Traitement du biais de sélection dans les enquêtes multimodes séquentielles

Odilon Saint-Cry DAKPAKETE, Marius DURAND-BARRIER,
Ilyass MESSABEL & TANO Anoumou Marc

2025-02-12

Contents

1 Introduction

Le processus de réalisation d'une enquête repose sur plusieurs étapes essentielles, telles que la formulation de la problématique, la conception des questionnaires, la mise en œuvre de la collecte des données et leur suivi. Chacune de ces phases peut générer des erreurs en l'absence d'une méthode rigoureuse. L'ensemble de ces erreurs constitue l'erreur totale de l'enquête.

Certaines erreurs proviennent de l'administration du questionnaire, c'est-à-dire des difficultés liées aux différents modes de collecte des données. Ce problème se manifeste particulièrement dans les enquêtes multimodes, qui combinent divers supports (Internet, téléphone, face-à-face, papier) afin d'optimiser le taux de réponse tout en maîtrisant les coûts.

Le développement rapide des technologies numériques a favorisé l'émergence des enquêtes multimodes, transformant ainsi les méthodes de collecte de données. L'utilisation simultanée de plusieurs modes permet de mieux répondre aux besoins d'une population de plus en plus diversifiée, tout en soulevant d'importants défis méthodologiques. Par exemple, le protocole séquentiel, souvent utilisé, propose d'abord un mode de collecte (généralement Internet), suivi d'un mode complémentaire (téléphone ou face-à-face) en cas de non-réponse. Bien que cette stratégie permette de maîtriser les coûts et d'augmenter le taux de réponse, elle expose l'enquête à deux types de biais majeurs :

- Le biais de sélection survient lorsque la probabilité de participation dépend directement de la variable d'intérêt (Y). Ainsi, le profil des répondants peut varier selon le mode de collecte, ce qui complique l'interprétation des résultats et limite l'efficacité des méthodes classiques de redressement.
- Le biais de mesure se traduit par des réponses divergentes d'un même individu selon le mode de collecte utilisé. Par exemple, lors d'une enquête téléphonique, un répondant peut être tenté d'ajuster ses réponses pour se conformer aux normes sociales, contrairement à ce qu'il ferait dans un questionnaire auto-administré.

Ce projet se concentre sur le biais de sélection dans le cadre d'une enquête multimode séquentielle combinant Internet et téléphone. Plus précisément, il vise à tester, au moyen de simulations de données, la validité des méthodes classiques de redressement en présence d'un biais de non-réponse non-ignorable. L'objectif est de déterminer jusqu'à quel point ces méthodes restent efficaces et à partir de quel seuil l'influence des variables non observables (U) compromet leur performance.

Pour ce faire, nous allons travailler sur les variables suivantes :

- X0 une variables auxiliaire connue pour les répondants et non-répondants ;
- X1 une variables sociodémographique connue uniquement pour les répondants ;
- Y la variable d'intérêt, connue uniquement pour les répondants : Y est expliquée par X0 et X1, mais aussi U

- U une variable non mesurée affectant la probabilité de réponse à Internet, au téléphone et Y, calculée pour les répondants et les non-répondants
- Pi la probabilité de répondre sur Internet
- Pt la probabilité de répondre au téléphone

2 Le concepte théorique

2.1 L'indicateur à estimer et les variables instrumentales

Pour commencer, nous faisons l'hypothése dans ce projet que Y est une variable numérique et nous nous intéressons à l'estimation de sa moyenne (μ). Dans le cas où nous connaissons toutes les valeurs des Y_i de la population, cette moyenne serait définit par la formule suivante :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Etant donné qu'il s'agit d'une enquête, nous ne connaissons pas toutes les valeurs de Y_i, nous devons donc estimer cette moyenne. Comme cela a été dit dans le document de travail (CASTELL, SILLARD, 2021), le plan de sondage est définie par le vecteur $S = (s_1, s_2, ..., s_N)$ avec s_i une variable aléatoire prenons la valeur 1 si l'individu i est sélectionné dans le plan de sondage et 0 sinon.

Ainsi,l'estimateur d'Horvitz-Thompson de la moyenne de Y est donné par la formule suivante :

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\pi_i} s_i$$

Cette estimateur est sans biais, car $\mathbb{E}[s_i] = \pi_i$ ce qui implique $\mathbb{E}[\hat{\mu}|y] = \mu$

Dans notre, il s'agit d'une enquête multimode séquentiel (internet, puis téléphone). Il nous faut donc définir deux variables instrumentale qui identifient les réponses par internet et les réponses par téléphone.

Soit z_i la variable aléatoire prenant la valeur 1 si l'individu i répond par internet et 0 sinon.

Soit w_i la variable aléatoire prenant la valeur 1 si l'individu i répond par téléphone et 0 sinon.

Nous disposons donc de deux vecteurs Z et W qui permettent respectivement d'identifier les réponses par internet et par téléphone. Ces vecteurs sont nos variables instrumentales.

Ainsi l'estimateur de la moyenne de Y peut être donné par la formule suivante :

$$\hat{\mu_1} = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\pi_i} s_i(z_i + w_i) = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{y_i}{\pi_i} s_i z_i + \sum_{i=1}^{N} \frac{y_i}{\pi_i} s_i w_i\right)$$

Cette d'estimateur est biaisé, car d'après le document de travail (CASTELL, SILLARD, 2021),$\mathbb{E}[s_i(z_i+w_i)] \neq \pi_i$ On pose $r_i = (z_i + w_i)$

Pour pouvoir corriger ce biais, nous introduisons $\hat{\rho}_1$ et $\hat{\rho}_2$ qui seront respectivement des modèles de z_i et de w_i de sorte que $\mathbb{E}[\hat{\mu_2}|y] = \mu$, c'est à dire nous allons estimer $\hat{\rho}_1$ et $\hat{\rho}_2$ de sorte que $\hat{\mu_2}$ soit sans biais.

$$\hat{\mu_2} = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{y_i}{\pi_i \hat{\rho}_1} s_i z_i + \sum_{i=1}^{N} \frac{y_i}{\pi_i \hat{\rho}_2} s_i w_i\right)$$