ADVANCED SCHOOL FOR STATISTICS AND SOCIO-ECONOMIC ANALYSES

Istat
National Institute of Statistics

Total nonresponse in sample surveys

Marco D. Terribili

What is the total nonresponse?

It is the situation where the expected response of a statistical unit called to participate in the survey is not obtained by the institute conducting the survey, for whatever reason (*Särndal and Lundström, 2005*).

The causes of total nonresponse

It can be determined by multiple causes:

- the statistical unit does not receive the questionnaire or is not contacted by the interviewer (***noncontact***),
- the contacted unit does not respond because unable , e.g. due to language problems (***inability***),
- he/she explicitly does not cooperate (***refusal***).

⇨ The distinction among the various components of nonresponse can be important both in the prevention of nonresponse and in the estimation of the parameters of the population under investigation.

Total nonresponse effects

The main effects of total nonresponse in a sample survey are:

- the reduction of the sample size and, consequently, of the degree of precision of the estimates (increasing the sampling variance)
- the introduction of bias effects that reduce the accuracy of the final estimates: the more non respondents differ systematically from respondents (with respect to the variables of interest), the greater is the bias.
  - Bias is a non-sampling or systematic error, independent from the sample size, defined as the difference between the expected value of the estimator and the real value of the parameter.

⇨ The occurrence of both effects leads to a loss in the reliability of the estimates.

Total nonresponse effects - *Bias effect*

Bias is the most important effect of nonresponse. To investigate its effects on estimates, treatment of nonresponse should be integrated in sampling and estimation theory.

One common approach is based on the *fixed response model*, which assumes that the population consists of two subpopulations or strata, exclusive and exhaustive: the stratum of the responding units that, with certainty, provide an answer and the stratum of the not responding units that, still with certainty, do not respond. In this framework, bias magnitude depends on both the differences in the characteristics under investigation between the two strata and the non-response rate:

$$B\left(\hat{\bar{Y}}_{s_R}\right) = \frac{N_{NR}}{N}\left(\bar{Y}_R - \bar{Y}_{NR}\right)$$

where

- N total number of units in the population
- N_{NR} number of not respondent units in the population
- $\bar{Y}_R$ means of the target variable on respondent units
- $\bar{Y}_{NR}$ means of the target variable on not respondent units

Total nonresponse dealing

To deal with the problem of non-response, in order to limit the reduction of the sample size and the bias effect, many actions can be undertaken, in every step of the survey process.

We can divide them in two kinds:

- **Prevention** during the phases of survey design and data collection.
- **Correction** during the estimation phase, modifying the sampling weights through adjustment factors defined on the basis of auxiliary information.

Total nonresponse dealing - *Prevention of the total nonresponse*

The aim of these actions is to prevent the effect of reduction of the sample size, according to the different types of nonresponse outcomes (*Bethlem, 2011*):

Noncontact	***Inability***	***Refusal***
• Increase the number of contact attempts • Longer fieldwork period • Lower the interviewer workload • Evening calls more profitable than daytime calls • Make contact in a different mode • Sending a reminder • Personalizing the letter of invitation • Mentioning the duration of the survey	• Multilingual interviewers • Translating questionnaires • Having a family member act as interpreter	• Advance letter • Incentive • Interviewer training • Mixed-mode data collection • Proxy respondents

Total nonresponse dealing - *Correction in the estimation phase*

In the estimation phase the compensation of nonresponse effects (bias) consists in adjusting the sampling weights associated to respondent units, to obtain that they represent also not respondent units. These techniques, called *weighting adjustment*, are based on auxiliary information, known for respondent and not respondent units or known totals referred to the complete sample or to the whole population of interest.

Using these information, nonresponse can be treated in two ways:

- **Implicit modeling,** based on the use of *calibration estimators*
- **Explicit modeling,** based on the assumption of the existence of a functional link between the response variable and a set of auxiliary variables known for all sample units, respondents and not respondents.

➢ The use of adjustment factors to reduce the bias produces generally, as a drawback, an increase in the variability of the weights, and so of the sampling variance.

Correction in the estimation phase - *Implicit modeling*

This approach, which adopts an implicit modeling of the response probability, is based on the use of *calibration estimators* (*Deville and Särndal, 1992*): the sample of respondent is weighted ("post-stratified") to reflect population distributions, based on known totals external to survey.

