Survey Methodology, December 1994
Vol. 20, No. 2, pp. 177–184
Statistics Canada

Nonparametric Estimation of Response Probabilities in Sampling Theory

THÉOPHILE NIYONSENGA[1]

ABSTRACT

We deal with the nonresponse problem by drawing on the model of selection in phases that was proposed by Särndal and Swenson (1987). To estimate response probabilities, we use the nonparametric approach first advanced by Giommi (1987). We define estimators according to the nonparametric estimation (NPE) model, and we study their general properties empirically. Inference is based on the concept of quasi-randomization (Oh and Scheuren 1983). The emphasis is on estimating the variance and constructing confidence intervals. We find, by way of a Monte Carlo study, that it is possible to improve the quality of the estimators considered by using a variant of the NPE approach. The latter also serves to confirm the performance of regression estimators in terms of variance estimation.

KEY WORDS: Weighting by phases; Regression estimator; Variance estimators.

1. INTRODUCTION

To counter the effect of nonresponse on the estimation of parameters of a finite population, we consider the phenomenon of nonresponse as a unit selection process in three phases. We therefore use weighting by phases. This adjustment procedure assigns to each unit observed a weight that is inversely proportional to the probability of appearing in the sample, to the unit response probability given the sample, and to the item response probability given the sample and the set of respondents per unit.

In practice, only the probabilities of inclusion in the sample are known. The problem facing us is to estimate individual response probabilities before incorporating them in formulas for the estimators of interest. The nonparametric estimation approach is one of the response probability estimation procedures. It is motivated by the use of auxiliary variables which are linked with unit and item response mechanisms (Giommi 1985, 1987), and which may be correlated with the variables of interest. This avoids assuming that nonresponse is independent of the variables being studied (Oh and Scheuren 1983). This approach also enables us to avoid postulating one or more parametric models governing response, such as the Logit and Tobit models (Grosbras 1987b; Chicoineau, Payen and Thélot 1985) or models of uniform response within subpopulations (Oh and Scheuren 1983; Särndal and Swenson 1985, 1987).

In the Monte Carlo study illustrating certain estimators according to the nonparametric approach, we consider the quite specific case in which the two response mechanisms are governed by the same auxiliary variables. The difference between items will reside in the degree of correlation between each item and the auxiliary variables.

2. NONRESPONSE: A THREE-PHASE SELECTION PROCESS

Consider a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$, of size N. Let s be a sample of fixed size n drawn from U according to a plan $\mathcal{P}(s)$ known and characterized by inclusion probabilities $\pi_k > 0, \forall\, k$ and $\pi_{k\ell} > 0\ \forall\, k \neq \ell$. We want to observe the units $k \in s$ in relation to a set of Q items $y_1, \ldots, y_q, \ldots, y_Q$ $(Q \geq 1)$, then estimate the total per item $t_q = \sum_U y_{qk}$, for every $q (q = 1, \ldots, Q)$. We assume that conditional on s, each unit k has a probability $\varphi_k > 0$ of participating in the survey and that the probability that two units k and ℓ participate is $\varphi_{k\ell} > 0$ with $\varphi_{kk} = \varphi_k$. We denote the set of units that agree to participate in the survey by r and the mechanism by which the set r was obtained by $\mathcal{P}(r \mid s)$. We further assume that conditional on s and r, each unit $k \in r$ responds to item y_q with probability $\psi_{qk} > 0$ and that the probability that two units k and $\ell \in r$ respond to item y_q is $\psi_{qk\ell} > 0$ with $\psi_{qkk} = \psi_{qk}$. We denote by r_q the set of units that, having agreed to participate in the survey, respond to item y_q and by $\mathcal{P}(r_q \mid s,r)$ the mechanism by which the set r_q is obtained for all $q (q = 1, \ldots, Q)$.

The sets s, r and r_q are obtained from three selection phases for which only the probabilities of inclusion in s are known. The composition of the unit selection mechanisms gives rise to probability outputs that we denote by $\pi_k \Theta_{qk}$ where $\Theta_{qk} = \varphi_k \psi_{qk}$ and $\Theta_{qk\ell} = \varphi_{k\ell} \psi_{qk\ell}$ with $\Theta_{qkk} = \Theta_{qk}$, which do not correspond to inclusion probabilities. Nor does the quantity Θ_{qk} correspond to an inclusion probability for the two response phases conditional on s. If we define the probabilities of inclusion in r_q by $\pi^*_{qk} = \mathbb{P}(k \in r_q)$ and the probabilities of inclusion in r_q given s by $\Theta^*_{qk} = \mathbb{P}(k \in r_q \mid s)$, then (i) $\pi^*_{qk} \neq \pi_k \Theta^*_{qk}$

[1] Théophile Niyonsenga, Ph.D., Researcher, Centre de Recherche Clinique, Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, QC, Canada, J1H 5N4.

and (ii) $\pi_k \Theta^*_{qk} \neq \pi_k \Theta_{qk}$. Furthermore, (iii) $\Theta^*_{qk} = \Theta_{qk}$ if probabilities ψ_{qk} are independent of r, and (iv) $\pi^*_{qk} = \pi_k \Theta_{qk}$ if the φ_k do not depend on s and if the ψ_{qk} do not depend on either r or s.

3. A FEW SPECIAL ESTIMATORS

Assume that there is an auxiliary variable x_q (for the q-th item) strongly correlated with the variable y_q and such that x_{qk} is known $\forall\, k \in s$ or $\forall\, k \in U$. We take the specific case in which $x_{qk} = x_k$, $\forall\, q (q = 1, \ldots, Q)$, and we assume the following linear model ξ

$$\begin{cases} \mathbb{E}_\xi (y_{qk} \mid x_k) = \beta_q x_k \\ \operatorname{Cov}_\xi (y_{qk}, y_{q\ell} \mid x_k, x_\ell) = \begin{cases} \sigma_q^2 x_k & \text{if } k = \ell, \\ 0 & \text{otherwise} \end{cases} \end{cases} \tag{3.1}$$

in which β_q and σ_q are unknown parameters. The following results are extensions of the findings of Särndal and Swenson (1987).

