

Traitement du biais de sélection dans les enquêtes multimodes séquentielles

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

2025-02-12

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 3 |
| 2 | Le concept théorique | 4 |
| 2.1 | L'indicateur à estimer et les variables instrumentales | 4 |

1 Introduction

Le processus d'une enquête nécessite plusieurs étapes. Parmi ces étapes, on compte la réflexion sur la problématique, la conception des questionnaires jusqu'à la mise en place de la collecte des données et leur suivi. Chacune de ces étapes peut générer des erreurs, si une méthode rigoureuse n'est pas appliquée. On parle dans ce cas de l'erreur totale de l'enquête.

Parmi les composantes qui forment l'erreur totale, nous avons celles qui sont générées par l'administration du questionnaire, c'est-à-dire les erreurs causées par les différents modes de collecte des données, plus précisément dans les enquêtes intégrant plusieurs modes de collecte des données.

L'émergence des enquêtes multimodes, facilitée par le développement rapide des technologies numériques, a entraîné une transformation majeure dans les méthodes de collecte de données. Ces enquêtes combinent plusieurs modes de collecte, tels qu'Internet, le téléphone, le face-à-face ou encore le papier, afin d'améliorer les taux de réponse tout en réduisant les coûts. Elles répondent à un besoin croissant de flexibilité pour s'adapter à des populations de plus en plus diversifiées, mais elles soulèvent aussi des défis méthodologiques importants. Un protocole courant dans ces enquêtes est celui dit séquentiel, où un mode de collecte initial est proposé (souvent Internet), suivi d'un autre (téléphone ou face-à-face) en cas de non-réponse. Bien que cette stratégie permette de minimiser les coûts tout en augmentant le taux de réponse, elle est également associée à deux types de biais majeurs : le biais de sélection et le biais de mesure. Le biais de sélection survient lorsque la probabilité de participation à l'enquête dépend directement de la variable d'intérêt (Y), c'est lorsque le profil des répondants au questionnaire diffère d'un mode de collecte à l'autre. Ce qui complique l'interprétation des résultats et rend difficile l'utilisation des méthodes classiques de redressement. Le biais de mesure, quant à lui, se manifeste lorsque les réponses d'un même individu diffèrent en fonction du mode de collecte utilisé, par exemple entre un questionnaire auto-administré et une enquête téléphonique où le répondant va donner des réponses pour paraître dans les normes sociales.

Ce projet se concentre sur le biais de sélection dans un contexte séquentiel Internet-Téléphone. Plus précisément, il vise à tester les méthodes classiques de redressement en présence d'un biais de non-réponse non-ignorable, à l'aide de simulations de données. La question centrale est de déterminer jusqu'à quel point ces méthodes restent valides, et à partir de quel seuil d'influence des variables non observables (U) elles deviennent inefficaces.

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

- X_0 une variables auxiliaire connue pour les répondants et non-répondants ;
- X_1 une variables sociodémographique connue uniquement pour les répondants ;
- Y la variable d'intérêt, connue uniquement pour les répondants : Y est expliquée par X_0 et X_1 , 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
- P_i la probabilité de répondre sur Internet
- P_t la probabilité de répondre au téléphone

2 Le concept théorique

2.1 L'indicateur à estimer et les variables instrumentales

Pour commencer, nous faisons l'hypothèse 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 définie 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 $S = (s_1, s_2, \dots, s_N)$ avec s_i une variable aléatoire prenant 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 $\mathbb{E}[\hat{\mu}] = \mu$

Dans notre cas, il s'agit d'une enquête multimode séquentiel (internet, puis téléphone). Il nous faut donc définir deux variables instrumentales 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.

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