Using these methods the distributions of the auxiliary variables, estimated on the sample of respondents, are constrained to the corresponding distributions estimated on the total sample or calculated on the entire population. This method can be very useful because:

- Often data for non-respondents are not available, whereas there might be some population data
- Sample weighting corrects only non-response bias; population weighting simultaneously corrects bias (nonresponse and frame imperfections) and it improves precision by reducing sampling errors.

Correction in the estimation phase - *Explicit modeling*

Theoretical framework

The theoretical basis on which the construction of a correction factor relies is *two-phase sampling* (*Särndal et al., 1992*):

I. The first phase sample is constituted by the initially selected sample,

II. The second phase sample is the set of respondent units.

Construction process of the adjustment factors

1) Modeling the relationship between the response variable and a set of auxiliary variables, known both for respondents and not respondents (*Nonresponse adjustments models*)
2) Estimation of response probabilities
3) Definition of adjustment factors

Explicit modeling - *Theoretical framework*

Denote by:

- U, the target population of interest
- s, the sample of size n ($i = 1, \ldots, n$) , selected following a given sample design
- s_R, the subsample of respondent units of s
- y_i the value of the y variable under investigation on the i-th unit of the population observable only for respondent units.

⇨ in the first phase, the units of the sample s are selected according to the chosen sampling design that assigns the *inclusion probability* of the first order to all units of the population U. The probability of the i-th unit to be included in the sample s, known and positive, is expressed as:

$$\pi_i = \mathrm{P}(i \in s), \quad (i \in U)$$

⇨ in the second phase, the units of the sample s are divided, on the basis of a *random unknown mechanism*, into two subsets, the respondents and the not respondents; this mechanism is summarized by assigning to each unit of the sample s the *response probability*, which represents the inclusion probability of the i-th unit in the second stage of the sampling, conditional to the first phase.

$$\theta_i = \mathrm{P}(i \in s_R \mid s), \quad (i \in s)$$

Explicit modeling - *Theoretical framework*

When the parameter of interest is the population mean of the generic variable y,

$$\bar{Y} = \frac{1}{N}\sum_{i \in U} y_i$$

an estimator for $\bar{Y}$ can be expressed as a linear function of the sampled units:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i \in s_R} d_i y_i \qquad (i = 1, \ldots, n_R),$$

In case of nonresponse, in the two-phase sampling framework, the unbiased estimator becomes

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i \in s_R} d_i \gamma_i y_i = \frac{1}{N}\sum_{i \in s_R} \frac{y_i}{\pi_i \theta_i} \qquad (i = 1, \ldots, n_R) \text{ and } s_R \subseteq s$$

where $d_i = 1/\pi_i$ is the initial direct weight in sample s, $\gamma_i = 1/\theta_i$ is the multiplicative factor in the sample of respondents s_R to correct for total nonresponse, assuming a response probability θ_i known for each respondent unit.

⇨ However, θ_i is unknown and has to be estimated according to a model specification (*Särndal and Lundström , 2005*). The estimator obtained by substituting θ_i by its estimate $\hat{\theta}_i$ is not unbiased.

Explicit modeling - *Nonresponse adjustment models*

Estimation of θ_i

The prediction of the response probability is carried out by modeling the relationship between the response variable and a set of auxiliary variables, known both for respondents and not respondents, using **parametric** and **non-parametric models**. As the dependent variable is a binary one, the models employed are:

Parametric model
Logistic regression, Logit (or Probit)
Non-parametric model
Classification trees, CART (or Chaid)

Parametric model - *Response propensity score*

The application of parametric models leads to the response propensity score method (Rosembaum and Rubin, 1983); it is based on two main assumptions:

- *Missing At Random*: response depends on a set of auxiliary variables, while the relationship between the variables of interest and the response variable is established in an indirect way
- *Matching:* response probability is defined within the interval (0,1)

This method exploits *Logistic* (or rarely *Probit*) regression model in which:

- the response variable is binomial and it assumes values $R_i = \begin{cases} 1 & \text{if } i\text{-th unit is respondent} \\ 0 & \text{if } i\text{-th unit is not respondent} \end{cases}$
- each i-th unit in the sample s is associated with a $p \times 1$ vector of auxiliary variables

$$\mathbf{X}_i = (X_{1i}, X_{2i}, \ldots, X_{pi})'$$

- the response probability for the i-th sample unit, conditionally on the value of the characteristics $\mathbf{X}_i$, is

$$\theta_i = \theta(\mathbf{X}_i) = P(Ri = 1 | \mathbf{X}_i)$$

Response propensity score - *Logit model*

The logistic model can be expressed as:

$$\text{Logit } P(R_i = 1 | \mathbf{X}_i) = \text{Log } \frac{P(R_i = 1 | \mathbf{X}_i)}{1 - P(R_i = 1 | \mathbf{X}_i)} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} \qquad (i = 1, \ldots, n)$$

The response propensity is expressed as:

$$\hat{\theta}_i = \hat{\theta}(\mathbf{X}_i) = \frac{exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi})}{1 + exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi})}$$

⇨ The parameters β of the model can be estimated by means the maximum likelihood method (MLE).

Response propensity score - *Definition of adjustment factors*

The estimated response probability (or response propensity), obtained through the Logit model, is used for non-response adjustment mainly in two ways:

- Response propensity weighting: the adjustment factor for i-th unit is defined as the reciprocal of the response propensity $\hat{\theta}_i$

$$\gamma_i = \frac{1}{\hat{\theta}(\mathbf{x}_i)} \quad i \in s_R$$

- Response propensity stratification: the response propensity $\hat{\theta}_i$ is used to define strata (or *cells*) and the correction factors is calculated at stratum level, as the reciprocal of the response rate observed in the stratum

$$\gamma_h' = \left(\hat{\theta}_h\right)^{-1} = \left(\frac{n_{R,h}}{n_h}\right)^{-1}$$

The response propensity distribution, for example, can be partitioned in strata of the same size, usually a small number of quantiles (quartiles or quintiles) is preferred.

Non-parametic model - *Classification and Regression Tree*

Another method to create strata to apply correction factors is *Classification And Regression Tree* (CART).

- CART is "a non-parametric decision tree learning technique" (*Breiman et al., 1984*) applied to obtain the optimal partition of a set of statistical units, based on the relationship between a dependent variable (in our case the response R_i) and a set of auxiliary variables.

- The main idea behind tree methods is to recursively partition the data into smaller and smaller groups in order to improve the fit at best.

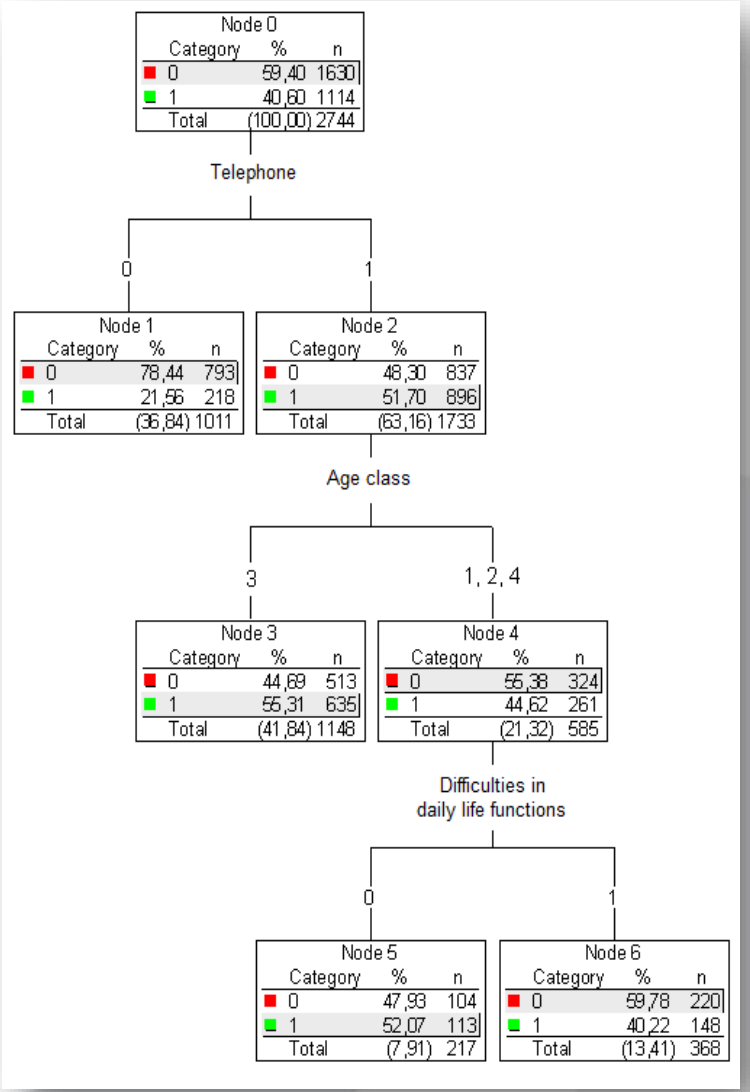
- Generally, the preferred strategy to find the optimal partition is to consider a full tree and to *prune* a large tree, using a "*cost-complexity*" function.

