La correction de la non-réponse par repondération

Thomas Deroyon

***Résumé* — L'objectif de cette note méthodologique est de décrire de façon rapide le principe de la correction de la non-réponse par repondération 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, $\sum_{s \subset U} p(s) \left[\hat{Y}_s - \sum_{i \in U} y_i\right]^2$ 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

A. Définition

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**.

B. Baisse de la précision

La variance 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 augmente de ce fait la variance 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.

C. 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 avec des poids non corrigés de la non-réponse surestimera 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 suite de cette note ;
- **les méthodes d'imputation**, décrites dans la note méthodologique sur la correction de la non-réponse par imputation.

III. La correction de la non-réponse par repondération

A. Principe

Le principe de la correction de la non-réponse par repondération (voir [2] et [9]) est d'augmenter les poids

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.

des répondants pour compenser le biais introduit par les non-répondants. Pour ce faire, la non-réponse est décrite comme un phénomène aléatoire. Tout se passe comme si chaque unité de l'échantillon avait une certaine probabilité, inconnue et non nulle, de répondre, ρ_i. Ainsi, la sélection des répondants dans l'échantillon peut être vue comme une phase additionnelle du plan de sondage (voir figure 1). Les répondants sont de fait sélectionnés dans la population totale en deux étapes : la sélection de l'échantillon S dans la population U, suivant un plan de sondage connu et maîtrisé ; puis la sélection des répondants dans l'échantillon, suivant un plan de sondage inconnu, que la répondération a pour objectif de décrire.

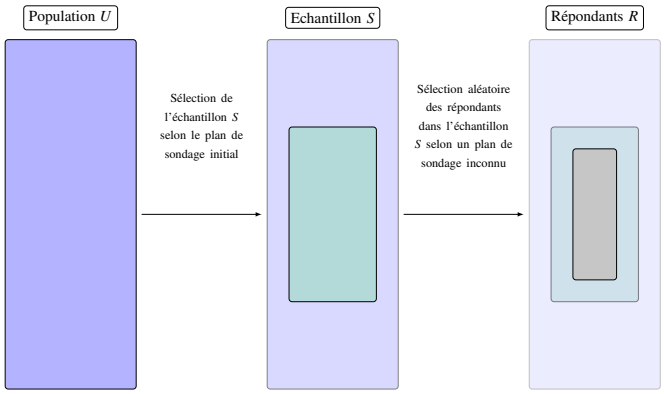

Fig. 1. La non-réponse comme phase additionnelle du plan de sondage

En effet, si l'on peut construire des estimateurs convergents des probabilités de réponse $\hat{\rho}_i$, l'estimateur corrigé de la non-réponse

$$\hat{Y}_R = \sum_{i \in R} \frac{y_i}{\pi_i \hat{\rho}_i} \tag{2}$$

est un estimateur asymptotiquement [2] sans biais sous le plan de sondage du total de y dans la population. Plusieurs méthodes sont fréquemment utilisées pour estimer les probabilités de réponse ρ_i. Nous n'évoquons dans la suite de cette note que les deux méthodes les plus utilisées dans les enquêtes de la statistique publique en France : la méthode des groupes de réponse homogène et le calage sur marges en une étape.

IV. Les groupes de réponse homogène (GRH)

A. Principe

Dans cette méthode (voir [3]), on suppose qu'il est possible de découper l'échantillon en parties disjointes, appelées groupes de réponse homogène, telles qu'à l'intérieur de ces groupes, toutes les unités de l'échantillon aient des comportements de réponse **indépendants** [3] et aient **la même probabilité de réponse**.

Dans chaque groupe, la probabilité de réponse commune est estimée soit comme le nombre d'unités répondantes divisé par le nombre total d'unités de l'échantillon appartenant au groupe, soit comme la somme des poids de sondage $1/\pi_i$ des unités répondantes divisée par la somme des poids des unités répondantes ou non-répondantes appartenant au groupe.

La méthode des groupes de réponse homogène est souvent considérée comme relativement robuste. En effet, l'estimateur corrigé de la non-réponse obtenu avec des groupes de réponse homogène peut être approximativement sans biais même si les hypothèses sur lesquelles repose la méthode, *i.e.* que toutes les unités d'un même groupe ont la même probabilité de réponse, est fausse.
En effet, on peut montrer (voir [1]) que le biais de l'estimateur obtenu avec des GRH est nul si la corrélation, dans chaque groupe, entre la variable d'intérêt dont on estime le total et la probabilité de réponse des unités est nulle.