Result 1. If x_k is known, $\forall\, k \in s$, then the regression estimator, denoted by $\hat{t}_{\text{Reg}}$ and defined by:

$$\hat{t}_{\text{Reg}} = \left(\sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \Big/ \sum_{r_q} \frac{x_k}{\pi_k \Theta_{qk}} \right) \sum_s \frac{x_k}{\pi_k}, \tag{3.2}$$

is approximately unbiased for t_q. Its approximate variance is a sum of three components V_1, V_2 and V_3 representing the respective portions of the variance due to the selection phases, that is:

$$V_1 = \sum \sum_U \Delta_{\pi_{k\ell}} (y_{qk}/\pi_k)(y_{q\ell}/\pi_\ell),$$

$$V_2 = \mathbb{E}\left\{ \sum \sum_s \Delta_{\varphi_{k\ell}} (E_{qk}/\pi_k \varphi_k)(E_{q\ell}/\pi_\ell \varphi_\ell) \right\},$$

$$V_3 = \mathbb{E}\,\mathbb{E}\left[\sum \sum_r \Delta_{\psi_{qk\ell}} (E_{qk}/\pi_k \Theta_{qk})(E_{q\ell}/\pi_\ell \Theta_{q\ell}) \mid s \right],$$

where the E_{qk} are theoretical residuals of model (3.1). An estimator of $V(\hat{t}_{\text{Reg}})$ is given by $\hat{V}(\hat{t}_{\text{Reg}}) = \hat{V}_1 + \hat{V}_2^+$ (where $\hat{V}_2^+ = \hat{V}_2 + \hat{V}_3$) with:

$$\hat{V}_1 = \sum \sum_{r_q} \frac{\Delta_{\pi_{k\ell}}}{\pi_{k\ell} \Theta_{qk\ell}} \left(\frac{y_{qk}}{\pi_k} \right) \left(\frac{y_{q\ell}}{\pi_\ell} \right), \tag{3.3}$$

and

$$\hat{V}_2^+ = \sum \sum_{r_q} \frac{\Delta_{\Theta_{qk\ell}}}{\Theta_{qk\ell}} \left(\frac{e_{qk}}{\pi_k \Theta_{qk}} \right) \left(\frac{e_{q\ell}}{\pi_\ell \Theta_{q\ell}} \right), \tag{3.4}$$

where $\Delta_{\pi_{k\ell}} = \pi_{k\ell} - \pi_k \pi_\ell$, $\Delta_{\varphi_{k\ell}} = \varphi_{k\ell} - \varphi_k \pi_\ell$, $\Delta_{\psi_{qk\ell}} = \psi_{qk\ell} - \psi_{qk}\psi_{q\ell}$ and $\Delta_{\Theta_{qk\ell}} = \Theta_{qk\ell} - \Theta_{qk}\Theta_{q\ell}$, the e_{qk} being the observed residuals obtained from model (3.1).

Result 2. If x_k is known, $\forall\, k \in U$, then the regression estimator, denoted by $\hat{t}_{\text{Reg1}}$ and defined by:

$$\hat{t}_{\text{Reg1}} = N \bar{x}_U \left(\sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \Big/ \sum_{r_q} \frac{x_k}{\pi_k \Theta_{qk}} \right), \tag{3.5}$$

is approximately unbiased for t_q. Its approximate variance is also a sum of three components V_1, V_2 and V_3. The expression of $V_1(\hat{t}_{\text{Reg1}})$ differs from that of $V_1(\hat{t}_{\text{Reg}})$ by the use of the theoretical residuals E_{qk} in place of the raw values y_{qk}, whereas the expressions of V_2 and V_3 are identical to those defined above for $\hat{t}_{\text{Reg}}$. An estimator of $V(\hat{t}_{\text{Reg1}})$ is given by $\hat{V}(\hat{t}_{\text{Reg1}}) = \hat{V}_1 + \hat{V}_2^+$ where:

$$\hat{V}_1 = \sum \sum_{r_q} \frac{\Delta_{\pi_{k\ell}}}{\pi_{k\ell} \Theta_{qk\ell}} \left(\frac{e_{qk}}{\pi_k} \right) \left(\frac{e_{q\ell}}{\pi_\ell} \right), \tag{3.6}$$

and where $\hat{V}_2^+ = \hat{V}_2 + \hat{V}_3$ is obtained by the formula (3.4).

Comment 1. If $x_k = 1$, $\forall\, k \in U$, the formula (3.5) defines an estimator, denoted by $\hat{t}_{\text{Exp}}$ where:

$$\hat{t}_{\text{Exp}} = N \sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \Big/ \sum_{r_q} \frac{1}{\pi_k \Theta_{qk}} = \frac{N}{\hat{N}} \sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}}. \tag{3.7}$$

The estimator $\hat{t}_{\text{Exp}}$ is called an "expansion estimator". An estimator of approximately unbiased variance for $V(\hat{t}_{\text{Exp}})$ is derived from formulas (3.4) and (3.6).

Comment 2. If we take $\Theta_{qk} = \Theta_q (0 < \Theta_q \leq 1)$, $\forall\, k \in U$, in formula (3.7), we obtain an estimator, denoted by $\hat{t}_{\text{Naive}}$, called a "naive estimator". Its expression is given by:

$$\hat{t}_{\text{Naive}} = N \sum_{r_q} \frac{y_{qk}}{\pi_k} \Big/ \sum_{r_q} \frac{1}{\pi_k}. \tag{3.8}$$

If the π_k are constant, the expression (3.8) becomes identical to formula (3.5) in which t is assumed that $\Theta_{qk} = \Theta_q (0 < \Theta_q \leq 1)$, $\forall\, k \in U$, and $x_k = 1$, $\forall\, k \in U$.

Comment 3. For the four estimators defined above, the underlying models are derived from model (3.1) and are the following: $y_{qk} = \beta_q x_k + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ and $V(\epsilon_{qk}) = \sigma_q^2 x_k$ for the first two, $y_{qk} = \beta_q + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ and $V(\epsilon_{qk}) = \sigma_q^2$ and N is known for the last two. For the naive estimator, it is necessary to add the uniform unit and item response model.

