Proceedings of Q2006
European Conference on Quality in Survey Statistics

Estimation of the Interviewer Effect by Generalized Interpenetrating Sampling[1]

Matteo Mazziotta, Fabrizio Solari[2]

Keywords: interpenetrating sampling, interviewer effect, replicated sampling

1. Introduction

Remote data collection systems are more and more widespread for official data collection and they are more frequently used than paper and pencil interviewing. A typical example of remote instrument for data collection is the telephone interviewing and in particular the computer assisted telephone interviewing (CATI), which is an interactive front-end computer system helping interviewers to ask questions over the telephone. The interviewers are more personalized, probing questing and use of historic data are standardized, and the questions can be more sophisticated than those on paper questionnaires.
It is important to monitor and supervise the interviewing system since it may have an impact on the variance of the estimates. For instance, it is well known that not skilled interviewers can influence the responses in the sense that responses within the interviewers may be correlated. If this is the case and if the interviewers affect the responses in a systematic different way, response variances increase and the raising is referred as interviewer effect. Standard variance estimators do not take into account the interviewer effect and they are usually negatively biased.
Miller and Cannell (1982) observed that the differences in the interviewing systems may have an impact on the magnitude of the interviewer effect. Therefore, interviewing methods have to be taken under consideration when assessing the interviewer effect. Furthermore, a typical feature of telephone interviewing is to have a different number of interviews for the interviewer and, as it will be shown in section 3.2, the interviewer 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 and the use of random or mixed ANOVA models (Kish, 1962; Fellegi, 1974; Hartley, Rao and La Motte, 1978; Biemer and Stokes, 1985). The literature usually focuses on the balanced case, that is when the number of interviews for interviewer is set to be constant. In this work it is proposed a re-examination of interpenetrating sampling techniques for the estimation of the interviewer effect in statistical surveys in

[1] This work is partially supported by MIUR grant 2005 “L’informazione statistica in agricoltura: necessità attuali e sviluppi futuri”.
[2] Matteo Mazziotta, Italian National Statistical Institute, via Magenta 4, 00185 Roma, Italy (mazziott@istat.it); Fabrizio Solari, Italian National Statistical Institute, via Magenta 4, 00185 Roma, Italy (solari@istat.it).

the case of unequal number of units for interviewer. Furthermore, an expression for an unbiased estimator of the total variance is derived. Finally, the proposed estimation method will be applied to the Italian agricultural survey.

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

$$y_{it} = y_i + e_{it}, \qquad (1)$$

where y_i is the true value and e_{it} is the error of measurement in *t*-th measurement, with $E(e_{it} \mid s) = \beta_i$ and $Var(e_{it} \mid s) = \sigma_i^2$, the term β_i representing the bias in the measurements for unit *i*, so that $\mu_i = y_i + \beta_i$ is the mean of variable *Y* for unit *i* in the *t*-th measurement. Furthermore, the covariance $Cov(e_{it}, e_{jt} \mid s)$ between unit *i* and *j* in the *t*-th measurement is supposed to be equal to σ_{ij}. The magnitude of the components β_i, σ_i^2 and σ_{ij} will depend on the nature of the item being measured and on the measuring instrument.

The error of measurement e_{it} can be split in two components, the first given by the systematic error β_i and the second by the fluctuating component of error $d_{it} = e_{it} - \beta_i$. Let π_i be the inclusion probability of unit *i*, considering the Horvitz-Thompson estimator $\hat{\bar{Y}}_t = \sum_{i=1}^{n} y_{it} / (N\pi_i)$ of the population mean $\bar{Y}$, standard calculations lead to the following expression of the MSE of $\hat{\bar{Y}}_t$

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

being $\hat{\bar{\mu}} = \sum_{i=1}^{n} \mu_i / (N\pi_i)$, $\hat{\bar{d}}_t = \sum_{i=1}^{n} d_{it} / (N\pi_i)$ and $\bar{\beta} = \sum_{i=1}^{N} \beta_i / N$. 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 term $Var\left(\hat{\bar{\mu}}\right)$ is the standard sampling variance of the Horvitz-Thompson estimator of $\bar{\mu} = \sum_{i=1}^{N} \mu_i / N$. The sum of the first two terms in (2) represent the total sampling variance $Var\left(\hat{\bar{Y}}_t\right)$, while $Var\left(\hat{\bar{d}}_t\right)$ is usually referred as the total response variance. Assuming the conditional independence of moments, a general expression of the total response variance is

