La correction de la non-réponse par imputation

Thomas Deroyon & Cyril Favre-Martinoz

***Résumé* — L'objectif de cette note méthodologique est de décrire de façon rapide le principe de la correction de la non-réponse par imputation et les méthodes les plus fréquemment utilisées pour la mettre en oeuvre.**

I. Rappels sur les sondages aléatoires

Les enquêtes de la statistique publique sont réalisées sur des parties de la population totale des ménages ou des entreprises, appelées échantillons, sélectionnées aléatoirement. Cette méthode présente en effet de bonnes propriétés statistiques. Elle consiste à associer à chaque partie s de la population une probabilité $p(s)$ d'être sélectionnée, et de choisir la partie de la population qui sera interrogée en respectant ces probabilités. Le plan de sondage ainsi défini conduit à associer à chaque individu i de la population une probabilité π_i d'être interrogé, appelée probabilité d'inclusion.

Dans ce cadre, si l'on souhaite estimer le total sur la population U d'une variable d'intérêt y à partir de l'échantillon interrogé S, alors l'estimateur par expansion classique, appelé également estimateur de Sen-Horvitz-Thompson, défini par

$$\hat{Y}_S = \sum_{i \in S} \frac{y_i}{\pi_i} \tag{1}$$

est un estimateur sans biais sous le plan de sondage. Cela veut dire que sa moyenne sur l'ensemble des échantillons possibles, pondérée par leur probabilité d'être choisis, $\sum_{S \subset U} p(s) \hat{Y}_S$, est égale au vrai total de y sur la population $\sum_{i \in U} y_i$.
De plus, la variance de l'estimateur sous le plan de sondage, $\sum_{S \subset U} p(s) [\hat{Y}_S - \sum_{i \in U} y_i]^2$ peut être estimée à partir des données disponibles sur l'échantillon S, plus ou moins aisément suivant la complexité du plan de sondage.

II. La non-réponse : définition et conséquences

Un individu de l'échantillon est non-répondant s'il n'a pas été possible d'obtenir une information exploitable sur tout ou partie du questionnaire pour cet individu. Si l'ensemble du questionnaire ou une trop grande partie du questionnaire est inexploitable, l'individu est en **non-réponse totale** : il n'a fourni aucune information réellement utilisable. Si seules certaines questions sont inexploitables, l'individu est en **non-réponse partielle** :

A. Baisse de la précision

La précision des estimateurs calculés sur des échantillons aléatoires est en général inversement proportionnelle au nombre d'unités disponibles dans l'échantillon. Or, la non-réponse fait baisser la taille de l'échantillon exploitable et diminue de ce fait la précision des estimateurs. Ce problème peut cependant être en partie traité en amont, en anticipant le taux de réponse à l'enquête et en augmentant la taille de l'échantillon sélectionné. De cette façon, le nombre de répondants à l'enquête sera suffisant pour que les estimateurs satisfassent les objectifs ou les contraintes de précision imposées à l'enquête.

B. Biais d'estimation

Le deuxième problème que pose la non-réponse est le plus important : l'estimateur par expansion calculé sur les seuls répondants R, $\sum_{i \in R} \frac{y_i}{\pi_i}$, est biaisé. Ce biais a deux origines :

- **défaut de couverture** : la somme des poids de sondage $\frac{1}{\pi_i}$ sur l'échantillon est, en moyenne, égale à la taille de la population U. La somme des poids des seuls répondants est, par contre, toujours inférieure à la taille de la population. Ceci tient au fait que chaque unité de l'échantillon représente un certain nombre d'unités de la population. La non-réponse entraîne ainsi qu'une partie de la population n'est pas représentée par l'échantillon ;
- **biais de sélection** : les répondants sont susceptibles de différer des non-répondants. Ainsi, dans une enquête comme l'enquête sur l'emploi en continu qui a pour but d'estimer le taux de chômage, si les personnes non-répondantes sont plus souvent des personnes en emploi, la part des chômeurs parmi les répondants sera supérieure à la part effective dans la population. L'estimateur du taux de chômage [1] calculé sur les répondants sur-estimera le taux de chômage dans la population.