4. ESTIMATORS WITH ESTIMATED RESPONSE PROBABILITIES

In practice, the response probabilities φ_k and ψ_{qk} as well as the probability outputs $\Theta_{qk} = \varphi_k\psi_{qk} (k \in U, q = 1, \ldots, Q)$ are actually parameters to be estimated. We estimate them by $\hat{\varphi}_k$, $\hat{\psi}_{qk}$ and $\hat{\Theta}_{qk} = \hat{\varphi}_k\hat{\psi}_{qk}$ respectively. We define estimators having the same form as the prototype estimators $\hat{t}_{\text{Exp}}$, $\hat{t}_{\text{Reg}}$ and $\hat{t}_{\text{Reg1}}$ seen in section 3, taking care to replace the unknown parameters by their respective estimates. We denote these estimators by $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg1}np}$ respectively. The variance estimators are obtained from the expressions (3.3), (3.4) and (3.6), in which the unknown parameters are replaced with their estimates.

4.1 Estimation of Response Probabilities

In theory, the probabilities φ_k and ψ_{qk} are functions of the auxiliary variables, that is, functions of the form $\varphi_k = f_1(\upsilon,z_k)$ and $\psi_{qk} = f_2(\mu_q,x_{qk})$ in which the quantities υ and $\mu_q (q = 1, \ldots, Q)$ are unknown parameters and where the pair of vectors (z,x_q), that is, $[(z_1,x_{q1}), \ldots, (z_k,x_{qk}), \ldots, z_N,x_{qN})]'$, contain the auxiliary information available for each item y_q. The nonparametric estimation approach uses only the information contained in (z,x_q) to estimate the φ_k and ψ_{qk}. We are considering here the specific case in which the $z_k = x_{qk} = x_k$, $\forall\, q$ $(q = 1, \ldots, Q)$, and $\forall\, k \in s$.

Let $x_s = \{x_k : k \in s\}$, all the auxiliary information relating to the sample. We specify $\tau_s = \{\tau_k : k \in s\}$, a set of functions such that $\tau_k : \mathbb{R}^n \to \mathbb{R}^1$, for all k in s. We denote by $g_k = \tau_k(x_s)$, $\forall\, k \in s$, the value of the k-th function evaluated in x_s. We subdivide s in n groups s_k not necessarily disjoint, the respective sizes of which are given by:

$$n_k = \sum_{j \in s} D(g_k - g_j), \ (k \in s),$$

$$D(g_k - g_j) = \begin{cases} 1 & \text{if } \mid g_k - g_j \mid \leq h_k, \\ 0 & \text{otherwise,} \end{cases}$$

for a given constant h_k which may depend on all the values g_k $(k \in s)$. The set $s_k = \{j : g_j \in [g_k \pm h_k]\}$, $\forall\, k \in s$, contains j units, whose values g_j vary little from one to another. This group is called the group whose unit k is the kernel, or simply the k-th group. In other words, s_k is a subset of s for which the values of x fall within the vicinity of $x = x_k$ in the sense of the Euclidian distance that specifies $d(k,j) = \mid \tau_k(x_s) - \tau_j(x_s) \mid \leq h_k = h(g_k)$, meaning that $s_k = \{j : d(k,j) \leq h_k\}$. Let $r_k = s_k \cap r$ and $r_{qk} = s_k \cap r_q$. The respective absolute frequencies of these sets are m_k and m_{qk} where:

$$m_k = \sum_{j \in r} D(g_k - g_j), \ (k \in r);$$

$$m_{qk} = \sum_{j \in r_q} D(g_k - g_j), \ (k \in r_q, q = 1, \ldots, Q).$$

Comment 4. In the general case in which nonresponse is governed by the pair of vectors (z, x_q) with $z \neq x_q$, the τ_k functions would be defined in terms of z in order to estimate the unit response probabilities φ_k and in terms of x_q to estimate the item response probabilities ψ_{qk}. Note that this kernel approach can be generalized to more than one auxiliary variable governing response. For two variables x_1 and x_2 governing nonresponse, we would specify the set $s_k = \{(j_1,j_2) : g_{j_1} \in [g_{k_1} \pm h_{k_1}] \text{ and } g_{j_2} \in [g_{k_2} \pm h_{k_2}]\}$.

Response probabilities φ_k and ψ_{qk} are estimated respectively by the rates:

$$\hat{\varphi}_k = \frac{m_k}{n_k}, \ \forall\, k \in r; \ \hat{\psi}_{qk} = \frac{m_{qk}}{m_k}, \ \forall\, k \in r_q, \qquad (4.1)$$

whereas the output $\Theta_{qk} = \varphi_k\psi_{qk}$ is estimated by the rate:

$$\hat{\Theta}_{qk} = \hat{\varphi}_k\hat{\varphi}_{qk} = m_{qk}/n_k, \ (k \in r_q, q = 1, \ldots, Q), \qquad (4.2)$$

which is nothing other than the response rate in the k-th group. This simplification of the estimated output $\hat{\Theta}_{qk} = \hat{\varphi}_k\hat{\psi}_{qk}$ is, however, possible only when the two response mechanisms are governed by the same auxiliary variables.

Two approaches are considered here: the one based on the values of the variable x (npx) and the one based on the ranks of the values of the variable x (npr). The NPE (npx), proposed by Giommi (1987), is obtained by taking $g_k = \tau_k(x_s) = x_k$ $(k \in s)$. To offset the possible effect of excessively large and excessively small values of x_s, we introduce a variant that consists in using the ranks of x_s, that is, NPE(npr). We consider the function u such that $u(z) = 1$ if $z \geq 0$ and $u(z) = 0$ if $z < 0$. For any unit k in s, let $u_k = \sum_s u(x_k - x_j)$ = the number of components of x_s that are less than or equal to x_k = the rank of x_k in s. The NPE(npr) is then equivalent to letting $g_k = \tau_k(x_s) = u_k (k \in s)$.

4.2 Selection of Interval Limits

The main problem in the NPE approach is the optimum choice of the h_k constants that determine the limits of the intervals $[g_k - h_k; g_k + h_k]$, $\forall\, k \in s$, that is, a choice of $h_k = h_k(g_s)$ that reduces the bias and mean square error of any estimator using the estimated outputs $\hat{\Theta}_{qk}$ specified in formula (4.2).