Enfin, chaque groupe doit contenir suffisamment d'unités, répondantes ou non-répondantes, pour que la probabilité de réponse commune soit estimée avec assez de précision. Il n'existe pas de règle autre qu'empirique concernant la taille minimale des groupes : on recommande en général que chaque groupe contienne au moins 100 unités, et d'éviter dans tous les cas les groupes contenant moins de 50 unités.

B. Les méthodes pour construire des groupes de réponse homogène

La propriété évoquée dans la section précédente IV-A et démontrée dans [1] guide les méthodes de construction des groupes de réponse homogène. Ceux-ci doivent être des groupes dans lesquels soit la variable d'intérêt est homogène, soit la probabilité de réponse des unités est proche, pour limiter la corrélation entre ces deux variables dans le groupe. Comme les enquêtes ont de nombreuses variables d'intérêt, les GRH sont le plus souvent construits de manière à regrouper des unités dont les probabilités de réponse diffèrent peu. Pour ce faire, de nombreuses méthodes sont disponibles. Nous nous limitons à celles utilisées dans la statistique publique en France :

α. La méthode par croisements

La méthode consiste à identifier dans un premier temps, les variables auxiliaires qualitatives [4] disponibles au niveau individuel pour les répondants et les non-répondants [5] corrélées au fait d'être répondant. Les GRH sont constitués en croisant les modalités de ces variables. Ainsi, ils regroupent des unités entre lesquelles on ne peut plus mettre en évidence de corrélation entre le fait d'être répondant et les variables auxiliaires disponibles. On suppose de ce fait qu'il n'y a pas non plus, dans ces groupes, de corrélation entre le comportement de réponse et les

2. *i.e.* quand les tailles de l'échantillon et de la population tendent vers l'infini. L'estimateur est de ce fait approximativement sans biais dès que la population et l'échantillon sont de taille raisonnable.

3. *i.e.* le fait qu'une unité réponde n'a aucune incidence sur le comportement de réponse d'une autre unité du groupe.

4. Les variables auxiliaires continues, comme le revenu pour un ménage ou le chiffre d'affaires pour une entreprise, doivent être préalablement discrétisées.

5. Ces variables peuvent venir de la base de sondage, de fichiers administratifs appariés avec la base de sondage. Il peut également s'agir de paradonnées décrivant le processus de collecte.

variables mesurées dans l'enquête.

En pratique, les variables auxiliaires corrélées au comportement de réponse sont identifiées à l'aide d'une première étape de modélisation, par exemple par un modèle de régression logistique, qui permet de les classer de la plus à la moins corrélée. Les GRH sont ensuite construits itérativement, soit en croisant les modalités de toutes les variables et en regroupant, quand les groupes ainsi obtenus sont de taille trop faible, les modalités des variables les moins significativement corrélées ; soit à l'inverse en découpant l'échantillon suivant les modalités de la variable auxiliaire la plus significativement corrélée au fait d'être répondant, puis en découpant itérativement les groupes ainsi obtenus suivant les modalités des autres variables par ordre d'intensité de la corrélation avec le fait d'être répondant, tant que les groupes obtenus sont de taille suffisante.

β. Les arbres de classification : l'algorithme CHAID

L'algorithme CHAID (*Chi-square Automatic Interaction Detection*, voir [6]) est assez proche de la méthode par croisements. Il consiste à découper itérativement l'échantillon en groupes sur la base des modalités de la variable auxiliaire la plus corrélée au fait d'être répondant, celle-ci étant identifiée cette fois sur la base de tests de corrélation du χ^2.

γ. La méthode des quantiles

La méthode des quantiles (voir [5]), comme la méthode de Haziza et Beaumont, sont des méthodes des scores. Ces méthodes supposent deux étapes. Dans un premier temps, on construit une estimation des probabilités de réponse $\hat{\rho}_i$ *via* un modèle de régression logistique expliquant le fait d'être répondant par les variables auxiliaires disponibles sur les répondants et les non-répondants [6]. Les GRH sont ensuite constitués en regroupant les unités, répondantes ou non-répondantes, dont les probabilités de réponse estimées $\hat{\rho}_i$, sont proches.
Dans la méthode des quantiles, les GRH sont construits en se basant sur les quantiles de la distribution des probabilités de réponse. Si l'on construit par exemple 10 GRH, le premier GRH est formé de l'ensemble des unités dont les probabilités de réponse estimées sont inférieures au premier décile de la distribution des $\hat{\rho}_i$. Le nombre de GRH peut être déterminé en fonction de la taille souhaitée pour ceux-ci, ou sur la base d'une procédure analogue à celle proposée par Haziza et Beaumont.

