Southampton Statistical Sciences Research Institute
University of Southampton

Maximum Likelihood Under Response Biased Sampling

Raymond Chambers, Alan Dorfman, Suojin Wang

Abstract

Informative sampling occurs when the probability of inclusion in sample depends on the value of the survey response variable. Response or size biased sampling is a particular case of informative sampling where the inclusion probability is proportional to the value of this variable. In this paper we describe a general model for response biased sampling, which we call array sampling, and develop maximum likelihood and estimating equation theory appropriate to this situation. The Missing Information Principle (MIP) (Orchard and Woodbury, 1972) yields one (indirect) approach to likelihood based survey inference (Breckling et al 1994). Some have questioned its applicability in the case of informative sampling, because of the way it conditions on the given sample. In this paper we describe a direct approach and show that it and the MIP-based approach lead to identical results under array sampling. Comparison is made to the weighted likelihood based approach described in Krieger and Pfeffermann (1992). Extensions to the theory are also explored.

S^{3}RI Methodology Working Paper M03/18

Maximum Likelihood Under Response Biased Sampling

By RAYMOND L. CHAMBERS[1]

University of Southampton, Southampton, UK

ALAN H. DORFMAN

Bureau of Labor Statistics, Washington, USA

and SUOJIN WANG

Texas A&M University, College Station, USA

April 2000

(1) Address for correspondence: Department of Social Statistics
University of Southampton
Highfield
Southampton SO17 1BJ
UK

SUMMARY

Informative sampling occurs when the probability of inclusion in sample depends on the value of the survey response variable. Response or size biased sampling is a particular case of informative sampling where the inclusion probability is proportional to the value of this variable. In this paper we describe a general model for response biased sampling, which we call array sampling, and develop maximum likelihood and estimating equation theory appropriate to this situation. The Missing Information Principle (MIP) (Orchard and Woodbury, 1972) yields one (indirect) approach to likelihood based survey inference (Breckling et al 1994). Some have questioned its applicability in the case of informative sampling, because of the way it conditions on the given sample. In this paper we describe a direct approach and show that it and the MIP-based approach lead to identical results under array sampling. Comparison is made to the weighted likelihood based approach described in Krieger and Pfeffermann (1992). Extensions to the theory are also explored.

Key Words: FINITE POPULATION, SUPERPOPULATION, INCLUSION PROBABILITIES, MISSING INFORMATION PRINCIPLE, SCORE FUNCTION, ESTIMATING EQUATION, WEIGHTED DISTRIBUTION LIKELIHOOD

1. Introduction

Consider the following situation. A population of N independent and identically distributed realisations of a random variable X, with density f(x) indexed by an unknown parameter θ, is sampled informatively. That is, the sampling method uses knowledge about the population realisations of X to randomly determine which population units to include in the sample. Unfortunately, the analyst (who wishes to make an inference about θ) only has access to the sample values of X, plus a limited amount of information about the sampling method, in the form of the value ξ of another parameter which, together with the (unknown) population values of X, completely determines the distribution of the outcomes of the sampling process. It is assumed that the analyst "knows" the functional form of the population density f, as well as enough about the sample design to specify the conditional distribution (given X) of the (random) sample inclusion indicator I.

The analyst wishes to estimate the value of θ on the basis of the available "data" (i.e. the sample values of X and the value of ξ). He/she can then proceed in one of two ways.

(a) On the basis of knowledge of the population density f of X and the distribution of I given X, the analyst can determine the "sampling density" $f_s(x)$ of X. That is, the distribution of sample values of X obtained by repeated draws from the JOINT population distribution of X and I. Operationally, this can be visualised as the distribution of sample values of X obtained by a two step process: (i) generate N iid population values of X from the density f(x) and index them as $x_1, x_2, .., x_N$; (ii) on the basis of these values, draw a sample s of n "labels" from the population index set {1, 2, ..., N} using the informative sampling method (which depends only on the known ξ and these values) and record the X-values for sample thus observed. Let $\mathbf{X}_N$ denote the random N-vector of population values of X, with $\mathbf{I}_N$ denoting the corresponding vector of realizations of I, and put $\mathbf{X}_S$ equal to the random n-vector of sample values obtained as a result of this two-stage process. Here S is used to signify the set-valued random variable whose outcome is s above. Note that knowledge of S is equivalent to knowledge of $\mathbf{I}_N$. Clearly the distribution of $\mathbf{X}_S$ will NOT, in general, be the same as the distribution of n randomly chosen values of X. The task facing the analyst therefore is to determine the distribution of $\mathbf{X}_S$. Suppose this distribution can also be modelled, and in such a way that the only unknown in this distribution is the value of another parameter, say ω, which is related, in a known way, to both θ and ξ. Provided the usual regularity conditions are satisfied, it is clear that the sample values in $\mathbf{X}_S$ can be used to determine the MLE for ω, and consequently the MLE for θ obtained via invariance and the known relationship between ω, θ and ξ. We refer to this approach as "sample-based" maximum likelihood.