According to Giommi (1985, 1987), the terms n_k, m_k and m_{qk} that are used to estimate the response probabilities are, apart from the standardization factors, estimators by the kernel method of the density function according to the

approach of Rosenblatt (1956) for the various series of values of g. As an example, it is easy to demonstrate that:

$$n_k = \sum_{j \in s} D(g_k - g_j) = 2nh(n)\hat{f}_n(g_k),$$

where $h(n) = h(g_k, k \in s)$ is a positive constant that converges toward zero at a quite appropriate rate. The theoretical optimum constant, according to the least mean square error criterion, is given by $h(n) = K_f n^{-1/5}$ where K_f, such as defined by Rosenblatt (1956) and Wegman (1972a and b), is obtained by the expression $K_f = [9f(x)/2 \mid f''(x)\mid^2]^{1/5}$.

In practice, $h(n)$ can be obtained only by simulation, since it depends on the density function to be estimated. Giommi (1985) used $h(n) = 2EI_s n^{-1/3}$ where EI_s is the interquartile range in the sample. Kraft, Lepage and van Eeden (1983) chose $h(n) = C(n)EI_s$ where $C(n) = (K_f/EIs)n^{-1/5}$. As our choice, we shall adopt $h(n) = C(n)S_{gs}$, where $C(n) = (K_f/S_{gs})n^{-1/5}$ and where S_{gs} is the corrected standard deviation of the values $g_k(k \in s)$. Basing ourselves on the study of Kraft, Lepage and van Eeden (1983), we will empirically determine a value $\hat{C}_n$ of C that is optimal according to the criterion of least bias and least mean square error of the estimator $\hat{t}^*_{\text{Exp}np}$ and compare the two versions of the NPE approach.

4.3 Expansion and Regression Estimators

Calculation of the approximate bias and variance of the estimators $\hat{t}_{\text{Exp}}$, $\hat{t}_{\text{Reg}}$ and $\hat{t}_{\text{Regl}}$ is simplified by the fact that the probabilities φ_k and ψ_{qk} are assumed to be known. For estimators $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg1}np}$, these probabilities are estimated by $\hat{\varphi}_k$ and $\hat{\psi}_{qk}$. These probability estimators do not respond to any probability model that would enable us to calculate the bias and the variance conditional on this model. In other words, the sets r_q are generated by unknown response mechanisms for which we estimate the response probabilities by an approach that does not allow for inference conditional on any model underlying the estimation of probabilities.

We would be tempted to resort to Taylor's serial development of the function $1/\hat{\Theta}_{qk}$ to justify the approximation of $1/\hat{\Theta}_{qk}$ by $1/\Theta_{qk}$. In this case, the bias and the variance of $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg1}np}$ would be approached by the approximate bias and variance of $\hat{t}_{\text{Exp}np}$, $\hat{t}_{\text{Reg}np}$ and $\hat{t}_{\text{Reg1}np}$. However, for sample sizes that are not sufficiently large, we are in danger of having $1/\hat{\Theta}_{qk} \not\equiv 1/\Theta_{qk}$ for the majority of the $k \in r_q$, and consequently:

$$V(\hat{t}^*_{\text{Exp}np}) \not\equiv V(\hat{t}_{\text{Exp}}),\ V(\hat{t}^*_{\text{Reg}np}) \not\equiv V(\hat{t}_{\text{Reg}}),\quad \text{and}$$
$$V(\hat{t}^*_{\text{Reg1}np}) \not\equiv V(\hat{t}_{\text{Reg1}}).$$

However, to construct confidence intervals based on $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg1}np}$, it is necessary to define estimators for their respective variances. Not having explicit expressions for these variances, it is difficult to define variance estimators and study their properties analytically. The choice of a given estimator is quite difficult to justify. The most natural way of obtaining variance estimators for the variances of $\hat{t}^*_{\text{Exp}np}$, $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg1}np}$ is to do a simple substitution of $\Theta_{qk}(=\varphi_k\psi_{qk})$, by $\hat{\Theta}_{qk}(=\hat{\varphi}_k\hat{\psi}_{qk})$, $\forall\, k \in r_q$, and of $\Theta_{qk\ell}$ by $\hat{\Theta}_{qk\ell}$, $\forall\, k \neq \ell \in r_q$ ($\hat{\Theta}_{qk\ell} = \hat{\varphi}_{k\ell}\,\hat{\psi}_{qk\ell}$), in all the formulas for variance estimators specified for the respective variance estimators of estimators $\hat{t}_{\text{Exp}np}$, $\hat{t}_{\text{Reg}np}$ and $\hat{t}_{\text{Reg1}np}$.

5. MONTE CARLO STUDY: COMPARISON OF ESTIMATORS

For simulation purposes, we assume that Bernoulli trials govern each of the response mechanisms (total or partial) and that a simple random sampling without replacement is the sample design used. We consider a vector $(y_1, y_2, y_3)'$ of three items ($Q = 3$) and a variable x containing the auxiliary information. We first generate the $x_k(k \in U)$ by a gamma distribution with parameters a_1 and a_2. The generation of items y_1, y_2, y_3 is based on the linear model (3.1) and the gamma distribution. More specifically, we generate the $y_{qk}(k \in U$ and $q = 1, 2, 3)$ according to a gamma distribution with parameters $a_{1q}(x_k)$ and $a_{2q}(x_k)$ defined by:

$$a_{1q}(x_k) = \frac{\beta_q^2 x_k}{\sigma_q^2},\quad a_{2q}(x_k) = \frac{\sigma_q^2}{\beta_q},$$

$$\sigma_q^2 = \beta_q^2 a_2 \left\{\frac{1}{\rho^2_{xy_q}} - 1\right\},\quad q = 1, 2, 3.$$

The choice of the gamma distribution is based on its general form, which gives rise to a great variety of distributions, and on the fact that it can represent the distribution of various types of populations (Johnson and Kotz 1970, p. 172). We establish *a priori* the parameters a_1, a_2, β_q and ρ_{xy_q} ($q = 1,2,3$), namely:

$$a_1 = 2,\quad a_2 = 10,\quad (\beta_1\beta_2\beta_3)' = (0.75\ 0.65\ 0.60)',$$
$$(\rho_{xy_1}\rho_{xy_2}\rho_{xy_3})' = (0.90\ 0.85\ 0.70)'.$$

To generate the unit and item response probabilities, we consider the following exponential forms:

$$\varphi_k = \exp\{-(\lambda_1 x_k + \lambda_2 v_k)\}\quad \text{and}$$
$$\psi_{qk} = \exp\{-(\lambda_{1q} x_k + \lambda_{2q} v_{qk})\},$$

where the v_k and the v_{qk} result from a uniform distribution $(0\,;1)$. The constants λ_1, λ_2, λ_{1q} and λ_{2q} are such that: $\lambda_1 = 0.15/\bar{x}_U$, $\lambda_{1q} = 0.15/\beta_q\bar{x}_U$ and $\lambda_2 = \lambda_{2q} = 0.45$ ($q = 1,2,3$). Such a parameterization makes it possible to have an average response rate (total or partial) of approximately 70%. We could have varied these constants or used other continuous functions.