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

It is to show that under SRS (3) can be rewritten as

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

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

Introducing the average intrasample coefficient $\rho = \sigma_d^2 \big/ \sigma_{dd'}$, (4) is equivalent to

$$Var\left(\hat{\bar{d}}_t\right) = \frac{\sigma_d^2}{n}\left[1+(n-1)\rho\right].$$

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

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

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

3. The interviewer effect

Large scale surveys usually make use of more than one interviewer. In this situation the errors of measurement within the same interviewer are likely to be positively correlated and the weight of the correlated component of the total response variance may not be negligible. If this is the case, it is well known that standard estimators of $Var\left(\hat{\bar{Y}}_t\right)$ are strongly biased since they do not take into account the correlated component of the total response variance.

From now on we will suppose that the correlated component of the total response variance is entirely due to the interviewer effect, that is, denoting by p_{ij} the probability that units *i* and *j* are assigned to the same interviewer

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

This is equivalent to put in (3)

$$Cov(y_i, y_j | s) = \begin{cases}\sigma_{ij} & \text{if } i \text{ and } j \text{ are assigned to the same interviewer}\end{cases}.$$

otherwise

In the next section we will describe the standard procedure based on interpenetrating sampling techniques adopted in SRS for the estimation of the correlated component of the total response variance for the classic case when the number of interviews for interviewer is constant, while in section 3.2 we will propose an extension of standard results to the case of unequal number of interviews for interviewer.

3.1 Interviewer effect in the classic case

In this context a general technique, proposed by Mahalanobis (1946), for the estimation of the interviewer effect is based, after randomly assigning subsamples of the same sizes to each interviewer, on linear mixed models, precisely on random effects ANOVA models. These methods allow unbiased estimates of the total variance and the correlated component of the total response variance.
If *k* interviewers are involved in the survey and *m* units are randomly assigned to each interviewer, under SRS the probability p_{ij} that unit *i* and *j* are interviewed by the same interviewer is $(m-1)/(n-1)$. It is easy to show that (6) can be written as follows

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

and, therefore, considering negligible the finite population correction, the variance $Var(\bar{y}_t)$ assumes the following expression

$$Var(\bar{y}_t)=\frac{1}{n}S_\mu^2+\frac{1}{n}\sigma_d^2+\frac{m-1}{n}\sigma_{dd'}\,. \qquad (7)$$

The subsamples of units referring to each interviewer can be used as interpenetrating samples. For the *l*-th interviewer $(l=1,\ldots,k)$ the variance $Var(\bar{y}_{lt})$ of the Horvitz-Thompson estimator $\bar{y}_{lt}$ is

$$Var(\bar{y}_{lt})=\frac{1}{m}S_\mu^2+\frac{1}{m}\sigma_d^2+\frac{m-1}{m}\sigma_{dd'}, \qquad l=1,\ldots,k$$

Applying the replicated sample technique to the interpenetrating samples, each of them related to one interviewer, an unbiased estimator of (7) is given by

$$\mathrm{var}(\bar{y}_t)=\frac{1}{k(k-1)}\sum_{l=1}^{k}(\bar{y}_{lt}-\bar{y}_t)^2\,.$$

Standard solutions to compute estimates of the correlated component of the total response variance require within and between interviewer variances

$$s_b^2 = n\,\text{var}(\bar{y}_t) \text{ and } s_w^2 = \frac{1}{k}\sum_{l=1}^{k} s_{lt}^2 ,$$

where $s_{lt}^2 = \sum_1^m \left(y_{ljt} - \bar{y}_{lt}\right)^2 \big/ (m-1)$. Since $E\left(s_{lt}^2\right) = S_\mu^2 + \sigma_d^2 + \sigma_{dd'}$ $(l = 1, \ldots, k)$, then it is easy to show that an unbiased estimator of $\sigma_{dd'}$ is

$$\hat{\sigma}_{dd'} = \frac{1}{m}\left(s_b^2 - s_w^2\right)$$

and, hence, $(m-1)\left(s_b^2 - s_w^2\right)/mn$ is an unbiased estimator of the correlated component of the total response variance.