δ. La méthode de Haziza et Beaumont

Les GRH sont construits (voir [7]) en appliquant un algorithme des centres mobiles, la distance entre unités étant définie comme le carré de la différence entre leurs probabilités de réponse estimées. Le nombre de GRH est déterminé en l'augmentant progressivement et en s'arrêtant au nombre le plus faible de GRH rendant compte d'une partie suffisante de la dispersion des probabilités de réponse estimées $\hat{\rho}_i$. Plus précisément :

- on construit d'abord deux GRH ;
- on estime ensuite la régression linéaire des probabilité de réponse estimées $\hat{\rho}_i$ sur les indicatrices d'appartenance aux GRH ;
- si le coefficient de détermination du modèle [7] est supérieur à un seuil fixé *a priori*, par exemple de 95 % ou 99 %, alors le modèle rend compte de 95 % ou 99 % de la dispersion des $\hat{\rho}_i$. On s'arrête donc à deux GRH. A l'inverse, si le R^2 du modèle est inférieur au seuil, on recommence le processus avec trois GRH ;
- on augmente le nombre de GRH jusqu'à obtenir des GRH rendant compte d'une part de la dispersion des $\hat{\rho}_i$ supérieure au seuil fixé *a priori*.

Les points de départ de l'algorithme peuvent être choisis au hasard, ou correspondre aux centres des groupes obtenus par la méthode des quantiles. Il est également possible d'appliquer l'algorithme avec plusieurs points de départ choisis aléatoirement et d'identifier les formes fortes, *i.e.* les ensembles d'unités qui appartiennent toujours aux mêmes groupes, quels que soient les points de départ de l'algorithme. Ces formes fortes sont ensuite regroupées en appliquant une classification ascendante hiérarchique.

V. Le calage sur marges

Le calage sur marges (voir la note méthodologique sur le calage sur marges) est en général appliqué à des poids permettant de construire des estimateurs sans biais. Si l'on connaît le total dans la population de variables, appelées variables de calage, mesurées dans l'enquête, le calage sur marges consiste à chercher les poids, appelés poids calés, les plus proches des poids d'origine et permettant d'estimer parfaitement les totaux des variables de calage. Les estimateurs construits avec les poids calés sont alors cohérents avec les informations déjà disponibles sur la population et plus précis pour les variables d'intérêt corrélées aux variables de calage.

Il est également possible d'utiliser le calage sur marges pour corriger la non-réponse (voir [10]). Cela revient à postuler que le fait d'être répondant dépend des variables de calage, *via* un modèle de régression linéaire généralisé dont la spécification dépend de la fonction de distance utilisée lors du calage. Pour que les poids calés permettent de construire des estimateurs sans biais, il faut alors que les variables expliquant le comportement de réponse fassent partie des variables de calage (voir [4]). Il faut également que la fonction de distance utilisée dans le calage sur marges corresponde au lien entre variables de calage et indicatrice de réponse. Haziza et Lesage (voir [8]) ont montré que, dans certains cas, notamment si l'une des variables de calage est une variable continue, l'utilisation d'un calage sur marges pour corriger la

6. D'autres techniques, par exemple de *machine learning*, comme le bagging, le boosting ou les forêts aléatoires, peuvent également être utilisées pour estimer les $\hat{\rho}_i$.

7. *i.e.* le ratio entre la variance expliquée par le modèle et la variance totale, parfois appelé R^2.

non-réponse pouvait conduire à une amplification du biais de non-réponse.

VI. Exemples

A. Enquêtes Sectorielles Annuelles

Les enquêtes sectorielles annuelles (ESA) servent à quantifier chaque année la décomposition des chiffres d'affaires des entreprises françaises suivant leurs différentes activités. Cette information permet de déterminer les comptes des entreprises par secteur, de réévaluer les secteurs auxquels appartiennent les entreprises répondantes et d'estimer enfin les matrices de passage secteur - branche essentielles pour la comptabilité nationale. L'échantillon comprend environ 160 000 entreprises, dont la moitié - les plus grandes - sont interrogées exhaustivement et l'autre moitié sélectionnée aléatoirement parmi les petites et moyennes entreprises françaises. Dans cette partie non exhaustive, le taux de réponse fluctue d'une année sur l'autre autour de 55 %.

