ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 10, 2006

The analysis of data from sample surveys under informative sampling

Abdulhakeem A.H. Eideh and Gad Nathan

Abstract. Sampling designs for surveys are often complex and informative, in the sense that the selection probabilities are correlated with the variables of interest, even when conditioned on explanatory variables. In this case conventional analysis that disregards the informativeness can be seriously biased, since the sample distribution differs from that of the population. In this paper we consider the relationships between the distribution of the sampled values and that of the population. Using different models for the conditional expectations of the inclusion probabilities, given the values of the variable of interest, we obtain the sample distributions and propose methods for estimation of their parameters. The results are applied to the analysis of longitudinal surveys, using an autoregressive model, and to surveys with two-stage cluster designs, with informative selection at each of the two stages.

1. Introduction

Some recent work, e.g., Pfeffermann, Krieger, and Rinott (1998), has considered the definition of a sample distribution under informative sampling. Survey data may be viewed as the outcome of two processes: the process that generates the values of a random variable for units in a finite population, often referred to as the superpopulation model, and the process of selecting the sample units from the finite population values, known as the sample selection mechanism. Analytic inference from survey data refers to the superpopulation model. When the sample selection probabilities depend on the values of the model response variable, even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process.

Received July 18, 2006.
2000 *Mathematics Subject Classification.* 62D05.
Key words and phrases. Informativeness, sample distribution, longitudinal surveys, autoregressive model, two-stage sample designs.

2. General methods of estimation for complex survey designs

Classical methods of estimation and inference for data from complex sample designs do not, in general, take informativeness into account. Thus, the usual Horvitz-Thompson estimator of a total, $T = \sum_{i=1}^{N} y_i$, is $\hat{T} = \sum_{i=1}^{N} y_i/\pi_i$, which weights the sample values by the reciprocals of the inclusion probabilities, π_i. The estimator is design-unbiased, i.e., $E_D[\hat{T}] = T$, where $E_D[.]$ denotes expectation under repeated sampling. However, if the sample design is informative, the estimator is model-biased.

An alternative, the pseudo-likelihood method (Binder, 1983), is based on solving the sample estimates of the population likelihood equations. Let $y_1, \ldots, y_N$ be the values of y in the finite population. These are considered as random variables with the pdf $f_p(y_i|\theta)$, where θ is the unknown super-population parameter. If all population units were observed, the MLE of θ is defined as the solution to the equations

$$U(\theta) = \sum_{i=1}^{N} \frac{\partial(\log f_p(y_i|\theta))}{\partial\theta} = 0. \tag{1}$$

The pseudo maximum likelihood estimator of θ is defined as the solution of the estimating equations, that is as the solution of equation (1), with a sample estimate of $U(\theta)$, i.e., $\hat{U}(\theta) = 0$.

We consider, instead, the relationships between the population distribution and the sample distribution. Let y_i be a random variable with a population pdf $f_p(y_i|\theta)$, where θ is an unknown parameter. Let $\pi_i = \Pr(i \epsilon s|\mathbf{y}, \mathbf{x})$ be the (conditional) inclusion probability of unit i, given $\mathbf{y}$, the variables of interest, and $\mathbf{x}$, the auxiliary variables. The sample distribution is given, under very general conditions, by

$$f_s(y_i) = \frac{E_p(\pi_i|y_i)}{E_p(\pi_i)} f_p(y_i) \tag{2}$$

(Pfeffermann, Krieger, and Rinott, 1998).

Equation (2) defines the relationship between the population and sample distributions, so that if π_i depends on y_i, then $E_p(\pi_i|y_i) \neq E_p(\pi_i)$ and $f_p(y_i) \neq f_s(y_i)$. In this case the population distribution differs from the sample distribution and the sample design is informative.