3.2 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 interviewer. 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 $(l = 1, \ldots, k)$, being $\sum_{l=1}^k m_l = n$.

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

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

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

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

$$Var\left(\hat{\bar{d}}_t\right) = \frac{1}{n}\sigma_d^2 + \frac{1}{n}\left(\frac{\sigma_m^2}{\overline{m}} + \overline{m} - 1\right)\sigma_{dd'} . \qquad (8)$$

Therefore, since $\sigma_m^2 > 0$, when different number of interviews are assigned to the interviewers, the coefficient multiplying $\sigma_{dd'}$ in (8) is larger than in the equal number of interviews case. Furthermore, it is worthwhile to observe that if σ_{ij} is constant for all *i* and *j*, $\sigma_{dd'}$ is minimum when all the interviewers are assigned the same number of interviews. Hence, in this case, the total response variance (8) is larger than in the equal number of interviews.

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

$$Var(\bar{y}_t) = \frac{1}{n} S_\mu^2 + \frac{1}{n}\sigma_d^2 + \frac{1}{n}\left(\frac{\sigma_m^2}{\bar{m}} + \bar{m} - 1\right)\sigma_{dd'} \ . \qquad (9)$$

Since the l-th subsample variance is $Var(\bar{y}_{lt}) = S_\mu^2 \big/ m_l + \sigma_d^2 \big/ m_l + \sigma_{dd'}(m_l - 1)\big/ m_l$, $(l = 1, ..., k)$, expression (9) is equivalent to $Var(\bar{y}_t) = \sum_{l=1}^{k} m_l Var(\bar{y}_{lt}) \big/ n^2$.
Then, by means of replicated sampling techniques it is easy to show that an unbiased estimator of (9) is given by

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

The estimation of the correlated component of the total response variance is based again based on the on $\tilde{s}_b^2 = n\,\text{var}(\bar{y}_t)$ and $s_w^2 = \sum_{l=1}^{k} s_{lt}^2 \big/ k$, with $s_{lt}^2 = \sum_{i=1}^{m_l} (y_{lit} - \bar{y}_{lt})^2 \big/ (m_l - 1)$. It must be observed that in this case $\tilde{s}_b^2$ does not denote the within interviewer variance, but it generalises expression of the s_b^2 estimator given in the previous section. 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

$$\hat{\sigma}_{dd'} = \frac{n(s_b^2 - s_w^2)}{\sum_{l=1}^{k} m_l^2} \ . \qquad (10)$$

In the next section it will be presented an application of the formulas derived above to the Italian survey on the crops product for which a different number of units are assigned to the interviewers.

4. An application: the crops product survey [3]

The Italian National Statistical Institute, ISTAT, carries out several agricultural surveys, among them the crops product survey. The survey, yearly carried out between November and December, collects data about the crops product sowed in the previous agricultural year and the intents of sowing for the current year. The survey, carried out with CATI technique, has a sample size of about 5,200 units and the results used in the application refer to the agricultural year 2003-2004.

[3] Data for the crops product survey have been supplied by the researchers in the Agricultural Statistical Office of Istat Laura Martino and Roberto Moro.

Table 1 – Number of interviews for interviewer.

