Statistica Sinica **8**(1998), 1087-1114

PARAMETRIC DISTRIBUTIONS OF COMPLEX SURVEY DATA UNDER INFORMATIVE PROBABILITY SAMPLING

Danny Pfeffermann, Abba M. Krieger and Yosef Rinott

Hebrew University, University of Pennsylvania and University of California San Diego

Abstract: The sample distribution is defined as the distribution of the sample measurements given the selected sample. Under informative sampling, this distribution is different from the corresponding population distribution, although for several examples the two distributions are shown to be in the same family and only differ in some or all the parameters. A general approach of approximating the marginal sample distribution for a given population distribution and first order sample selection probabilities is discussed and illustrated. Theoretical and simulation results indicate that under common sampling methods of selection with unequal probabilities, when the population measurements are independently drawn from some distribution (superpopulation), the sample measurements are asymptotically independent as the population size increases. This asymptotic independence combined with the approximation of the marginal sample distribution permits the use of standard methods such as direct likelihood inference or residual analysis for inference on the population distribution.

Key words and phrases: Design variables, independence, likelihood, mixtures, PPS sampling, weighted distribution.

1. Introduction

Survey data may be viewed as the outcome of two random processes: The process generating the values in the finite population, often referred to as the 'superpopulation model', and the process selecting the sample data from the finite population values, known as the 'sample selection mechanism'. Analytic inference from survey data relates to the superpopulation model, but when the sample selection probabilities are correlated with the values of the model response variables even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process.

In this article, we propose a general method of inference on the population distribution (model) under informative sampling that consists of approximating the parametric distribution of the sample measurements. The sample distribution is defined as the distribution of measurements corresponding to the units in the sample. Let Y_i denote the value of a response variable Y, associated with unit i that belongs to a finite population $U = \{1, \ldots, N\}$. We assume that the population values are independent realizations from a distribution with probability density function (pdf) $f_p(y_i|\theta)$ which depends on parameters θ. In Section 3 and onward we allow the pdf to depend also on known values of concomitant variables, as in regression or logistic regression models. The population pdf may be either discrete or continuous.

The (marginal) sample pdf of Y_i is defined here as $f(y_i|i \in s)$ where S denotes the selected sample and is obtained by application of Bayes theorem as

$$f_s(y_i|\theta^*) = f(y_i|i \in s) = \Pr(i \in s|y_i) f_p(y_i|\theta) / \Pr(i \in s), \tag{1.1}$$

where θ^* is a function of θ and any parameters indexing $\Pr(i \in s|y_i)$. In Section 3 we derive an alternative representation to (1.1) and justify the use of the sample pdf (1.1) for inference. Note that unless $\Pr(i \in s|y_i) = \Pr(i \in s)$ for all y_i, the sample and population pdfs are different, in which case the sampling scheme is informative. A simple example giving rise to such informative sampling is the case where $f_p(y_i|\theta)$ is Gamma(α, β) and $\Pr(i \in s|y_i)$ is proportional to y_i. It is easy to see (e.g., Patil and Rao (1978)), that the sample pdf is in this case $f_s(y_i|\theta^*) = \text{Gamma}(\alpha + 1, \beta)$. Several practical examples are studied in Sections 4 and 5.

The main theme of this paper is that it is possible in general to approximate the parametric distribution of the sample measurements and use this approximation for inference on the corresponding population distribution, exploiting the relationship between the two distributions. In particular, since the parameters of the sample pdf include the parameters of the population pdf, the parameters of the population pdf can be estimated by applying maximum likelihood or other estimation methods to the sample measurements, employing their sample distribution. The hypothesis that the population distribution belongs to a given family of distributions can be studied by testing that the sample distribution belongs to another, derived family of sample distributions, and so forth. The main advantage of basing the inference on the sample distribution is that it permits the use of standard efficient inference procedures. See Section 5 for illustrations and references.

It should be emphasized that under standard sampling methods, the sampled observations are not independent. However, in Section 6, we consider sampling schemes in common use and establish conditions under which for independent population measurements, the sample measurements are asymptotically independent. This allows us to construct the sample likelihood from the marginals and apply standard inference procedures. The implications of these results to cluster sampling are also discussed. Simulation results illustrate the asymptotic independence of the sample measurements for other sampling schemes as well. The theoretical proofs are given in the Appendix.

In survey sampling practice, the sample selection probabilities are defined by the values of a set of design variables like strata and cluster indicators, size measures, etc. When the population values of all the design variables used for the sample selection or adequate approximations of them are known, an alternative method of coping with the informativeness of the sampling scheme is by conditioning on these values. See Section 2. This modeling paradigm, however, is often very complicated and may be of little intrinsic interest. It is not feasible when some or all of the design variables are known only for the sample units or when the sample selection probabilities depend directly on the values of the response variables. A different approach in wide use to deal with the effects of informative sampling is to replace the ordinary sample estimates or estimating equations by weighted analogues obtained by weighting the sample observations inversely proportional to the sample selection probabilities. The use of this approach is restricted in general to point estimation and does not permit the use of standard inference tools such as likelihood based inference or residual analysis. Probabilistic statements require large sample normality assumptions. See Pfeffermann (1993) for discussion and references.

2. Ignorable and Informative Sampling Schemes

In this section we review conditions under which the sampling scheme can be ignored for inference and discuss their limitations. To simplify notation, we suppress the parameters in the densities below. Denote by $\boldsymbol{I}$ the $(N \times 1)$ sample indicator (vector) variable such that $I_j = 1$ if unit $j \in U$ is selected to the sample and $I_j = 0$ otherwise. The sample S is defined accordingly as $S = \{j | j \in U, I_j = 1\}$. Note that by definition of probability sampling, $\Pr(I_j = 1) > 0$ for all $j \in U$. Let $Z = [Z_1, \ldots, Z_N]'$ define an $(N \times q)$ matrix of population values of q design variables $Z(1), \ldots, Z(q)$ (different from Y, see below) employed for the sample selection process. The design variables may include indicator variables determining stratum and cluster membership or quantitative size variables. The values in Z may be regarded as random realizations although in what follows we condition on Z.

The joint pdf of $\boldsymbol{Y} = (Y_1, \ldots, Y_N)'$ and $\boldsymbol{I}$, given Z, can be written as

$$f_p(\boldsymbol{y}, \boldsymbol{i}|z) = f_p(\boldsymbol{y}|z) \Pr(\boldsymbol{i} \mid \boldsymbol{y}, z), \tag{2.1}$$

where $f_p(\boldsymbol{y}, \boldsymbol{i}|z)$ is the conditional pdf of $(\boldsymbol{Y}, \boldsymbol{I})|Z = z$ at $(\boldsymbol{y}, \boldsymbol{i})$ and similarly for $\Pr(\boldsymbol{i}|\boldsymbol{y}, z)$.

Drawing the sample partitions the population values of $\boldsymbol{Y}$ and Z into the sets $[\boldsymbol{Y}_s, Z_s] = \{(Y_j, Z_j), j \in s\}$ and $[\boldsymbol{Y}_{\tilde{s}}, Z_{\tilde{s}}] = \{(Y_\ell, Z_\ell), \ell \notin s\}$. Note that $S = S(\boldsymbol{I})$. Suppose first that both Z_s and $Z_{\tilde{s}}$ are known. This is normally the case when selecting the sample, but not necessarily so in a secondary analysis based on files released for 'public use'. For known Z, the data consist of the triple $(\boldsymbol{Y}_s, Z, \boldsymbol{I})$, and the joint pdf of $(\boldsymbol{Y}_s, \boldsymbol{I})$ given Z is obtained by integrating (2.1) over the nonsampled items $\boldsymbol{Y}_{\tilde{s}}$ with the sample units held fixed, i.e.,

$$f(\boldsymbol{y}_s, \boldsymbol{i}|z) = \int \Pr(\boldsymbol{i}|\boldsymbol{y}_s, \boldsymbol{y}_{\tilde{s}}, z) f_p(\boldsymbol{y}_s, \boldsymbol{y}_{\tilde{s}}|z) d\boldsymbol{y}_{\tilde{s}}. \tag{2.2}$$

The formulation in (2.2) is very general and imposes no restrictions on the sample selection mechanism. Ignoring the sampling mechanism, however, means that $\Pr(\boldsymbol{i}|\boldsymbol{y}_s, \boldsymbol{y}_{\tilde{s}}, z)$ is omitted from the right hand side of (2.2) and inference conditioned on Z is based on the pdf

$$f(\boldsymbol{y}_s|z) = \int f_p(\boldsymbol{y}_s, \boldsymbol{y}_{\tilde{s}}|z) d\boldsymbol{y}_{\tilde{s}}. \tag{2.3}$$

Clearly inference based on (2.2) is not generally the same as inference based on (2.3) because of possible sample selection effects. Suppose, however, that the selection probabilities only depend on Z in the sense that

$$\Pr(\boldsymbol{i}|\boldsymbol{y}_s, \boldsymbol{y}_{\tilde{s}}, z) = \Pr(\boldsymbol{i}|z). \tag{2.4}$$

Under this condition, inference on the conditional pdf of $Y|Z$, postulating (2.3), is equivalent to inference based on (2.2) and the sampling mechanism is *ignorable*. In particular, $f(\boldsymbol{y}_s|\boldsymbol{i}, z) = f(\boldsymbol{y}_s|z)$. Thus, by conditioning on the population values of all the design variables that determine the selection probabilities, the sampling mechanism can be ignored and standard inference procedures apply.

Weaker conditions for sample ignorability are given in Rubin (1976) and Little (1982), depending on whether the inference is based on repeated sampling theory, direct likelihood or the posterior distribution under the Bayesian formulation. Sugden and Smith (1984) establish conditions for sample ignorability for the case where the sampling condition (2.4) is satisfied but only proxy design variables $W = W(Z)$ are known at the inference stage. A special case of a proxy design variable is the vector $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_N)$ of the first order sample inclusion probabilities $\pi_j = \Pr(j \in s) = \Pr(I_j = 1|z)$. If $\Pr(\boldsymbol{i}|z) = \Pr(\boldsymbol{i}|\boldsymbol{\pi})$, the vector $\boldsymbol{\pi}$ is an 'adequate summary' of Z and the sampling mechanism can be ignored for inference that conditions on $\boldsymbol{\pi}$. Rubin (1985) shows that $\boldsymbol{\pi}$ is in fact the coarsest possible adequate summary of Z, although it may be too coarse.