  - The nonresponse adjustment factor is defined as the reciprocal of the response rate observed in each group, similarly to the response propensity stratification.

Non-parametic model - *Classification and Regression Tree*

The resulting diagram of the algorithm represents how all the units, initially included in a starting group (called *root*), are divided bipartition by bipartition in a small number of groups, called *leaves*.

At each step a bipartition variable is chosen to maximize the homogeneity (in terms of entropy or Gini index) in the two resultant leaves, regarding the relationship between the response variable R_i and the auxiliary variables $\mathbf{X}_i$.



Nonresponse adjustment models - *Model selection*

In the model studying phase, several alternative models (both parametric and non-parametric) are generally fitted, on varying the auxiliary variables considered in the models, possibly recoding some of them, for example from continuous to classification variables, with the aim of obtaining the "best" model to correct the bias.

Model assessment is performed following different methods to evaluate the goodness of fit of the models to data:

Logistic models
AIC (*Akaike Information Criterion*)
BIC (*Bayesian Information Criterion*)
Log-likelihood
CART
Cost-Complexity function

Definition of adjustment factors - *Methods evaluation*

The considered different methods are evaluated in terms of bias reduction and increase of the weights variability. Two indicators can be utilised:

- Concordance Index (CI), to evaluate the goodness of fit of the model

$$CI = 1 - \frac{\sum_{i=1}^{n} \left| R_i - \hat{\theta}_i \right|}{n}$$

- The statistic $1+CV^2(w)$ *(Kish, 1992),* to evaluate the impact of the variability of final weights on the variance of the estimates. This statistic which represents an approximated increase of the variance of the estimates due to the application of correction factors to the direct sampling weights.

  - This synthetic measure permits to appreciate the extent to which the proposal of adjustment factors produces an increase of the variability in the weights and thereby a reduction of the precision of the survey estimates, with respect to the situation of equal weights.

Alternative methods for nonresponse adjustment

The different types of nonresponse

In recent years, alternative approaches for nonresponse adjustment were applied considering the different types of nonresponse

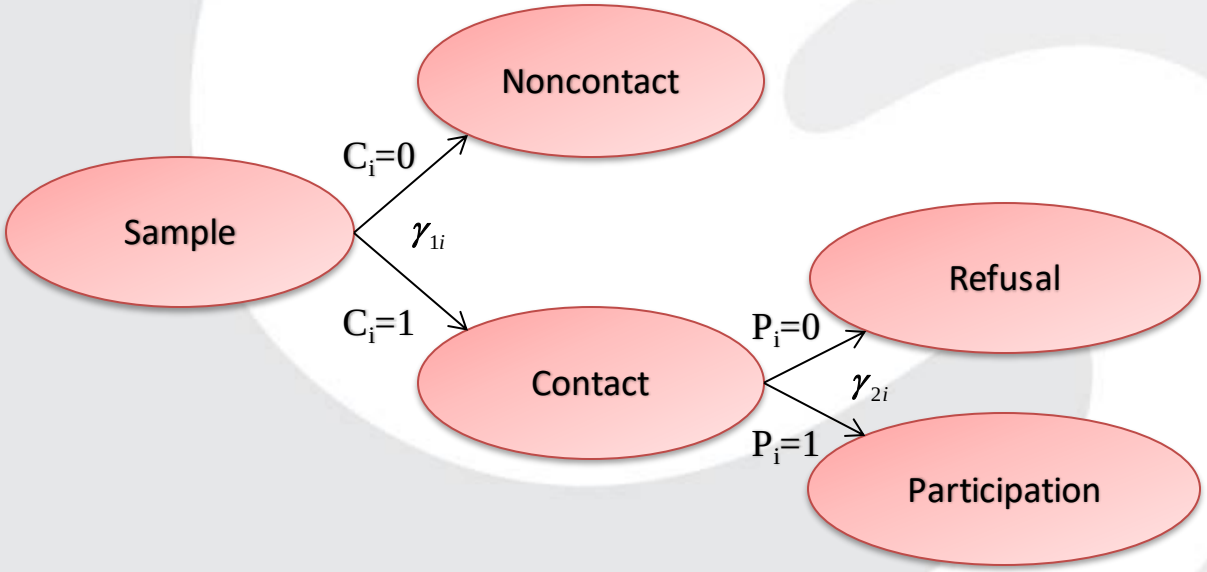
- There are situations where distinct causes of nonresponse have different influence on the survey variables.
- If the response types show a different relationship with the survey variable, the effect on the nonresponse bias is different.
- In this case it could be important to model and to take into account the different kinds of nonresponse.

