# 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.

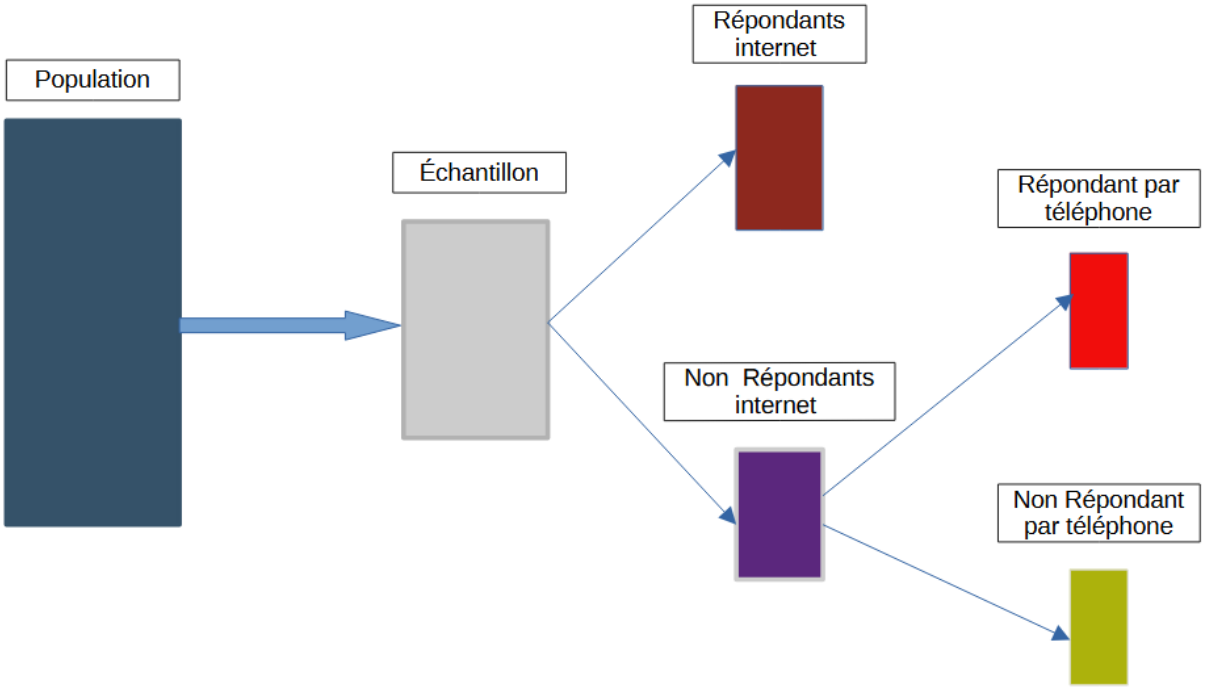



Figure 1: Le processus schématisé de l'enquête

## 2 Concept théorique

Dans ce projet, nous faisons l'hypothèse que Y est une variable numérique et nous cherchons à estimer sa moyenne ($\mu$). Étant donné que nous sommes dans un cadre de simulation, nous disposons de l'ensemble des valeurs $Y_i$ de la population simulée. La moyenne réelle de Y est alors définie par la formule suivante :

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

Nous réalisons ensuite plusieurs simulations à partir de cet échantillon total, en tenant compte de la présence de non-réponse. Dans ces cas, certaines valeurs $Y_i$, restent inconnues, ce qui nous oblige à estimer la moyenne. Comme mentionné dans Castell & Sillard (2021), le plan de sondage est défini par le vecteur $S = (s_1, s_2, ..., s_N)$ , où chaque $s_i$ est une variable aléatoire prenant la valeur 1 si l'individu i est sélectionné dans l'échantillon, et 0 sinon.

L'estimateur d'Horvitz-Thompson de la moyenne de Y est alors donné par :

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

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

Toutefois, dans notre étude, l'enquête est réalisée en mode séquentiel (Internet, puis téléphone). Si l'on applique une pondération classique sans distinction des modes de collecte, l'estimateur s'écrit :

$$\hat{\mu_1} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{\pi_i} s_i r_i$$

où $r_i$ est une variable indiquant si l'individu a répondu à l'enquête. Cependant, si l'on utilise une approche multiplicative qui distingue les modes de collecte, nous devons introduire deux indicatrices :

- $z_i$ : variable aléatoire valant 1 si l'individu i répond par Internet, 0 sinon.
- $w_i$ : variable aléatoire valant 1 si l'individu i répond par téléphone, 0 sinon.