There are three major problems associated with conditioning on Z or $W(Z)$ for securing sampling ignorability.

(1) It requires in principle that the population values of all the design variables, or at least adequate summaries of them, are known. As already mentioned, this is often not the case in a secondary analysis of public use data.

(2) Modelling $f_p(\boldsymbol{y}|z)$ or even $f_p[\boldsymbol{y}|w(z)]$ could be complicated and, perhaps more importantly, be of little interest. For example, in epidemiological studies sampling probabilities are often determined by a preliminary, inexpensive but inaccurate screening test, but there is no interest in conditioning on the screening test results when modeling the more accurate diagnostic measurements obtained later for the sampled units.

(3) Conditioning on Z to control for the effects of the sampling mechanism is not sufficient when Condition (2.4) is not satisfied as happens, for example, in retrospective studies where the selection probabilities are determined directly by the response variable values. Several actual sampling designs and inference problems giving rise to such situations are reviewed in Pfeffermann (1996). See also Section 4.

In the rest of this paper we consider situations, often occurring in practice, where the only design information available to the analyst is the vector $\boldsymbol{\pi}_s$ of the first order sample selection probabilities for units in the sample, and possibly also the sample values of some or all the design variables. In such cases conditioning on Z is no longer plausible. These situations fall under cases (iv)-(vi) in Table 1 of Sugden and Smith (1984), with $\boldsymbol{w}_s = \boldsymbol{\pi}_s$. Unlike in Sugden and Smith, however, we permit the selection probabilities to depend on the values of the response variable, thus violating also the sampling ignorability condition (2.4).

3. Marginal Distribution of Sample Observations

In what follows, we allow the population pdf to depend on known values of concomitant variables $\boldsymbol{x}_i$ such that $Y_i \sim f_p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$. The vectors $\boldsymbol{x}_i$ may include some of the design variables as well as other auxiliary variables. For example, in a regression model with y measuring hypertension, the $\boldsymbol{x}$-variables may include strata indicator variables and age, used to define sampling rates, as well as blood lead levels and other health measurements known only after sampling. In this and the next two sections we consider the marginal distribution of the sample measurements. In Section 6 we define conditions under which the sample measurements are asymptotically independent so that the joint sample distribution can be approximated by the product of the marginal distributions.

For the case where the population pdf depends on concomitant variables, the marginal sample pdf of Y_i is defined analogously to (1.1) as

$$f_s(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1) = \Pr(I_i = 1|y_i, \boldsymbol{x}_i) f_p(y_i|\boldsymbol{x}_i)/\Pr(I_i = 1|\boldsymbol{x}_i). \quad (3.1)$$

Comment 1. $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$ is generally not the same as the sample selection probability $\pi_i = \Pr(I_i = 1|\boldsymbol{y}, z)$ where $\boldsymbol{y}$ and z denote the realized population values of Y and Z.

Comment 2. It follows from (3.1) that the marginal sample pdf is different from the marginal population pdf (before sampling), unless $\Pr(I_i = 1|y_i, \boldsymbol{x}_i) = \Pr(I_i = 1|\boldsymbol{x}_i)$ for all y_i, in which case the sampling scheme is noninformative conditional on $\boldsymbol{x}_i$.

Comment 3. The sample pdf can be viewed as a special case of the family of 'weighted distributions' introduced by Rao (1965). Several models and selection procedures giving rise to such distributions are discussed in Patil and Rao (1978).

In what follows, we regard the probabilities $\pi_i = \Pr(I_i = 1|\boldsymbol{y}, z)$ as realizations of random variables (Smith (1988)). As mentioned before, in general $\pi_i \neq \Pr(I_i = 1|y_i, \boldsymbol{x}_i)$. Nonetheless, the following relationship holds,

$$\Pr(I_i = 1|y_i, \boldsymbol{x}_i) = \int \Pr(I_i = 1|y_i, \boldsymbol{x}_i, \pi_i) f_p(\pi_i|y_i, \boldsymbol{x}_i) d\pi_i = E_p(\pi_i|y_i, \boldsymbol{x}_i) \tag{3.2}$$

since $\Pr(I_i = 1|y_i, \boldsymbol{x}_i, \pi_i) = \pi_i$. Substituting (3.2) into (3.1) yields the alternative expression for the marginal sample pdf,

$$f_s(y_i|\boldsymbol{x}_i) = E_p(\pi_i|y_i, \boldsymbol{x}_i) f_p(y_i|\boldsymbol{x}_i) / E_p(\pi_i|\boldsymbol{x}_i), \tag{3.3}$$

where the expectation in the denominator follows by an argument similar to (3.2).

The prominent feature of (3.3) is that for a given population pdf, the corresponding sample pdf is fully determined by the conditional expectation $E_p(\pi_i|y_i, \boldsymbol{x}_i)$. Files of survey data released for secondary analysis ordinarily contain the selection probabilities in the form of the sampling weights $w_i = \pi_i^{-1}$ (possibly modified to account for unit nonresponse) so that the expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$ can be estimated in principle from the sample data. (See Section 5 for details.) Note that the use of (3.3) does not require the specification of the full distribution of the π_i which is often intractable.

Comment 4. When the population pdf does not depend on concomitant variables, so that $Y_i \sim f_p(y_i|\theta)$, the relationship (3.3) reduces to

$$f_s(y_i) = E_p(\pi_i|y_i) f_p(y_i) / E_p(\pi_i). \tag{3.4}$$

The denominator is now a fixed number, although it may depend on unknown parameters.

We conclude this section by justifying the consideration of the conditional pdfs (3.1) and (3.3), or more generally $f(y_s|X_s, s)$ where $X_s = \{\boldsymbol{x}_i, i \in s\}$. Suppose that the population measurements are independent such that $Y_i \sim f_p(y_i|\boldsymbol{x}_i)$

where the $\boldsymbol{x}_i$ are fixed values of concomitant variables, and assume for now that the sampling units are selected to the sample independently with probabilities π_i having expectations $\pi(y_i, \boldsymbol{x}_i) = E(\pi_i|y_i, \boldsymbol{x}_i)$. (Asymptotic independence of the sample measurements is discussed in Section 6.) Simple calculations imply that the full distribution of $(\boldsymbol{Y}_s, \boldsymbol{i})$ is in this case

$$f(\boldsymbol{y}_s, \boldsymbol{i}|X) = \Pi_{i\in s}[\pi(y_i, \boldsymbol{x}_i)f_p(y_i|\boldsymbol{x}_i)/\pi_{0i}]\Pi_{i\in s}(\pi_{0i})\Pi_{i\notin s}(1-\pi_{0i}), \tag{3.5}$$

where $\pi_{0i} = \int \pi(y_i, \boldsymbol{x}_i)f_p(y_i|\boldsymbol{x}_i)dy_i = \Pr(i \in s|\boldsymbol{x}_i)$ and $X = \{\boldsymbol{x}_i, i \in U\}$. Note that the pdf in the square brackets is the conditional marginal sample pdf as defined by (3.1).

Let the population pdf be indexed by the unknown parameter θ such that $f_p(y|\boldsymbol{x}) = f_p(y|\boldsymbol{x};\theta)$, and suppose that the vectors $\boldsymbol{x}_i$ are only known for units in the sample. If the target of inference is, for example, to estimate θ by use of maximum likelihood, it is clear that the likelihood obtained from (3.5) is not operational because the product $\Pi_{i\notin s}(1-\pi_{0i})$ depends on all the individual vectors $\boldsymbol{x}_i, i \notin s$ and these vectors are usually not part of the data. Thus, data availability dictates basing the inference in this case on the more limited pdf $f(\boldsymbol{y}_s|X_s, s;\theta) = \Pi_{i\in s}[\pi(y_i, \boldsymbol{x}_i)f_p(y_i|\boldsymbol{x}_i;\theta)/\pi_{0i}]$.

A different approach for handling the case where the $\boldsymbol{x}$-values are unknown for units outside the sample, is to consider the $\boldsymbol{x}_i$ as random, and model their distribution. However, modelling the distribution of concomitant variables may be a formidable task, and the resulting likelihood function $f(\boldsymbol{y}_s, X_s, \boldsymbol{i})$ involves integrals with respect to this distribution which tend to be cumbersome. Note also that inference on θ can be sensitive to the specified model (see, e.g., Rotnitzky and Robins (1997)).

Comment 5. Modelling the full sample distribution (the joint distribution of the observed response and covariates values, and the selected sample) becomes even more complicated when elements of the $\boldsymbol{x}$ vectors are missing for sample units as well. Therefore, recent work on missing data has focused on the use of semi-parametric regression models for estimating the parameter θ in $f_p(y|\boldsymbol{x};\theta)$. See, e.g., Rotnitzky and Robins (1997). An alternative approach proposed by Robins (1997) is to model the response probabilities given the observed data and use the estimated probabilities to construct Horvitz-Thompson type estimators based on cases with complete data. This approach requires certain conditions on the non-response mechanism. The application of both approaches is computationally intensive and is restricted to estimation problems, unlike the use of the conditional sample distribution that we propose, derived as a product of the marginals (3.1) or (3.3). However, the latter distribution is not operational for inference when elements of the covariates are missing.

4. Examples

4.1. Preface

In this section, we assume that selection to the sample is carried out with unequal selection probabilities, independently between units. This is known as 'Poisson sampling' in the sampling literature (Hajek (1981), ch.6). This method is not often used, a major reason being that the sample size is random. Nevertheless, in Section 6 we show that many of the sampling methods in common use for selection with unequal probabilities produce asymptotically the same sample distribution as obtained under the Poisson sampling scheme.

The common feature of the examples in Sections 4.2-4.4 is that, while the parameters of the population pdf and the sample pdf are different as a result of the sample selection, the two distributions are in the same family. Generalizations of this property are considered in Section 4.5.

