Non-Response Correction through Re-weighting

Thomas Deroyon

***Résumé* — The purpose of this methodological note is to provide a brief description of the principle of non-response correction through re-weighting and the methods most frequently used to implement that principle.**

I. Reminders Concerning Random Samples

Official statistics surveys are carried out on parts of the total population of households or businesses, known as samples, selected at random. In fact, this method offers good statistical properties. It consists of assigning to each part s of the population a probability $p(s)$ of being selected, and of selecting the part of the population to be surveyed in accordance with those probabilities. The sampling method thus defined results in assigning to each individual i in the population a probability π_i of being surveyed, known as the probability of inclusion.

In this context, if the aim is to estimate the total within the population U for a variable of interest y using the sample surveyed S, then the traditional expansion estimator, also known as the Sen-Horvitz-Thompson estimator, defined by

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

is an unbiased estimator under the sampling plan. This means that its average over all possible samples, weighted by their probability of being selected, $\sum_{s \subset U} p(s)\hat{Y}_s$, is equal to the true total of y across the population $\sum_{i \in U} y_i$.
In addition, the variance of the estimator under the sampling plan, $\sum_{s \subset U} p(s)[\hat{Y}_s - \sum_{i \in U} y_i]^2$ can be estimated based on the data available on the sample S, more or less easily depending on the complexity of the sampling plan.

II. Non-response : Definition and Consequences

A. Definition

An individual in the sample is classed as a non-respondent if it has not been possible to obtain usable information on all or part of the questionnaire for that individual. If the entire questionnaire or too large a part of the questionnaire is unusable, the individual is deemed to be a **total non-response** :he or she did not provide any information that is actually usable. If only certain questions are unusable, the individual is deemed to be a **partial non-response**.

B. Reduction in Accuracy

The variance of the estimators computed on random samples is generally inversely proportional to the number of units available in the sample. Non-responses decrease the size of the usable sample and thereby increase the variance of the estimators. However, this problem can be partly handled upstream, by anticipating the survey response rate and increasing the size of the sample selected. This will ensure that the number of respondents to the survey will be sufficient for the estimators to satisfy the accuracy constraints or objectives imposed on the survey.

C. Estimation Bias

The second problem posed by non-responses is the most significant : the expansion estimator based only on respondents R, $\sum_{i \in R} \frac{y_i}{\pi_i}$, is biased. This bias has two origins :

- **lack of coverage :** : the sum of the survey weights $\frac{1}{\pi_i}$ across the sample is, on average, equal to the size of the population U.The sum of the weights of respondents alone, however, is always less than the size of the population. This is due to the fact that each unit in the sample represents a certain number of units in the population. Non-responses therefore result in part of the population not being represented by the sample ;
- **selection bias** : respondents are likely to differ from non-respondents. Therefore, in a survey such as the continuous employment survey, the aim of which is to estimate the unemployment rate, if non-respondents are more often those in employment, the proportion of unemployed people among the respondents will be higher than the actual proportion within the population. An unemployment rate estimator [1] calculated based on respondents with non-response weights that have not been corrected will overestimate the rate of unemployment within the population.

The various methods of non-response correction are intended to limit or even eliminate the bias introduced by non-responses. There are two main method types :

- **re-weighting methods**, described below in this note ;
- **imputation methods**, described in the methodological note describing the correction of non-responses through imputation.

III. Non-response Correction through Re-weighting

A. Principe

The principle of the correction of non-responses through re-weighting (see [2] and [9])aims to increase the weight of respondents by compensating for the bias introduced by non-respondents. In order to achieve this, non-response is described as a random phenomenon. Each unit within the sample is considered to have a certain probability

1. Defined as the number of unemployed people within the active workforce, *i.e.* the sum of the number of unemployed and the number of people in employment.

(unknown but not zero) of responding, ρ_i. The selection of the respondents within the sample can therefore be viewed as an additional phase of the sampling design (see Figure 1). In fact the respondents are selected from the total population in two stages : the selection of the sample S from the population U, , in accordance with an existing and well-understood sampling plan, followed by the selection of respondents from within the sample in accordance with an unknown sampling plan, which the re-weighting aims to describe.

