

Estimation of Interviewer Effect by Generalized Interpenetrating Sampling

La Stima dell'Effetto Intervistatore attraverso una Generalizzazione del Metodo dei Campioni Compenetranti

Piero Demetrio Falorsi, Matteo Mazziotta, Fabrizio Solari

Istat, via Magenta 4, 00185 Roma,
{falorsi, mazziott, solari}@istat.it

Riassunto: Nel presente lavoro viene proposta una rivisitazione della tecnica dei campioni compenetranti per la stima dell'effetto intervistatore in una indagine statistica. Le indagini reali condotte con tecnica telefonica non garantiscono che il numero di interviste per rilevatore sia costante rendendo, di conseguenza, non applicabili le usuali tecniche di stima dell'effetto intervistatore. Vengono proposte nuove espressioni per la stima della varianza totale e della componente correlata della varianza di risposta legata all'effetto intervistatore, permettendo in tale modo di individuare le fonti dell'errore statistico.

Keywords: interpenetrating sampling, interviewer effect, replicated sampling

1. Introduction

Istat carries out many surveys by means of Computer Assisted Telephone Interviewing (CATI) techniques. A typical feature of CATI is to have a non uniform distribution of the interviews among the interviewers. As shown in section 3, this event may cause a raising in non sampling error due to the interviewer effect. This effect tends to be higher for large departures of the interviews from uniform distribution among the interviewers. The interviewer effect is usually estimated by means of interpenetrating sampling techniques even if in literature they have only been applied for an equal number of interviews for interviewers. In this work, the authors propose a generalized approach of these techniques for the estimation of the interviewer effect in real situations.

2. Model for errors of measurement and interpenetrating sampling

We briefly describe the standard model used to specify errors of measurement (see, for instance, Cochran, 1977). It is supposed that for each unit a large number of independent measurements are possible. Let y_{it} be the value observed for the target variable Y in the t -th measurement. Then, for the i -th unit in the sample

$$y_{it} = y_i + \beta_i + d_{it}, \quad (1)$$

where y_i is the true value, β_i is the bias and d_{it} is the error of measurement with mean equal to zero, $Var(d_{it} | s) = \sigma_i^2$ and $Cov(d_{it}, d_{jt} | s) = \sigma_{ij}$.

Considering the Horvitz-Thompson estimator $\hat{Y}_t = \sum_{i=1}^n y_{it} / (N\pi_i)$ of the population mean, standard calculations lead to the following expression of the MSE of $\hat{Y}_t$

$$MSE(\hat{Y}_t) = Var(\hat{\mu}) + Var(\hat{d}_t) + 2Cov(\hat{\mu}, \hat{d}_t) + \bar{\beta}^2, \quad (2)$$

being $\hat{\mu} = \sum_{i=1}^n \mu_i / (N\pi_i)$, $\hat{d}_t = \sum_{i=1}^n d_{it} / (N\pi_i)$, $\bar{\beta} = \sum_{i=1}^N \beta_i / N$ and $\mu_i = y_i + \beta_i$. The covariance term in (2) is usually assumed to be zero but Fellegi (1964), Koch (1973), Sarndal *et al.* (1992) and Lessler and Kalsbeek (1992) show real situations where this condition is not satisfied. The first two terms are the total sampling variance $Var(\bar{y}_t)$, while $Var(\hat{d}_t)$ is usually referred as response variance. Assuming the conditional independence of moments, a general expression of the response variance is

$$Var(\hat{d}_t) = \frac{1}{N^2} \sum_{i=1}^N \frac{\sigma_i^2}{\pi_i} + \frac{2}{N^2} \sum_{i=1}^N \sum_{j>i}^N \sigma_{ij}^2 \frac{\pi_{ij}}{\pi_i \pi_j}. \quad (3)$$

Under SRS (3) can be written as:

$$Var(\hat{d}_t) = \frac{1}{n} \sigma_d^2 + \frac{n-1}{n} \sigma_{dd'}, \quad (4)$$

where $\sigma_d^2 = \sum_{i=1}^N \sigma_i^2 / N$ and $\sigma_{dd'} = 2 \sum_{i=1}^N \sum_{j>i}^N \sigma_{ij} / (N(N-1))$. The first term in (4) is known as simple response variance, while the second term is the correlated component of the total response variance.

Since under SRS $\hat{Y}_t = \bar{y}_t$, (2) can be expressed as:

$$MSE(\bar{y}_t) = \frac{1-f}{n} S_\mu^2 + \frac{1}{n} \sigma_d^2 + \frac{n-1}{n} \sigma_{dd'} + \bar{\beta}^2, \quad (5)$$

with $S_\mu^2 = \sum_{i=1}^N (\mu_i - \bar{\mu})^2 / (N-1)$.

It is known that standard estimators of $Var(\bar{y}_t)$ are usually biased since they do not take into account the correlated component of the total response variance. This is the case when more interviewers are utilised for data collection and they affect responses in the sense that there may be some correlation between errors of measurement within the same interviewer.

From now on we will suppose that the correlated component of the total response variance is entirely due to the interviewer effect, that is $Cov(y_{it}, y_{jt} | s) = \sigma_{ij}$ if i and j are assigned to the same interviewer and 0 otherwise.