4.2. Logistic regression models

Let Y be a categorical response variable taking the values $0, 1, \ldots, L-1$. Let X define a set of explanatory variables and suppose that $\Pr(Y = \ell|\boldsymbol{x})$ can be modeled using logistic regression such that for unit $i \in U$

$$\Pr(Y_i = \ell|\boldsymbol{x}_i) = \exp(\alpha_\ell + \boldsymbol{x}_i'\boldsymbol{\beta}_\ell) / \sum_{j=0}^{L-1} \exp(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j), \tag{4.1}$$

where $\alpha_0 = 0$, $\boldsymbol{\beta}_0 = \mathbf{0}$ for uniqueness. The model (4.1) defines the population pdf, prior to sampling. Assume that the sample is selected by Poisson sampling with probabilities $\Pr(I_i = 1|Y_i = \ell, \boldsymbol{x}_i) = P_\ell$, $\ell = 0, \ldots, L-1$. By (3.1) and (4.1) and after canceling $\sum_{j=0}^{L-1} \exp(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j)$ in the numerator and the denominator, the sample pdf is seen to be,

$$\begin{aligned}\Pr(Y_i = \ell|\boldsymbol{x}_i, I_i = 1) &= P_\ell \exp(\alpha_\ell + \boldsymbol{x}_i'\boldsymbol{\beta}_\ell) / \sum_{j=0}^{L-1} P_j \exp(\alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j) \\ &= \exp(\alpha_\ell^* + \boldsymbol{x}_i'\boldsymbol{\beta}_\ell) / \sum_{j=0}^{L-1} \exp(\alpha_j^* + \boldsymbol{x}_i'\boldsymbol{\beta}_j), \ell = 0, \ldots, L-1, \end{aligned} \tag{4.2}$$

where $\alpha_\ell^* = [\log(P_\ell/P_0) + \alpha_\ell]$ so that $\alpha_0^* = 0$.

It follows from (4.2) that the sample pdf is again logistic with the same slope coefficients as in the population pdf, but with different intercepts. Clearly, when $P_\ell = P_0$ for all ℓ, $\alpha_\ell^* = \alpha_\ell$, and the population and sample pdfs coincide.

Comment 6. The distribution (4.2) does not apply to the case of retrospective studies, in which the numbers n_l of observations for which $Y = \ell$ are determined in advance. In that case the probabilities $\Pr(I_i = 1|Y_i = \ell, \boldsymbol{x}_i)$ have a rather complicated structure that depends on all the $\boldsymbol{x}$-values in the population. In order to estimate the parameters of the model in such studies, Prentice and Pyke (1979) derive the density of the covariates conditioned on the category. This leads to a different likelihood function that involves nuisance parameters.

4.3. Linear regression models

Let the population distribution be

$$Y_i|\boldsymbol{x}_i \sim N(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2). \tag{4.3}$$

Suppose that the sample inclusion probabilities have expectations

$$E_p(\pi_i|y_i, \boldsymbol{x}_i) = \exp[A_1 y_i + g(\boldsymbol{x}_i)] \tag{4.4}$$

for some function $g(\boldsymbol{x})$. By (3.3), (4.3) and (4.4),

$$\begin{aligned} f_s(y_i|\boldsymbol{x}_i) &= \exp(A_1 y_i) f_p(y_i|\boldsymbol{x}_i) \Big/ \int \exp(A_1 y_i) f_p(y_i|\boldsymbol{x}_i) dy_i \\ &= \exp(A_1 y_i) f_p(y_i|\boldsymbol{x}_i) / \exp[A_1(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}) + \frac{1}{2}A_1^2\sigma^2] \\ &= \frac{1}{\sigma}\phi[(y_i - (\beta_0 + A_1\sigma^2) - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma], \end{aligned} \tag{4.5}$$

where $\phi(\cdot)$ is the standard normal pdf. Hence the regression of Y on $\boldsymbol{x}$ in the sample is the same as in the population, except for the intercept term which changes to $(\beta_0 + A_1\sigma^2)$.

A special case of the model defined by (4.3) and (4.4) arises under the following structure,

$$Y_i = \gamma_0 + \boldsymbol{x}_i'\boldsymbol{\gamma} + \gamma_\pi \log(\pi_i) + \epsilon_i;\ \log(\pi_i) = \alpha_0 + \boldsymbol{x}_i'\boldsymbol{\alpha} + \delta_i, \tag{4.6}$$

where ϵ_i and δ_i are independent normal disturbances. The model defined by (4.6), with the right hand side equation postulated for the sample units is studied by Skinner (1994).

An interesting extension of the relationship (4.4) is obtained by adding a quadratic term $(A_2 y_i^2, A_2 < 0)$ to the right hand side equation. Simple algebra yields

$$f_s(y_i|\boldsymbol{x}_i) = N[(\beta_0 + A_1\sigma^2 + \boldsymbol{x}_i'\boldsymbol{\beta})/C, \sigma^2/C], \tag{4.7}$$

where $C = (1 - 2\sigma^2 A_2)$. Thus, while the normality of the population pdf is still preserved after sampling, the slope coefficients and the residual variance change in this case in proportion to the fixed factor C.

The effect of informative sampling on linear regression models is studied by Goldberger (1981). Assuming that the X-variables are multinormally distributed along with Y, and that the selection to the sample is "explicit on Y", Goldberger shows that the vector of coefficients defining the *linear* regression of Y on X in the sample is a scalar multiple of the vector of regression coefficients in the population. Note that under this formulation, the conditional marginal sample pdf is not necessarily normal and in particular, the conditional expectation $E_s(Y_i|\boldsymbol{x}_i)$ is not necessarily linear. On the other hand, the sampling scheme studied by Goldberger is more flexible and does not require the specification of (4.4).

4.4. Gamma models under probability proportional to size sampling

Let the population pdf be gamma with shape parameter α and mean μ_i so that

$$f_p(y_i) \propto y_i^{\alpha-1} \exp(-\alpha y_i/\mu_i). \tag{4.8}$$

Let the sample inclusion probabilities have expectations $E_p(\pi_i|y_i) \propto y_i$. By (3.3), the sample pdf of Y_i is again gamma with shape parameter $(\alpha+1)$ and mean $\mu_i(\alpha+1)/\alpha$. This result generalizes a familiar result on sampling from a gamma distribution with probabilities proportional to y ($\pi_i \propto y_i$ as opposed to $E_p(\pi_i|y_i) \propto y_i$ in the present case). If, following McCullagh and Nelder (1989), ch. 8, it can be assumed that $\log(\mu_i) = (\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta})$ for given vectors $\boldsymbol{x}_i$,

$$E_s(Y_i|\boldsymbol{x}_i) = \exp\{\beta_0 + \log[(\alpha+1)/\alpha] + \boldsymbol{x}_i'\boldsymbol{\beta}\}, \tag{4.9}$$

implying that in this case again the slope coefficients of the sample and population pdfs are the same.

4.5. General invariance relationships

In this section, we identify more general structures for which the population and the sample distributions are in the same family.

Let the population pdf belong to the exponential family

$$f_p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_i) = a_i(\boldsymbol{\theta}_i)\exp\Big[\sum_{k=1}^{K}\theta_{ki}b_{ki}(y_i) + c_i(y_i)\Big], \tag{4.10}$$

where $\boldsymbol{\theta}_i = (\theta_{1i},\ldots,\theta_{Ki})'$ defines the natural parameterization of the family, taking values in the parameter space $\boldsymbol{\Theta} \subset R^K$, and $b_{ki}(\cdot)$ and $c_i(\cdot)$ are known functions. The dependence on $\boldsymbol{x}_i$ operates via the parameters θ_{ki}, see below.

Suppose that the sample inclusion probabilities have expectations

$$E_p(\pi_i|y_i,\boldsymbol{x}_i) = r_i \exp\Big[\sum_{k=1}^{K} d_{ki}b_{ki}(y_i)\Big], \tag{4.11}$$

where r_i and $\{d_{ki}\}$ are constants which may depend on $\boldsymbol{x}_i$, but not on y_i. The following proposition provides a general distribution invariance property.

Proposition 1. *If the population pdf of Y_i belongs to the exponential family defined by* (4.10) *and the sample inclusion probabilities obey* (4.11)*, then the sample pdf of y_i belongs to the same exponential family with parameters $\theta^*_{ki} = \theta_{ki} + d_{ki}$,* (*provided $\boldsymbol{\theta}^*$ lies in $\boldsymbol{\Theta}$*)*.*

The proof follows directly from (3.3).

The problem considered in this section and the result stated in Proposition 1 resemble the familiar issue of the identification of conjugate prior distributions in Bayesian inference. Interestingly, Cox and Hinkley (1974) call the family of prior distributions for which the posterior distributions are in the same family *closed under sampling.* This terminology is perfectly suited to the present context.

The dependence on $\boldsymbol{x}_i$ in the equations (4.10) and (4.11) operates in a very general way via θ_{ki} and d_{ki} respectively. This dependence may be made more explicit for a class of regression models of Y on $\boldsymbol{x}$ if the following linear relationships are assumed,

$$\theta_{ki} = \phi_{0k} + \boldsymbol{x}_i'\boldsymbol{\phi}_k; \quad d_{ki} = \Psi_{0k} + \boldsymbol{x}_i'\boldsymbol{\Psi}_k. \tag{4.12}$$

Another invariance result is obtained by taking $(\phi_{0k}, \boldsymbol{\phi}_k)$, $k = 1, \ldots, K$, as parameters defining a more restricted family of population pdfs relating Y_i to $\boldsymbol{x}_i$.

Corollary 1. *Under the conditions of Proposition* 1 *and Assumption* (4.12)*, the sample pdf belongs to the same restricted family with ϕ_{0k} and $\boldsymbol{\phi}_k$ replaced by* $(\phi_{0k} + \Psi_{0k})$ *and* $(\boldsymbol{\phi}_k + \Psi_k)$ *respectively. In particular, if the functions d_{ki} do not depend on $\boldsymbol{x}_i$, i.e., $\Psi_k = 0$, the coefficients of $\boldsymbol{x}_i$ in the natural parameterization of the sample pdf are the same as for the population pdf.*

5. Sample Distributions under More General Sampling Schemes

5.1. Selection with probabilities proportional to size

In the discussion so far, we assumed a known form for the conditional expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$. This clearly need not be the case in practice, particularly when the π_i's depend also on design variables not contained among the concomitant X-variables.

The prominent advantage of the use of probability sampling is that, except in cases of nonresponse (not considered here), the sample selection probabilities are known. Assuming they are available to the analyst for at least the units in the sample, the form of the expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$ can be identified and estimated in principle from the sample data.