La correction de la non-réponse totale dans la partie non exhaustive de l'échantillon des ESA est effectuée chaque année par la méthode des groupes de réponse homogène[8], construits par application de la méthode par croisements. Les variables les plus corrélées au comportement de réponse sont identifiées à l'aide d'un modèle de régression logistique, parmi un ensemble assez large de variables auxiliaires issues du répertoire d'entreprises (année de création, région d'implantation du siège, secteur, effectif, catégorie juridique) et des déclarations fiscales des entreprises (chiffre d'affaires, investissement brut ...), et classées par ordre décroissant sur la base de la variation du critère d'information d'Akaike que leur retrait du modèle entraîne. Les GRH sont ensuite constitués sur la base de la procédure itérative décrite plus haut. La méthode conduit à construire chaque année environ 500 GRH contenant au moins 50 entreprises.

B. Enquête Emploi en Continu

L'enquête emploi en continu (EEC) permet de décrire le marché du travail en France et notamment d'estimer le taux de chômage au sens du Bureau International du Travail (BIT). Depuis 2003, l'enquête est réalisée en continu sur l'année : chaque semaine, un échantillon de ménages est interrogé sur le statut au regard de l'activité de ses occupants au cours de la semaine. L'enquête interroge environ 100 000 personnes chaque trimestre. Le taux de réponse fluctue d'un trimestre à l'autre autour de 80 %.

La correction de la non-réponse dans l'Enquête Emploi est réalisée chaque trimestre par calage sur marges en une étape. Les variables de calage sont de deux types :

- des marges portant sur les logements : nombre total de logements, nombre de logements neufs, nombre de logements par type (maison, appartement), par nombre de pièces, par type de zone urbaine,...
- des marges portant sur la population, *i.e.* la pyramide des âges par sexe et par région[9] fournies par les données d'état civil et le recensement de la population.

VII. Conclusion : quelle méthode utiliser ?

La repondération ne peut être utilisée pour corriger la non-réponse partielle : elle pourrait conduire à un poids corrigé de la non-réponse différent par variable d'intérêt de l'enquête. De fait, la correction de la non-réponse partielle est effectuée par imputation.
Elle est par contre privilégiée pour la correction de la non-réponse totale par rapport à l'imputation, bien qu'il n'existe pas de supériorité théorique d'une méthode sur l'autre. La repondération ne nécessite cependant que de décrire le mécanisme de réponse, alors que, pour corriger la non-réponse totale par imputation, il faut définir un modèle d'imputation pour chacune des variables mesurées dans l'enquête. De plus, les calculs de précision des estimateurs sont plus simples lorsque la correction de la non-réponse totale a été réalisée par repondération.

Le calage sur marges en une étape peut présenter des risques, aussi la pratique recommandée est d'appliquer la procédure en deux étapes décrite par [8] : commencer par une correction de la non-réponse totale par repondération suivant la méthode des groupes de réponse homogène, puis appliquer un calage sur marges aux poids corrigés de la non-réponse pour améliorer la précision des estimateurs et diminuer les biais résiduels. Il n'existe pas de méthode de constitution de groupes de réponse homogène théoriquement supérieure aux autres. Il est ainsi recommandé, pour chaque enquête, de tester différentes méthodes et de choisir celle qui aboutit à la description du comportement de réponse la plus conforme à l'observé. Cela peut se faire, par exemple, en sélectionnant une fraction aléatoire (par exemple 2/3) de l'échantillon (appelée échantillon d'apprentissage) sur laquelle sont construits des GRH, puis d'appliquer les GRH ainsi obtenus sur le reste de l'échantillon (appelé échantillon de test) pour voir dans quelle mesure la méthode réussit à attribuer des probabilités de réponse élevées aux répondants et faibles aux non-répondants.

8. hormis pour les plus grandes entreprises, qui sont interrogées exhaustivement chaque année, et dont la correction de la non-réponse est réalisée par imputation.

9. avec un niveau de détail dans l'information mobilisée différent suivant les régions.

References

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

[2] Brick, J. M. (2013) : Unit non-response and weighting adjustment - a critical review, Journal of Official Statistics, 29, 329-353.

[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] Dupont F. (1993).; Calage et redressement de la non-réponse totale - Validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989, Actes des Journées de Méthodologie Statistique, 1993.

[5] Eltinge, J.L. et Yansaneh, I.S. (1997). : Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey, Survey Methodology, 23, 33-40.

[6] Kass, G. (1980). : A exploratory technique for investigating large quantities of categorical data , Applied Statistics, 29, 119-127.

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

[8] Haziza, D. et Lesage, E. (2014). : A discussion of weighting procedures for unit nonresponse, Journal of Official Statistics, 32, 129-145.

[9] Kalton, G. et Flores-Cervantes, I. (2003). : Weighting Methods, Journal of Official Statistics, 19, 81-97.

[10] Särndal, C.E. et Lundström, S. (2005). : Estimation in Surveys with Nonresponse, New York : John Wiley and Sons.

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