Introduction à la théorie des sondages - Cours 6

Antoine Rebecq
antoine.rebecq@insee.fr

INSEE, direction de la méthodologie

14 mars 2016

Sommaire I

Chapitre 1

La non-réponse

Partie 1

Non-réponse : définition et conséquences

Non-réponse : définition et conséquences

Non-réponse : incapacité d'obtenir des réponses utilisables, pour tout ou partie des variables d'intérêt.

On distingue 2 types de non-réponse :

- **la non-réponse totale :** on ne dispose d'aucune information sur l'unité sélectionnée autre que celles présentes dans la base de sondage
- **la non-réponse partielle :** l'unité sélectionnée répond seulement à une partie de l'enquête mais pas à l'ensemble des variables d'intérêt.

Causes de la non-réponse partielle

- Refus de répondre à certaines questions (jugées indiscrètes)
- L'individu enquêté ne comprend pas la question
- L'enquêteur ne comprend pas la réponse de l'enquêté
- Abandon du questionnaire en cours d'enquête

Causes de la non-réponse totale

- Refus de l'individu de réponde à l'ensemble de l'enquête
- L'individu n'a pas pu être contacté (déménagement, absence)
- Incapacité de répondre à l'enquête
- Abandon au tout début du questionnaire
- Charge statistique

Non-réponse : définition et conséquences

En pratique, la non-réponse entraîne pour les estimateurs relatifs aux variables d'intérêt :

- l'introduction d'un biais
- une diminution de la précision.

Non-réponse : définition et conséquences

Principal problème = biais.

Ne rien faire revient à supposer que les non-répondants ont un comportement identique à celui des répondants.

Or les refus se répartissent rarement de manière aléatoire dans la population, et les répondants présentent généralement des caractéristiques différentes de celles des non-répondants $\rightarrow$ estimateur biaisé.

Non-réponse : définition et conséquences

Également, perte de précision de l'estimation :

- en effet, le fichier exploitable in fine est de taille plus faible que le fichier tiré
- possibilité de pallier cette perte de précision en anticipant le taux de non-réponse et en gonflant la taille de l'échantillon

Non-réponse : définition et conséquences

Il est donc impératif de traiter la non-réponse :

- En amont, méthodes pour minimiser le phénomène
- En aval, méthodes de correction de la non-réponse

Partie 2

Correction de la non-réponse par imputation

Principe

Remplacer les valeurs manquantes par des valeurs “plausibles” (approche modèle).

Utilité

Méthode préférablement utilisée pour traiter la non-réponse partielle, mais peut également servir à corriger la non-réponse totale.

Diverses méthodes - déterministes

- Méthode déductive
- Cold-deck
- Moyenne, ratio, régression, tendance unitaire, etc.
- Plus proche voisin

Diverses méthodes - aléatoires

- Hot-deck aléatoire
- Imputations avec résidus

Imputation déterministe vs. aléatoire

Les imputations déterministes faussent la distribution de la variable d'intérêt.

Les imputations aléatoires préservent la distribution des variables d'intérêt, au prix d'une augmentation de la variance due à l'imputation.

Le choix se fait en fonction des objectifs de l'enquête.

Classes d'imputation

En pratique, on forme des classes avant d'imputer afin de réduire le biais dû à la non-réponse.

Les classes doivent être homogènes par rapport aux probabilités de réponse et/ou à la variable d'intérêt.

Partie 3

Correction de la non-réponse par repondération

Observation : en cas de non-réponse totale

L'estimation des totaux est biasée à la baisse :

$$\hat{N}_{NR} = \sum_{k \in R} \frac{1}{\pi_k} < \sum_{k \in s} \frac{1}{\pi_k} = \hat{N}_{\pi}$$

Correction de la non-réponse par repondération

La repondération est une technique qui consiste à augmenter les poids des unités répondantes pour compenser les unités défaillantes. C'est une méthode utilisée pour corriger la **non-réponse totale**.

Utilité

Méthode uniquement utilisée pour traiter la non-réponse totale.
La non-réponse totale peut également être traitée par imputation, mais en général on préfère la repondération :

- Perturbe moins les liaisons entre variables
- Ne donne pas l'illusion qu'on a affaire à des données complètes
- Calculs de variance plus simples