Suppose first that the π_i's are measures of size, regarded as continuous measurements from some pdf $g(\pi)$. Under some regularity conditions, the expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$ may then be approximated by low order polynomials in y_i and the components of $\boldsymbol{x}_i$, or by exponentials of such polynomials, via the Taylor series approximation. Thus, letting $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$, for the first case

$$E_p(\pi_i|y_i, \boldsymbol{x}_i) \approx \sum_{j=0}^{J} A_j y_i^j + h(\boldsymbol{x}_i), \tag{5.1}$$

where $h(\boldsymbol{x}_i) = \sum_{p=1}^{m} \sum_{k=1}^{K(p)} B_{kp} x_{ip}^k$ and $\{A_j\}$ and $\{B_{kp}\}$ are unknown parameters to be estimated from the sample (see below). Substituting (5.1) in (3.3) and assuming the existence of the moments $E^{(j)} = E_p(Y_i^j|\boldsymbol{x}_i)$, the sample pdf can be approximated accordingly as

$$f_s(y_i|\boldsymbol{x}_i) \approx \frac{\sum_{j=1}^{J}(A_j E^{(j)}) f_p^{(j)}(y_i|\boldsymbol{x}_i) + [A_0 + h(\boldsymbol{x}_i)] f_p(y_i|\boldsymbol{x}_i)}{\sum_{j=1}^{J}(A_j E^{(j)}) + [A_0 + h(\boldsymbol{x}_i)]}, \tag{5.2}$$

where $f_p^{(j)}(y_i|\boldsymbol{x}_i) = y_i^j f_p(y_i|\boldsymbol{x}_i)/E^{(j)}$. It follows from (5.2) that under the approximation (5.1), the sample pdf is a mixture of the densities $f_p^{(j)}(y_i|\boldsymbol{x}_i)$, $j = 0, \ldots, J$. Note that changing the function $h(\boldsymbol{x})$ to another function $h^*(\boldsymbol{x})$ only affects the mixture coefficients.

Comment 7. The vector parameter θ^* indexing the sample pdf $f_s(y_i|\boldsymbol{x}_i; \theta^*)$ in (5.2) consists of the vector parameter θ indexing $f_p(y_i|\boldsymbol{x}_i; \theta)$ and the coefficients $\{A_j\}$ and $\{B_{kp}\}$. Thus, the sample pdf may depend on many more parameters than the population pdf.

Examples. Suppose there are no concomitant variables and let $E_p(\pi_i|y_i) = Ay_i$. Then, by (5.2), $f_s(y_i) = y_i f_p(y_i)/E_p(Y)$. In this case the sample pdf depends on the same parameters as the population pdf. As a second example, let $E_p(\pi_i|y_i) = A_0 + \sum_{j=1}^{J} A_j y_i^j$ and suppose that the population pdf is Gamma (α, β). In this case,

$$f_s(y_i) = \sum_{j=0}^{J} C_j \text{ Gamma}(\alpha + j, \beta) / \sum_{j=0}^{J} C_j, \tag{5.3}$$

where $C_0 = A_0$ and $C_j = A_j \alpha(\alpha + 1) \cdots (\alpha + j - 1)/\beta^j$, $j = 1, \ldots, J$. Hence the sample pdf is a mixture of gamma densities, with shape parameters $(\alpha + j)$ and a common scale parameter β. The vector parameter θ^* contains in this case $\theta = (\alpha, \beta)$ and the mixture coefficients $\{C_j\}$. Krieger and Pfeffermann (1997) use standard goodness of fit test statistics applied to the density (5.3) with estimated mixture coefficients $\hat{C}_j$ (see below) for testing the hypothesis that the population pdf is Gamma(α, β).

Next consider the approximation

$$E_p(\pi_i|y_i, \boldsymbol{x}_i) \approx \exp\Big[\sum_{j=0}^{J} A_j y_i^j + h(\boldsymbol{x}_i)\Big]. \tag{5.4}$$

As discussed by Skinner (1994), this approximation is appealing in the common situation where the selection to the sample is carried out in several stages so that the ultimate inclusion probabilities are products of the selection probabilities at the various stages. If the vectors $\boldsymbol{x}_i$ contain the design variables used at the various stages, then it is natural to express the expectations $E_p(\pi_i|y_i,\boldsymbol{x}_i)$ as a multiplicative function of y_i and $\boldsymbol{x}_i$. Under (5.4),

$$f_s(y_i|\boldsymbol{x}_i)=\Big[\exp(\sum_{j=1}^{J}A_jy_i^j)\Big]f_p(y_i|\boldsymbol{x}_i)/E_p\Big[\exp(\sum_{j=1}^{J}A_jy_i^j)|\boldsymbol{x}_i\Big]. \qquad (5.5)$$

Comment 8. The pdf (5.5) does not depend on A_0 and $h(\boldsymbol{x})$. The approximation (5.4) was used for the examples in Sections 4.2-4.4 and the invariance relationships of Section 4.5.

The disadvantage of the approximation (5.4) and the resulting representation (5.5) is that some of the parameters of the population pdf may no longer be identifiable from the sample observations of Y and X alone. For example, in the linear regression case considered in Section 4.3, the population regression intercept β_0 cannot be separated from the sample regression intercept $(\beta_0+A_1\sigma^2)$ unless the 'sampling coefficient' A_1 is estimated separately, employing for example the relationship (4.4), see Comment 9 below.

The identifiability problems associated with the use of the approximation (5.4) are circumvented in general when the approximation (5.1) is used. Assuming independence of the sample observations, see Section 6, the parameters of the population pdf and the regression coefficients $\{A_j\}$ and $\{B_{kp}\}$ parameterizing the conditional expectations $E_p(\pi_i|y_i,\boldsymbol{x}_i)$ can be estimated simultaneously by standard techniques applied to the sample pdf (5.2). For example, Krieger and Pfeffermann (1997) use the EM algorithm for estimating the parameters of the sample density defined by (5.3).

Comment 9. When the number of parameters indexing the sample pdf is large, it is often computationally much easier to estimate these parameters in two steps. In the first step, the coefficients $\{A_j\}$ and $\{B_{kp}\}$ are estimated from the observed probabilities π_i, employing the relationships (5.1) or (5.4). In the second step the parameters indexing the population pdf are estimated by maximum likelihood or other standard methods, with the estimates of A_j and B_{kp} held fixed. As discussed before, the use of this two step procedure may become necessary when employing the approximation (5.4). Note that the estimation in the first step requires the use of probability weighted regression or other methods that account for sample selection effects, since the π_i play the dual role of determining the sample selection probabilities and the values of the regression dependent variable. See Pfeffermann (1993) for review of several such methods and Krieger and Pfeffermann (1992) for an illustration of the use of this two-step estimation procedure.

5.2. Stratified and multistage sampling designs

The discussion in Section 5.1 assumes that the conditional expectation $E_p(\pi_i|y_i, \boldsymbol{x}_i)$ is continuous in y and $\boldsymbol{x}$ which, under some regularity conditions, permits the use of Taylor series approximations. In stratified sampling, the sample selection probabilities are constant within strata so that the approximations (5.1) or (5.4) are no longer valid.

In Section 4.2 we consider the logistic regression example where the response variable takes $(L+1)$ values which define the sample selection probabilities directly. Below we consider an example of a stratified sample with the strata defined by the ascending values of a continuous size measure. See, for example, Hausman and Wise (1981) and Korn and Graubard (1995) for surveys employing such designs.

Let $Z_i = q(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})$ be a random size variable with X_i either random or fixed, where q is of known form and $\boldsymbol{\alpha}$ represents a fixed (but possibly unknown) vector of parameters. Let $a^{(0)} < a^{(1)} < \cdots < a^{(L)}$ define L percentiles of the distribution of Z with $a^{(0)} = -\infty$ and $a^{(L)} = \infty$. These percentiles define a division of the population values into L strata, $U_1, \ldots, U_L$, of sizes $N_1, \ldots, N_L$, based on the realized values $z_1, \ldots, z_N$. The division is such that unit i belongs to stratum U_ℓ iff $a^{(\ell-1)} \le z_i \le a^{(\ell)}$. Assuming simple random stratified sampling with sample sizes $n_\ell = N_\ell P_\ell$, $\ell = 1, \ldots, L$, where the P_ℓ's are fixed proportions, the sample pdf of $Y_i|\boldsymbol{x}_i$ is in this case

$$f_s(y_i|\boldsymbol{x}_i) = \begin{cases} P_1 f_p(y_i|\boldsymbol{x}_i)/w & \text{if } z_i \le a^{(1)} \\ P_2 f_p(y_i|\boldsymbol{x}_i)/w & \text{if } a^{(1)} < z_i \le a^{(2)} \\ \quad\vdots & \\ P_L f_p(y_i|\boldsymbol{x}_i)/w & \text{if } a^{(L-1)} < z_i, \end{cases} \tag{5.6}$$

where $w = \Pr(I_i = 1|\boldsymbol{x}_i) = \sum_{\ell=1}^{L}[P_\ell \int_{a^{(\ell-1)}}^{a^{(\ell)}} f_p(z|\boldsymbol{x}_i)dz]$. Hausman and Wise (1981) and Krieger and Pfeffermann (1992) study maximum likelihood estimation of the parameters θ indexing the pdf $f_p(y_i|\boldsymbol{x}_i, \theta)$ under this sampling design, for the case where $(Y_i, \boldsymbol{X}_i)$ is multivariate normal and the function q is linear. The difference between the two studies is in the assumptions regarding the knowledge of the P_ℓ's or $a^{(\ell)}$'s.

Another family of sampling designs in common use is hierarchical multistage designs in which each stage, except the last, involves the selection of clusters that are nested within clusters selected in the previous stage. In a typical two stage cluster sample for example, with C clusters and N_c units within cluster

c, the clusters are selected with probabilities proportional to a size measure Z, whereas selection within the clusters is with equal probabilities. The sample sizes within the clusters are usually determined in such a way that the ultimate sample inclusion probabilities are constant across the population, ensuring that the sampling scheme is 'self-weighting'.