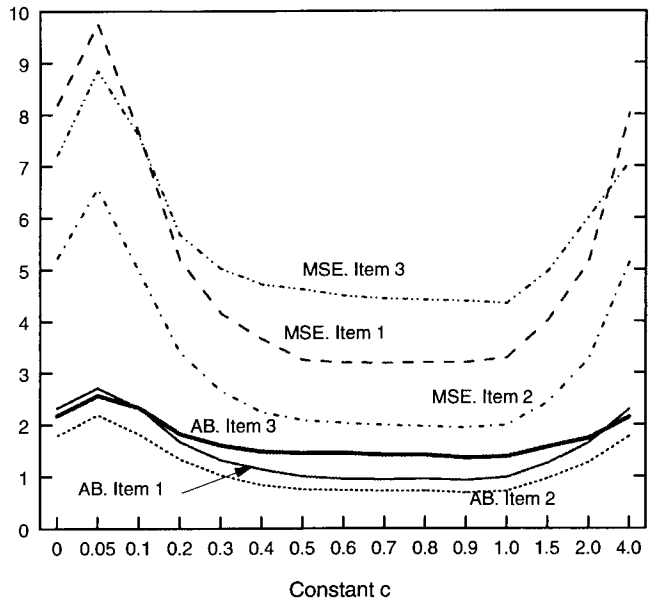

Figure 5.1 Absolute bias and MSE: the estimator $\hat{t}^*_{\text{Exp}npx}$ for $n = 60$

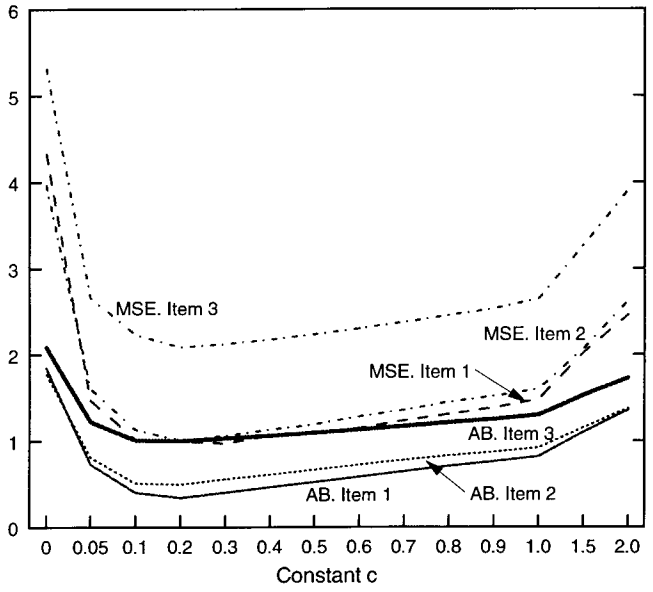

Figure 5.2 Absolute bias and MSE: the estimator $\hat{t}^*_{\text{Exp}npr}$ for $n = 200$

5.1 Comparison of the Two Variants of the NPE Approach

We consider a population of size $N = 100$ and draw a sample s of size $n = 60$, which we subject to the response mechanisms. We repeat the sampling $\mathbb{K}$ times and calculate the bias $\mathbb{B}(\hat{t}^*_{\text{Exp}np})$ and the mean-square error $\text{MSE}(\hat{t}^*_{\text{Exp}np})$, for different values of $C (C \geq 0)$. Next we repeat this experiment with $N = 1{,}000$ and $n = 200$. The results of this empirical study are illustrated by the diagrams of $\mathbb{B}(\hat{t}^*_{\text{Exp}np})$ et $\text{MSE}(\hat{t}^*_{\text{Exp}np})$ as a function of the constant C. From this brief study we observe, firstly, that the value $\hat{C}_n$ of the optimal constant C is in the interval $[0\,;1]$, depends on the size of the sample and decreases as the sample size increases (Figures 5.1 and 5.2).

We also observe that the estimator $\hat{t}^*_{\text{Exp}npr}$ is still better in terms of less bias and mean square error than the estimator $\hat{t}^*_{\text{Exp}npx}$ in the interval $[0;1]$ as illustrated as an example in Figure 5.3 for item 3, the item the least correlated with the auxiliary variable. A very important fact to be noted is that for the estimator $\hat{t}^*_{\text{Exp}npr}$ we more quickly reach the values of the bias and the mean square error of the estimator $\hat{t}_{\text{Naive}}$ in $[0;1]$ at $C = 0.05$ and outside this interval at $C = 4$. Unlike with the estimator $\hat{t}^*_{\text{Exp}npr}$, the values of the bias and the mean square error of the estimator $\hat{t}^*_{\text{Exp}npx}$ first reach maximum values at $C = 0.05$ before taking on the values of the bias and mean-square error of $\hat{t}_{\text{Naive}}$ at $C = 0$. We also note that for a fairly large size n and for any value of C in the interval $[0;1]$, the variation is hardly perceptible (Figure 5.3). For this reason, we suggest that a compromise value be used: $C = 0.5$ (that is, $h = 0.5S_{gs}$).