(b) The analyst can use methods described in Breckling et al (1994) which, in the sampling context, apply the Missing Information Principle (MIP) (Orchard and Woodbury 1972); see also Chambers, Dorfman, and Wang (1998). That is, he/she can write down the score function for θ generated by $\mathbf{X}_N$ and S, and then obtain the sample-based score function for θ by taking the conditional expectation of this population-based score given S and the sample values of X. The MLE for θ is obtained by setting this score to zero and solving for θ. The essence of this approach is that, in taking the above conditional expectation, there is no need to "model" the distribution of $\mathbf{X}_S$, since this is replaced by a function dependent on the joint distribution of $\mathbf{X}_s$ and S, where s is the index set corresponding to the non-missing values in the population [i.e. s is the realization of S]. We refer to this approach as "MIP-based" maximum likelihood.

Both (a) "sample-based" likelihood and (b) "MIP-based" maximum likelihood are firmly grounded in probability theory, and therefore must in principle lead to the same maximum likelihood estimator, irrespective of whether the sampling method is informative or not. However, this equivalence is not especially clear in the case of informative sampling. Consequently it is of interest to explicitly demonstrate it for a method of informative sampling that has wide applicability.

In the following section we therefore introduce a simple method of informative sampling, which we refer to as *array sampling*, and develop both the sample-based and MIP-based ML estimating equations that arise. We show that under array sampling both sample-based and MIP approaches lead to the same MLE for the parameter θ of interest. By way of comparison, in Section 3 we develop the weighted distribution maximum likelihood estimator (WDMLE) for this case. This estimator was introduced by Krieger and Pfeffermann (1992) specifically for informative sampling. We show that under array sampling the WDMLE in fact corresponds to an approximation to the MLE. Furthermore it is an approximation that can be inefficient in some situations. In Section 4 we go on to generalise the definition of the MIP approach to arbitrary estimating equations. In particular, we show that under array sampling a MIP-based estimating equation for a population mean based on an unbiased population level estimating equation for this parameter remains unbiased. Section 5 then concludes the paper with some discussion of extensions and potential further research in this area.

2. MLE under Array Sampling

The following sampling method, which we refer to as *array sampling* from now on, provides a context for the theory developed in this paper. It can be taken as a first order model for draw by draw with replacement informative sampling from a finite population of size N. It also approximates the behaviour of systematic sampling applied to a randomly ordered population. With an extension to “ragged array sampling” (see Section 5), it approximates the without replacement sampling scheme proposed in Rao, Hartley and Cochran (1962) where the size measure is in fact the survey variable of interest.

Consider a population which can be represented as an n × M array $[X_{ij}]$ of iid realisations of a nonnegative random variable X, with density f(x) indexed by an unknown parameter θ. The overall population size is thus N = nM. For simplicity we assume θ is scalar, although the generalisation to the vector case is straightforward as illustrated in Example 2 below. Array sampling from this population is defined as follows: Generate n independently distributed random variables $\{A_i, i = 1, 2, .., n\}$, each taking values over the integers between 1 and M, and such that, for j = 1, .., M

$$\mathrm{pr}\left(A_i = j \middle| X_{ik}; k = 1, 2, \cdots, M\right) = \frac{\xi + X_{ij}}{M\xi + T_i} \qquad (1)$$

where $\xi > 0$ is a known constant and T_i is the sum of X-values in row i of the array.

Our sample data then consist of the observed values of A_i, together with the values $\{X_{iA_i}; i = 1, 2, \cdots, n\}$. That is, the values of A_i determine the outcome of an informative sampling mechanism. Our objective is to compute the MLE for θ on the basis of these sample data.

