Optimisation des plans de sondage pour les enquêtes auprès des personnes et des ménages par l'utilisation de prédicteurs de non-réponse

Lionel Qualité et Clément Chevalier

Office Fédéral de la Statistique / Université de Neuchâtel

13es Journées de méthodologie statistique de l'Insee (JMS) | 12 juin 2018

Motivation

- La majorité des enquêtes de l'OFS sont réalisées (essentiellement) par téléphone,
- Un numéro de téléphone "connu" pour 75% de la population,
- Taux de réponse 2 à 3 fois plus bas chez les personnes que l'on doit contacter en premier par courrier,
- $\Rightarrow$ poids d'extrapolation 2 à 3 fois plus élevés pour celles-ci, et une précision dégradée.

Proposition

- Anticiper ces différences de taux de réponse,
- Calculer une allocation “optimale” de l’échantillon (sous contraintes de coût, etc.),
- $\rightarrow$ Compenser la non-réponse au moment du tirage en sur-échantillonnant les personnes qui répondent moins souvent,
- Comment ? Quels risques ?

Calcul rapide

- Taux de réponse avec téléphone $\approx 60\%$, sans téléphone $\approx 20\%$,
- Coûts : 1CHF pour la lettre d'annonce, 1CHF pour la carte-réponse, 100CHF par réponse, 0CHF pour les refus,
- On suppose la variable d'intérêt autant dispersée pour les personnes avec et pour les personnes sans téléphone connu,
- Plan de sondage et non-réponse simple à deux strates : avec ou sans téléphone connu,
- $\rightarrow$ allocation optimale $\approx 20\%$ plus efficace que sondage aléatoire simple.

Enquêtes de l'OFS - 1

- Sélectionnées depuis fin 2010 dans une liste de population et ménages (avant dans le répertoire téléphonique),
- On connaît l'âge, sexe, état-civil, nationalité, identifiant du bâtiment/logement... on retrouve un numéro de téléphone pour ≈75% de la population,
- Plusieurs enquêtes par téléphone (au moins pour une partie des questions ou pour une partie des répondants),
- Une grande enquête papier/internet reprend le questionnaire de recensement.

Enquêtes de l'OFS - 2

- Déroulement enquêtes téléphoniques : Lettre d'annonce puis appel téléphonique ou bien lettre d'annonce avec carte réponse pour fournir un numéro de téléphone, puis appel téléphonique,
- "E-survey" progressivement mis en avant quand c'est possible,
- Plans "simples" : taux de sondage uniforme dans les cantons pour les unités finales (personnes pour les enquêtes personnes, ménages pour les enquêtes ménages).
- Tirages indépendants des ménages (plan de Poisson ou Bernoulli), puis le cas échéant sélection d'une seule personne par ménage.

Pondération

- Démarche usuelle :
  1. Estimation de probabilités de réponse (modèle logistique ou recherche de "groupes homogènes de réponse") à l'aide des variables de la base de sondage, puis dilatation des poids de tirage,
  2. Calage sur marges fournies par la ou les bases de sondage (Deville et Särndal, 1992),
  3. *Combinaison des poids pour les enquêtes en plusieurs vagues.*
- Estimateur d'un total (simplifié) : $\widehat{Y} = \sum_{i \in r} w_i y_i$, où $w_i = \frac{1}{\pi_i} \times \frac{1}{\widehat{r}_i} \times g_i$; π_i est la probabilité de tirage, $\widehat{r}_i$ est l'estimation d'une probabilité de réponse r_i et g_i est le facteur d'ajustement résultat du calage.

Variance

- Approximation : Plan de Poisson et prise en compte du calage par linéarisation.
- Si l'on connaissait les r_i et les résidus de régression e_i des y_i sur les variables de calage :

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

- Souvent, l'estimateur $\widehat{Y}$ avec les $\widehat{r}_i$ a une variance plus faible que si l'on utilisait $\widehat{Y}_r$ et les "vrais" r_i (Kim et Kim, 2007),
- Les e_i dépendent des variables d'intérêt. On suppose ici qu'ils sont "échangeables" dans la population.

Optimisation - 1

- Variance (moyenne sous le modèle de super-population) proportionnelle à :

$$V = \sum_{i \in U} \frac{1}{\pi_i r_i} - N,$$

où N est la taille de la population U.

- Coût moyen pour une unité i sélectionée : $C_i = r_i C_i^r + (1 - r_i) C_i^{nr}$ (C_i^r si réponse, C_i^{nr} si non réponse).
- Contrainte de coût : $C = \sum_{i \in U} \pi_i C_i$.