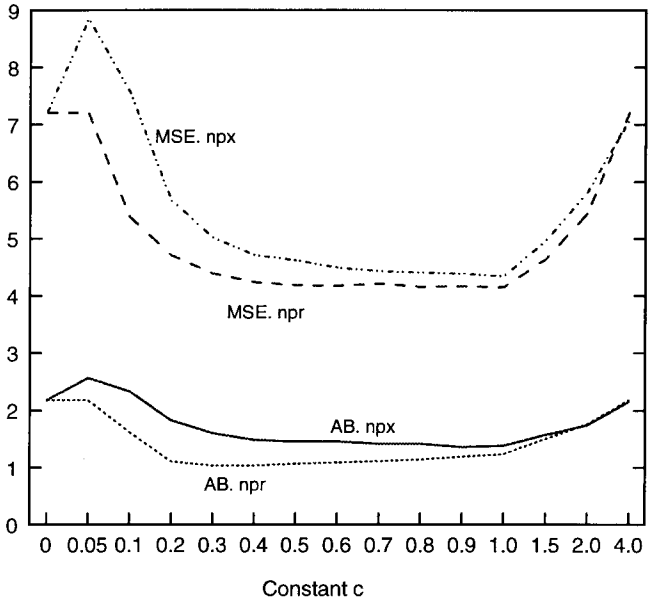

Figure 5.3 Absolute bias and MSE: the estimators $\hat{t}^*_{\text{Exp}npx}$ and $\hat{t}^*_{\text{Exp}npr}$ for item 3

5.2 Overall Comparison of Estimators

The complete operation of the simulation consists in (i) first, drawing the sample s of size $n = 200$ of the population of size $N = 1{,}000$, (ii) then applying the unit and item response mechanisms to obtain sets $r_q (q = 1,2,3)$, and (iii) lastly, calculating, for each estimator, the values

of $\hat{t}$ and $\hat{V}(\hat{t})$. We repeat this operation $\mathbb{K}$ times. Once the experiment is completed, we calculate, as performance measurements, (i) the bias $\mathbb{B}(\hat{t}) = \mathbb{E}(\hat{t}) - t_q$, (ii) the mean square error $\mathrm{MSE}(\hat{t}) = \mathbb{E}(\hat{t} - t_q)^2$, (iii) the expectation of the variance estimator $\mathbb{E}(\hat{V}(\hat{t}))$ and (iv) the theoretical recovery rate $P_o(\hat{t}) = \mathbb{P}\{\mid \hat{t} - t_q \mid \leq Z_{\alpha/2}[V(\hat{t})]^{1/2}\}$. We can also calculate, for each given estimator, (v) the relative error $\mathrm{RE}(\hat{t})\,[= \mathbb{B}(\hat{t})/t]$, (vi) the variance $V(\hat{t})\,[= \mathrm{MSE}(\hat{t}) - (\mathbb{B}(\hat{t}))^2]$, (vii) the relative bias $\mathrm{RB}(\hat{t})\,[= \mid \mathbb{B}(\hat{t}) \mid /(V(\hat{t}))^{1/2}]$ as well as (viii) the relative error of the variance estimator $\mathrm{RE}(\hat{V}(\hat{t}))$ $[= \mathbb{B}(\hat{V}(\hat{t}))/V(\hat{t})]$ in order to examine the sensitivity of the variance estimators to nonresponse.

5.3 Interpretation of the Results of the Global Simulation

I. The Prototype Estimators

The simulation results confirm the theory. For these estimators, we make the following observations, based on Tables 5.1 to 5.4:

(i) $\hat{t}_{\mathrm{Exp}}$, $\hat{t}_{\mathrm{Reg}}$ and $\hat{t}_{\mathrm{Reg1}}$ are approximately unbiased;

(ii) $\mathrm{MSE}(\hat{t}_{\mathrm{Reg1}}) < \mathrm{MSE}(\hat{t}_{\mathrm{Reg}}) < \mathrm{MSE}(\hat{t}_{\mathrm{Exp}})$;

(iii) $V(\hat{t}_{\mathrm{Reg1}}) < V(\hat{t}_{\mathrm{Reg}}) < V(\hat{t}_{\mathrm{Exp}})$ and $\mathbb{E}[\hat{V}(\hat{t}_{\mathrm{Reg1}})] < \mathbb{E}[\hat{V}(\hat{t}_{\mathrm{Reg}})] < \mathbb{E}[\hat{V}(\hat{t}_{\mathrm{Exp}})]$.

For these estimators, we also expected that:

(i) $\mathbb{E}\hat{V}(\hat{t}_{\mathrm{Exp}}) \approx V(\hat{t}_{\mathrm{Exp}})$, $\mathbb{E}\hat{V}(\hat{t}_{\mathrm{Reg}}) \approx V(\hat{t}_{\mathrm{Reg}})$ and $\mathbb{E}\hat{V}(\hat{t}_{\mathrm{Reg1}}) \approx V(\hat{t}_{\mathrm{Reg1}})$;

(ii) Negligible relative bias $[\mathrm{RB}(\hat{t}) < 0.10]$; the recovery rates are close to the theoretical rates. The relative errors $\mathrm{RE}(\hat{t})$ and $\mathrm{RE}(\hat{V}(\hat{t}))$ are negligible, and are in part due to the simulation (errors due to the limited number of repetitions of the experiment).

Table 5.1
The Values of $\mathbb{B}(\hat{t})$, $\mathrm{MSE}(\hat{t})$

	y_1		y_2		y_3	
$\hat{t}_{\mathrm{Exp}}$	−0.036	1.690	−0.052	1.525	−0.056	2.299
$\hat{t}_{\mathrm{Reg}}$	−0.020	0.735	−0.019	0.744	−0.030	1.446
$\hat{t}_{\mathrm{Reg1}}$	−0.012	0.319	−0.012	0.431	−0.021	1.202
$\hat{t}_{\mathrm{Naive}}$	−2.037	5.069	−1.937	4.535	−2.220	5.911
$\hat{t}^*_{\mathrm{Exp}npx}$	−0.690	1.345	−0.777	1.407	−1.228	2.604
$\hat{t}^*_{\mathrm{Exp}npr}$	−0.601	1.175	−0.709	1.249	−1.140	2.345
$\hat{t}^*_{\mathrm{Reg}npr}$	−0.293	0.785	−0.414	0.830	−0.895	1.834
$\hat{t}^*_{\mathrm{Reg1}npr}$	−0.285	0.376	−0.407	0.520	−0.886	1.621