Interviewer Code	Variable						
	Durum Wheat	Common Wheat	Barley	Grain Maize	Sugar Beet	Soya Bean	Temporary Fodder
003	52	38	48	47	26	12	61
011	35	29	29	38	30	9	43
015	54	51	19	47	33	9	24
016	69	72	60	67	53	21	82
017	109	87	84	125	76	39	126
020	145	101	78	146	110	49	148
023	106	89	68	97	69	31	92
024	215	129	132	198	138	69	181
027	202	156	120	185	139	54	206
028	205	130	98	189	136	62	184
030	144	94	89	132	86	46	156
031	53	36	33	57	34	13	51
032	2	0	0	1	1	0	1
033	48	32	29	47	29	20	42
036	50	47	29	59	46	18	50
037	91	71	71	86	62	35	100
038	64	51	47	78	52	25	54
040	157	90	103	140	96	55	114
041	79	51	44	81	53	35	67
042	65	61	68	80	49	28	86
043	89	58	55	75	50	22	77
044	44	36	26	40	27	11	44
045	96	86	80	93	65	34	85
046	54	46	35	45	35	21	60
047	79	44	41	70	51	27	64
048	66	50	51	62	41	18	64
049	80	46	55	70	51	27	62
050	6	2	2	7	4	3	3
051	128	109	86	132	93	46	128
052	85	55	53	78	56	32	74
053	53	32	26	45	32	10	49
054	60	33	38	59	42	14	63
055	67	54	47	57	45	19	65
057	105	56	62	99	60	34	101
058	86	42	53	77	50	32	65
059	70	36	40	64	43	24	55
201	1	2	0	1	0	0	0
Total	3,114	2,202	1,999	2,974	2,063	1,004	2,927

In CATI surveys, as in all the surveys using telephone interviewing, it is usual that the number of interviews is different among the interviewers and this the case for the Italian crops product survey as shown in table 1 reporting, for all the variables used in the application study, the number of interviews for interviewer. Table 1 shows that, for every variable, the number of interviews for interviewer is significantly different so that it is not

appropriate to use the standard solution presented in section 3.1 and, hence, estimators presented in section 3.2 need to be applied.
In order to achieve reliable results, in this application we have taken into account only variables with sample size larger or equal to 1,000. To obtain reliable within interviewer variance estimates, only interviewers with number of interviews $m_i \geq 10$ have been used for computation. Moreover, variables related to the intents of sowing have been excluded from computation.
Table 2 displays the percentage of the correlated component of the total response variance over the total variance. Exploratory data analysis have been performed for each variable in order to identify and remove outliers.

Table 2 – Percentage of the correlated component of the total response variance on the total variance

Variable	Percentage of the correlated component of response over the total variance
Durum Wheat	35,7
Common Wheat	30,4
Barley	39,6
Grain Maize	34,3
Sugar Beet	27,4
Soya Bean	20,2
Temporary Fodder	14,8

The results show that the average weight of the correlated component of the total response variance over the total variance is about 30%. Nonetheless, the values are not entirely due to the interviewer effect since other sources of measurement errors may occur (coding, registration, etc.). In fact, the values in table 2 have been obtained under the hypothesis that all the covariances σ_{ij} between the sampling units are due to the interviewer effect, but may not be the case for the crops product survey data.
No estimates of the simple response variance can be computed since there are no repeated measurements on the sampling units. In this context, only when a quality survey is carried out to estimate the measurement errors, we can obtain estimates for all the components of the error.

References

Biemer P.P., Stokes S.L. (1985), "Optimal Design of Interviewer Variance Experiments in Complex Survey", *JASA*, 80, pp. 158-166.
Cochran W.G. (1977), *Sampling Techniques*, New York: Wiley.
Fellegi I. (1964), "Response Variance and its Estimation", *JASA*, 59, pp. 1016-1041.
Fellegi I. (1974), "An Improved Method of Estimating the Correlated Response Variance", *JASA*, 69, pp. 496-501.

Hartley H.O., Rao J.N.K., La Motte L.R. (1978), “A Simple Synthesis–Based Method of Variance Component Estimation”, *Biometrics*, 34, pp. 233-242.
Kish L. (1962), “Studies of Interviewer Variance for Attitudinal Variables”, *JASA*, 57, pp. 92-115.
Koch G. (1973), “An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means”, *JASA*, 68, pp. 906-913.
Lessler J.T., Kalsbeek W.T., (1992), *Nonsampling errors in surveys*, New York: Wiley.
Mahalanobis P.C. (1946), “Recent Experiments in Statistical Sampling in the Indian Statistical Institute”, *JRSS*, 109, pp. 325-370.
Miller V.P., Cannell C.F. (1982), “A Study of Experimental Techniques for Telephone Interviewing”, *Public Opinion Quarterly*, 46, pp. 250-269.
Särndal C.E., Swensson B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.