Optimisation - 2

- Minimiser V :

$$\pi_i = \frac{C}{S}\frac{1}{\sqrt{r_i C_i}}, \text{ où } S = \sum_{k\in U}\sqrt{\frac{C_k}{r_k}}.$$

- Si des π_i dépassent 1, les bloquer à 1 et recommencer (ou trouver la solution directement).
- Sinon : $V_{opt} = S^2/C - N$
- pour le plan de Bernoulli de même coût, $V_{Bern} = A \times B/C - N$ où $A = \sum_{k\in U} C_k$, $B = \sum_{k\in U} 1/r_k$.

Estimation du gain possible

- Si l'on avait utilisé ces π_i dans une enquête passée : estimer A, B et S avec les données d'enquête,
- Options :
  - Sur l'échantillon de répondants avec les poids calés ou non,
  - Sur l'échantillon sélectionné, avec les poids de tirage (ou après calage),
  - Sur toute la base de sondage si l'on peut prédire des $\widehat{r_i}$ pour toutes les unités.
- Pour une nouvelle enquête, sélectionnée dans une base de sondage mise à jour : prédire des $\widehat{r_i}$ pour toutes les unités.

Quelques résultats

- Pour l'enquête sur la formation MZB, et en utilisant le modèle de réponse "de production" : V_{opt}/V_{Bern} estimé à 0.83 en 2011, à 0.86 en 2016 (les taux de réponse avec/sans téléphone se sont rapprochés en 5 ans),
- Pour l'enquête de recensement (enquête papier/internet avec un taux de réponse proche de 90%), gain estimé $< 3\%$ en 2016,
- $\Rightarrow$ intéressant s'il y a des groupes d'unités avec des taux de réponse faibles *et si on a confiance dans les estimations*.

Quelques limites

- Résultats dépendent beaucoup du modèle de réponse choisi, des $\widehat{r}_i$ (estimer $B = \sum_{k \in U} 1/r_k$!),
- Prédiction des $\widehat{r}_i$ pour une nouvelle enquête nécessite d'utiliser exclusivement des prédicteurs dans la base de sondage,
- Erreur ou aléa de prédiction : perte de précision si les probabilités de réponse estimées post-enquêtes sont différentes des probabilités prédites,
- Unités avec des probabilités de réponse prédites très faibles sont beaucoup plus échantillonnées : ça peut être bénéfique pour une enquête mais on risque de les lasser et faire encore baisser leur propension à répondre.

Deuxième proposition

- Le “calcul rapide” du début était plein de promesses, essayer d’y revenir :
- Chercher des probabilités de sélection fonctions (avec paramètres) de certaines variables choisies disponibles dans la base de sondage,
- Par exemple en prenant un modèle logistique “emboité” dans celui utilisé pour la non-réponse ou en arrêtant la procédure de segmentation plus tôt,
- Choisir les paramètres qui minimisent V (estimé sur l’échantillon précédent ou sur la base de sondage).

Application simple - 1

- Un seul prédicteur : ne pas connaître de numéro de téléphone ($i \in A$), ou en connaître un ($i \in NA$),
- Deux probabilités de sélection,

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

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

Application simple - 2

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

- MZB 2016 : $C \approx 1.2$mioCHF, téléphones connus $\approx 78\%$, $\bar{r}_{NA} \approx 0.51$, $\bar{r}_A \approx 0.255$,
- $\pi_A/\pi_{NA} \approx 1.95$, $V_A/V_{Bern} \approx 0.93$.

Variables d'intérêt

- Si la dispersion de la variable d'intérêt, ou bien $\sum e_i^2$, est plus faible parmi les personnes qui répondent mieux alors on peut perdre en précision,
- Avec le "calcul simple", possibilité de prévoir une tolérance : gain en précision par rapport à des probabilités égales tant que le rapport des dispersions est dans un intervalle choisi,
- Même possibilité pour un modèle différent sur les e_i^2 ?
- Optimisation pour une variable d'intérêt : facile à intégrer dans les calculs (laisser les e_i) mais on n'a des $\widehat{e}_i$ que pour les répondants $\rightarrow$ seule la "deuxième proposition" est utilisable.

Taux de réponse - 1

- Les calculs conduisent à sur-échantillonner les unités qui répondent moins,
- Mécaniquement cela fait baisser le taux de réponse apparent,
- Pour l'enquête MZB, avec les paramètres de 2016, il passerait de 45% à 39%,
- Biais ?

Taux de réponse - 2

- Approximation :

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

  $\mathrm{E}(\widehat{r}_i)$ est l'espérance de $\widehat{r}_i$ sous le plan de sondage.
- Ne dépend pas des π_i ou seulement au travers de $\mathrm{E}(\widehat{r}_i)$!
- Morale : on ne réduit pas le biais en sélectionnant les unités qui répondent le plus volontiers.

Références

- Deville, J.-C., and Särndal, C.-E., 1992, Calibration estimators in survey sampling. *Journal of the American Statistical Association 87 pp. 376-382.*

- Kim, J. K., and Kim, J. J., 2007, Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics 35, 4, pp. 501-514.*