Despite the self-weighting property of the sample, the sample distribution may still be different in such cases from the population distribution, as illustrated by the following simple example. Let the population model be

$$Y_{ci} = \boldsymbol{x}'_{ci}\boldsymbol{\beta} + \gamma Z_c + e_{ci}; \quad Z_c \sim \text{ Gamma}(\alpha, \beta)\ , e_{ci} \sim N(0, \sigma^2_\ell)\ {}^{c=1,\ldots,C}_{i=1,\ldots,N_c} \tag{5.7}$$

and suppose that the clusters are selected with probabilities π_c such that $E_p(\pi_c|Z_c) = A_0 + A_1 Z_c$. For example, Y_{ci} may represent the salary of employee i working for establishment c of a certain branch, $\boldsymbol{x}_{ci}$ may represent personal characteristics (profile), and Z_c the size of the corresponding establishment as measured by the total number of employees. It follows that

$$f_s(z_c) = (A_0 + A_1 z_c) \text{ Gamma}(\alpha, \beta)/[A_0 + A_1 E_p(Z_c)] \tag{5.8}$$

which, by (5.3), is a mixture of Gamma densities. Thus the distribution of the sample observations of Z, and hence of Y, is different in this case from the population distribution despite the self-weighting property of the sample. This follows from the fact that even though the inclusion probabilities $\pi_{ci} = \Pr(ci \in S)$ are constant, $\Pr(ci \in S|y_{ci}, \boldsymbol{x}_{ci})$ depends on y_{ci} and hence $f_p(y_{ci}|\boldsymbol{x}_{ci}) \neq f_s(y_{ci}|\boldsymbol{x}_{ci})$. This example illustrates that when deriving the sample distribution under multistage sampling designs, it might be necessary to model the conditional expectations of the sample selection probabilities at each stage of the selection process, and not just model the ultimate inclusion probabilities.

6. Independence of Sample Measurements under Common Sampling Schemes

6.1. Theoretical results

Having derived the marginal sample distribution, the question that arises is whether under commonly used sampling methods the sample measurements are approximately independent. This question is fundamental as many of the standard inference procedures assume independence of the observations like, for example, (classical) likelihood-based inference by which the sample likelihood is computed as a product of the marginal densities. The examples of Section 4, and the more general formulation in Section 5, assume, at least implicitly, that selection to the sample is carried out independently using Poisson sampling. Under this sampling scheme it is clear that when the population values are independent, so are the sample values. Common sampling methods for selection with unequal probabilities involve, however, joint selection of the sampling units such that the second and higher order joint selection probabilities are no longer simple products of the corresponding first order selection probabilities. Also, the selection probabilities ordinarily depend on the population means of the design variables. When the latter are regarded as random, it introduces another source of 'sample dependence'.

In the Appendix we prove the following asymptotic independence theorem related to PPS sampling *with replacement*. The same independence property holds for various sampling schemes of selection without replacement via a similar proof. The implications of these results to cluster sampling are discussed at the end of this section. As before, for convenience, we suppress parameter symbols from the various densities. Also, equalities and inequalities related to conditional densities or expectations are assumed to hold a.s.

Let $\boldsymbol{Z} = \{Z_1, \ldots, Z_N\}$ consist of independent realizations of a positive random size variable Z. Suppose that a sample S of fixed size n is selected with replacement with probabilities proportional to Z: on each draw, unit j is drawn with probability $P_j = Z_j/(N\bar{Z})$, $j = 1, \ldots, N$, where $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$. Let $(Y_i, \boldsymbol{X}_i)$ define a random response variable and random concomitant variables respectively associated with unit i, $i = 1, \ldots, N$.

Consider the following conditions:

(a) $\Pr(Z_i \geq 0) = 1$; $E(Z_i^k) < \infty$ for $k = 1, 2, \ldots$, and $f(\boldsymbol{z}|X) = \Pi_{i=1}^{N} f(z_i|\boldsymbol{X}_i)$ where $X = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$.

(b) For each $m \geq 1$, there exists $K = K_m$ such that $E(\frac{K}{Z_1 + \cdots + Z_K})^m < \infty$.

(c) $f_p(\boldsymbol{y}|X, Z) = \Pi_{i=1}^{N} f_p(y_i|\boldsymbol{X}_i, Z_i)$ and $f_p(y_i|\boldsymbol{X}_i, Z_i) < B$ for some B, all i, y_i, and for (X_i, Z_i) in a set having probability 1.

(d) $E(Z_i|\boldsymbol{X}_i) > c$ for some $c > 0$, and $E(Z_i^2|\boldsymbol{X}_i) < B$ for some B.

Let $X_s = \{\boldsymbol{X}_i, i \in S\}$, $\boldsymbol{Y}_s = \{Y_i, i \in S\}$ and let $f(\boldsymbol{y}_s|X_s, s)$ denote the conditional density of the Y values for units in the sample.

Theorem 1. (i) *If S consists of n distinct units, then under Conditions (a)-(d), as $N \to \infty$ (with n fixed),*

$$f_s(\boldsymbol{y}_s|X_s) = f(\boldsymbol{y}_s|X_s, s) = \frac{\Pi_{i \in S} E(Z_i|y_i, \boldsymbol{X}_i) f_p(y_i|\boldsymbol{X}_i)}{\Pi_{i \in S} E(Z_i|\boldsymbol{X}_i)} + O(\frac{1}{N^{1/2}}), \tag{6.1}$$

where the expectations in the numerator and the denominator are with respect to the conditional distributions of $Z_i|Y_i, \boldsymbol{X}_i$ and $Z_i|\boldsymbol{X}_i$, respectively.

(ii) *When S contains repetitions with multiplicities τ_i, such that $\sum_{i \in S^*} \tau_i = n$ where S^* consists of the distinct elements in S, then under Conditions (a)-(d),*

$$f_s(\boldsymbol{y}_s|X_s) = \frac{\Pi_{i \in S^*} E(Z_i^{\tau_i}|y_i, \boldsymbol{X}_i) f_p(y_i|\boldsymbol{X}_i)}{\Pi_{i \in S} E(Z_i|\boldsymbol{X}_i)} + O(\frac{1}{N^{1/2}}). \tag{6.2}$$

The terms $O(N^{-1/2})$ in (6.1) *and* (6.2) *depend on n and the bounds on the moments defined under the conditions* (*a*)-(*d*)*, but not on S and X_s.*

Comment 10. The sample densities $E(Z_i|y_i, \boldsymbol{X}_i)f_p(y_i|\boldsymbol{X}_i)/E(Z_i|\boldsymbol{X}_i)$ in the right hand side of (6.1) correspond to the marginal sample pdfs defined by (3.3) with π_i in (3.3) replaced by $Z_i/[NE_p(Z)]$ in the numerator and the denominator of (6.1).

Examination of the conditions (a)-(d) reveals that they are not very restrictive. Sampling with replacement is often applied in practice because it permits simple variance estimators. Furthermore, once the asymptotic independence of the sample measurements under PPS sampling with replacement is established, it allows us to derive the relationship (6.1) for the following sampling schemes *without replacement*. See the Appendix for details.

A - *Successive Sampling.* Draw one unit each time with replacement, with probabilities $P_j = Z_j/N\bar{Z}$, until n distinct units have been selected (Hajek (1981), Ch. 9). This sampling scheme is equivalent to drawing the units in succession without replacement, such that on the $(r+1)$st draw, the selection probability for unit i, not previously selected, is $P(i \in S) = Z_i/\sum_{j \notin S_r}^{N} Z_j$ with S_r denoting the units selected on the first r draws.

B - *Rejective Sampling.* Draw one unit each time with replacement, with probabilities proportional to $\alpha_i = P_i/(1-P_i)$ but reject the sample already selected (and hence start the whole sampling process again) if a repetition occurs. This sampling scheme is equivalent to the use of Poisson sampling conditional on having n distinct units (Hajek (1981), Ch. 7).

C - *Sampford's Method.* Draw the first unit with probabilities P_i and the remaining $(n-1)$ units with probabilities proportional to $P_j/(1-nP_j)$, with replacement. As in B, a sample is rejected once a unit is selected twice, in which case the whole sampling process is started again (Hajek (1981), Ch. 8).

The results stated so far and proved in the Appendix assume that the population measurements are generated independently. This assumption is violated in clustered populations where measurements within the same cluster are ordinarily correlated. When fitting models to clustered populations, it is customary to assume independent cluster effects (possibly represented by an observable variable), and independent residuals given the cluster effects. See, for example, the model defined by (5.7) in Section 5.2.

It follows that if the clusters and ultimate sampling units are selected by one of the methods considered in this section (see next section for simulation results for other methods), the cluster effects corresponding to the sampled clusters are asymptotically independent and so are the unit level residuals. Thus, the model holding for the sample has the same form as the population model, and it can be extracted, in principle, and used for inference.

6.2. Simulation study

In order to illustrate the theoretical results of Section 6.1 and examine the independence of the sample measurements for other sampling methods in common use, we performed a simulation study to compare the empirical distribution of various sample statistics under the different sampling methods with the distribution of the same statistics under independent sampling from the marginal sample distribution(3.3). Among the statistics considered is the "product likelihood", defined as the product of the corresponding marginals. If the distribution of the product likelihood under the various sampling methods is close to its distribution under independent sampling (in which case it is the true likelihood), it suggests that the sample measurements as obtained under these methods are approximately independent, and hence inference based on independence assumptions is justified.

The simulation study consists of several steps. In Steps I-III data were generated according to models corresponding to some of the examples in this paper, while in Step IV independent observations were generated according to the product of the marginals as in (3.1) or (3.3). (See below for details.) Steps I-IV were repeated R times. In Step V we use these simulations to study the sample distributions of the statistics considered in order to assess the independence of the sample measurements under the various sampling methods.

Step I. Generate N independent population measurements $\{Y_1,\ldots,Y_N\}$ from pdf $f_p(y|x)$ with the x-values generated from Gamma(α_x,β_x). This was done in two different ways: I_1 - logistic distribution, i.e., $P(Y_i=j|x_i)=\exp(a_j+b_jx_i)/\sum_{\ell=0}^{L-1}\exp(a_\ell+b_\ell x_i)$, $i=1,\ldots,N$, $j=0,\ldots,(L-1)$; I_2 - Normal distribution, i.e., $Y_i=a+bx_i+\epsilon_i$, $\epsilon_i\sim N(0,\sigma_\epsilon^2)$, $i=1,\ldots,N$.

Step II. Generate population values for a design variable $Z_i=h(Y_i,u_i)$ where the $\{u_i\}$ are random. The Z values were generated in two different ways: II$_1$ - $Z_i=B_0+B_1Y_i+u_i$, $u_i\sim$ Gamma(α_u,β_u); II$_2$ - $Z_i=\exp(B_0^*+B_1^*Y_i+u_i^*)$, $u_i^*\sim$ Gamma(α_u^*,β_u^*), $i=1,\ldots,N$.