Les différentes méthodes de correction de la non-réponse ont pour but de limiter, voire supprimer, le biais qu'introduit la non-réponse. Il existe deux principales familles de méthodes :

- **les méthodes de répondération**, décrites dans la note méthodologique décrivant la correction de la non-réponse par repondération
- **les méthodes d'imputation**, décrites dans la suite de cette note

III. La correction de la non-réponse par imputation

A. Principe

Le principe des méthodes d'imputation est simple : il consiste à remplacer les valeurs manquantes des variables d'intérêt de l'enquête par des valeurs plausibles, construites

1. défini comme le nombre de chômeurs sur le nombre d'actifs, *i.e.* la somme du nombre de chômeurs et du nombre de personnes en emploi

en mobilisant une information externe à l'enquête, en s'appuyant sur les réponses données par les répondants à l'enquête ou en combinant informations fournies par les répondants et données extérieures. L'estimateur corrigé de la non-réponse du total de la variable y sur la population U est alors égal à

$$\hat{Y}_R^I = \sum_{i \in R} \frac{y_i}{\pi_i} + \sum_{i \in S-R} \frac{y_i^*}{\pi_i} \quad (2)$$

avec R l'ensemble des répondants et y_i^* la valeur imputée pour la variable y à l'individu i.

Les valeurs imputées sont construites en supposant qu'il existe un modèle, déterministe ou aléatoire, reliant dans la population les valeurs de la variable d'intérêt aux valeurs d'autres variables, appelées variables auxiliaires, disponibles pour les répondants et les non-répondants.[2] On suppose ainsi que les valeurs observées dans la population sont issues de ce modèle (appelé parfois aussi superpopulation). Les répondants à l'enquête sont utilisés pour estimer les paramètres du modèle (voir figure 1). Les valeurs imputées aux non-répondants sont ensuite obtenues en appliquant le modèle avec comme valeurs des variables auxiliaires celles observées pour les non-répondants et comme paramètres du modèle ceux estimés sur les répondants.

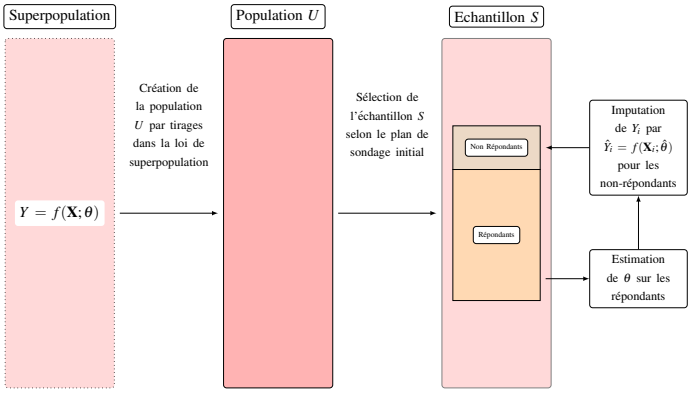

Fig. 1. La correction de la non-réponse par imputation

Sous ce principe général, il existe de nombreuses méthodes différentes d'imputation, que nous allons à présent lister rapidement.

IV. Les méthodes d'imputation

A. Classifications des méthodes

Il est possible de distinguer les différentes méthodes d'imputation existantes suivant deux classifications distinctes.

La première classification oppose :

- **les méthodes déterministes** : si l'on applique plusieurs fois la méthode d'imputation, la valeur imputée est toujours la même. Appartiennent à ce groupe l'imputation déterministe, le *colddeck*, l'imputation par la moyenne, par la médiane, par le ratio, par la régression, par la tendance unitaire ou encore par le plus proche voisin ;
- **les méthodes aléatoires** : la valeur imputée diffère à chaque application de la méthode. Relèvent de ce groupe les méthodes d'imputation avec résidu et le *hotdeck* aléatoire.

Il est également possible de classer les méthodes d'imputation en distinguant :

- **les méthodes par donneur** : la valeur utilisée pour l'imputation est la réponse fournie par un répondant à l'enquête. Figurent dans ce groupe l'imputation par le plus proche voisin et l'imputation par *hotdeck* ;
- **les méthodes par valeur prédite** : la valeur imputée n'est pas basée sur la réponse d'un seul répondant, mais mélange informations externes à l'enquête et réponses de plusieurs répondants.

B. Les méthodes déterministes

α. L'imputation déterministe

L'imputation déterministe consiste à exploiter les relations existant entre variables du questionnaire pour en déduire, de manière certaine ou sous des hypothèses raisonnables, la valeur à imputer. Elle permet par exemple d'imputer avec certitude un total non déclaré mais dont la ventilation suivant une typologie est renseignée. Cette méthode ne s'applique que dans peu de cas et uniquement pour la correction de la non-réponse partielle.