To start, we adopt the sample-based likelihood approach. Clearly, the sample values $\{X_{iA_i}; i = 1, 2, \cdots, n\}$ are independently distributed. Let Δ be a small positive increment, with $\Delta_x = (x, x + \Delta)$. Then

$$\begin{aligned}\mathrm{pr}\left(X_{iA_i} \in \Delta_x\right) &= \sum_{j=1}^{M} \mathrm{pr}\left(A_i = j, X_{ij} \in \Delta_x\right) \\ &= \mathrm{pr}\left(X_{ij} \in \Delta_x\right) \sum_{j=1}^{M} \mathrm{pr}\left(A_i = j \middle| X_{ij} \in \Delta_x\right) \\ &= \mathrm{pr}\left(X_{ij} \in \Delta_x\right) \sum_{j=1}^{M} \int \mathrm{pr}\left(A_i = j \middle| X_{ij} \in \Delta_x, \sum_{k \neq j}^{M} X_{ik} = u\right) f(u)du \\ &= \mathrm{pr}\left(X_{ij} \in \Delta_x\right) M \int \frac{\xi + x}{M\xi + x + u} f(u)du + o(\Delta) \\ &= M\mathrm{pr}\left(X_{ij} \in \Delta_x\right)(\xi + x) E\left(\frac{1}{M\xi + x + U}\right) + o(\Delta)\end{aligned}$$

where U is the sum of M - 1 independent realisations of X, and we use f(.) as generic notation for a density (here f(u) denotes the density of the random variable U). Dividing both sides by Δ and letting Δ tend to zero leads to an expression for the "sample density" of X under array sampling:

$$f_s(x) = Mf(x)(\xi + x)E\left(\frac{1}{M\xi + x + U}\right). \qquad (2)$$

Clearly the different sample values of X are independently and identically distributed under array sampling. Consequently the sample-based score function for θ is

$$\mathrm{sc}_s(\theta) = \sum_{i=1}^{n} \left\{ \frac{\partial_\theta f(X_{is})}{f(X_{is})} + \frac{\partial_\theta E\left(\dfrac{1}{M\xi + X_{is} + U}\right)}{E\left(\dfrac{1}{M\xi + X_{is} + U}\right)} \right\}. \tag{3}$$

Here ∂_θ denotes partial differentiation with respect to θ, X_{is} denotes a sample value of X and the expectation is with respect to U for fixed X_{is}.

On the other hand, under the MIP approach, the score function for θ is

$$\mathrm{sc}_s(\theta) = \sum_{i=1}^{n} \left\{ \partial_\theta \log f(X_{is}) + (M-1)E\left(\partial_\theta \log f(X_{ir}) \middle| X_{is}, A_i\right) \right\} \tag{4}$$

where X_{ir} denotes a "generic" non-sample value of X from the i^{th} "row" of the population. Now

$$\begin{aligned} E\left(\partial_\theta \log f(X_{ir}) \middle| X_{is}, A_i\right) &= \int \partial_\theta \log f(x) f\left(x \middle| X_{is}, A_i\right) dx \\ &= \int \partial_\theta \log f(x) \frac{\mathrm{pr}\left(A_i \middle| X_{is}, X_{ir} = x\right)}{\mathrm{pr}\left(A_i \middle| X_{is}\right)} f(x) dx \\ &= \frac{\int \partial_\theta f(x) E\left(\dfrac{1}{M\xi + X_{is} + x + V}\right) dx}{E\left(\dfrac{1}{M\xi + X_{is} + U}\right)}. \end{aligned} \tag{5}$$

But

$$\begin{aligned} \partial_\theta E\left(\frac{1}{M\xi + X_{is} + U}\right) &= \partial_\theta \int \cdots \int \frac{f(x_1)\cdots f(x_{M-1})}{M\xi + X_{is} + x_1 + \cdots x_{M-1}} dx_1 \cdots dx_{M-1} \\ &= (M-1) \int \cdots \int \frac{\left(\partial_\theta f(x_1)\right) f(x_2) \cdots f(x_{M-1})}{M\xi + X_{is} + x_1 + \cdots x_{M-1}} dx_1 \cdots dx_{M-1} \\ &= (M-1) \int \partial_\theta f(x) E\left(\frac{1}{M\xi + X_{is} + x + V}\right) dx\ . \end{aligned} \tag{6}$$

Substituting (5) into the MIP score function (4) and (6) into the sample-based score function (3), we see that these score functions are identical.