Paragraphe 1

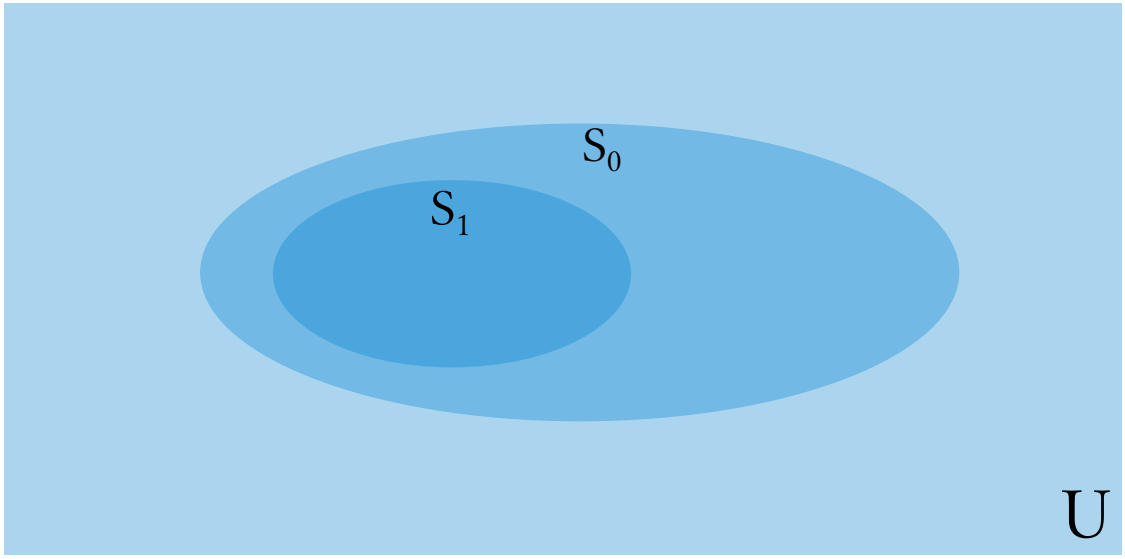
Plan en deux phases

Correction de la non-réponse par repondération

Un plan en deux phases est un échantillon issu d'un échantillon de la population.

La première phase consiste à tirer un échantillon s_0 dans la population $\mathcal{U}$, et la deuxième consiste à tirer l'échantillon $s_1 = s$ au sein de l'échantillon de première phase s_0.

Correction de la non-réponse par repondération



$$\mathcal{U} \xrightarrow[\pi_{0k}]{} s_0 \xrightarrow[\pi_{1k}]{} s_1 = s$$

Correction de la non-réponse par repondération

Théorème (Estimateur en expansion)

$$\hat{T}_{Y,2\phi} = \sum_{k \in s_1} \frac{y_k}{\pi_{0k}\pi_{1k}}$$

est un estimateur sans biais du total $T(Y)$.

Correction de la non-réponse par repondération

Démonstration.

$$\begin{aligned}\mathbb{E}(\hat{T}_{Y,2\phi}) &= \mathbb{E}\mathbb{E}\left(\sum_{k\in s_1}\frac{y_k}{\pi_{0k}\pi_{1k}}|s_0\right)\\ &= \mathbb{E}(\sum_{k\in s_0}\frac{y_k}{\pi_{0k}})\\ &= T(Y)\end{aligned}$$

Correction de la non-réponse par repondération

Attention : $\pi_{1k} = \Pr(k \in s | k \in s_0)$, donc en général $\pi_{0k} \cdot \pi_{1k} \neq \Pr(k \in s)$.

L'estimateur $\hat{T}_{Y,2\phi}$ n'est donc pas un estimateur de Horvitz-Thompson.

Correction de la non-réponse par repondération

Utilité du plan en deux phases (exemple des enquêtes VQS [1] et CARE [2]) :

- S'il est difficile de constituer une base de sondage
- Si l'on s'intéresse à une sous-population spécifique