On définit alors $r_i = (z_i + w_i)$, et les vecteurs Z et W permettent d'identifier les répondants selon leur mode de réponse. L'estimateur devient :

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

Cet estimateur est biaisé, car selon Castell & Sillard (2021), $\mathbb{E}[s_i(z_i + w_i)] \neq \pi_i$

Pour corriger ce biais, nous introduisons $\hat{\rho}_1$ et $\hat{\rho}_2$, qui modélisent respectivement $z_i$ et de $w_i$, de manière à garantir que $\mathbb{E}[\hat{\mu_2}|y] = \mu$. Nous estimons donc $\hat{\rho}_1$ et $\hat{\rho}_2$ de sorte que l'estimateur corrigé soit :

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

# 3 Simulation

## 3.1 Variables simulées

La mise en place d'un cadre de simulation rigoureux repose sur la maîtrise des corrélations entre les différentes variables. Trois types de variables sont simulées, à savoir:

– **Variables représentant la probabilité de réponse selon le mode de collecte**

- Pi: probabilité de répondre sur Internet
- Pt: probabilité de répondre au téléphone

– **Variable d'intérêt de l'enquête**

- Y : variable cible de l'enquête, observée uniquement pour les répondants.Elle est expliquée par X0 et X1, mais aussi U.

– **Variables explicatives**

- X0: variable auxiliaire disponible pour l'ensemble des répondants et non-répondants.
- X1: variable sociodémographique disponible uniquement pour les répondants.
  - U: variable instrumentale non observée influençant simultanément la probabilité de réponse par Internet, par téléphone et la valeur de Y. Cette variable est simulée aussi bien pour les répondants que pour les non-répondants.

Nous faisons ensuite varier la corrélation entre ces variables afin d'évaluer l'impact du biais dans l'estimation de Y.

La matrice de corrélation utilisée pour la simulation est la suivante :

....

Enfin, nous représentons les distributions des variables simulées à l'aide d'histogrammes pour mieux visualiser leurs caractéristiques.

### 3.2 Histogramme des variables simulées

### 3.3 Simulation de l'échantillon et des répondants

Une population simulée d'un million d'individus sert de base à l'étude. Un échantillon aléatoire de 10 000 individus est ensuite extrait pour l'enquête.

La réponse par internet est déterminée en comparant, pour chaque individu, sa probabilité de réponse $p_{\text{internet}}$ à un seuil aléatoire : si $p_{\text{internet}}$ dépasse ce seuil, l'individu répond, sinon il est reclassé comme non-répondant. Ces non-répondants sont ensuite sollicités par téléphone, suivant le même principe avec un seuil distinct. Ceux qui ne répondent à aucun des deux modes sont considérés comme absents de l'enquête.

Chaque individu se voit attribuer un seuil tiré d'une loi uniforme $U(0,1)$. Il est considéré comme répondant si sa probabilité de réponse dépasse ce seuil. Cette approche assure que le comportement simulé reflète fidèlement les probabilités de réponse définies.

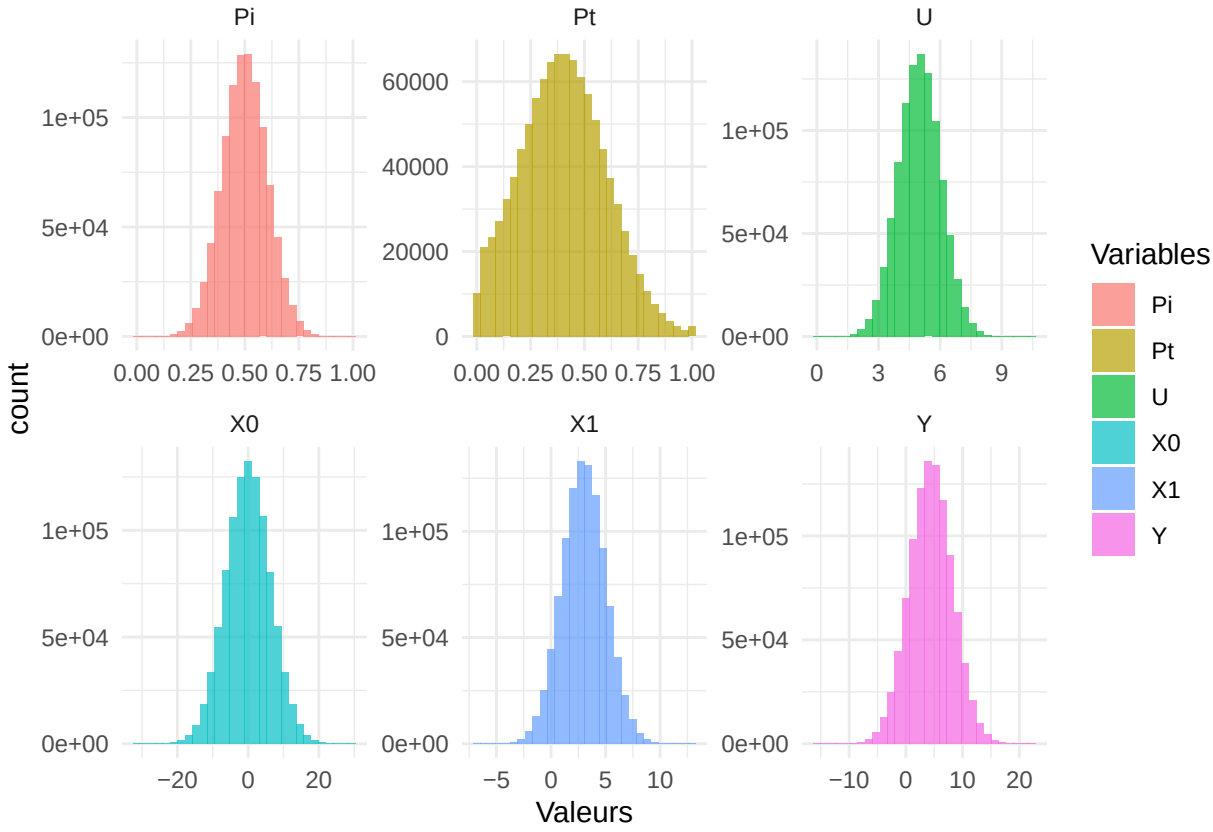
Le taux de réponse par mode est obtenu en faisant la moyenne des probabilités de réponse : pour Internet, sur l'ensemble de l'échantillon ; pour le téléphone, sur les non-répondants Internet.