Thus application of the missing information principle, originally defined and developed outside the sampling context, and an approach based directly on getting the sample based density, yield the same maximum likelihood estimator in the informative sampling scheme we have described. This is not truly surprising of course, since MIP is just one way of doing maximum likelihood in certain circumstances, but it is perhaps illuminating to see how the two rather different approaches converge.

Example 1. Let X follow a gamma(θ_1, θ_2) distribution with density

$$f(x) = \frac{\theta_1^{\theta_2} x^{\theta_2 - 1} \exp(-\theta_1 x)}{\Gamma(\theta_2)} I(x > 0)$$

where I denotes the indicator function and $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are the parameters of interest. It can be seen that U is distributed as gamma(θ_1, (M-1)θ_2). Moreover

$$\partial_{\theta_1} E\left\{\frac{1}{M\xi + X_{is} + U}\right\} = E\left\{\frac{(M-1)\theta_2/\theta_1 - U}{M\xi + X_{is} + U}\right\}$$

$$\partial_{\theta_2} E\left\{\frac{1}{M\xi + X_{is} + U}\right\} = (M-1)E\left\{\frac{\log(\theta_1) + \log(X) - G(\theta_2)}{M\xi + X_{is} + X + V}\right\}$$

where G(x) = $\Gamma'(x)/\Gamma(x)$, V is the sum of M-2 iid realisations of X, X is independent of V and U = X + V. The sample-based score function for $\boldsymbol{\theta}$ is therefore the vector

$$\begin{pmatrix} nM\left[\dfrac{\theta_2}{\theta_1} + \xi - \dfrac{1}{nM}\displaystyle\sum_{i=1}^{n}\left(E\left\{\dfrac{1}{M\xi + X_{is} + U}\right\}\right)^{-1}\right] \\ n\left[M\log(\theta_1) + \overline{Y}_s - MG(\theta_2) + \dfrac{M-1}{n}\displaystyle\sum_{i=1}^{n}\left(E\left\{\dfrac{1}{M\xi + X_{is} + U}\right\}\right)^{-1} E\left\{\dfrac{\log(X)}{M\xi + X_{is} + U}\right\}\right] \end{pmatrix}$$

where $\overline{Y}_s$ denotes the mean of the logarithms of the X_{is} values. Monte Carlo simulation as well as numerical integration may be used to solve for the MLE for $\boldsymbol{\theta}$.

Although we do not show it here, the argument that shows the equivalence under array sampling of the sample-based and MIP-based approaches when X is a discrete-valued non-negative random variable proceeds on essentially the same lines as above.

Example 2: Suppose X is distributed as Poisson with parameter θ. Then the distribution of U is also Poisson, with parameter (M - 1)θ. Hence

$$\partial_\theta E\left\{\frac{1}{M\xi + X_{is} + U}\right\} = \frac{1}{\theta}E\left\{\frac{U-(M-1)\theta}{M\xi + X_{is} + U}\right\}$$

and so the sample-based score function for θ is

$$\begin{aligned}
sc_s(\theta) &= \frac{n}{\theta}\left[\overline{X}_s - \theta + \frac{1}{n}\sum_{i=1}^{n}\frac{E\left\{\frac{U}{M\xi + X_{is} + U}\right\}}{E\left\{\frac{1}{M\xi + X_{is} + U}\right\}} - (M-1)\theta\right] \\
&= \frac{n}{\theta}\left[\overline{X}_s - \theta + \frac{1}{n}\sum_{i=1}^{n}\frac{E\left\{1 - \frac{M\xi + X_{is}}{M\xi + X_{is} + U}\right\}}{E\left\{\frac{1}{M\xi + X_{is} + U}\right\}} - (M-1)\theta\right] \\
&= \frac{n}{\theta}\left[\overline{X}_s - \theta + \frac{1}{n}\sum_{i=1}^{n}\left(E\left\{\frac{1}{M\xi + X_{is} + U}\right\}\right)^{-1} - M\xi - \overline{X}_s - (M-1)\theta\right] \\
&= \frac{nM}{\theta}\left[\frac{1}{nM}\sum_{i=1}^{n}\left(E\left\{\frac{1}{M\xi + X_{is} + U}\right\}\right)^{-1} - \xi - \theta\right].
\end{aligned}$$

Setting this score to zero and solving for θ leads to the estimating equation for the MLE for this parameter

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\left(E\left\{\frac{1}{\xi + \frac{X_{is} + U}{M}}\right\}\right)^{-1} - \xi.$$

This is a nonlinear estimating equation, since the expectation term also depends on θ. Again, the solution may be found by numerical methods such as Monte Carlo simulation or numerical integration. See the simulation study in the next section.