Table 5.2
The Values of $V(\hat{t})$, $\mathbb{E}[\hat{V}(\hat{t})]$ and $100*\mathbb{E}[\hat{V}_1(\hat{t})]/\mathbb{E}[\hat{V}(\hat{t})]$

	y_1			y_2			y_3		
$\hat{t}_{\mathrm{Exp}}$	1.689	1.683	29.8	1.525	1.485	29.1	2.296	2.235	26.9
$\hat{t}_{\mathrm{Reg}}$	0.734	0.697	72.2	0.744	0.702	61.5	1.445	1.391	42.6
$\hat{t}_{\mathrm{Reg1}}$	0.319	0.293	34.0	0.431	0.402	32.7	1.201	1.130	29.3
$\hat{t}_{\mathrm{Naive}}$	0.918	0.911	43.3	0.784	0.766	43.5	0.983	0.958	44.2
$\hat{t}^*_{\mathrm{Exp}npx}$	0.869	1.403	32.0	0.804	1.173	32.3	1.097	1.322	35.4
$\hat{t}^*_{\mathrm{Exp}npr}$	0.814	1.291	35.1	0.746	1.089	35.2	1.046	1.285	37.1
$\hat{t}^*_{\mathrm{Reg}npr}$	0.700	0.627	73.9	0.658	0.588	66.6	1.033	0.955	50.5
$\hat{t}^*_{\mathrm{Reg1}npr}$	0.294	0.259	36.7	0.355	0.315	37.6	0.836	0.751	37.1

Table 5.3
The Values of $\mathrm{RE}(\hat{t})$ and $\mathrm{RE}(\hat{V}(\hat{t}))$

	y_1		y_2		y_3	
$\hat{t}_{\mathrm{Exp}}$	−0.0024	−0.0015	−0.0040	−0.0242	−0.0045	−0.0267
$\hat{t}_{\mathrm{Reg}}$	−0.0014	−0.0510	−0.0015	−0.0556	−0.0024	−0.0373
$\hat{t}_{\mathrm{Reg1}}$	−0.0008	−0.0812	−0.0009	−0.0684	−0.0017	−0.0596
$\hat{t}_{\mathrm{Naive}}$	−0.1377	−0.0083	−0.1474	−0.0230	−0.1787	−0.0260
$\hat{t}^*_{\mathrm{Exp}npx}$	−0.0466	0.6141	−0.0591	0.4582	−0.0988	0.2046
$\hat{t}^*_{\mathrm{Exp}npr}$	−0.0406	0.5860	−0.0540	0.4591	−0.0917	0.2282
$\hat{t}^*_{\mathrm{Reg}npr}$	−0.0198	−0.1038	−0.0315	−0.1077	−0.0720	−0.0752
$\hat{t}^*_{\mathrm{Reg1}npr}$	−0.0193	−0.1191	−0.0310	−0.1124	−0.0713	−0.1015

Table 5.4
The Levels $P_o(\hat{t})$ at 90%, 95% and the $\mathrm{RB}(\hat{t})$

	y_1			y_2			y_3		
$\hat{t}_{\mathrm{Exp}}$	0.873	0.922	0.027	0.870	0.914	0.042	0.852	0.904	0.037
$\hat{t}_{\mathrm{Reg}}$	0.881	0.929	0.024	0.876	0.929	0.022	0.870	0.917	0.025
$\hat{t}_{\mathrm{Reg1}}$	0.866	0.926	0.021	0.873	0.923	0.018	0.860	0.914	0.019
$\hat{t}_{\mathrm{Naive}}$	0.322	0.427	2.126	0.298	0.405	2.187	0.287	0.389	2.239
$\hat{t}^*_{\mathrm{Exp}npx}$	0.851	0.906	0.740	0.800	0.874	0.866	0.667	0.758	1.172
$\hat{t}^*_{\mathrm{Exp}npr}$	0.872	0.925	0.666	0.830	0.893	0.820	0.700	0.789	1.114
$\hat{t}^*_{\mathrm{Reg}npr}$	0.839	0.908	0.350	0.806	0.878	0.510	0.712	0.789	0.880
$\hat{t}^*_{\mathrm{Reg1}npr}$	0.804	0.871	0.526	0.767	0.844	0.683	0.678	0.763	0.969

II. The Naive Estimator

The naive estimator registers absolute values of $\mathbb{B}(\hat{t})$ and $\mathrm{RE}(\hat{t})$ that are very high in relation to the other estimators (Tables 5.1 and 5.3). The same is true for the values of $\mathrm{MSE}(\hat{t})$ (Table 5.1). The values of the observed recovery rates $P_o(\hat{t})$ as well as those of the relative bias $\mathrm{RB}(\hat{t})$ are hardly surprising, considering the size of the point estimate bias (Table 5.4).

The behaviour, in terms of variance and variance estimator (Table 5.2) of $\hat{t}_{\text{Naive}}$, is due to the fact that it constitutes a particular case of $\hat{t}_{\text{Exp}}$, assuming uniform response mechanisms. In a sense, this amounts to assuming that the data are missing randomly.

III. The Adjusted Estimators

The reduction of the bias and the mean square error resulting from the use of the adjusted estimators (Table 5.1) is quite significant, in comparison with the naive estimator, especially for the regression estimators (the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$). In terms of variance (Table 5.2), we have the following inequalities:

$$V(\hat{t}^*_{\text{Regl}npr}) < V(\hat{t}^*_{\text{Reg}npr}) < V(\hat{t}^*_{\text{Exp}npr}) < V(\hat{t}^*_{\text{Exp}npx}),$$

which are analytically difficult to demonstrate. Little variation [in terms of $V(\hat{t})$ and $\mathbb{E}(\hat{V}(\hat{t}))$] is observed between items y_1 and y_2 in light of the little variation between the correlations (0.05). On the other hand, the effect of the correlation with the auxiliary variable on $V(\hat{t})$ and of $\mathbb{E}(\hat{V}(\hat{t}))$ may be observed by comparing items y_1 and y_3, then y_2 and y_3: the variations between the correlations are greater in these two cases (0.20 and 0.15 respectively).

