Non-Response Correction through Imputation

Thomas Deroyon & Cyril Favre-Martinoz

***Résumé* — The purpose of this methodological note is to provide a brief description of the principle of non-response correction through imputation 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

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

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 Imputation

A. Principe

The principle behind imputation methods is simple : it involves replacing the missing values for the variables of interest in the survey with plausible values, created using information external to the survey, relying on the responses given by the respondents to the survey or combining information provided by respondents and external data. The corrected

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.

non-response estimator for the total variable y across the population U is then equal to

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

where R is all respondents and y_i^* is the imputed value for the variable y for the individual i.

The imputed values are created by assuming that there is a model, either deterministic or random, linking within the population the values of the variable of interest to the values of other variables, known as auxiliary variables, available for respondents and non-respondents.[2] It is thus assumed that the values observed in the population are derived from this model (sometimes also called a superpopulation model). Survey respondents are used to estimate the parameters of the model (see Figure 1). The values imputed to non-respondents are then obtained by applying the model, using the values observed for non-respondents as the values of the auxiliary variables and using the parameters estimated for respondents as the parameters of the model.

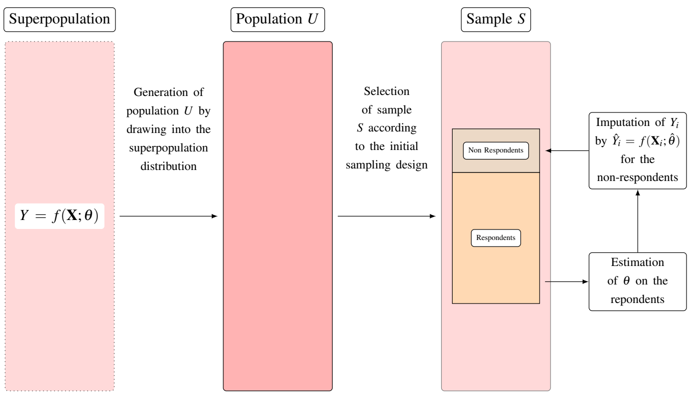

Fig. 1. Non-Response Correction through Imputation

Under this general principle, there are many different methods of imputation, which we will now quickly list.

IV. Imputation Methods

A. Classification of Methods

It is possible to make a distinction between the various existing imputation methods in accordance with two different classifications.

The first classification compares :

- **deterministic methods** : if the imputation method is applied several times, the value imputed is always the same. This group includes deterministic imputation, cold-deck imputation, mean imputation, median imputation, ratio imputation, regression imputation, unit trend imputation or even nearest-neighbour imputation ; ;

2. These variables are derived from the sampling frame from which the survey sample is drawn or from administrative sources matched to the sampling frame. Paradata describing the survey collection process may also be used.

- **random methods** : the imputed value differs each time the method is applied. This group includes the methods for imputation with residuals and random hot-deck imputation.

It is also possible to classify the imputation methods making a distinction between :

- **donor methods** :the value used for the imputation is the response provided by a survey respondent. This group includes nearest-neighbour imputation and hot-deck imputation ;
- **predicted value methods** : the imputed value is not based on the response of a single respondent, but mixes information external to the survey and the responses of multiple respondents.

B. Deterministic Methods

α. Deterministic Imputation

Deterministic imputation involves exploiting the relationships between the variables of the questionnaire to deduce from them, with certainty or under reasonable assumptions, the value to be imputed. It makes it possible, for example, to impute with certainty a total that is undeclared, but which has a breakdown provided according to a typology. This method only applies in a few cases and only for the correction of partial non-responses.

β. Cold-Deck Imputation

Cold-deck imputation involves replacing the missing value with a value from an external source. It is often used in business surveys to impute the number of employees of a business or the sector, where that number is requested at the beginning of a questionnaire as framing data, using the values entered in the SIRENE Business Register. This method assumes the existence of a reliable external source, in which the variable to be imputed is available and measured over a period of time and in accordance with a method and concepts close to those of the survey.