β. Le *colddeck*

L'imputation par *colddeck* consiste à remplacer la valeur manquante par une valeur issue d'une source externe. Elle est souvent utilisée dans les enquêtes auprès des entreprises pour imputer l'effectif salarié d'une entreprise ou le secteur lorsque ceux-ci sont demandés au début d'un questionnaire comme données de cadrage, en utilisant les valeurs renseignées dans le répertoire d'entreprises Sirene. Cette méthode suppose l'existence d'une source externe fiable, dans laquelle la variable à imputer est disponible et mesurée sur une période et suivant une méthode et des concepts proches de ceux de l'enquête.

γ. L'imputation par la moyenne ou par la médiane

Cette méthode consiste à remplacer la valeur manquante par la moyenne ou la médiane des réponses à cette variable fournies par les répondants. La méthode est en général appliquée en découpant la population en groupes disjoints, appelés classes d'imputation. Les réponses des répondants de chaque classe sont alors utilisées pour construire les valeurs à imputer aux non-répondants de la classe. L'imputation par la moyenne permet de créer des imputations qui respectent les relations linéaires existant entre variables (par exemple des contraintes comptables) mais les moyennes imputées sont sensibles aux réponses atypiques. A l'inverse, l'imputation par la médiane est robuste aux

2. Ces variables sont issues de la base de sondage dans laquelle est tiré l'échantillon de l'enquête, ou de sources administratives appariées avec la base de sondage. Il peut également s'agir de paradonnées décrivant le processus de collecte de l'enquête.

réponses atypiques, mais conduit à des imputations qui ne respectent pas les relations linéaires pouvant exister entre variables à imputer. La méthode est efficace si les classes d'imputation sont homogènes en termes de valeurs de la variable d'intérêt.

δ. L'imputation par le ratio

L'imputation par le ratio nécessite de disposer d'une variable auxiliaire quantitative pour les répondants et les non-répondants. Dans ce cas, la méthode consiste à calculer le ratio moyen ou médian entre la variable d'intérêt et la variable auxiliaire observé sur les répondants et à imputer pour remplacer la valeur manquante le produit de la valeur de la variable auxiliaire et de l'estimateur du ratio calculé sur les répondants. La méthode est le plus souvent appliquée à l'intérieur de classes d'imputation. Elle est efficace si la variable d'intérêt et la variable auxiliaire sont fortement corrélées, et si leur ratio est homogène à l'intérieur des classes d'imputation.

ε. L'imputation par la régression

L'imputation par la régression est une forme de généralisation de l'imputation par le ratio. Si l'on dispose de variables auxiliaires pour les répondants et les non-répondants, la méthode consiste à estimer sur les répondants un modèle de régression linéaire ou linéaire généralisée, suivant la nature de la variable à imputer, expliquant la variable à imputer par les variables auxiliaires. Les valeurs des variables auxiliaires pour les non-répondants et les paramètres du modèle estimés sur les répondants servent ensuite à construire des valeurs prédites pour chaque non-répondant qui remplacent les valeurs manquantes de la variable d'intérêt. Cette méthode suppose de disposer d'informations auxiliaires riches et est d'autant plus efficace que la relation est forte entre variable à imputer et variables auxiliaires et que la forme retenue pour le modèle d'imputation est correcte.

ζ. L'imputation par la tendance unitaire

L'imputation par la tendance unitaire s'applique aux enquêtes répétées dans le temps. La valeur imputée est obtenue en multipliant la réponse donnée par l'entreprise ou le ménage lors de l'édition précédente de l'enquête par un facteur d'actualisation, qui peut être calculé sur les répondants appartenant à la même classe d'imputation que le non-répondant, ou qui peut être le taux de croissance observé pour le non-répondant entre les deux périodes pour une variable auxiliaire corrélée à la variable à imputer.

η. L'imputation par le plus proche voisin

La méthode (voir [7]) consiste à définir une distance entre observations sur la base de variables auxiliaires disponibles sur les répondants et les non-répondants. La valeur imputée est alors la réponse donnée par le répondant le plus proche du non-répondant sur la base de cette distance. La méthode dépend fortement du choix de la distance retenue. La méthode par appariement sur la valeur prédite (*predictive mean matching*) est une forme d'imputation par le plus proche voisin qui nécessite deux étapes. Dans un premier temps, on construit sur les répondants un modèle explicatif de la variable à imputer en fonction des variables auxiliaires. Ce modèle est utilisé pour calculer une valeur prédite de la variable d'intérêt pour les répondants et les non-répondants. La distance entre observations est alors calculée comme le carré de la différence entre valeurs prédites de la variable à imputer. Il est également possible de construire un modèle expliquant le fait d'être répondant, pour la variable à imputer, en fonction des variables auxiliaires. La distance entre observations est alors égale au carré de l'écart entre probabilités de réponse prédites par le modèle.