## 4 correction des biais de selection

Dans notre cas, il s'agit de traiter le biais de non-réponse non-ignorable. Pour cela, nous allons utiliser les méthodes de redressement suivantes :

### 4.1 simulation des données de la population avec une forte influence de U sur Y, Pi et Pt

$$\begin{pmatrix} & X_0 & X_1 & U & P_i & P_t \\ X_0 & 1 & 0.3 & 0.2 & 0.9 & 0.8 \\ X_1 & 0.3 & 1 & 0.4 & 0.2 & 0.1 \\ U & 0.2 & 0.4 & 1 & 0.8 & 0.7 \\ P_i & 0.9 & 0.2 & 0.8 & 1 & 0.4 \\ P_t & 0.8 & 0.1 & 0.7 & 0.4 & 1 \end{pmatrix}$$



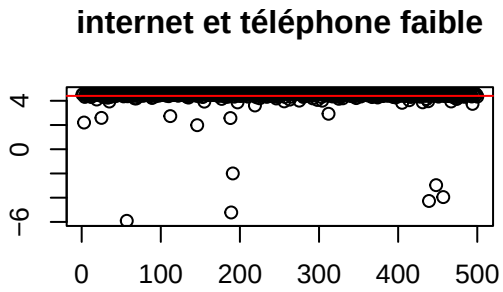

```
## [1] 4.343482
```

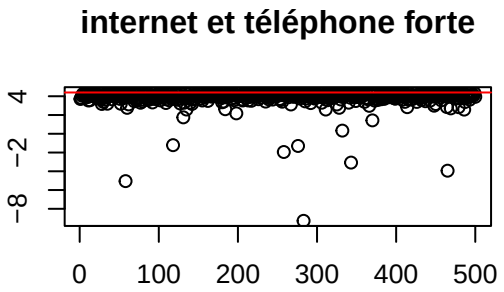
```
## [1] -5.909801
```

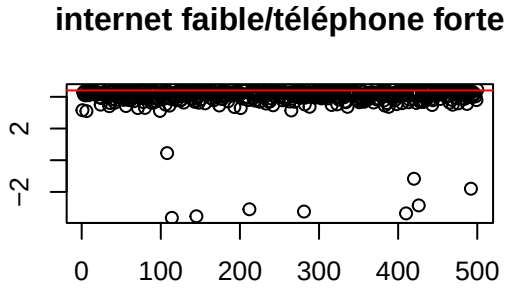
```
## [1] 4.698287
```

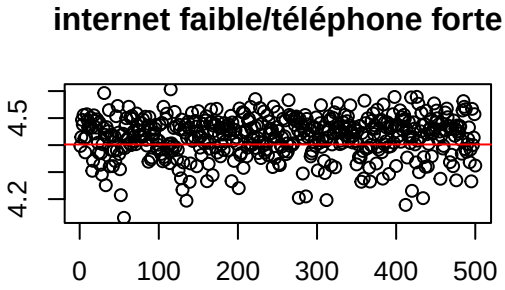
```
## [1] 0.9717
```

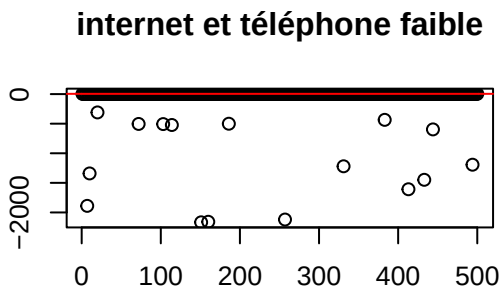
### 4.2 simulation des données de la population avec une forte influence de U sur Y