In order to evaluate the sample distribution, we consider the following four different models for the relationships between the conditional expectations of the inclusion probabilities, π_i, and the values of y_i:

The exponential model $E_p(\pi_i|y_i) = \exp(a_0 + a_1 y_i)$.
The linear model $E_p(\pi_i|y_i) = b_0 + b_1 y_i$.
The logit model $E_p(\pi_i|y_i) = \frac{exp(c_0+c_1 y_i)}{1+exp(c_0+c_1 y_i)}$.
The Probit model $E_p(\pi_i|y_i) = \Phi(d_0 + d_1 y_i)$.

Applying equation (2) to these models gives the following results:
For the exponential model:

$$f_s(y_i|\mathbf{x}_i) = \frac{\exp(a_1 y_i) f_p(y_i|\mathbf{x}_i)}{M_p(a_1)}, \tag{3}$$

where $M_p(a_1)$ is the moment generating function (mgf) of the population pdf of y_i and $\mathbf{x}_i$ is a vector of auxiliary variables.

For the linear model:

$$f_s(y_i|\mathbf{x}_i) = b_0^* f_p(y_i|\mathbf{x}_i) + b_1^* \frac{y_i f_p(y_i|\mathbf{x}_i)}{E_p(y_i|\mathbf{x}_i)}, \tag{4}$$

where $b_0^* = \frac{b_0}{b_0 + b_1 E_p(y_i|\mathbf{x}_i)}$ and $b_1^* = 1 - b_0^*$. Similar results are obtained for the logit and for the probit models.

To illustrate these results we consider, as an example, the situation under a linear regression model. Let the population distribution be given by

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

and assume that the conditional expectations of the inclusion probabilities follow the exponential model

$$E_p(\pi_i|y_i) = \exp(a_0 + a_1 y_i).$$

Then the sample distribution is given by

$$y_i|x_i \sim N(\beta_0 + a_1\sigma^2 + \beta_1 x_i, \sigma^2).$$

Thus the sample distribution remains normal with a shift of the mean by $a_1\sigma^2$ and the same variance as that of the population distribution. It can be seen, therefore, that the sample design is not informative, i.e., the sample distribution is the same as the population distribution, if and only if $a_1 = 0$.

3. Estimation of the parameters

Next we consider the possibilities for estimation of the unknown parameters. In practice, the conditional expectations of the sample inclusion probabilities, required to evaluate the sample distribution (2) are not known. Assuming that the only available data to the analyst are the sample weights, w_i, and sample values of y_i, which is the case in secondary analysis, the question that arises is: how can we identify and estimate, $E_p(y_i)$, based only on the sample data? The answer is provided by the following relationships between the sample expectations of the sample weights, w_i, and the population expectations of the inclusion probabilities, π_i:

$$E_s(w_i|y_i) = \frac{1}{E_p(\pi_i|y_i)} \tag{5}$$

(Pfeffermann and Sverchkov, 1999).

Under the exponential model $E_p(\pi_i|y_i) = \exp(a_0 + a_1 y_i)$ for the relationships between the conditional expectations of the inclusion probabilities and y_i, the application of (5) implies that $\log(E_s(w_i|y_i) = -(a_0 + a_1 y_i)$. Thus the first stage of the estimation procedure is the estimation, by ordinary least squares, of a_0 and of a_1 from the simple regression relationship of $W_i = -log(w_i)$ on y_i. Alternatively, we could also estimate a_1 directly, using non-linear regression techniques. In the second stage we substitute, $\hat{a}_1$, the estimator of a_1, in the sample distribution (3). Differentiation of the resulting sample likelihood yields a system of non-linear equations, whose numeric solution provides the required estimates.

