**JOINT EU/OECD WORKSHOP ON RECENT DEVELOPMENTS IN BUSINESS AND CONSUMER SURVEYS**

**Methodological session II**: Task Force & UN Handbook on conduct of surveys – response rates, weighting and accuracy

# UN Handbook Ch. 7 'Managing sources of non-sampling error': recommendations on response rates

**Mauro Politi – Roberto Gismondi (Istat – Italian National Institute of Statistics)**

Bruxelles - 14 November 2013

Istat

# List of topics

1. Introduction
2. Non sampling errors: Coverage
3. Non sampling errors: Measurement
4. Non sampling errors: Processing
5. Non sampling errors: Non responses - Introduction
    - 5.1 Response rates
    - 5.2 Tackling non responses
    - 5.3 Imputation criteria
6. Conclusions

# 1. Introduction

If the population total - concerning a generic target variable *y* - is estimated through a sample survey, the total estimation error (Mean Squared Error) is given by the sum of sampling and non sampling errors:

**$MSE = \sigma^2 + B$**

- $\sigma^2$ is the variance of the estimates for the universe based on a random sample
- B is the bias of the estimate

If random sampling is used, an estimate of $\sigma^2$ can be computed from the sample

The bias is the deviation between the true value and the expected value of the estimates and is the net effect of all the non-sampling errors mentioned in the following list

# 1. Introduction

**Non-sampling errors arise from many sources:**

- defects in the sampling frame because the business register is incomplete or out of date
- improper selection of the units to be sampled
- refusal be some selected units to provide information (total or partial refuse)
- mistakes when collecting and editing the answers or entering them into the data base (codification, registration, revision)

**The various kinds of errors:**

- Sampling error
- Non sampling errors
  - Coverage
  - Measurement and processing
  - Non response (Total or Partial)

# 2. Non sampling errors: Coverage

There are 4 main blocks of possible coverage errors:

- **Not completeness**: population includes some units which do not belong to the list from which the sample is drawn
- **Clusters of units**: the same name in the list is associated to more than one unit in the population
- **Unknown or not existing names**: the list contains some names that do not correspond to any unit in the population
- **Replicated names**: the population includes units to which correspond more than one name in the list

The main consequence of these errors is that they influence the real inclusion probabilities respect to the original sampling design

Not completeness: the bias depends on the share of units not included in the list and the difference between the *y*-means in the two subpopulations (belonging and not belonging to the list)

# 2. Non sampling errors: Coverage Telephone surveys

The list used for drawing the sample may be the national list of household which have a fixed telephone number

1) **over-coverage**: telephone numbers which correspond to second houses and professional activities; sample size lower than the desired one (solution: increase the theoretical sample size)
2) **under-coverage**: families without a fixed telephone, or with a fixed telephone but not present in the list

**Recommendations in both cases:**

- to estimate bias comparing average profiles (kind of municipality, age, sex,..) of the effective sample units and population units
- To merge the actual list with a second list (preliminary evaluation of the bias affecting the second list is recommended)
- To reduce bias through calibration estimators, able to reproduce given population totals (ISTAT: calibration for CATI/CAPI data)

# 3. Non sampling errors: Measurement (response) errors

**The observed value is different from the true one.** It can be due to:

1) behavior of the respondent unit (lack of capability to report correctly: enterprise instead of KAU, household instead of consumer,…)
2) The instrument used to get information (ambiguous phrasing of questions; unclear layout of questionnaire,…)
3) The effect to the interviewer and, in general, the kind of survey technique (insufficient knowledge to answer correctly; lack of motivation to report correctly,…)

**Detection of measurement errors**

- Comparing observed and true values (of course, when available)
- Replication of interviews using more expert interviewers

# 4. Non sampling errors: Processing errors

Processing errors may be introduced during:

- **data entry**
- **data editing**
- **data tabulation**

Errors at the data entry stage depend on the data collection method used. CAPI and CATI methods guarantee logical consistency and immediate controls. Questionnaire: simple and pre-tested!