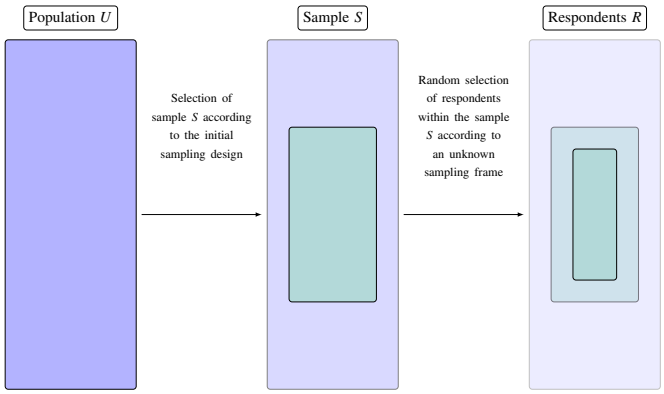

Fig. 1. Non-response as an additional phase of the sampling design

Indeed, if we are able to establish convergent estimators of the probabilities of response $\hat{\rho}_i$, the corrected non-response estimator

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

is an asymptomatically [2] unbiased estimator under the sampling plan for the total y within the population. Several methods are frequently used in order to estimate the probabilities of response ρ_i. In the remainder of this note, we will only discuss the two methods that are most frequently used for official statistical surveys in France : the homogeneous response groups method and one-step margin calibration.

IV. Homogeneous Response Groups (HRG)

A. Principe

With this method (see [3]), he assumption is made that it is possible to divide the sample into disjointed parts, known as homogeneous response groups, in such a way that all of the units from the sample within these groups display **independent** [3] response behaviour and have **the same probability of response**.

Within each group, the joint probability of response is estimated as either the number of responding units divided by the total number of units within the sample that belong to that group, or as the sum of the survey weights $1/\pi_i$ of the responding units, divided by the sum of the weights of the responding or non-responding units belonging to the group.

The homogeneous response groups method is often considered to be relatively robust. Indeed, the corrected non-response estimator obtained with homogeneous response groups may be more or less unbiased, even if the hypotheses on which the method is based, *i.e.* that all of the units within a single group have the same probability of response, is incorrect.
It is possible to demonstrate (see [1]) that the bias of the estimator obtained with HRGs is zero if the correlation, within each group, between the variable of interest for which the total is being estimated and the probability of response for the units is zero.

Finally, each group must contain a sufficient number of units, responding or non-responding, to ensure that the joint probability of response can be estimated with an adequate degree of precision. The only rules concerning the minimum size of the groups are empirical : it is generally recommended that each group contains at least 100 units and that groups containing fewer than 50 units are avoided under all circumstances.

B. Methods for Constructing Homogeneous Response Groups

The property referred to in the previous section IV-A and demonstrated in [1] guides the methods used to establish homogeneous response groups. These must be groups in which either the variable of interest is homogeneous or for which the probability of response for the units is similar, in order to limit the correlation between these two variables within the group. As the surveys have numerous variables of interest, HRGs are more often than not constructed in order to group together units for which there is little variation in the probability of response. There are many different methods available for this. We will only discuss those used for official statistics in France :

α. The cross tabulation method

The method consists of identifying, in the first instance, the qualitative auxiliary variables 1 available at the level of the individual for the respondents and the non-respondents [4] available at the level of the individual for the respondents and the non-respondents [5] correlated in accordance with whether or not they responded. The HRGs are established by cross tabulating the modes of these variables. They therefore group together the units for which the correlation between having responded and the auxiliary variables available are no longer evident. It is therefore assumed that there is no longer any correlation between the response behaviour and the variables measured by the survey within these groups.

In practice, the auxiliary variables correlated with the response behaviour are identified during an initial mo-

2. *i.e.* when the population sample sizes approach infinity. As a result, the estimator is more or less unbiased since the size of the population and the sample is reasonable.

3. *i.e.* the fact that one unit has responded does not have any influence over the response behaviour of another unit within the group.

4. Continuous auxiliary variables, such as household income or turnover in the case of a business, must be discretised in advance.