Similarly, under the linear model $E_p(\pi_i|y_i) = b_0 + b_1 y_i$, in the first stage we estimate b_0 and b_1 from the simple regression relationship of π_i on y_i. In the second stage, $\hat{b}_0$ and $\hat{b}_1$, the estimators of b_0 and b_1, are substituted in the sample distribution (4). Again, differentiation of the resulting sample likelihood yields a system of non-linear equations, whose numeric solution provides the required estimates. An alternative method of estimation is that based on pseudo-likelihood. The census log-likelihood is given by

$$\ell_c(\theta|y_1, \ldots, y_N) = \log \prod_{i=1}^{N} f_p(y_i|\theta) = \sum_{i=1}^{N} \log[f_p(y_i|\theta)], \tag{6}$$

under the assumption of super-population independence between $y_1, \ldots, y_N$. The census maximum likelihood estimators of θ are defined as the solutions of the equations

$$U_c(\theta) = (\frac{\partial}{\partial\theta})\ell_c(\theta|y_1, \ldots, y_N) = \sum_{i=1}^{N} \frac{\partial[\log f_p(y_i|\theta)]}{\partial\theta} = 0. \tag{7}$$

The sample based pseudo maximum likelihood estimators are the solutions to the weighted estimators of the census likelihood equations

$$\hat{U}_w(\theta) = (\frac{\partial}{\partial\theta})\ell_w(\theta) = \sum_{i\epsilon s} w_i \frac{\partial[\log f_p(y_i|\theta)]}{\partial\theta} = 0. \tag{8}$$

4. Tests of informativeness

A sample design is non-informative if and only if $f_s(y_i) = f_p(y_i)$ for all y_i. In this section we examine possible tests of the hypothesis that the design is non-informative, based on the sample data, and a measure of the degree of informativeness. An alternative formulation of the hypothesis of non-informativeness is:

$$\frac{E_s(w_i y_i^k)}{E_s(w_i)} = E_s(y_i^k), \tag{9}$$

for all $k = 1, 2, \ldots$ and for all $i \epsilon s$. Thus the test of the hypothesis that the sample design is non-informative can be represented by the series of tests:

$$H_{0k} : corr_s(y_i^k, w_i) = 0, k = 1, 2, \ldots, \tag{10}$$

which can be tested by standard methods (Pfeffermann and Sverchkov, 1999).

While in theory this requires an infinite number of tests, in practice only a limited number is required. However, there still remains the problem of multiple testing and an alternative based on the Kullback-Leibler information measure of the distance between two distributions (Kullback and Leibler, 1951) is proposed. This measure of minimal discrimination between the sample distribution, f_s, and the population distribution, f_p, is defined as:

$$I(f_s; f_p) = E_s[\log f_s(y_i) - \log f_p(y_i)]. \tag{11}$$

Using the relationships (2) and (5), it can be shown that this measure may be written as:

$$I(f_s; f_p) = E_s[\log E_s(w_i)] - E_s[\log E_s(w_i|y_i)]. \tag{12}$$

Notice that the expected value is taken under the sample distribution, f_s, which means that we are assuming that y_i has pdf f_s and the hypothesis that this is equal to the population distribution, f_p, can be tested on the basis of the sample values. For example, assume that the population distribution is exponential with parameter θ, i.e., $y_i \sim \exp(\theta)$ and that the conditional expectations of the sample selection probabilities follow the exponential model $E_p(\pi_i|y_i) = \exp(a_0 + a_1 y_i)$. In this case the Kullback-Leibler information measure can be shown to be

$$I(f_s; f_p) = E_s[\log \frac{f_s(y_i)}{f_p(y_i)}] = \log(\frac{\theta - a_1}{\theta}) + \frac{a_1}{\theta - a_1}. \tag{13}$$

Substituting the estimates of θ and of a_1 in (13), the statistic for testing non-informativeness can be written as:

$$\hat{I}(f_s; f_p) = \log(\frac{\hat{\theta} - \hat{a}_1}{\hat{\theta}}) + \frac{\hat{a}_1}{\hat{\theta} - \hat{a}_1}, \tag{14}$$

which has the asymptotic χ_1^2 distribution under the null hypothesis (Kullback, 1978, Section 5.5).

5. Application to longitudinal surveys

Recently there is increasing interest in longitudinal surveys - those for which variables or characteristics are measured for the same units at different points of time (occasions or waves). Each series of observations for a unit can be viewed therefore as a time series, usually of short length. Longitudinal data can be collected prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on the same individual from a panel survey or from historical records. Often the sample

units selected at the first point in time are retained for observation at subsequent points in time and the selection may be informative.