γ. Mean or Median Imputation

This method involves replacing the missing value with the mean or median of the responses for that variable provided by the respondents. The method is generally applied by dividing the population into separate groups, known as imputation classes. The responses of the respondents in each class are then used to create the values to be imputed for the non-respondents in the class. Mean imputation makes it possible to create imputations that respect the linear relationships existing between the variables (accounting constraints, for example), however the means imputed are sensitive to atypical responses. Conversely, median imputation is robust to atypical responses, but results in imputations that do not respect the linear relationships that may exist between variables to be imputed. The method is effective if the imputation classes are homogeneous in terms of the values of the variable of interest.

δ. Imputation by the Ratio

Imputation by the Ratio requires the availability of a quantitative auxiliary variable for respondents and non-respondents. In this case, the method involves calculating the mean or median ratio between the variable of interest and the auxiliary variable observed for the respondents and replacing the missing value by imputing the product of the value of the auxiliary variable and the ratio estimator calculated for the respondents. The method is most often applied within imputation classes. It is effective if the variable of interest and the auxiliary variable are closely correlated, and if their ratio is homogeneous within the imputation classes.

ε. Regression Imputation

Regression imputation is a form of general application of ratio imputation. If auxiliary variables are available for the respondents and non-respondents, the method involves estimating a linear or generalised linear regression model for the respondents, in accordance with the nature of the variable to be imputed, using the auxiliary variables to explain the variable to be imputed. The values of the auxiliary variables for the non-respondents and the parameters of the models estimated based on the respondents are then used to create predicted values for each non-respondent that replace the missing values for the variable of interest. This method requires the availability of a wealth of auxiliary information and is all the more effective the stronger the relationship is between the variable to be imputed and the auxiliary variables and the more the type of imputation model selected is correct.

ζ. Unit Trend Imputation

Regression imputation is a form of general application of ratio imputation. If auxiliary variables are available for the respondents and non-respondents, the method involves estimating a linear or generalised linear regression model for the respondents, in accordance with the nature of the variable to be imputed, using the auxiliary variables to explain the variable to be imputed. The values of the auxiliary variables for the non-respondents and the parameters of the models estimated based on the respondents are then used to create predicted values for each non-respondent that replace the missing values for the variable of interest. This method requires the availability of a wealth of auxiliary information and is all the more effective the stronger the relationship is between the variable to be imputed and the auxiliary variables and the more the type of imputation model selected is correct.

η. Nearest-Neighbour Imputation

The method (see [7]) involves defining a distance between observations based on the auxiliary variables available for the respondents and non-respondents. The imputed value is then the response given by the closest respondent to the non-respondent based on this distance. The method is highly dependent on the distance selected. The method involving matching to the predicted value (*predictive mean matching*) is a form of nearest-neighbour imputation that requires two stages. First, an explanatory model of the variable to be imputed is constructed for the respondents, in accordance with the auxiliary variables. This model is used to calculate a predicted value for the variable of interest for respondents and non-respondents. The distance between observations is then calculated as the square of the difference between predicted values of the variable to be imputed. It is also possible to construct a model explaining being a respondent, for the variable to be imputed, in accordance with the auxiliary variables. The distance between observations is then equal to the square of the difference between the probabilities of responding predicted by the model.

C. Random Methods

α. Residual Methods

Residual methods involve using a deterministic imputation method and adding a random residual to the imputed value created through this method. This residual can be determined in two ways :

- either the residuals are drawn in a parametric law that is fixed, a priori, for example a standard normal variance law , with the parameters of the law (in the example σ^2) being estimated for the respondents ;
- or the residuals are randomly drawn from the prediction errors of the deterministic imputation method observed for the respondents. The prediction errors are determined as follows : for each respondent, the difference between the value of the variable to be imputed actually observed and the value that would be imputed for the respondent is calculated by applying the deterministic imputation method.

β. Hot-Deck Imputation

Hot-deck imputation (seer [1]) involves randomly selecting a respondent whose response is used to impute the missing value. The method is generally applied within imputation classes.