Defining the size variable as a direct function of the response variable is clearly an extreme case of informative sampling. Note that I and II define four different combinations of population distributions and design size variables.

Step III. Select PPS samples of size n with Z as the size variable using seven different sampling methods (one sample per population × method) as follows:

III$_1$ - *Sampford's method* (method C in Section 6.1); III$_2$ - *Chao's* (1982) *method*; III$_3$ - *Rejective sampling* (method B in Section 6.1); III$_4$ - *Successive* *sampling* (method A in Section 6.1); III$_5$ - *Systematic PPS sampling* (Hajek (1981), Ch. 10); III$_6$ - PPS sampling *with replacement*; III$_8$ - *Simple random sampling without replacement.*

Methods III$_1$, III$_3$, III$_4$ and III$_6$ are covered by Theorem 1 and Corollary 1 of the Appendix. Method III$_8$ is used as a benchmark for assessing the sampling effects on the distribution of the sample measurements.

Step IV. Generate n *independent* values from the marginal sample pdf defined by (3.3) with $f_p(y|x)$ as in I and $\pi_i \propto Z_i$. Measurements for the four marginal sample distributions corresponding to the two population distributions defined in I, and the two design variables defined in II, were obtained in two stages. In the first stage values (y, x) were generated from the population distribution $f_p(y, x) = f_p(y|x)f_p(x)$ and the x-values selected to the sample with probability $\pi = Z/N\bar{Z}$, where Z is defined as in Step II above and $\bar{Z}$ is the corresponding mean of the population Z-values obtained in Step II. This process was repeated independently until n sample values, $x_1, \ldots, x_n$ were obtained. Note that the y_i's were only generated in order to enable us to generate the Z-values needed for computing the π's. In the second stage values y_i, $i = 1, \ldots, n$, were generated independently from the corresponding marginal sample pdf as obtained from (3.3). The four sample pdfs are defined below where we use the notation (I_a, II_b), $a, b = 1, 2$ to denote the population pdf defined by I_a (Step I) and the design variable defined by II_b (Step II).

Case 1: (I$_1$,II$_1$), $\Pr(Y_i = j|x_i, i \in s) = \exp(\tilde{a}_j + b_j x_i)/\sum_{\ell=0}^{L-1} \exp(\tilde{a}_\ell + b_\ell x_i)$, where $\tilde{a}_j = \{a_j + \ln[B_1 j + B_0 + E(u_i)]\}$, $j = 0, \ldots, (L-1)$.

Case 2: (I$_1$,II$_2$), $\Pr(Y_i = j|x_i, i \in s) = \exp(a_j^* + b_j x_i)/\sum_{\ell=0}^{L-1} \exp(a_\ell^* + b_\ell x_i)$, where $a_j^* = (a_j + B_1^* j)$, $j = 0, \ldots, (L-1)$.

Case 3: (I$_2$,II$_1$), $f_s(y_i|x_i) = [\tilde{B}_0 f_p(y_i|x_i) + B_1(a + bx_i)f_p^{(1)}(y_i|x_i)]/[\tilde{B}_0 + B_1(a + bx_i)]$ where $f_p(y_i|x_i) = N(a + bx_i, \sigma_\epsilon^2)$, $f_p^{(1)}(y_i|x_i) = y_i f_p(y_i|x_i)/(a + bx_i)$ and $\tilde{B}_0 = B_0 + E(u_i)$.

Case 4: (I$_2$,II$_2$), $f_s(y_i|x_i) = N(a + B_1^*\sigma_\epsilon^2 + bx_i, \sigma_\epsilon^2)$, (see Example 4.3).

Step V. Compute the empirical deciles of each of eight different sample statistics as obtained under Steps I - III for the seven sampling methods listed in Step III, and compare them with the empirical deciles obtained in step IV.

6.3. Results

The results obtained for the various statistics are very similar, therefore we restrict attention to the comparison of the "product likelihood" statistics, computed as the product of the marginal densities (3.3) as defined for Cases 1-4 in Section 6.2.

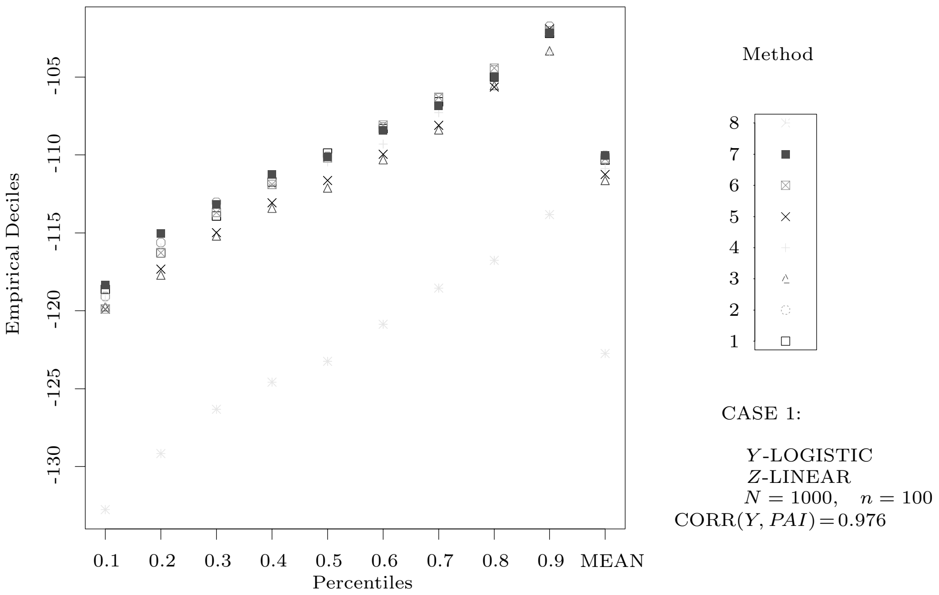

Figure 1. Empirical deciles and mean value of LOG-LIKELIHOOD under different sampling schemes and IID sampling from sample distribution.

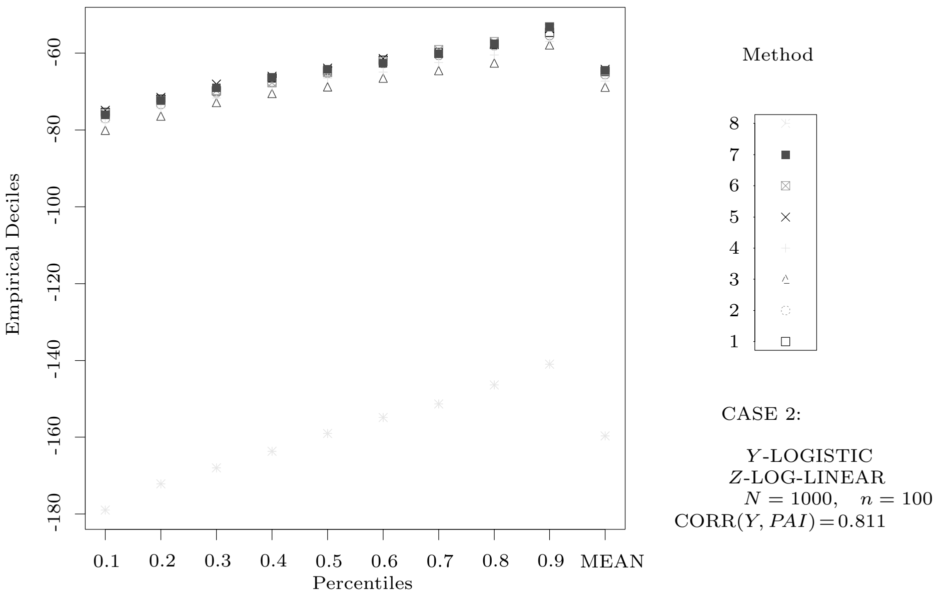

Figure 2. Empirical deciles and mean value of LOG-LIKELIHOOD under different sampling schemes and IID sampling from sample distribution.

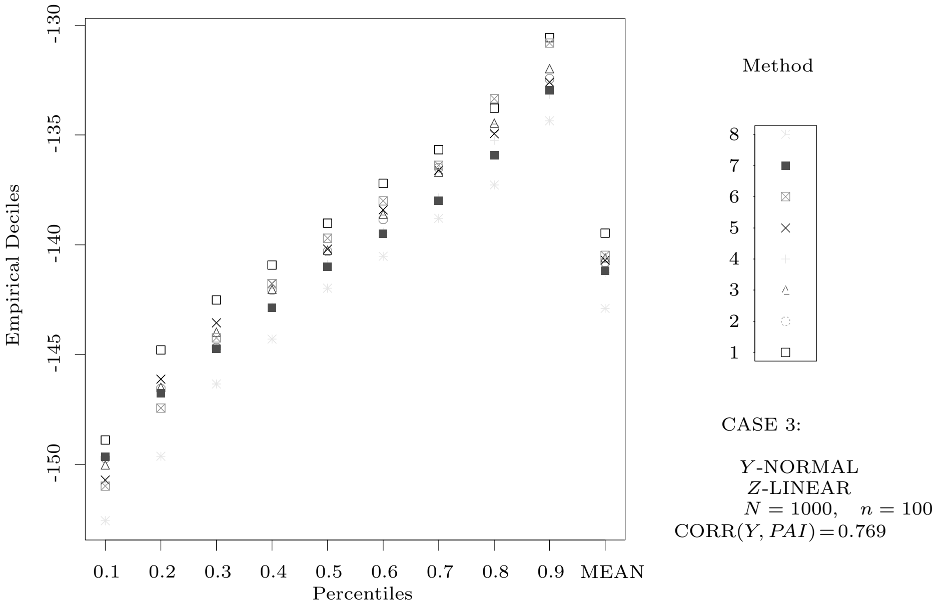

Figure 3. Empirical deciles and mean value of LOG-LIKELIHOOD under different sampling schemes and IID sampling from sample distribution.