We consider the following typical situation for a longitudinal survey. Observations $y_{i1}, \ldots, y_{iT}$ for T consecutive periods are obtained for each sampled unit, i. We assume the first-order autoregressive model - AR(1):

$$y_{it} = \mu + \phi(y_{i,t-1} - \mu) + \varepsilon_{it}; i = 1, \ldots, N; t = 2, \ldots, T, \tag{15}$$

where $y_{i1} \sim N(\mu, \frac{\sigma^2}{1-\phi^2}), \varepsilon_{it} \sim_{ind} N(0, \sigma^2)$ and $|\phi| < 1$ to ensure stationarity. We assume that units selected for the first period remain in the sample over all T periods.

We assume that the vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ are independently distributed with the population distribution $f_p(\mathbf{y}_i) = f_p(\mathbf{y}_i|\theta)$, depending on the unknown parameters θ, which is given by

$$f_p(\mathbf{y}_i) = f_p(y_{i1}) \prod_{t=2}^{T} f_p(y_{it}|H_{i,t-1}), \tag{16}$$

where $H_{i,t-1}$ are the observations on unit i until and including time $t-1$. Applying the autoregressive model (15), we obtain for the population distribution

$$f_p(\mathbf{y}_i) = (\frac{2\pi\sigma^2}{1-\phi^2})^{-\frac{1}{2}} \exp[-\frac{1-\phi^2}{2\sigma^2}(y_{i1} - \mu)^2](2\pi\sigma^2)^{-\frac{T-1}{2}}$$
$$\exp[-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\{(y_{it} - \mu) - \phi(y_{i,t-1} - \mu)\}^2]. \tag{17}$$

Thus for the exponential model $E_p(\pi_i|y_{i1}) = \exp(a_0 + a_1 y_{i1})$, we obtain for the sample distribution

$$f_s(\mathbf{y}_i) = (\frac{2\pi\sigma^2}{1-\phi^2})^{-\frac{1}{2}} \exp[-\frac{1-\phi^2}{2\sigma^2}(y_{i1} - \mu - a_1\frac{\sigma^2}{1-\phi^2})^2](2\pi\sigma^2)^{-\frac{T-1}{2}}$$
$$\exp[-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\{(y_{it} - \mu) - \phi(y_{i,t-1} - \mu)\}^2]. \tag{18}$$

It can easily be seen that this differs from the population distribution (17) only in the change of the mean of y_{i1} from μ to $\mu + a_1\frac{\sigma^2}{1-\phi^2}$.

Similarly for the linear model $E_p(\pi_i|y_{i1}) = b_0 + b_1 y_{i1}$, the sample distribution is given by

$$f_s(\mathbf{y}_i) = \frac{b_0 + b_1 y_{i1}}{b_0 + b_1\mu} f_p(y_{i1}) \prod_{t=2}^{T} f_p(y_{it}|H_{i,t-1}). \tag{19}$$

Note that the sample design is non-informative in this case if and only if $b_1 = 0$ and $b_0 \neq 0$.

6. Simulation study for longitudinal survey application

In order to demonstrate the above results, a small simulation study was carried out by generating $N = 5000$ population values for the first period by $y_{i1} \sim N(\mu, \frac{\sigma^2}{1-\phi^2})$. For the remaining periods, population values of y_{it}, $t = 2, \ldots, T$, were generated by the autoregressive model (15), with $T = 3$ and $T = 10$. Five hundred samples of size $n = 500$ were selected by systematic *PPS* (probability proportional to size) sampling, with inclusion probabilities $\pi_i = \frac{nz_i}{\sum_{j=1}^{N} z_j}$, where values of z_i were determined by each of three alternative models:

The exponential model	$z_i = \exp(0.5 + 0.2y_i + u_i)$;	$u_i \sim U(0, 1)$.
The linear model	$z_i = 4 + 5y_i + u_i$;	$u_i \sim U(0, 25)$.
Non-informative sampling	$z_i = \exp(u_i)$;	$u_i \sim U(0, 4)$.