5. These variables can come from the sampling frame or from administrative files linked to the sampling frame. Paradata describing the collection process may also be used.

delling stage, for example using a logistic regression model, which allows them to be ranked from the most to the least correlated. The HRGs are then constructed by means of an iterative process, either by cross tabulating the modes of all of the variables and, where the groups obtained are too small, by grouping together the modes of the variables that display the least correlation ; or by dividing the sample in accordance with the modes of the auxiliary variable displaying the greatest correlation with the fact of having responded, then by iteratively dividing the groups obtained in this manner in accordance with the modes of the other variable in order of the degree to which they are correlated with the fact of having responded, until such time as the size of the groups obtained is adequate.

β. Classification trees : the CHAID algorithm

The CHAID (*Chi-square Automatic Interaction Detection*, see [6]) is fairly similar to the cross tabulation method. It involves dividing the sample into groups by means of an iterative process based on the modes of the auxiliary variable displaying the greatest degree of correlation with the fact of having responded, which, this time, is identified on the basis of tests of the correlation of χ^2.

γ. The quantiles method

Like the Haziza and Beaumont method, the quantiles method (see [5])is a scoring method. These methods are performed in two stages. Firstly, an estimate of the probabilities of response $\hat{\rho}_i$ *via* is established via a logistic regression model demonstrating the fact of having responded by means of the auxiliary variables available for the respondents and non-respondents [6]. HRGs are then established by grouping together the units, both respondents and non-respondents, for which the estimated probabilities of response $\hat{\rho}_i$, are similar. With the quantiles method, the HRGs are constructed on the basis of the quantiles of the distribution of the probabilities of response. If, for example, we establish 10 HRGs, the first HRG is formed from all of the units for which the estimated probabilities of response are below the first decile of the distribution of the $\hat{\rho}_i$. The number of HRGs can be determined on the basis of their desired size or on the basis of a procedure similar to that proposed by Haziza and Beaumont.

δ. The Haziza and Beaumont method

HRGs are established (see [7]) by applying an algorithm to the mobile centres, with the distance between units being defined as the square of the difference between their estimated probabilities of response. The number of HRGs is determined by increasing it progressively and stopping at the smallest number of HRGs taking account of an adequate share of the dispersion of the estimated probabilities of response $\hat{\rho}_i$. More specifically :

- we start by establishing two HRGs ;
- we then estimate the linear regression of the estimated probabilities of response $\hat{\rho}_i$ for the binary indicators associated with the HRGs ;
- if the coefficient of determination of the model [7] exceeds a threshold set *a priori*, for example of 95 % or 99 %, the model will take account of 95 % or 99 % of the dispersion of the $\hat{\rho}_i$. We therefore stop at two HRGs. Conversely, if the R^2 of the model falls below the threshold, we start the process again with three HRGs ;
- the number of HRGs is increased until we obtain HRGs taking account of a share of the dispersion of the $\hat{\rho}_i$ that is greater than the threshold set *a priori*.

The starting points for the algorithm can be selected at random, or they can correspond to the centres of the groups obtained by means of the quantiles method. It is also possible to apply the algorithm with multiple starting points selected at random and to identify strong trends *i.e.* groups of units that always fall into the same groups, regardless of the starting points used for the algorithm. These strong trends are then grouped together by applying an ascending hierarchical classification.

V. Margin Calibration

Margin calibration (see the methodological note on margin calibration) is generally applied to the weights, allowing unbiased estimators to be established. If the total within the population of variables, referred to as calibration variables, which are measured during the survey, is known, margin calibration consists of finding the weights, referred to as calibrated weights, that most closely match the original weights and that allow precise estimates to be made of the totals for the calibration variables. The estimators established using the calibrated weights are therefore consistent with the information that is already available within the population and more precise for the variables of interest correlated with the calibration variables.

It is also possible to use margin calibration to correct non-responses (see [10]). This amounts to assuming that the fact of having responded depends on the calibration variables, via a generalised linear regression model, the specification for which depends on the distance function used during calibration. To ensure that the calibrated weights allow for the establishment of unbiased estimators, it is essential that the variables explaining the response behaviour are included in the calibration variables (see [4]). The distance function used for margin calibration must also correspond to the link between the calibration variables and the response indicator. Haziza and Lesage (see [8]) demonstrated that, in some cases, particularly where one of the calibration variables is an ongoing variable, the use of margin calibration in order to correct non-responses could lead to an amplification of the non-response bias.