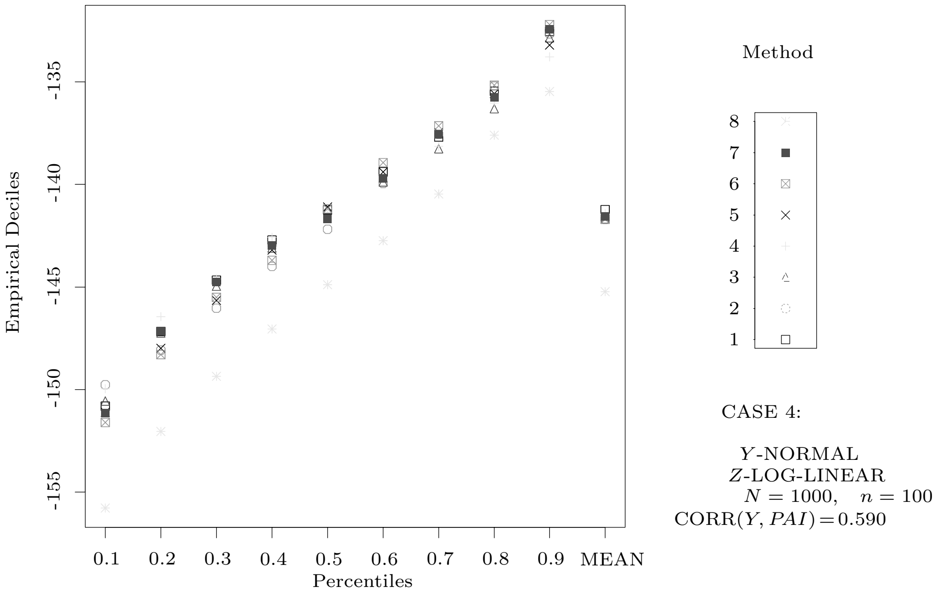

Figure 4. Empirical deciles and mean value of LOG-LIKELIHOOD under different sampling schemes and IID sampling from sample distribution.

Figures 1-4 display the empirical deciles of the loglikelihood as obtained for each sampling method under the two population distributions and the two design variables. When simulating from the logistic distribution we set $L = 4$, so there are four levels for the response variable. The deciles were computed based on $R = 300$ simulations with populations of size $N = 1000$ and samples of size $n = 100$. Also shown on the figures are the mean values of the loglikelihood over the 300 samples, and the mean population correlation, CORR(Y, PAI), between the response $\boldsymbol{y}$ and the selection probabilities $\boldsymbol{\pi}$. The numbers in the legend of each graph correspond to the numbers defining the various sampling methods in the description of Step III. Method 7 is the independent selections from the marginal sample distribution and Method 8 is simple random sampling without replacement.

The graphs indicate that the six sampling methods 1-6 for selection with unequal probabilities yield very similar deciles, despite the relatively large sampling fraction, $n/N = 0.1$. These deciles are indeed similar to the deciles obtained under independent drawing from the marginal sample distribution (Method 7), suggesting that the sample measurements obtained under the six sampling methods are not dependent in a way that significantly affects the distribution of the loglikelihood.

It should be emphasized that the two design variables considered in this study define extreme cases of informative sampling. This can be inferred from the high correlations between the response variable and the selection probabilities, (CORR(Y,PAI)), and by comparison of the deciles obtained under the first seven sampling methods with the deciles obtained under simple random sampling. (Notice in particular Figures 1 and 2 for the logistic model.) The results of the simulation study suggest therefore that whereas ignoring the sampling mechanism in the inference process (i.e., assuming simple random sampling) may be very misleading, basing the inference on the marginal sample distribution, assuming independence of the sample measurements, appears to be a sound procedure.

7. Summary

In this article, we show how the distribution holding for the sample measurements may be derived from the distribution holding in the population and the first order selection probabilities for units in the sample. Theoretical and simulation results indicate that for large populations and small sampling fractions, when the population values are independent so are, asymptotically, the sample values, even when employing sampling methods that involve joint selection of the sample units. Extracting the sample distribution as proposed here has the advantage of permitting the use of classical inference tools, without the need to condition on the population values of design variables that determine the selection probabilities. As discussed in Section 2, the latter may not be possible because of data availability or may not be of scientific interest.

In the approach considered in this article, the sample selection probabilities for units in the sample are only used for modeling the expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$. It could be argued that a more explicit use of these probabilities would increase the efficiency of the inference process. This could be done in principle by considering the joint sample distribution of $(Y_i, \pi_i)|X_i$, i.e.,

$$f_s(y_i, \pi_i|\boldsymbol{x}_i) = f(y_i, \pi_i|\boldsymbol{x}_i, i \in s) = \frac{\pi_i f_p(\pi_i|y_i, \boldsymbol{x}_i) f_p(y_i|\boldsymbol{x}_i)}{E_{Y_i|\boldsymbol{x}_i}[E_p(\pi_i|y_i, \boldsymbol{x}_i)]}. \tag{7.1}$$

Unlike the sample pdf defined by (3.3), which only requires the specification of the expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$, the use of (7.1) requires the specification of the full pdf $f_p(\pi_i|y_i, \boldsymbol{x}_i)$. This is in general much more complicated.

The next obvious step in this research is the application of the proposed approach to real data sets. The most interesting and important question in this respect is the modeling of the expectations $E_p(\pi_i|y_i, \boldsymbol{x}_i)$ and the sensitivity of the resulting sample distribution to the model specification. Work in this direction is currently in progress.

Acknowledgement

D. Pfeffermann's work was supported by a grant from the Israel Science Foundation, and by the UK Economic and Social Research Council's programme for the Analysis of Large and Complex Data sets at the University of Southampton. A. Krieger's work was supported by a summer grant from the Research Foundation of the University of Pennsylvania. Y. Rinott was supported in part by NSF grant DMS-95-04616. The authors would like to thank Chris Skinner, an associate editor and a referee for useful comments and suggestions.

Appendix

A. Proof of Theorem 1

We need several simple lemmas for the proof.

Lemma 1. *For any $m > 0$, there exists $D_m < \infty$ such that for $N \geq K_m$, $E(1/\bar{Z}^m) < D_m$.*

Proof. It suffices to show that for $k \geq K_m$, $E[(k+1)/(Z_1 + \cdots + Z_{k+1})]^m \leq E[k/(Z_1 + \cdots + Z_k)]^m$. The latter inequality follows from the relation $E((Z_1 + \cdots + Z_k)/k \,|\, Z_1 + \cdots + Z_{k+1}) = (Z_1 + \cdots + Z_{k+1})/(k+1)$, and from Jensen's inequality applied to the convex function $g(x) = 1/x^m$, yielding

$$\begin{aligned} Eg\Big(\frac{Z_1+\cdots+Z_{k+1}}{k+1}\Big) &= Eg\Big(E[\frac{Z_1+\cdots+Z_k}{k}\mid Z_1+\cdots+Z_{k+1}]\Big)\\ &\le E\Big\{E[g(\frac{Z_1+\cdots+Z_k}{k})\mid Z_1+\cdots+Z_{k+1}]\Big\}\\ &= Eg\Big(\frac{Z_1+\cdots+Z_k}{k}\Big). \end{aligned}$$

Lemma 2. *Let $\mu = EZ_i$. For any fixed n, As $N \to \infty$, $E[\frac{1}{\bar{Z}^n} - \frac{1}{\mu^n}]^2 = O(\frac{1}{N})$.*

Proof. By the Cauchy-Schwarz inequality

$$E\Big[\frac{1}{\bar{Z}^n} - \frac{1}{\mu^n}\Big]^2 = E\Big(\frac{\bar{Z}^n-\mu^n}{\bar{Z}^n\mu^n}\Big)^2 \le \Big(E\frac{1}{\bar{Z}^{4n}\mu^{4n}}E(\bar{Z}^n-\mu^n)^4\Big)^{1/2}.$$

By Lemma 1, $E[1/(\bar{Z}^{4n}\mu^{4n})]$ is bounded. Write $(\bar{Z}^n-\mu^n)^4$ as a product of $(\bar{Z}-\mu)^4$ and $(\bar{Z}^{n-1}+\bar{Z}^{n-2}\mu+\cdots+\mu^{n-1})^4$, and note that the second expression has a finite second moment. Also note that for i.i.d. random variables Z_i satisfying $EZ_i^{2k} < \infty$, $E(\bar{Z}-\mu)^{2k} = O(\frac{1}{N^k})$. The latter property (with $k = 4$) and another application of the Cauchy-Schwarz inequality yields the desired result.

Lemma 3. *As $N \to \infty$, $E(\prod_{i\in S} Z_i(|\frac{1}{\bar{Z}^n} - \frac{1}{\mu^n}|) \mid X_s) = O(\frac{1}{N^{1/2}})$.*

Proof. Suppose for notational convenience that $S = \{1,\ldots,n\}$. By Jensen's inequality, the positivity of the Z_i's and the last part of Assumption (a), we have

$$\begin{aligned} &E\Big(\frac{E(Z_1|\boldsymbol{X}_1)+\cdots+E(Z_n|\boldsymbol{X}_n)}{N}+\frac{Z_{n+1}+\cdots+Z_N}{N}\Big)^{-n}\\ &\le E(\bar{Z}^{-n}|X_s) \le E\Big(\frac{Z_{n+1}+\cdots+Z_N}{N}\Big)^{-n}. \end{aligned}$$

It is easy to see that both the left hand side and the right hand side of the latter relation are within $O(\frac{1}{N})$ of $E(\frac{1}{\bar{Z}^n})$. The Cauchy-Schwarz inequality, and Lemma 2 now imply the result.