D. How are the Imputation Classes Created ?

The imputation classes (see [6]) must be such that the values of the variable to be imputed for mean imputation, median imputation or hot-deck imputation, or the ratio between the variable to be imputed and the auxiliary variable for ratio imputation, are homogeneous and have little correlation with the probability of responding to each observation. The imputation classes can thus be constructed so that the values of the variable to be imputed observed for the respondents are homogeneous within them, or based on principles similar to those for the creation of homogeneous response groups (see the methodological note on re-weighting), by seeking to construct groups within which the assumption that all observations, for respondents or non-respondents, have the same probability of response is credible.

V. Examples

A. The Information System for new business companies

The Information System for new business companies (Sine) is a survey carried out every two years, in which a sample of newly created businesses is surveyed three times over a five-year period : the first time after a few months, the second time after three years of existence and the last time after five years. This survey makes it possible to study the characteristics of business creators, the channels through which they financed their creation and the difficulties they face. It also makes it possible to estimate the three- and five-year survival rates for new businesses.
In the SINE surveys, the correction of non-responses, both total and partial, is performed via hot-deck imputation (except for the variables available in the SIRENE business register, for which cold-deck imputation is used). For the correction of partial non-responses, each variable to be imputed is assigned an auxiliary variable to which it is very closely correlated ; the imputation classes are defined as the set of the observations all having the same responses for the auxiliary variable. For the correction of total non-responses, the imputation classes are constructed from auxiliary variables correlated to being a respondent.

B. Household Wealth Survey

The Household Wealth Survey is carried out every six years and aims to measure the material and financial wealth of a sample of French households in detail. The questionnaire thus details all the investments and accounts that a household may have and asks each time if the members of the household surveyed have any. In addition, the questionnaire seeks to ascertain the extent of the risk associated with each investment, so as to be able to study the investment behaviour of households in accordance with their other characteristics (income, level of educational attainment, social class, etc.) and the development of the risks assumed by the households in accordance with the economic situation.

The imputation procedures for the correction of partial non-responses in the Household Wealth Survey (see [4]) must respect the correlations between the different qualitative variables measured in the survey, between household ownership of a securities account and an equity savings plan, for example, and the levels of risk associated with each of them. In order to do this, various imputation methods have been tested in the survey : hot-deck methods, in which a single donor is used to impute several variables simultaneously, and joint imputation methods, in which each variable is imputed in its law subject to the other variables of interest. For example, securities account ownership is first imputed, then its level of risk is imputed in the distribution observed among respondents with a securities account. The equity savings plan indicator is then imputed in the distribution observed among respondents with the same values for the securities account ownership indicator and the same level of associated risk, where applicable.

VI. Conclusion : Which Method Should Be Used ?

Imputation methods are used to correct partial non-responses. They can also be used to correct total non-responses, but re-weighting methods are generally preferred for that purpose (see the methodological note on the Non-response correction through re-weighting).

Random imputation methods respect the distributions of the imputed variables, or the relationships between imputed and auxiliary variables, but they create more variance that the deterministic methods. Balanced random imputation methods have recently been developed by Chauvet et al. (see [5]) in order to preserve the distribution of the imputed variable while limiting imputation variance. It is also important to note that implementation of imputation methods using auxiliary variables is only necessary if the selection process leading to the respondent sample is also explained by the auxiliary variables used in the imputation model. In other words, if the selection of individuals is completely independent of the auxiliary variables used in the imputation model, not only will the correction of selection bias be very weak, but the final estimate could be less accurate due to the variance caused by the imputation mechanism, in the case of random imputation. Deterministic methods lead to more accurate estimators of total variables of interest or means for the variables of interest than random methods, but they distort the distributions of the variables of interest or their correlations with the auxiliary variables used for imputation or the correlations between imputed variables. Donor methods make it possible to easily generate potential imputed values for the variables that cannot take any value (qualitative variables, for example). In general, the selection of an imputation method depends on the variable under consideration, the number of observations to be imputed, the auxiliary information available and how the survey data is used. For this reason, it is essential to identify the observations and imputed variables in the individual data files.

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^o 1, diffusée le 10 octobre 2017 .