C. Les méthodes aléatoires

α. Les méthodes à résidu

Les méthodes à résidu consistent à partir d'une méthode d'imputation déterministe, et à ajouter à la valeur imputée construite par cette méthode un résidu aléatoire. Ce résidu peut être déterminé de deux manières :

- soit les résidus sont tirés dans une loi paramétrique fixée *a priori*, par exemple une loi normale centrée de variance σ^2, les paramètres de la loi (dans l'exemple σ^2) étant estimés sur les répondants ;
- soit les résidus sont tirés aléatoirement parmi les erreurs de prédiction de la méthode d'imputation déterministe observées sur les répondants. Les erreurs de prédiction sont déterminées de la manière suivante : pour chaque répondant, on calcule l'écart entre la valeur de la variable à imputer effectivement observée et la valeur qui serait imputée au répondant en appliquant la méthode d'imputation déterministe.

β. Le *hotdeck*

Le *hotdeck* (voir [1]) consiste à choisir aléatoirement un répondant dont la réponse est utilisée pour imputer la valeur manquante. Cette méthode est en général appliquée à l'intérieur de classes d'imputation.

D. Comment construire les classes d'imputation ?

Les classes d'imputation (voir [6]) doivent être telles que les valeurs de la variable à imputer pour l'imputation par la moyenne, la médiane ou par *hotdeck*, ou le ratio entre la variable à imputer et la variable auxiliaire pour l'imputation par le ratio, sont homogènes et peu corrélées à la probabilité de répondre de chaque observation. Les classes d'imputation peuvent ainsi être construites de façon à ce que les valeurs de la variable à imputer observées pour les répondants y soient homogènes, ou sur des principes analogues à ceux présidant à la construction des groupes de réponse homogène (voir note méthodologique sur la répondération), en cherchant à construire des groupes à l'intérieur desquels l'hypothèse que toutes les observations, répondantes ou non-répondantes, ont la même probabilité de réponse est crédible.

V. Exemples

A. Système d'Information sur les Nouvelles Entreprises

Le système d'information sur les nouvelles entreprises (Sine) est une enquête réalisée tous les deux ans, dans laquelle un échantillon d'entreprises venant d'être créées est interrogé trois fois sur une période de cinq ans : la première fois au bout de quelques mois, la deuxième fois au bout de trois ans d'existence et la dernière fois au bout de cinq ans. Cette enquête permet d'étudier les caractéristiques des créateurs d'entreprises, les canaux par lesquels ils ont financé leur création et les difficultés qu'ils rencontrent. Elle permet également d'estimer les taux de survie à trois et cinq ans des nouvelles entreprises.

Dans les enquêtes Sine, la correction de la non-réponse, totale comme partielle, est réalisée par *hotdeck* (sauf pour les variables disponibles dans le répertoire d'entreprises Sirene, qui sont imputées par *coldeck*). Pour la correction de la non-réponse partielle, à chaque variable à imputer est associée une variable auxiliaire à laquelle elle est très corrélée ; les classes d'imputation sont définies comme l'ensemble des observations ayant les memes réponses pour la variable auxiliaire. Pour la correction de la non-réponse totale, les classes d'imputation sont construites à partir des variables auxiliaires corrélées au fait d'être répondant.

B. Enquête Patrimoine

L'enquête Patrimoine est réalisée tous les six ans et a pour but de mesurer de manière détaillée le patrimoine, matériel et financier, d'un échantillon de ménages français. Le questionnaire détaille ainsi l'ensemble des placements et comptes que peut détenir un ménage et demande à chaque fois si les membres du ménage interrogé en détiennent un. De plus, le questionnaire cherche à déterminer le caractère plus ou moins risqué de chaque placement, de manière à pouvoir étudier les comportements de placement des ménages en fonction de leurs autres caractéristiques (revenus, diplôme, catégorie sociale, ...) et l'évolution des risques assumés par les ménages en fonction de la conjoncture.