Alternative methods for nonresponse adjustment

The sequential response process

The components of nonresponse can be hierarchically distinct. The participation of an individual in the survey constitutes the last stage of a process characterized by a sequence of events, namely the different stages of the response process, each of which is nested in the previous.

- **An example in which the components of nonresponse are «noncontact» and «refusal»**



Adjustment methods for different nonresponse types

Among the methods which take into account the different nonresponse types, an interesting one considers the sequential nature of the response process: it is the **sequential weight adjustment method**:

- The adjustment procedure is developed in two successive steps: in the first the initial sampling weights are adjusted to account for noncontact, while in the second they are further adjusted to correct for nonparticipation.
- Nested models can be used for the construction of the adjustment factors considering two different models, one for each stage of response process: the first estimating the probability of being contacted and the second estimating the probability for contacted individuals to participate in the survey.

⇨ The two stages of the response process are independent conditionally on a set of auxiliary variables (fundamental assumption of the method)

Sequential weight adjustment method - *Nested models*

In the first model the binary dependent variable assumes values

$$C_i = \begin{cases} 1 & \text{if } i\text{-th unit is contacted} \\ 0 & \text{if } i\text{-th unit is not contacted} \end{cases} \qquad i = (1,\ldots,n)$$

In the second model the binary dependent variable assumes values

$$P_i = \begin{cases} 1 & \text{if } i\text{-th unit is respondent} \\ 0 & \text{if } i\text{-th unit is not respondent} \end{cases} \qquad i = (1,\ldots,n_c)$$

⇨ Therefore, in the estimator the sampling weight is modified by two adjustment factors: the first one based on the estimated contact probability , the second one based on the estimated participation probability, conditionally on the contact.

A case study

The ISTAT survey "Social integration of people with disabilities"

- It is a specific follow-up of the Multipurpose survey *"Health conditions and use of health services"*
- A peculiarity of the Disability survey is that all the survey variables collected in the first occurrence (Health survey) were available in the Disability survey for the whole list of individuals: this information corresponds to the set of auxiliary variables known for respondent and non respondent units, necessary for the construction of the weights
- The Disability survey suffered from a high nonresponse rate, due more to the lack of contact of the disabled individuals highlighted by the Health survey rather than to the refusal to cooperate when contacted:

Outcomes	Frequency	Percentage
Not contacted units	1290	36,8%
Contacted units, whereof:	1454	45,6%
Not respondents	*340*	*9,7%*
Respondents	*1114*	*31,8%*
Total	2744	100,0%

A case study

Methodological framework

- Preliminary analysis (bivariate, multivariate) of nonresponse types
- Parametric and nonparametric model fitting in the traditional approach
- Parametric and nonparametric model fitting in the sequential approach
- Result assessment and choice of the adjustment correction factor

Preliminary analysis of nonresponse types

Distribution of disabled individuals by response group, age class and geographic area (row %)

Group		Age class								
		6-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	Total
Traditional approach perspective										
Respondent		1.08	1.35	2.96	5.12	6.82	10.5	23.25	48.92	100%
Not respondent units (not contacted + refusals)		4.97	2.7	2.64	5.71	4.66	8.9	20.8	49.63	100%
Sequential approach perspective										
First stage	Non-contacts	5.66	2.48	2.79	6.05	4.88	9.53	18.99	49.61	100%
	Contacted units	1.38	1.86	2.75	4.95	6.12	9.56	24.28	49.11	100%
Second stage	Participant units	1.08	1.35	2.96	5.12	6.82	10.5	23.25	48.92	100%
	Refusals	2.35	3.53	2.06	4.41	3.82	6.47	27.65	49.71	100%

