à leur estimation.

Naturellement, on va choisir les variables $\mathbf{u}_i$ parmi les variables qui semblent le mieux déterminer les probabilités de réponse. Dans le cas des enquêtes thématiques de l'OFS, le prédicteur le plus puissant est l'indicatrice de disponibilité d'un numéro de téléphone. On s'aperçoit qu'une bonne partie du gain en efficacité est déjà obtenu en choisissant simplement pour π_i la fonction de cet indicateur qui minimise V. Notons A l'ensemble des unités pour lesquelles on ne connaît pas de numéro de téléphone et $NA = U \backslash A$. On trouve la solution du problème de minimisation

sous contrainte :

$$\pi_A = \frac{C}{\sum_{i\in A} C_i} \frac{\left(\sum_{i\in A} 1/r_i \sum_{i\in A} C_i\right)^{\frac{1}{2}}}{\left(\sum_{i\in A} 1/r_i \sum_{i\in A} C_i\right)^{\frac{1}{2}} + \left(\sum_{i\in NA} 1/r_i \sum_{i\in NA} C_i\right)^{\frac{1}{2}}}, \text{ et}$$

$$\pi_{NA} = \frac{C}{\sum_{i\in NA} C_i} \frac{\left(\sum_{i\in NA} 1/r_i \sum_{i\in NA} C_i\right)^{\frac{1}{2}}}{\left(\sum_{i\in A} 1/r_i \sum_{i\in A} C_i\right)^{\frac{1}{2}} + \left(\sum_{i\in NA} 1/r_i \sum_{i\in NA} C_i\right)^{\frac{1}{2}}}.$$

La valeur de V pour ce choix est :

$$V_A = \frac{1}{C}\left[\left(\sum_{i\in A} 1/r_i \sum_{i\in A} C_i\right)^{\frac{1}{2}} + \left(\sum_{i\in NA} 1/r_i \sum_{i\in NA} C_i\right)^{\frac{1}{2}}\right]^2 - N.$$

Pour l'enquête MZB de 2016, la réduction de variance obtenue avec ce choix de probabilités de sélection par rapport à un plan de Bernoulli est estimée à un peu plus de 7%.

Les sommes dont dépendent π_A et π_{NA}, et dans le cas général le paramêtre β doivent être estimés. Si les $\widehat{r}_i$ ne peuvent pas être calculés pour toute la base de sondage, et si la population a notablement évolué depuis la derniére enquête, on peut imaginer de se rabattre sur un calage un peu hasardeux des données de cette enquête passée sur les marges de la nouvelle population.

5 Variables d'intérêt

Les calculs réalisés jusqu'ici ne nécessitent pas d'information sur les variables d'intérêt de l'enquête mais seulement sur les probabilités de réponse attendues.

L'optimisation proposée pourrait faire perdre de la précision pour une variable d'intérêt dans une configuration défavorable. En effet, la procédure conduit à sur-représenter 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.

Dans le cas de probabilités de sélection déterminées en fonction de la connaissance d'un numéro de téléphone, comme dans la section 4, on peut proposer des compromis entre réduction du budget, gains d'efficacité et tolérance à des rapports de dispersion défavorables entre la population A et la population NA.

Par ailleurs, la variance (1) peut aisément être estimée pour des variables d'intérêt dans le cas d'enquêtes répétées. On peut 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. Ceci a été fait pour l'Enquête Familles et Générations de 2018.

6 Taux de réponse

Le fait de sur-représenter 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 si l'on retenait l'allocation avec les probabilités d'inclusion "optimales".

Se pose alors la question du risque de biais d'estimation. Si la non-réponse résulte réellement d'un mécanisme aléatoire avec probabilités r_i, et en omettant le calage, on peut donner l'approximation au premier ordre du biais suivante :

$$\mathrm{B}(\widehat{Y}) \approx \sum_{i\in U} y_i \left[\frac{r_i}{\mathrm{E}(\widehat{r}_i)} - 1\right],$$

où $\mathrm{E}(\widehat{r}_i)$ désigne l'espérance sous le plan de sondage de l'estimateur de la probabilité de réponse.

Ce biais ne dépend donc pas des probabilités de sélection π_i, ou alors seulement au travers de leur importance pour une bonne estimation des probabilités individuelles de réponse.

Il nous faut convaincre nos collègues que ce n'est pas un bon taux de réponse global à l'enquête qui garantit des estimations (presque) sans biais, mais des probabilités de réponse individuelles r_i élevées (et une bonne estimation de ces r_i). Or, à l'OFS comme ailleurs certainement, le taux de réponse est un des premiers éléments relevés pour apprécier la qualité d'une enquête...

Références

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

[2] Graf, É., and Qualité, L. Sondage dans des registres de population et de ménages en Suisse : coordination d'échantillons, pondération et imputation. *Journal de la Société Française de Statistique 155*, 4 (2014), 95–133.

[3] Kim, J. K., and Kim, J. J. Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics 35*, 4 (2007), 501–514.