6. Other techniques, for example machine learning, such as bagging, boosting or random forests, can also be used to estimate the $\hat{\rho}_i$.

7. *i.e.* the ratio between the variance demonstrated by the model and the total variance, sometimes called R^2.

VI. Examples

A. Annual Sectoral Surveys

Annual Sectoral Surveys (ESAs) are used to provide a breakdown of annual turnover figures for French businesses by business area. This information makes it possible to determine the accounts of the businesses per sector, to re-evaluate the sectors to which the responding businesses belong and finally to estimate the sector–industry transfer matrices, which are essential to the national accounts. The survey involves approximately 160,000 businesses, of which half - the largest - are surveyed in detail, while the other half are selected at random from among small and medium-sized French businesses. Within this non-exhaustive segment, the response rate fluctuates at around 55 % from one year to the next.

The total non-response within the non-exhaustive segment of the sample for the Annual Sectoral Survey is corrected each year using homogeneous response groups [8], established by applying the cross tabulation method. The variables demonstrating the greatest degree of correlation with the response behaviour are identified using a logistic regression model from among a relatively large set of auxiliary variables taken from the business register (year of creation, region in which the head office is located, sector, workforce, legal category) and the tax returns submitted by the businesses (turnover, gross investment, etc.), and ranked in descending order on the basis of the variation in the Akaike information criterion resulting from their removal from the model. The HRGs are then established on the basis of the iterative procedure described above. The method results in the establishment of approximately 500 HRGs each year, each containing at least 50 businesses.

B. Labor Force Survey

The Labor Force Survey (EEC) allows the labour market in France to be described and in particular allows for an estimate to be made of the unemployment rate, as defined by the International Labour Office (ILO). Since 2003, the survey has been performed continuously throughout the year : each week, a sample of households is surveyed with regard to its status in view of the activity of its occupants during the course of the week. Approximately 100,000 people are surveyed each quarter. The response rate fluctuates at around 80 % from one quarter to the next..

The correction of non-responses within the Employment Survey is performed each quarter by means of single-step margin calibration. There are two types of calibration variables :

- margins relating to dwellings : total number of dwellings, number of new dwellings, number of dwellings by type (house, apartment), by number of rooms, by urban zone type, etc.
- margins relating to the population, *i.e.* the pyramid of ages by gender and by region [9] provided by civil registration and the population Census.

8. Except for the largest businesses, which are surveyed in detail each year and for which the correction for non-responses is performed by means of imputation.

9. With a different level of detail within the information used depending on the region.

VII. Conclusion : Which Method Should be Used ?

Re-weighting can only be used to correct partial non-responses : it could result in a different corrected non-response weight for each of the variables of interest of the survey. Partial non-responses are actually corrected by means of imputation.

By contrast, it is preferred over imputation when it comes to correcting total non-responses, even though, in theory, neither of the methods is considered to be superior. Nevertheless, re-weighting only requires the response mechanism to be described, whereas, in order to correct total non-response by means of imputation, it is necessary to define an imputation model for each of the variables measured by the survey. Furthermore, calculations to establish the precision of the estimators are more simple where total non-responses are corrected by means of re-weighting.

Single-step margin calibration can present risks, so the recommended approach is to apply the two-step procedure described by [8] : start by correcting the total non-response by re-weighting in accordance with the homogeneous response groups method, then apply margin calibration to the corrected non-response weights in order to improve the precision of the estimators and reduce the residual bias. In theory, none of the methods used to establish homogeneous response groups is superior to the others. It is therefore recommended that different methods are tested for each survey in order to choose the one that results in a description of the response behaviour that most closely matches that observed. This could be done, for example, by selecting a random fraction (for example 2/3) of the sample (referred to as the learning sample) on the basis of which the HRGs are established and then applying the HRGs obtained in this manner to the rest of the sample (referred to as the test sample) to see the extent to which the method has succeeded in assigning high probabilities of response to the respondents and low probabilities of response to the non-respondents..

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