In terms of variance estimators (Table 5.2), we observe that:

$$\hat{V}(\hat{t}^*_{\text{Regl}np}) < \hat{V}(\hat{t}^*_{\text{Reg}np}) < \hat{V}(\hat{t}^*_{\text{Exp}np}),$$

as such is the case for the estimators $\hat{t}_{\text{Reg}}$, $\hat{t}_{\text{Regl}}$ and $\hat{t}_{\text{Exp}}$. What is surprising, and is of course due to the effect of the auxiliary variables on the variance components relative to the response mechanisms, is the fact that the estimators $\hat{t}^*_{\text{Exp}np}$ overestimate the variance with very large absolute values of $\text{RE}(\hat{V}(\hat{t}))$, while the regression estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$ underestimate the variance with absolute values of $\text{RE}(\hat{V}(\hat{t}))$ that are smaller in relation to those of $\hat{t}^*_{\text{Exp}np}$ (Table 5.3). For the estimators $\hat{t}^*_{\text{Exp}np}$, not only is the total variance high in relation to that of the regression estimators, but also the relative contribution of the sampling variance is low (Table 5.2).

In terms of recovery rate (Table 5.4), the estimators $\hat{t}^*_{\text{Exp}np}$ yield observed rates that are closer to theoretical rates than the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$. However, the values of the relative bias $\text{RB}(\hat{t})$ are higher for $\hat{t}^*_{\text{Exp}np}$ than for $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$, which makes the confidence intervals less reliable.

IN CONCLUSION

(i) If the goal of the estimation is to reduce bias and mean square error, all the estimators adjusted for nonresponse perform well in relation to the uniform response mechanism (which basically amounts to doing nothing about nonresponse). The rate of reduction of the bias of each estimator in relation to the naive estimator is at least 66%. The regression estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$ are the most promising of the various estimators considered (Table 5.1).

(ii) If the goal is to construct confidence intervals, we need a pair of estimators $[\hat{t}, \hat{V}(\hat{t})]$ that simultaneously minimize the absolute biases $|\ \mathbb{B}(\hat{t})\ |$ and $|\ \mathbb{B}(\hat{V}(\hat{t}))\ |$. Tables 5.1 and 5.2 clearly show that the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Regl}np}$ are the best. These estimators are less sensitive to nonresponse if we consider the values of $\text{RE}(\hat{t})$ and $\text{RE}(\hat{V}(\hat{t}))$ (Table 5.3). Nevertheless the criterion of reliability of the confidence intervals ($\text{RB}(\hat{t}) < 0.10$) is never met (Table 5.4).

(iii) The behaviour of the estimators adjusted (i) for item y_1, which is the item the most highly correlated with the auxiliary variable, compared to item y_3, then (ii) for item y_2 compared to item y_3 (y_3 being the item that is least correlated with the auxiliary variable), shows that with very strong explanatory variables (for y_q and for Θ_{qk}), better results can be achieved not only in terms of less bias $|\ \mathbb{B}(\hat{t})\ |$ and $|\ \mathbb{B}(\hat{V}(\hat{t}))\ |$ but also in terms of less mean square error (a gain in precision in relation to the naive estimator) and a better recovery rate for the confidence intervals (Tables 5.1 to 5.4).

(iv) The behaviour of the estimators $\hat{t}^*_{\text{Reg}np}$ and $\hat{t}^*_{\text{Reg}np}$, in terms of bias, variance and variance estimation, is consistent with the studies conducted by Särndal and Hui (1981), Särndal and Swenson (1985, 1987), Bethlehem (1988) and Kott (1987) on the usefulness of regression estimators in nonresponse situations and the importance of having good predictor variables for the items of interest and the response mechanisms.

ACKNOWLEDGEMENTS

I wish to express my thanks to Carl-Erik Särndal for his support in every sense of the word in the writing of my Ph.D. thesis, on which this article is based. Despite his many responsibilities and the other demands on his time, he taught me a great deal in this field of sampling, which he masters so well and in which he has become a figure of international prominence through his many published works (articles and books) and collaborative efforts.

I would also like to thank the referees and the Associate Editor for their constructive comments. On the one hand, their observations and suggestions improved the original version of this article. On the other, they provided ideas for subsequent studies.

REFERENCES

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

CHICOINEAU, F., PAYEN, J.F., and THÉLOT, C. (1985). Modélisation et redressement des non-réponses: le cas du salaire. *Bulletin of the International Statistical Institute*, LI-3, 15.3, 1-23.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd Ed.). New York: Wiley.

GIOMMI, A. (1985). On the estimation of the individual response probabilities. *Bulletin of the International Statistical Institute*, 2, 577-578.

GIOMMI, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology*, 13, 127-134.

GROSBRAS, J.-J. (1987b). Les réponses manquantes. In *Les sondages*. (Eds. J.-J. Droesbeke, B. Fichet and F. Tassi). Paris: Economica.

JOHNSON, N.L., and KOTZ, S. (1970). *Continous univariate distributions-1*. New York: Houghton.

KOTT, P.S. (1987). Nonresponse in a periodic sample survey. *Journal of Business and Economic Statistics*, 5, 287-293.

KRAFT, C.H., LEPAGE, Y., and VAN EEDEN, C. (1983). Some finite-sample size properties of Rosenblatt density estimates. *The Canadian Journal of Statistics*, 11, 95-104.

OH, H.L., and SCHEUREN, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin and D.B. Rubin), 2, 143-184. New York: Academic Press.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of the density function. *Annals of Mathematical Statistics*, 27, 832-837.

SÄRNDAL, C.-E., and HUI, T-K. (1981). Estimation for non-response situations: to what extent must we rely on models? In *Current Topics in Survey Sampling*. (Eds. D. Krewski, R. Platek and J.N.K. Rao), 227-246. New York: Academic Press.

SÄRNDAL, C.-E., and SWENSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Bulletin of the International Statistical Institute*. LI-3, 15.2, 1-16.

SÄRNDAL, C.-E., and SWENSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

WEGMAN, E.J. (1972a). Nonparametric probability density estimation: A summary of available methods. *Technometrics*, 14, 533-546.

WEGMAN, E.J. (1972b). Nonparametric probability density estimation: A comparison of density estimation methods. *Journal of Statistical Computations and Simulations*, 1, 225-245.