The estimation methods tested were:

- Unweighted maximum likelihood, as if sampling was non-informative (UML)
- Weighted maximum (pseudo-)likelihood (WML)
- Sample maximum likelihood - exponential model (SMLE)
- Sample maximum likelihood - linear model (SMLL)

The relative mean square errors (RMSE - the empirical MSE divided by θ) of each of the three parameters, for each of the three models and for each of the four methods of estimation are given in Table 1, for $T = 3$.

		Estimation Method			
The true model	**Parameter**	**UML**	**WML**	**SMLE**	**SMLL**
Exponential	μ	0.0600	0.0422	0.0184	0.0188
	σ^2	0.0422	0.0453	0.0423	0.0423
	ϕ	0.0157	0.0161	0.0165	0.0165
Linear	μ	0.0562	0.0184	0.0185	0.0174
	σ^2	0.0407	0.0451	0.0406	0.0408
	ϕ	0.0185	0.0185	0.0175	0.0177
Non informative	μ	0.0174	0.0308	0.0237	0.0222
	σ^2	0.0404	0.0754	0.0405	0.0405
	ϕ	0.0165	0.0277	0.0163	0.0164

TABLE 1. Relative Mean Square Errors (T=3)

The main results are as follows:

- If the true model is informative, the unweighted (UML) and the weighted maximum pseudo-likelihood (WML) estimators of μ have

much larger mean square errors than the estimators based on methods (SMLE, SMLL).

- If sampling is non-informative the errors of the model-based estimators of μ (SMLE and SMLL) do not differ much from those of the unweighted estimator (UML) but are much smaller than those of the weighted estimator (WML).
- The differences between the two sample maximum likelihood estimators based on different methods (SMLE, SMLL) are small.
- The model-based estimators, SMLE and SMLL, are relatively robust to departures from the assumed models.
- The differences between the errors of the estimators of the other parameters, σ^2 and ϕ, are small. Similar results (not shown) were obtained for the case of T=10.

7. Application to two-stage cluster sampling

A very common sample design is the two-stage cluster sample design. This is usually the suitable design chosen for sampling from a population which has a hierarchical structure, e.g., households within localities or pupils within schools, where the costs of investigating each higher level unit (e.g., locality) are high compared to the marginal cost of investigating a second-level unit, e.g., household. We assume the following hierarchical population model (with random intercepts):

$$\begin{array}{ll} \text{First level:} & \mu_i = \mathbf{z}_i'\gamma + \eta_i \\ \text{Second level:} & y_{ij}|\mu_i = \mu_i + \mathbf{x}_{ij}'\beta + e_{ij}, \end{array} \tag{20}$$

where: $i = 1, \ldots, N; j = 1, \ldots, M_i; \eta_i \sim N(0, \sigma_\eta^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$ are independent; $\beta = (\beta_1, \ldots, \beta_p)'$, $\gamma = (\gamma_1, \ldots, \gamma_q)'$ are vectors of unknown fixed regression parameters; and $\mathbf{z}_i$, $\mathbf{x}_{ij}$ are vectors of known auxiliary variables, at the first and second level, respectively.

The sample design is assumed to be a two-stage cluster sampling design with (possibly) informative sampling for the first and second stages. Let $g_i, g_{ij}, i = 1, \ldots, N; j = 1, \ldots, M_i$ be random design variables, used for the sample selection, but not included in the working model under consideration.