1. Vie Quotidienne et Santé
2. Capacités, Aides et REssources des seniors

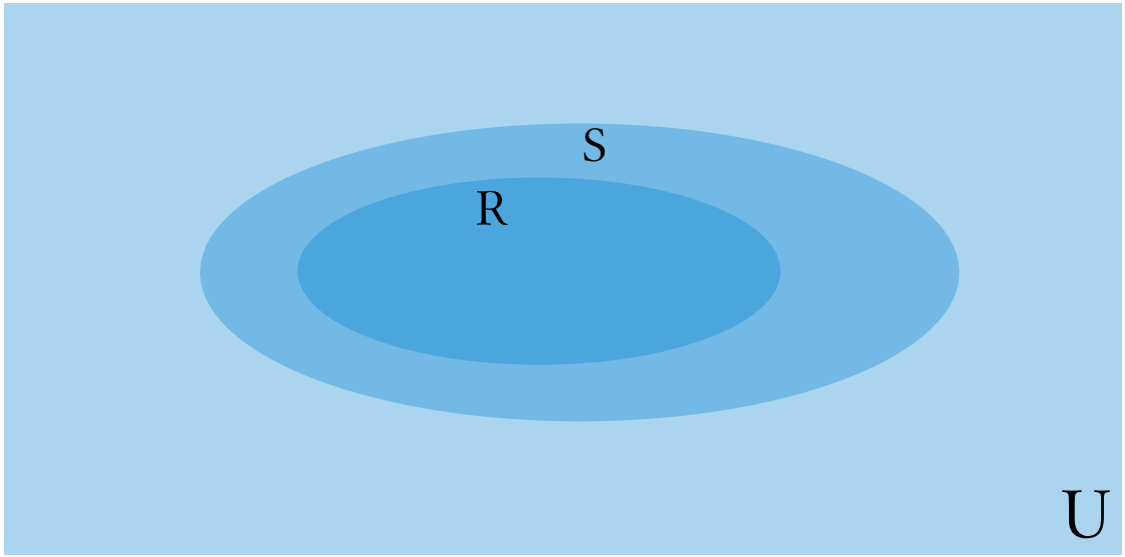
Paragraphe 2

Repondération

Plan en deux phases

La non-réponse peut-être considérée comme un tirage en deux phases (on note p_k la propension à répondre de l'unité k).

Plan en deux phases



$$\mathcal{U} \xrightarrow[\pi_k]{} s \xrightarrow[p_k]{} R$$

Plan en deux phases

Problème : Les p_k sont inconnues.

Principe de la repondération

On cherche à estimer le total : $T(Y)$.

En l'absence de non-réponse, on utilise l'estimateur d'Horvitz-Thompson : $\hat{T}_{Y\pi}$

En présence de non-réponse, un estimateur sans biais est donné par : $\sum_{k \in R} \frac{y_k}{\pi_k p_k}$, avec p_k estimées par un modèle.

Remarque : Comme on a $0 < p_k < 1$, cela conduit bien à augmenter les poids initiaux (de Horvitz-Thompson $\frac{1}{\pi_k}$)

Modélisation

Les p_k estimées par un modèle sont notées $\hat{p}_k$

- Modélisation par classe
- Régression logistique
- Méthodes d'apprentissage (supervisé) ...

Partie 4

Aspects théoriques de la non-réponse

Aspects théoriques de la non-réponse

On peut modéliser la non-réponse (totale) par plusieurs mécanismes.

Mécanisme MCAR

CMAR = Missing Completely At Random. Mécanisme de réponse globalement homogène.

Chaque individu a la même probabilité de répondre (taux de réponse empirique) : $p = \frac{n_r}{n}$. Dans le cas d'un SAS, l'estimateur d'une moyenne est inchangé : $\hat{\bar{Y}}_{CNR} = \frac{1}{n}\sum_{k\in s}\frac{y_k}{\frac{n_r}{n}} = \frac{1}{n_r}\sum_{k\in r} y_k$

On retient que "pour l'estimation des moyennes, ne pas corriger de la non-réponse revient à postuler un mécanisme globalement uniforme".

Mécanisme MAR

MAR = Missing At Random. Mécanisme de non-réponse ignorable.

Le mécanisme est uniforme par classes de variables $X_1 \dots X_j$.

Aspects théoriques de la non-réponse

Le calcul de la probabilité de réponse estimée $\hat{p}_k$ revient à modéliser le mécanisme de non-réponse.

L'estimation repondérée est sans biais dès lors que le mécanisme est MAR et que les variables déterminantes de la non-réponse sont correctement identifiées.

Mécanisme NMAR

NMAR = Non Missing At Random. Mécanisme de non-réponse non ignorable.

Même après conditionnement par toutes les variables auxiliaires disponibles, la probabilité de réponse reste dépendante de la variable d'intérêt Y de l'enquête. Ceci cause un **biais**.

Le problème est difficile à identifier et à traiter.

Chapitre 2

Autres aspects de la théorie des sondages

Sondage non probabiliste

Utilisé par les instituts marketing (sondage d'opinion, mesure d'audience)

Méthodes de collecte mises en œuvre ("échantillon opportuniste") :

- Méthode des quotas (marginaux)
- Méthode des quotas croisés

Sondage non probabiliste

Exemple d'une feuille de quotas :
`http://nesstar.ined.fr/quest/IE0186_Q4.pdf`

Sondage non probabiliste

Estimation : poids égaux (on est obligé de supposer que les unités sont tirées à probabilités égales)

Des techniques de redressement sont généralement utilisées.

Sondage non probabiliste

Le biais est composé de deux termes :

- Un terme dépendant de la différence de structure entre l'échantillon et la population (peut être annulé avec des quotas croisés)
- Un terme dépendant de la corrélation entre Y et la vraie probabilité d'inclusion

Sondage non probabiliste

En pratique, pour des échantillons de taille faible (< 1000), le sondage par quotas est préférable au sondage probabiliste :

- Erreur limitée dès lors que les variables sont bien choisies
- Coûts de mise en œuvre plus faible.

Mais nécessite des enquêteurs formés à la méthode.

Sondage non probabiliste

Estimation d'erreur obligatoire pour tout sondage marketing.
Fondée sur la variance d'un sondage aléatoire simple.

Exemple :
`http://www.odoxa.fr/barometre-politique-de-fevrier/`