In any data editing process, the main risk is that errors may be introduced by making the wrong adjustment. Editing should be done at the same time as the data are entered in the database. In general, the need for editing of tendency survey information is significantly less than that required for quantitative surveys

The risk of error at the data tabulation stage arises due to use of incorrect estimation criteria, or incorrect programs for processing the individual records

# 5. Non sampling errors: Non responses - Introduction

Non-response can be divided in two categories:

1) Item non-response (missing values): respondents don't complete the whole survey. Sometimes respondents are not willing or able to answer a certain question
2) Unit non-response (missing records): respondents are not able or willing to cooperate with the survey

Non responses have two main consequences:

a) increase of the sampling error, since the estimate variance increases if the number of respondent units decreases
b) effects on the non sampling error, due to the potential bias derived from the fact that the average profiles of respondent and not respondent units are different

Bias depends on the share of not respondents (in the population) and the difference between the *y*-means of respondents and not respondents. The bias is not dependent on the sample size

# 5.1 Non responses – Response rates

- Preventing unit non-response is a crucial element of data collection
- During the fieldwork period, a constant monitoring of the response rates is necessary
- Both weighted and un-weighted response rates are watched
- In order to improve response rates, re-interviews can be managed (large firms); non responding firms/consumers can be substituted at random
- The simplest non–response rate is the % of sample units from which no information on the target variable is available (units which have ceased to belong to the target universe - terminated, switched to another kind of activity - are not part of non–response). If $n_{obs}$ is the number of responding units, it is:

$$NR_1 = \frac{n - n_{obs}}{n}$$

# 5.1 Non responses – Response rates

- For sample surveys with different inclusion probabilities ($\pi_i$) for different enterprises/household, and for surveys where answers are weighted according to the size of the reporting units ($w_i$), the correct measure of the non respondent weight is given by:

$$NR_2 = \frac{\sum_{i=1}^{n-n_{obs}} \frac{y_i}{\pi_i} w_i}{\sum_{i=1}^{n} \frac{y_i}{\pi_i} w_i}$$

- The minimum response rate should be at least 50%. Without the use of a fixed panel, however, the response rate will need to be somewhat higher – 60% or 70%
- The combination of $NR_1$ with $NR_2$ can provide indication on the distribution of the collected responses across companies. When the un-weighted non-response rate ($NR_1$) is lower than the weighted $NR_2$, some of the largest enterprises did not respond

# 5.2 Non responses – Tackling non responses

1) A first **tool** for tackling non responses consists in **increasing the number of responses using call backs**. Call backs have a cost and they must be carried out within a limited time. Non respondents are often substituted by units chosen at random

2) As second methodology is **post-stratification**

   We suppose that propensity to respond is connected to levels of the variable object of interest

   We suppose to split population in *L* strata (*adjustment cells*), with quite different rates of response and average levels of the variable *y* inside. The post-stratified estimator will be:

$$\hat{\bar{y}}_{ps} = \sum_{h=1}^{L} (N_h / N)\, \hat{\bar{y}}_{obs,h}$$

   where $\hat{\bar{y}}_{obs,h}$ is an unbiased estimator of the respondents' mean in the population adjustment cell *h*

# 5.2 Non responses – Tackling non responses

Bias is given by: $B(\hat{\bar{y}}_{ps}) = \sum_{h=1}^{L}(N_h / N)\, NR_{1h}(\hat{\bar{y}}_{obs,h} - \hat{\bar{y}}_{notobs,h})$

where $NR_{1h}$ is the non-response rate in stratum *h*. Bias is minimized if: a) in the same stratum means of respondents and non respondents are quite the same; b) response rates are quite different among strata

3) A third tool is **re-weighting of respondents**

$$\hat{\bar{y}}_{rw} = \frac{\sum_{i=1}^{n_{obs}} y_i / \pi_i \alpha_i}{\sum_{i=1}^{n_{obs}} 1 / \pi_i \alpha_i}$$

where $\alpha_i$ is the response probability (estimated using l*ogit* or *probit* models), and $\pi_i$ is the inclusion probability. Their product is the estimated *final response probability*

# 5.3 Non responses – Imputation criteria

Imputations are predictions for the missing values. The goals are:

1. To obtain a completely filled data file: tabulation is easier
2. To increase the quality of the micro file and/or of the parameter estimates. However, this goal is not guaranteed: data variability may be reduced; non-response bias may not be reduced; covariance between two variables may be reduced

In deductive or logical imputation, you examine whether it is possible to derive the value of one or more of the missing variables from the values that were observed. For deductive imputation, it is not necessary to specify or to estimate models

Deductive imputation is then the most logical subsequent step

**Model-based and donor methods** can be used afterwards

For estimating the parameters, these methods can profit from the values already filled in deductively

# 5.3 Non responses – Imputation criteria

In **mean imputation**, a missing value is replaced by the mean value of respondents. Mean imputation leads to a peak in the distribution:

$$\tilde{y}_i = \bar{y}_{obs} \equiv \frac{\sum_{k \in obs} y_k}{n_{obs}}$$

where $y_k$ is the observed score of the *k-th* respondent and $n_{obs}$ is the number of item respondents for variable *y*