- *First stage sampling*: A sample, s, of size n of primary sampling units (*PSU*'s or *clusters*), is selected, with inclusion probabilities

$$\pi_i = Pr(i \epsilon s | \mu_i, \mathbf{z}_i, g_i) = h_1(\mu_i, \mathbf{z}_i, g_i).$$

- *Second stage sampling*: A sample, s_i, of size m_i of secondary sample units, is selected from the i-th selected PSU, with conditional inclusion probabilities

$$\pi_{j|i} = Pr(j \epsilon s_i | i \epsilon s, y_{ij}, \mathbf{x}_{ij}, g_{ij}) = h_2(y_{ij}, \mathbf{x}_{ij}, g_{ij}).$$

The sample distributions of the first stage means, μ_i, depend on the model assumed for the conditional expectations of the first stage selection probabilities:

- *Under the exponential model,* $E_p(\pi_i|\mu_i, \mathbf{z}_i) = g_e(\mathbf{z}_i)\exp[(b_0 + b_1\mu_i)]$:

$$\mu_i \sim N(\mathbf{z}_i'\gamma + b_1\sigma_\mu^2, \sigma_\mu^2),$$

  so that the sample distribution of μ_i, is independent of b_0 and of $g_e(\mathbf{z}_i)$ and only its mean is shifted by $b_1\sigma_\mu^2$ from that of the population distribution.
- *Under the linear model,* $E_p(\pi_i|\mu_i, \mathbf{z}_i) = g_\ell(\mathbf{z}_i)(a_0 + a_1\mu_i)$:

$$f_s(\mu_i|\mathbf{z}_i) = \frac{[a_0 + a_1\mu_i + g_\ell(\mathbf{z}_i)]f_p(\mu_i|\mathbf{z}_i)}{a_0 + a_1\mathbf{z}_i\gamma + g_\ell(\mathbf{z}_i)},$$

  so that the sample distribution can be expressed as a mixture of the normal and of the weighted normal population distribution of μ_i, given $\mathbf{z}_i$, and is non-informative if and only if $a_1 = 0$.

The sample distributions of the second stage observations y_{ij} can be shown to be as follows for the two models assumed for the conditional expectations of the second stage selection probabilities:

- *Under the exponential model,*
  $E_p(\pi_{j|i}|\mathbf{x}_{ij}, y_{ij}, \mu_i) = k_e(\mathbf{x}_{ij}, \mu_i)\exp(d_0 + d_1 y_{ij})$:

$$y_{ij}|\mathbf{x}_{ij}, \mu_i \sim N(\mu_i + \mathbf{x}_{ij}'\beta + d_1\sigma_e^2, \sigma_e^2),$$

  so that the conditional sample distribution of y_{ij} is independent of d_0 and of $k_e(\mathbf{x}_{ij}, \mu_i)$ and only its mean is shifted by $d_1\sigma_e^2$ from that of the population distribution.
- *Under the linear model,* $E_p(\pi_{j|i}|\mathbf{x}_{ij}, y_{ij}, \mu_i) = k_\ell(\mathbf{x}_{ij}, \mu_i) + (c_0 + c_1 y_{ij})$:

$$f_s(y_{ij}|\mathbf{x}_{ij}, \mu_i) = \frac{[c_0 + k_\ell(\mathbf{x}_{ij}, \mu_i) + c_1 y_{ij}]f_p(y_{ij}|\mathbf{x}_{ij}, \mu_i)}{c_0 + k_\ell(\mathbf{x}_{ij}, \mu_i) + c_1\mu_i + c_1\mathbf{x}_{ij}'\beta},$$

  so that the sample distribution can be expressed as a mixture of the normal and of the weighted normal population distribution of y_{ij}, given $\mathbf{x}_{ij}$ and μ_i, and is non-informative if and only if $c_1 = 0$.

To estimate the unknown parameters, we can use two-stage parametric estimation as before. However there is a problem, for the first stage, in estimating b_1 from the relationship $E_s(w_i|\mu_i) = [g_e(\mathbf{z}_i)]^{-1}\exp[-(b_0 + b_1\mu_i)]$, under the exponential model, say, since the values of μ_i are not observable. Possible solutions are to replace μ_i by the sample cluster mean, $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$, or to use the "errors in variables methods" (Fuller, 1987).