Group		Geographic Area					
		North-West	North-East	Centre	South	Islands	Total
Traditional approach perspective							
Respondent		18.58	16.97	17.5	33.93	13.02	100%
Not respondent units (not contacted + refusals)		17.42	13.37	16.93	37.98	14.29	100%
Sequential approach perspective							
First stage	Non-contacts	15.89	12.64	17.29	39.22	14.96	100%
	Contacted units	19.67	16.78	17.06	33.77	12.72	100%
Second stage	Participant units	18.58	16.97	17.5	33.93	13.02	100%
	Refusals	23.24	16.18	15.59	33.24	11.76	100%

Parametric and nonparametric model fitting

Response, contact and participation: logit and CART models assessment

	Traditional approach		Sequential approach			
	Response		Contact		Participation	
Model	Covariates	Index	Covariates	Index	Covariates	Index
Logit Model AIC	telephone	3.388	Telephone	3.347	age in 5 classes	1.564
	age in 4 classes		age in 2 classes			
	marital status		marital status			
	disability level		motor disability			
	motor difficulty		number of invalidity			
	number of invalidities		number of disability			
CART $\Phi_\alpha(T)$	telephone	0.406	Telephone	0.325	age in 3 classes	0.249
	age in 4 classes					
	difficulties in daily life functions					

Result assessment

Summaries of the distributions of final weights

Model	Traditional Approach	Technique	Average	Max	Min	$CV(w)$	$1+CV^2(w)$
LOGIT	Response propensity stratification	Quartiles	1046.72	7692.57	98.83	0.825	1.680
		Quintiles	1037.98	8861.92	99.02	0.821	1.673
		Deciles	1037.62	9781.18	89.22	0.855	1.731
	Response propensity weighting	Individual propensity	1022.55	7235.38	94.09	0.784	1.615
CART		Terminal nodes	1035.76	6796.77	94.09	0.753	**1.567**

Model	Sequential Approach	Technique	Average	Max	Min	$CV(w)$	$1+CV^2(w)$
LOGIT	Response propensity stratification	Quintiles	1028.87	7081.31	104.13	0.732	1.555
	Response propensity weighting	Individual propensity	1027.73	7350.38	101.51	0.732	1.555
CART		Terminal nodes	1026.71	7003.45	102.98	0.729	**1.531**

Condordance Index

Model	Approach	Technique	Concordance Index (CI)		
			Response	Contact	Participation
LOGIT	Response propensity stratification	Quartiles	0.569	0.574	
		Quintiles	0.569	0.581	0.645
		Deciles	0.573	0.584	
	Response propensity weighting	Individual propensity	0.565	0.569	0.647
CART		Terminal nodes	0.574	0.583	0.648

REFERENCES

- Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in household surveys.* Wiley, New York.
- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification Regression Trees*. Wadsworth International Group, Belmont.
- Cochran, W.G. (1977). *Sampling techniques*. Wiley, New York.
- Groves, R.M., Couper, M.P. (1998). *Nonresponse in household interview surveys*. Wiley, New York.
- Iannacchione, V.G. (2003). Sequential weight adjustments for location and cooperation propensity for 1995 national survey of family growth. *Journal of Official Statistics*, 19: 31-43.
- Kalton, G., Flores-Cervantes, I. (2003). *Weighting methods*. Journal of Official Statistics, 19: 81-97.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54: 139-157.
- Oh, H.L. , Scheuren, F. (1983). Weighting adjustments for unit nonresponse. Weighting adjustments for unit nonresponse, in *Incomplete Data in Sample Surveys, Vol. 2, Theory and Bibliographies*, W.G. Madow, I. Olkin & D.B. Rubin, eds., Academic Press, New York. pp. 143-184.
- Olson, K. M. (2006). *Survey participation, non response bias, measurement error bias, and total bias*, Sociology Department, Faculty Publications, University of Nebraska. Lincoln.
- Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse, *Survey Methodology*, 22: 43-53.
- Rosenbaum, P.R. and Rubin, D.B. (1984) Reducing the bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association,* 79, 516-524.
- Särndal, C.E., Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley, New York.
- Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*, Cap XV, Springer. New York.