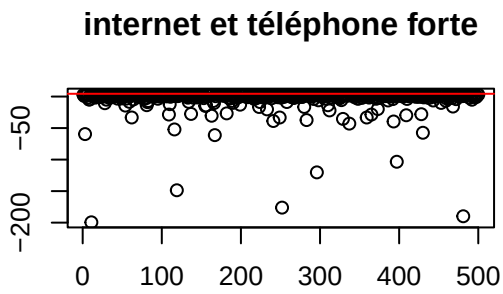


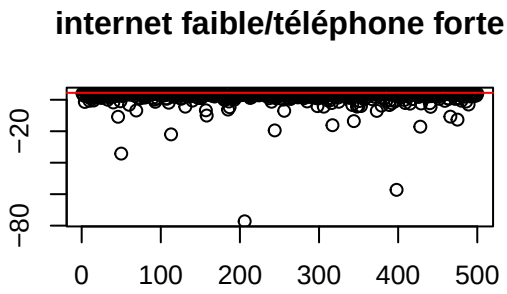



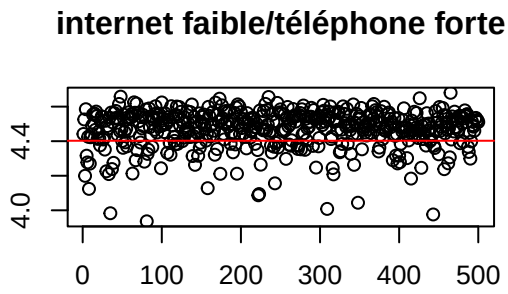



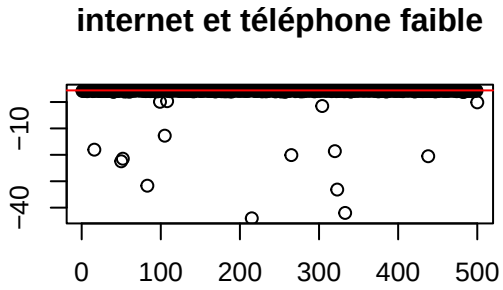



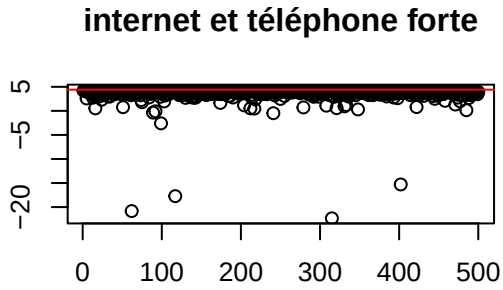

internet et téléphone faible
0
−2000
0
100
200
300
400
500

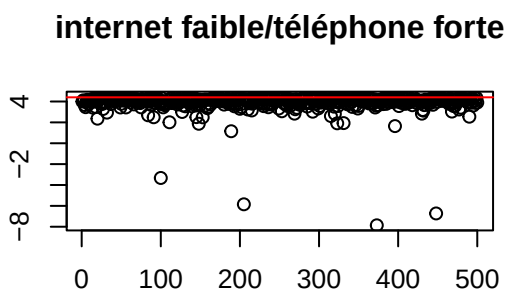

internet et téléphone forte
−50
−200
0
100
200
300
400
500

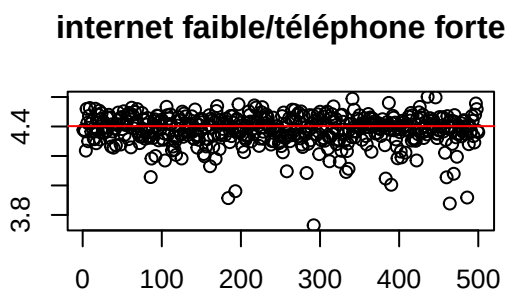

internet faible/téléphone forte
−20
−80
0
100
200
300
400
500


internet faible/téléphone forte
4.4
4.0
0
100
200
300
400
500




### 4.3 simulation des données de la population avec une forte influence de U sur Y et Pi