8. Simulation study for two-stage cluster sampling application

In order to check the performance of the methods of estimation, a small simulation study was carried out in which a hierarchical population was generated. For the first level, $N = 10,000$ independent normal values of the cluster-specific-effects were generated from $\mu_i \sim N(0, 0.25)$. For the second level, $M_i = 100$ independent normal values of the final units were generated from $y_{ij}|\mu_i \sim N(\mu_i, 0.49)$. In the first stage of sampling, $n = 100$ clusters were selected by systematic PPS sampling, with the size variable z_i defined by the exponential model $z_i = \exp(0.8 + \mu_i)$, so that the first stage inclusion probabilities were defined as: $\pi_i = (100z_i) \setminus \sum_{k=1}^{N} z_k$. For the second stage all final units in sampled PSU's were selected, so that $m_i = M_i$ and $\pi_{j|i} = 1$, for all j and $i \epsilon s$.

The estimation methods used were:

- Unweighted maximum likelihood, as if sampling was non-informative (UML)
- Weighted maximum (pseudo-)likelihood (WML)
- Sample maximum likelihood - exponential model (SMLE)

Relative biases (RB - the empirical bias divided by θ) and relative root mean square errors (RRMSE) are given in Table 2 for each of the parameters estimated, μ, σ_μ^2, and σ_e^2, for each of the three estimation methods considered.

		Estimation Method		
Parameter	**Indicator**	**UML**	**WML**	**SMLE**
μ	RB	0.0008	0.0013	0.0004
	RRMSE	0.0119	0.0126	0.0126
σ_μ^2	RB	-0.0412	-0.0126	-0.0412
	RRMSE	0.1484	0.1728	0.1484
σ_e^2	RB	0.0020	0.0027	0.0020
	RRMSE	0.0157	0.0177	0.0157

TABLE 2. Relative Biases and Relative Root Mean Square Errors of three estimation methods

The main results are as follows:

- The UML and WML estimators of μ are slightly biased. For the pseudo maximum likelihood estimator based on the sample distribution (SMLE) the bias is reduced substantially.
- The UML estimator of μ has a somewhat smaller RRMSE, than the WML and SMLE estimators which are the same.

- The RB and RRMSE of the estimators of the variances are the same under the UML and SMLE methods (since under exponential sampling the variances and covariances of measurements within clusters do not change).
- The WML estimators of the variances have higher RRMSE than the UML and the SMLE estimators.

9. Conclusions

Overall we have shown that the bias in estimating means, when the sampling design is informative, can be reduced by use of sample distribution-based estimators. The performance of the estimates based on the sample distribution is fairly robust to the choice of model. Thus an important finding from the simulation results relates to the sensitivity analysis of the estimators to departures from the assumed model. We find that the sample distribution is not too sensitive to the modelling of the conditional expectation of the first order sample inclusion probabilities.

However mean square errors are not always reduced and in some cases, at least, the unweighted estimator may be overall more efficient. For the variances, unweighted or sample distribution based estimates seem to perform better than the weighted estimators. There is no doubt that much further work in this area is required, but the basic idea of basing estimation on the sample distribution, in the case of informative designs, does work and can reduce the biases inherent in ignoring the informativeness of the sample design.

References

Binder, D. A. (1983), *On the variances of asymptotically normal estimators from complex surveys*, Internat. Statist. Rev., **51**, 279–292.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley & Sons, Inc., New York.

Kullback, S. (1978), *Information Theory and Statistics*, Peter Smith, Gloucester.

Kullback, S., and Leibler, R. A. (1951), *On information and sufficiency*, Ann. Math. Statistics, **22**, 79–86.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998), *Parametric distributions of complex survey data under informative probability sample*, Statist. Sinica, **8**, 1087–1114.

Pfeffermann, D., and Sverchkov, M. (1999), *Parametric and semi-parametric estimation of regression models fitted to survey data*, Sankhyā Ser. B, **61**, 166–186.

Department of Mathematics, Alquds University, Palestine

Department of Statistics, Hebrew University, 91905 Jerusalem, Israel
E-mail address: msabdul@palnet.com
E-mail address: gad@huji.ac.il