*y*-values can be weighted ($w_k$), for example due to differences in the inclusion probability. The resulting imputation is a less biased estimator of the population mean:

$$\tilde{y}_i = y_{obs}^{(w)} \equiv \frac{\sum_{k \in obs} w_k \cdot y_k}{\sum_{k \in obs} w_k}$$

This method is only recommended if no auxiliary information is available or when the available auxiliary variables are only marginally associated with the imputation variable

# 5.3 Non responses – Imputation criteria

In **group mean imputation**, a missing value is replaced by the mean *y*-value of units that have a valid value and are in the same stratum subpopulation as the item non-respondent. Consequently, in group mean imputation there are a number of smaller peaks:

$$\tilde{y}_{hi} = \bar{y}_{h;obs} \equiv \frac{\sum_{k \in h;obs} y_{hk}}{n_{h;obs}}$$

where $y_{hk}$ is the observed score of the *k-th* respondent in group *h* and $n_{h,obs}$ the number of item respondents for variable *y* in group *h*

Auxiliary information is used in group mean imputation, which involves classification into groups (subpopulations, imputation classes) based on one or more qualitative variables

The more homogeneous the subpopulations with respect to the variable to be imputed, the better the imputation (if the classification in subpopulations discriminates respondents and non-respondents)

# 5.3 Non responses – Imputation criteria

**Longitudinal imputation instead of the cross-sectional methods**

1) Earlier or later observations of the same object are very good predictors for the missing value. This means that the quality of the imputation can be strongly improved
2) To correctly estimate changes, it is important that the imputation takes into account previous and future values

**Last observation carried forward** (*LOCF*) is frequently applied because it is very easy. The last observed value of an individual is used for the values of all later periods that must be imputed:

$$\tilde{y}_i^t = y_i^{t-1}$$

This method is mainly applicable to categorical variables which change very little or not at all over time. For other variables, it often wrongly produces an overly stable picture of the situation

# 5.3 Non responses – Imputation criteria

**Ratio imputation** is frequently used for longitudinal data for which it is reasonable to assume that the observation at period *t* is proportional to the observation at period *t*-1

This method, which is frequently used in economic statistics, can be considered a refinement of last observation carried forward

$$\tilde{y}_i^t = \frac{\sum_{obs} y_i^t}{\sum_{obs} y_i^{t-1}} \cdot y_i^{t-1}$$

Just as in mean imputation, ratio imputation can be applied separately per subpopulation (imputation class)

This is done mainly if the ratios between the subpopulations vary strongly

Ratio imputation can be applied for quantitative variables

# 5.3 Non responses – Imputation criteria

## Donor estimation

Sample units are stratified, so that each stratum will contain only a small number of responding and not responding units

For each non respondent a donor unit is selected. If in the stratum no donor can be found, strata can be hierarchical enlarged until one donor is found. The donor can be selected:

- at random among respondents;
- minimizing distance between non respondent unit and donor. A common distance between units *i* and *j* is the Minkowsky one:

$$d_{ij} = \left( \sum_{k=1}^{p} \alpha_k \left| x_{ik} - x_{jk} \right|^s \right)^{\frac{1}{s}}$$

where $x_{ik}$ is the value of the *k-th* variable on the *i-th* unit, $\alpha_k$ is a weight used to standardize the *k-th* variable, *s* is a positive integer

# 5.3 Non responses – Imputation criteria

**Recommendations**

- For qualitative questions, the recommended method is to assume the same distribution over the response alternatives [(+), (=) and (–)] as the responding report units in that industry. For questions requiring answers in percentages or numbers, assume that the non–responding report units have the mean value of responding report units in that industry
- Institutes should describe in their metadata the nature of the procedures used in the treatment of missing data
- It is recommended that institutes closely monitor the impact of missing data (especially for large firms) and develop a clear set of strategies to minimize non-response
- The use of imputation methods for the treatment of missing data should be considered with care, in order to avoid possible bias

# 6. Conclusions

- The **quality report** on tendency surveys is not optional, but is itself a **basic part of the whole production process**

- Which are / should be the **population lists** from which the current samples may be drawn?

- Which are the **most common best practices** in the EU and in the World as regards **prevention** (*ex-ante*) of non responses?

- Which are the **most common best practices** in the EU and in the World as regards **re-weighting and imputation** for tackling non responses (*ex-post*)?

# Thank you for the attention

mauro.politi@istat.it
roberto.gismondi@istat.it