3. Approximate MLE and Weighted Distribution Likelihood

Observe that for large M an approximation to (1) is

$$\mathrm{pr}\left(A_i = j \middle| X_{ij}\right) \approx \frac{\xi + X_{ij}}{M(\xi + \mu)} \tag{7}$$

where $\mu = E(X)$. Following the same line of development as the one that led to equation (2), but now introducing the approximation (7), we see that

$$\begin{aligned}\mathrm{pr}\left(X_{iA_i} \in \Delta_x\right) &= \sum_{j=1}^{M} \mathrm{pr}\left(A_i = j, X_{ij} \in \Delta_x\right) \\ &= \mathrm{pr}\left(X_{ij} \in \Delta_x\right) \sum_{j=1}^{M} \mathrm{pr}\left(A_i = j \middle| X_{ij} \in \Delta_x\right) \\ &\approx \mathrm{pr}\left(X_{ij} \in \Delta_x\right) \frac{(\xi + x)}{(\xi + \mu)}\end{aligned}$$

so the sample density of X under array sampling can be approximated by

$$\tilde{f}_s(x) = f(x) \frac{\xi + x}{\xi + \mu}. \tag{8}$$

This is the weighted distribution (WD) approximation to the actual sample distribution of X suggested by Krieger and Pfeffermann (1992). The approximate score function defined by it is

$$sc_{WD}(\theta) = \sum_{i=1}^{n} \frac{\partial_\theta f\left(X_{is}\right)}{f\left(X_{is}\right)} - n \frac{\partial_\theta \mu}{\xi + \mu}. \tag{9}$$

In what follows, we refer to the solution to (9) as the weighted distribution estimator.

To illustrate the use of (9) we again suppose that X is distributed as Poisson(θ). Then $\mu = \theta$ and (9) becomes

$$sc_{WD}(\theta) = \frac{n}{\theta}\left(\overline{X}_s - \theta - \frac{\theta}{\xi + \theta}\right).$$

The weighted distribution estimator for θ is the positive root of the estimating equation defined by setting this expression to zero and solving for θ. This root is given by

$$\tilde{\theta} = \frac{\overline{X}_s - \xi - 1 + \sqrt{(\overline{X}_s - \xi - 1)^2 + 4\overline{X}_s\xi}}{2}$$

which can be seen to be always less than the sample mean $\overline{X}_s$.

The approximation (8) which underlies the weighted distribution approach is based on the assumption that the average T_i/M for the i^{th} row of the array can be replaced by the unknown population mean μ without significant loss of efficiency. On the surface, this assumption seems innocuous; the values making up this row average are independent and identically distributed with common mean μ, and so standard asymptotic theory applies provided M is large. However, this substitution is not as straightforward as it seems. Comparing (2) and (8) we see that the actual approximation being made is

$$\frac{1}{\xi + \mu} \approx E\left(\frac{1}{\xi + \frac{x + U}{M}}\right).$$

If ξ is "small" relative to μ (the situation that will typically be of interest), this approximation should be reasonable provided M is large and x is "close" to μ. In other cases the approximation may not be very good and the weighted distribution estimator will be biased and/or inefficient. In particular, this will be the case where X is strongly skewed to large positive values and the overall sampling fraction is high.

To illustrate these points, we carried out a small simulation study, consisting of nine related experiments, in each of which we generated 500 arrays of size $n \times M$ from a Poisson(θ) distribution, and then sampled one unit per row. The sample size was fixed at $n = 20$; M was allowed to vary from experiment to experiment. In all experiments we took $\xi = 0.5$. In each experiment, the value of θ was fixed, and we calculated the empirical bias and root mean square error of three estimators of θ: the sample mean $\overline{X}_s$, the weighted distribution estimator $\tilde{\theta}_{wd}$, and the maximum likelihood estimator. The last was calculated iteratively using the last equation of Example 2, and we give it in three versions. The first, a "one step estimator" $\hat{\theta}_1$, was calculated by plugging in $\tilde{\theta}_{\mathrm{wd}}$ on the right side of the equation, using a series expansion of 1001 terms to calculate the expectation. Succeeding iterations plug in the result of the preceding iteration. We give the result of ten iterations $\hat{\theta}_{10}$, and of twenty, $\hat{\theta}_{20}$.