Proof of Theorem 1. Let $\boldsymbol{Z} = \{Z_1,\ldots,Z_N\}$. By Bayes' theorem

$$f(\boldsymbol{y}_s|X_s,s) = \frac{\int f(\boldsymbol{y}_s|X_s,s,\boldsymbol{Z})\Pr(S|X_s,Z)dF(Z|X_s)}{\int \Pr(S|X_s,Z)dF(Z|X_s)}. \tag{A.1}$$

Next note that for PPS sampling with replacement with Z as the size variable,

$$\Pr(S|X_s,\boldsymbol{Z}) = \Pr(S|\boldsymbol{Z}) = \prod_{i\in S}(Z_i/N\bar{Z}), \tag{A.2}$$

where the sample S is viewed as a multiset, so that repetitions are taken into account if they occur. By Condition (c), if S has n distinct elements,

$$f(\boldsymbol{y}_s|X_s, s, \boldsymbol{Z}) = \prod_{i\in S} f_p(y_i|\boldsymbol{X}_i, Z_i). \tag{A.3}$$

Hence, by Lemma 3, Condition (a) and the boundedness of $f_p(y_i|\boldsymbol{X}_i, Z_i)$ (Condition (c)), we have that the numerator of (A.1) equals $\frac{1}{N^n}\{\prod_{i\in S} E[f_p(y_i|\boldsymbol{X}_i, Z_i) \frac{Z_i}{\mu}|\boldsymbol{X}_i]+O(\frac{1}{N^{1/2}})\}$. In the same way, the denominator of (A.1) equals $\frac{1}{N^n}\{\prod_{i\in S} E[\frac{Z_i}{\mu}|\boldsymbol{X}_i] + O(\frac{1}{N^{1/2}})\}$. Since by Condition (d), $E(Z_i|\boldsymbol{X}_i) > c$, it follows that

$$f(\boldsymbol{y}_s|X_s, s) = \frac{\prod_{i\in S} E[f_p(y_i|\boldsymbol{X}_i, Z_i)Z_i|\boldsymbol{X}_i]}{\prod_{i\in S} E(Z_i|\boldsymbol{X}_i)} + O(\frac{1}{N^{1/2}}), \tag{A.4}$$

from which the relationship (6.1) of Theorem 1 follows straightforwardly. The relationship (6.2) corresponding to the case where S contains repetitions is obtained in the same way, thus completing the proof.

B. Asymptotic Sample Independence under Successive and Rejective Sampling

In Section 6.1 we describe three sampling methods of selection without replacement labeled as A, B and C. Under successive sampling (method A), the probability of drawing a given sample S of size n, conditioned on the vector $\boldsymbol{Z}$ is,

$$\Pr(S|\boldsymbol{Z}) = (\prod_{i\in S} P_i) \sum_{r_1,\ldots,r_n} [(1 - P_{r_1}) \cdots (1 - P_{r_1} - \cdots - P_{r_{n-1}})]^{-1}, \tag{B.1}$$

where $P_i = Z_i/N\bar{Z}$ and the summation is over all permutations of members of S (Hajek (1981), Ch. 9).

Under Rejective sampling (method B), the probability of obtaining a sample S of size n is given by

$$\Pr(S|\boldsymbol{Z}) = c\prod_{i\in S} P_i/(1 - P_i), \tag{B.2}$$

where c is the normalizing constant (Hajek (1981), Ch. 7).

Under Sampford's sampling scheme (method C),

$$\Pr(S|\boldsymbol{Z}) = (P_1 \cdots P_n)[1 - (P_1 + \cdots + P_n)]/[(1 - nP_1) \cdots (1 - nP_n)]. \tag{B.3}$$

In view of (B.1), (B.2) and (B.3), the result of Theorem 1 can be extended as follows.

Corollary 1. *The relationship* (6.1) *holds for successive and rejective sampling under the same conditions* (*a*)-(*d*). *It holds also for Sampford's method if* $E(e^{tZ_i}|\boldsymbol{X}_i) < B < \infty$ *for some* $B, t > 0$.

Proof. We provide the details of the proof for Sampford's method; the other two cases being simpler or similar. Note that for all three sampling schemes S contains n distinct elements.

Let G_s denote the set $\{\boldsymbol{Z} : \frac{1}{1-nP_i} < 1 + \frac{1}{N^{1/2}},\ P_i < \frac{1}{N^{1/2}};\ i \in S\}$. It is easy to see from (B.3) that for $\boldsymbol{Z} \in G_s$

$$\Pr(S|\boldsymbol{Z}) = (P_1 \cdots P_n)[1 + O(\frac{1}{N^{1/2}})]. \tag{B.4}$$

We first show that for every sample S and sufficiently large N, $\Pr(G_s|X_s) \geq (1 - \frac{1}{N})$. The notation below suppresses the conditioning on X_s, however, all probabilities and expectations are conditioned on X_s. Note that the left hand side condition in G_s implies the second. Therefore, it suffices to prove that for large N, $\Pr(\frac{1}{1-nP_i} < 1 + \frac{1}{N^{1/2}}) \geq 1 - (\frac{1}{N})$. Indeed, direct calculations show that for large N,

$$\Pr\Big(\frac{1}{1-nP_i} < 1 + \frac{1}{N^{1/2}}\Big) \geq \Pr\Big(\sum_{j=1}^{N} Z_j > N^{3/4} Z_i\Big) \geq 1 - \frac{1}{N}, \tag{B.5}$$

where the second inequality follows from standard large deviation (or Bernstein) bounds, (see the proof at the end of this Appendix).

We now derive the relationship (6.1). By (B.4) and (B.5), the numerator of (A.1) can be written as

$$\begin{aligned}
&E\prod_{i\in S}\Big\{f_p(y_i|\boldsymbol{X}_i, Z_i)\frac{Z_i}{N\bar{Z}}\Big\}\Big[1{+}O(\frac{1}{N^{1/2}})\Big]\mathbb{I}_{G_s}{+}E\prod_{i\in S}\{f_p(y_i|\boldsymbol{X}_i, Z_i)\}P(S|\mathbf{Z})(1{-}\mathbb{I}_{G_s})\\
&= E\prod_{i\in S}\Big\{f_p(y_i|\boldsymbol{X}_i, Z_i)\frac{Z_i}{N\bar{Z}}\Big\}\Big[1 + O(\frac{1}{N^{1/2}})\Big]\\
&\quad +E\Big\{-\prod_{i\in S}\{f_p(y_i|\boldsymbol{X}_i, Z_i)\frac{Z_i}{N\bar{Z}}\}[1{+}O(\frac{1}{N^{1/2}})]{+}\prod_{i\in S}\{f_p(y_i|\boldsymbol{X}_i, Z_i)\}P(S|\mathbf{Z})\Big\}(1{-}\mathbb{I}_{G_s}),
\end{aligned}$$

where $\mathbb{I}_{G_s}$ denotes the indicator of the set G_s, and all expectations are taken with respect to $\mathbf{Z}$, conditioned on X_s. By the boundedness of $f_p(y_i|\boldsymbol{X}_i, Z_i)$ and the result that $1 - \Pr(G_s) = O(\frac{1}{N})$, we obtain by a standard Cauchy-Schwarz argument that the numerator of (A.1) equals $E\prod_{i\in S}\{f_p(y_i|\boldsymbol{X}_i, Z_i)\frac{Z_i}{N\bar{Z}}\}[1 + O(\frac{1}{N^{1/2}})]$. A similar (but easier) calculation shows that the denominator of (A1) equals $E\prod_{i\in S}\{\frac{Z_i}{N\bar{Z}}\}[1 + O(\frac{1}{N^{1/2}})]$. Application of Lemmas 1-3 (which do not depend on the sampling plan) yields the relationship (6.1).

Finally, we show that $\Pr(\sum_{j=1}^{N} Z_j < N^{3/4} Z_i) \leq \frac{1}{N}$, proving the right hand side inequality in (B.5). In fact, here we derive a stronger result: the right hand side bound can be shown to be of order $e^{-cN^{1/4}}$ for some positive constant c. Bernstein's type inequality for i.i.d. variables U_i with $EU_i < 0$ states that $\Pr(\sum_{i=1}^{N} U_i > N\epsilon) \leq e^{-N\epsilon t}$ for some $t > 0$.

Now, $\Pr(N^{3/4} Z_i > N\delta) = \Pr(e^{tZ_i} > e^{t\delta N^{1/4}}) \leq Ee^{tZ_i}/e^{t\delta N^{1/4}} = O(e^{-t\delta N^{1/4}})$ since $E(e^{tZ_i}|\boldsymbol{X}_i)$ is assumed to be bounded. Applying this relationship with $\epsilon < EZ_i$, we obtain $\Pr(-\sum_{j=1}^{N} Z_j > -N^{3/4} Z_i) = \Pr(-\sum_{j=1}^{N} Z_j + N\epsilon > N\epsilon - N^{3/4} Z_i) \leq O(e^{-t\delta N^{1/4}}) + \Pr(-\sum_{j=1}^{N} Z_j + N\epsilon > N(\epsilon - \delta)) \leq O(e^{-\delta t N^{1/4}}) + e^{-N(\epsilon-\delta)t}$; the latter inequality follows from the Bernstein inequality, with $U_i = -(Z_i - \epsilon)$ and $\delta < \epsilon$.

References

Chao, M. T. (1982). A general purpose unequal probability sampling plan. *Biometrika* **69**, 653-656.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Goldberger, A. S. (1981). Linear regression after selection. *J. Econom.* **15**, 357-366.

Hajek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Hausman, J. A. and Wise, D. A. (1981). Stratification on endogenous variables and estimation: The gary income maintenance experiment. In *Structural Analysis of Discrete Data with Econometric Applications* (Edited by C. F. Mansky and D. McFadden), 366-391. MIT Press, Cambridge MA.

Korn, E. L. and Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *Amer. Statist.* **49**, 291-295.

Krieger, A. M. and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology* **18**, 225-239.

Krieger, A. M. and Pfeffermann, D. (1997). Testing of distribution functions from complex sample surveys. *J. Official Statist.* **13**, 123-142.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **77**, 237-250.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics* **34**, 179-189.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Internat. Statist. Rev.* **61**, 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Resarch* **5**, 239-261.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions* (Edited by A. C. Atkinson and S. E. Fienberg), 320-332. Springer-Verlag, New York.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine* **16**, 21-37.

Rotnitzky, A. and Robins, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine* **16**, 81-102.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics* **2** (Edited by J. M. Bernardo, M. H. Degroot, D. V. Lindley and A. F. M. Smith), 463-472. North-holland, Amsterdam.

Skinner, C. J. (1994). Sample models and weights. In *Proceedings of the Section on Survey Research Methods*, 133-142. American Statistical Association, Alexandria, VA.

Smith, T. M. F. (1988). To weight or not to weight, that is the question. In *Bayesian Statistics* **3** (Edited by J. M. Bernardo, M. H. Degroot, D. V. Lindley and A. F. M. Smith), 437-451. Oxford University Press, Oxford.

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **74**, 495-506.

Department of Statistics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel.

E-mail: msdanny@olive.mscc.huji.ac.il

Department of Statistics, University of Pennsylvania, The Wharton School, Philadelphia, PA 19104-6302, U.S.A.

E-mail: abba@stat.wharton.upenn.edu

Department of Mathematics, University of California San Diego, La Jolla, CA 92093, U.S.A.

E-mail: yrinott@ucsd.edu

(Received January 1997; accepted February 1998)