Let n denote the sampling size and k the number of interviewers; it is known that if the same number $m = n/k$ of units is assigned to each interviewer, by means of interpenetrating sampling techniques, it is possible to derive unbiased estimates of the correlated component of the total response variance. The standard solution is given by $(m-1)(s_b^2 - s_w^2)/(mn)$, where $s_b^2 = n \sum_{l=1}^k (\bar{y}_{lt} - \bar{y}_t)^2 / (k(k-1))$ and $s_w^2 = \sum_{l=1}^k s_{lt}^2 / k$ are the between and within interviewers variance respectively and $s_{lt}^2 = \sum_{i=1}^m (y_{lit} - \bar{y}_{lt})^2 / (m-1)$ (see for instance Mahalanobis, 1946).

3. Interviewer effect with generalized interpenetrating sampling

Literature usually deals with the estimation of the correlated component of the total response variance when the same number of units is assigned to each interviewers. In this section we propose a generalisation of the standard results presented above. It is supposed that m_l sampling units have randomly assigned to the l -th interviewer ($l = 1, \dots, k$), being $\sum_{l=1}^k m_l = n$.

In this situation, the expression of the response variance (3) is

$$Var(\hat{d}_t) = \frac{1}{N^2} \sum_{i=1}^N \frac{\sigma_i^2}{\pi_i} + \frac{2}{N^2} \sum_{i=1}^N \sum_{j>i}^N \sigma_{ij}^2 \frac{\pi_{ij}}{\pi_i \pi_j} p_{ij}, \quad (6)$$

where p_{ij} is the probability that units i and j are assigned to the same interviewer. Referring to the SRS case, simple algebra shows that

$$\begin{aligned} p_{ij} &= \sum_{l=1}^k \frac{m_l}{n} \frac{m_l - 1}{n - 1} \\ &= \frac{1}{n-1} \left(\frac{\sigma_m^2}{\bar{m}} + \bar{m} - 1 \right), \end{aligned}$$

being $\bar{m}$ and σ_m^2 mean and variance of the m_l values respectively.

Therefore, the variance response (6) can be re-expressed as

$$Var(\hat{d}_t) = \frac{1}{n} \sigma_d^2 + \frac{1}{n} \left(\frac{\sigma_m^2}{\bar{m}} + \bar{m} - 1 \right) \sigma_{dd'}. \quad (7)$$

It is worthwhile to observe that, when different number of interviews are assigned to the interviewers, the response variance (7) is larger than in the equal number of interviews case since $\sigma_m^2 > 0$.

Furthermore, ignoring the finite population correction, the total sampling variance $Var(\bar{y}_t)$ of $\bar{y}_t = \sum_{l=1}^k m_l \bar{y}_{lt} / n$ assumes the following expression

$$Var(\bar{y}_t) = \frac{1}{n}S_\mu^2 + \frac{1}{n}\sigma_d^2 + \frac{1}{n}\left(\frac{\sigma_m^2}{\bar{m}} + \bar{m} - 1\right)\sigma_{dd'}, \quad (8)$$

Since the l -th sub-sample variance is $Var(\bar{y}_{lt}) = S_\mu^2/m_l + \sigma_d^2/m_l + \sigma_{dd'}(m_l - 1)/m_l$, ($l = 1, \dots, k$), expression (8) is equivalent to $Var(\bar{y}_t) = \sum_{l=1}^k m_l Var(\bar{y}_{lt})/n^2$.

Then, by means of replicated sampling techniques it is easy to show that an unbiased estimator of (8) is given by

$$var(\bar{y}_t) = \frac{1}{n^2 - \sum_{l=1}^k m_l^2} \left[\sum_{l=1}^k m_l (\bar{y}_{lt} - \bar{y}_t)^2 \right].$$

The estimation of the correlated component of the total response variance is based on $s_b^2 = n var(\bar{y}_t)$ and $s_w^2 = \sum_{l=1}^k s_{lt}^2/k$, with $s_{lt}^2 = \sum_{i=1}^{m_l} (y_{lit} - \bar{y}_{lt})^2/(m_l - 1)$. Since $E(s_b^2) = nV(\hat{\bar{y}}_t)$ and $E(s_w^2) = S_\mu^2 + \sigma_d^2 - \sigma_{dd'}$, simple algebra shows that an unbiased estimator of the correlated component of the total response is

$$\hat{\sigma}_{dd'} = \frac{n(s_b^2 - s_w^2)}{\sum_{l=1}^k m_l^2}. \quad (9)$$

The estimator (9) has been successfully applied to the re-interview survey of 2003 Farm Structure Survey (Falorsi *et al.*, 2005). The results show that for most of variables taken into account the correlated component of the total response variance results to be larger than the other components, displaying an clear interviewer effect.

References

- Cochran W.G. (1977) *Sampling Techniques*, Wiley, New York.
- Falorsi P.D., Mazziotta M., Solari F. (2005) La stima degli errori di misura nelle indagini sull'agricoltura: uno studio sull'indagine di controllo della SPA 2003, in: *Atti del convegno AGRI@STAT*, Firenze 30-31 maggio 2005.
- Fellegi I. (1964) Response variance and its estimation, *JASA*, 59, 1016-1041.
- Koch G. (1973) An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means, *JASA*, 68, 906-913.
- Lessler J.T., Kalsbeek W.T., (1992) *Nonsampling errors in surveys*, Wiley, New York.
- Mahalanobis P.C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute, *JRSS*, 109, 325-370.
- Särndal C.E., Swensson B., Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.