Les procédures d'imputation pour la correction de la non-réponse partielle dans l'enquête Patrimoine (voir [4]) doivent respecter les corrélations entre les diverses variables qualitatives mesurées dans l'enquête, par exemple entre la possession par les ménages d'un compte de titres et d'un plan d'épargne action, et les degrés de risque associés à chacun d'entre eux. Pour ce faire, différentes méthodes d'imputation ont été testées dans l'enquête : des méthodes par *hotdeck* où un même donneur est utilisé pour imputer plusieurs variables simultanément et des méthodes d'imputation jointe, où chaque variable est imputée dans sa loi conditionnellement aux valeurs des autres variables d'intérêt. Par exemple, la possession d'un compte titre est d'abord imputée, puis son degré de risque est imputé dans la distribution observée parmi les répondants ayant un compte-titre. L'indicatrice de possession d'un plan d'épargne action est ensuite imputée dans la distribution observée parmi les répondants ayant les mêmes valeurs de l'indicatrice de possession d'un compte titre et le même degré de risque pour celui-ci, le cas échéant.

VI. Conclusion : quelle méthode utiliser ?

Les méthodes par imputation servent à la correction de la non-réponse partielle. Elles peuvent également servir à la correction de la non-réponse totale, mais on leur préfère en général les méthodes par repondération (voir la note méthodologique sur la correction de la non-réponse par repondération).

Les méthodes d'imputation aléatoires respectent les distributions des variables imputées, ou les relations entre variables imputées et variables auxiliaires, mais génèrent un surcroît de variance par rapport aux méthodes déterministes. Des méthodes d'imputation aléatoires équilibrées ont récemment été developpées par Chauvet et al. (voir [5]) afin de préserver la distribution de la variable imputée tout en limitant la variance d'imputation. Il est également important de noter que la mise en œuvre de méthodes d'imputation mobilisant des variables auxiliaires n'est nécessaire que si le processus de sélection conduisant à l'échantillon de répondants est également expliqué par les variables auxiliaires mobilisées dans le modèle d'imputation. Autrement dit si la sélection des individus est complètement indépendante des variables auxiliaires mobilisées dans le modèle d'imputation, non seulement la correction du biais de sélection sera très faible, mais l'estimation finale pourrait être moins précise du fait de la vaiance induite par le mécanisme d'imputation dansle cas d'une imputation aléatoire.
Les méthodes déterministes conduisent à des estimateurs des totaux ou moyennes des variables d'intérêt plus précis que les méthodes aléatoires, mais déforment les distributions des variables d'intérêt ou leurs corrélations avec les variables auxiliaires utilisées pour l'imputation ou entre variables imputées. Les méthodes par donneur permettent de générer aisément des valeurs imputées possibles pour les variables qui ne peuvent pas prendre n'importe quelle valeur (par exemple pour les variables qualitatives). De manière générale, le choix d'une méthode d'imputation dépend de la variable considérée, du nombre d'observations à imputer, de l'information auxiliaire disponible et de l'utilisation faite des données d'enquête. C'est pourquoi il est essentiel d'identifier dans les fichiers de données individuelles les observations et les variables imputées.

References

[1] Andridge, R., Little R. (2010) : A review of hot deck imputation for survey nonresponse, International Statistical Review, 78, 40-64.

[2] Bethlehem, J. (1988) : Reduction of non-response biais through regression estimation, Journal of Official Statistics, 4, 251-360.

[3] Caron, N. (2005) : La correction de la non-réponse par repondération et par imputation, Document de travail de la Direction des Statistiques Démographiques et Sociales de l'Insee, n°M0502.

[4] Chaput, H., Chauvet G., Haziza D., Salembier L., Solard J. (2012) : Procédure d'imputation jointe pour les variables catégorielles - une application à l'enquête Patrimoine 2010, Actes des Journées de Méthodologie Statistique, 2012.

[5] Chauvet, G., Deville, J.C., Haziza, D. (2011) : On balanced random imputation in surveys, Biometrika, 98, 459-471.

[6] Haziza, D. et Beaumont, J.-F. (2007). : On the construction of imputation classes in surveys, International Statistical Review, 75, 25-43.

[7] Vandershelden, M. (2005) : Homogamie et choix du conjoint - Traitement de la non-réponse, Imputation de variables qualitatives corrélées, Document de travail de la Direction des Statistiques Démographiques et Sociales de l'Insee, n°F0505.

Département des méthodes statistiques
Version n° 1, diffusée le 10 octobre 2017 .