Table 1 shows the ratio of average value of the 500 estimates to the target θ and the ratio of their empirical mean square error to that of $\hat{\theta}_{20}$. Not surprisingly, the sample mean is biased high. The weighted distribution esimator is biased low, the more so the greater the skewness (small θ, large M). The downward bias can be confirmed theoretically; see the Appendix. It leads to higher mean square error for the weighted distribution estimator than for the maximum likelihood estimator, although the difference is not particularly serious except at small M.

Table 1. Simulation results for array sampling of Poisson(θ) random variables (500 simulations with n = 20, $\xi = 0.5$)

θ	M	$\overline{X}_s$	$\tilde{\theta}_{wd}$	$\hat{\theta}_1$	$\hat{\theta}_{10}$	$\hat{\theta}_{20}$
			Mean/θ			
1	3	1.42	0.81	0.90	1.02	1.02
1	5	1.50	0.87	0.91	1.00	1.00
1	10	1.58	0.94	0.95	0.99	1.00
2	3	1.25	0.86	0.91	0.99	0.99
2	5	1.31	0.92	0.94	0.99	1.00
2	10	1.36	0.97	0.97	0.99	1.00
3	3	1.12	0.93	0.96	0.99	0.99
3	5	1.15	0.97	0.97	0.99	1.00
3	10	1.16	0.98	0.98	0.99	0.99
			MSE/MSE($\hat{\theta}_{20}$)			
1	3	2.41	1.33	1.06	1.00	1
1	5	2.81	1.15	1.06	0.99	1
1	10	3.27	1.03	1.02	0.99	1
2	3	2.02	1.36	1.14	1.00	1
2	5	2.31	1.12	1.07	1.00	1
2	10	2.63	1.02	1.01	1.00	1
3	3	1.55	1.20	1.09	1.00	1
3	5	1.80	1.05	1.03	1.00	1
3	10	1.89	1.02	1.02	1.00	1

In some circumstances therefore, the weighted distribution estimator may be a convenient alternative to the maximum likelihood estimator. However, a general delineation of these circumstances is an open question. The following artificial example illustrates that it is not just a question of large M.

Example 3. Consider the estimating equation for a population of M normal random variables with variance known (or one could consider a simple random sample)

$$\sum_{j=1}^{M}\left(X_j - \mu\right) = 0\,,$$

which can be written

$$X_i + \sum_{j \neq i} X_j = M\mu .$$

Approximating the second term on left by (M-1)μ (as in (8)) gives the “approximate” estimating equation

$$X_i + (M-1)\mu = M\mu$$

yielding the estimator $\hat{\mu} = X_i$, which, although unbiased, is considerably less efficient than $\hat{\mu} = \overline{X}$.

4. A Generalisation to Unbiased Estimating Equations for the Population Mean under Array Sampling

Consider estimation of the population mean μ of X under array sampling. Given the expression (2) for the sample density of X, it is straightforward to show

$$E(X_{is}) = ME\left(X\left\{\frac{\xi + X}{M\xi + X + U}\right\}\right)$$

where X and U are independent and U is the sum of M-1 independent and identically distributed realisations of X. Hence an unbiased sample-based estimating equation for μ is

$$\sum_{i=1}^{n}\left\{X_{is} - ME\left(X\left\{\frac{\xi + X}{N\xi + X + U}\right\}\right)\right\} = 0. \qquad (10)$$

Given a model for the population distribution of X (which will depend on the unknown parameter θ), this equation can be solved for θ, and hence for $\mu = \mu(\theta)$, by Monte Carlo simulation of the independent random variables X and U.

A generalisation of the MIP-based approach to estimation of μ is based on the following unbiased population-based estimating equation for this parameter:

$$L(\mu) = \sum_{i=1}^{n}\sum_{j=1}^{M}\left(X_{ij} - \mu\right) = 0.$$

Let U_{ir} denote the sum of the nonsampled X-values in the i^{th} row of the population. The MIP-based approach replaces the preceding estimating equation by

$$
\begin{aligned}
L_s(\mu) &= E\left(L(\mu) \middle| \text{sample data}\right) \\
&= \sum_{i=1}^{n}\left\{(X_{is}-\mu)+E\left(U_{ir}-(M-1)\mu \middle| X_{is}, A_i\right)\right\} \\
&= \sum_{i=1}^{n}\left\{X_{is}+(M-1)\frac{E\left(\dfrac{X}{M\xi+X_{is}+U}\right)}{E\left(\dfrac{1}{M\xi+X_{is}+U}\right)}\right\}-Mn\mu \\
&= 0 \qquad\qquad (11)
\end{aligned}
$$

where the third step follows from (5) with $\partial_\theta \log f(X_{ir})$ replaced with U_{ir}. This equation can be solved using Monte Carlo simulation of the independent random variables X and V, with U = X + V. Note that since L(μ) coincides with the ML estimating equation under a Poisson model, the solution to (11) is the MIP estimate of $\mu = \theta$ and will be the same as the expression for the sample-based MLE of θ derived in Example 2.

The estimator of μ defined by the solution of (11) is consistent provided $E(L_s(\mu)) = 0$. This follows directly from the fact that since L(μ) is an unbiased estimating equation,

$$E\left(L_s(\mu)\right) = E\left(E\left(L(\mu) \middle| \text{sample data}\right)\right) = E\left(L(\mu)\right) = 0\,.$$

That is, a generalised MIP approach based on an unbiased population level estimating equation for μ also leads to an unbiased estimating equation for this parameter.

5. Extensions

An immediate (and straightforward) extension of the theory developed in the preceding sections is to “ragged array sampling” where the i^{th} “row” of the population array is of size M_i, i = 1, .., n. Of more interest, perhaps, is an extension to alternative methods of sampling within each “row”, and in particular to cases where we are unsure whether the sampling method is informative or not. One way of modelling this situation (within the array sampling framework) is to assume the availability of known constants ξ_{ij} which (suitably rescaled) represent our “best guess" about the probability that unit j in row i is selected, given the "characteristics" of the units in row i. However, there is the chance that these ξ_{ij} do not "capture" the true probability of selection, which may in fact be related to the actual X-values in the row. We can model this scenario by a straightforward extension of the array sampling probability model (1):

$$\mathrm{pr}\left(A_i = j \middle| \xi_{ik}, X_{ik}; k = 1,2,\cdots,M_i\right) = \frac{\xi_{ij} + \gamma X_{ij}}{T_i(\xi) + \gamma T_i(X)}. \qquad (12)$$

Here γ is an unknown nonnegative parameter measuring the extent of "informativeness" of the sampling method, $T_i(\xi)$ is the total of the ξ_{ij} in row i and $T_i(X)$ is the corresponding X-total. Exactly the same argument as that leading to (2) can then be used to show that the sample density for X for a "draw" from row i of the population is

$$f_{is}(x) = f(x)\left(T_i(\xi) + \gamma M_i x\right)E\left(\frac{1}{T_i(\xi) + \gamma(x + U_i)}\right) \qquad (13)$$

where U_i denotes the random variable corresponding to the sum of M_i-1 independent realisations of X. Independence of draws from different rows implies the score function for θ and γ is then defined by the component score functions

$$sc_s(\theta) = \sum_{i=1}^{n}\left\{\frac{\partial_\theta f(X_{is})}{f(X_{is})} + \frac{\partial_\theta E\left(\frac{1}{T_i(\xi) + \gamma(X_{is} + U_i)}\right)}{E\left(\frac{1}{T_i(\xi) + \gamma(X_{is} + U_i)}\right)}\right\}. \qquad (14)$$

and

$$sc_s(\gamma) = \sum_{i=1}^{n}\left\{\left(\gamma + \frac{\bar{\xi}_i}{X_{is}}\right)^{-1} + \frac{\partial_\gamma E\left(\frac{1}{T_i(\xi) + \gamma(X_{is} + U_i)}\right)}{E\left(\frac{1}{T_i(\xi) + \gamma(X_{is} + U_i)}\right)}\right\}. \qquad (15)$$

where $\bar{\xi}_i$ is the average of the ξ_{ij} in row i.

In theory, equations (14) and (15) can be solved numerically to get the MLE's for θ and γ, using, for example, Monte Carlo simulation and numerical differentiation to evaluate the second "offset" terms in these expressions. In any case, solving (14) for a range of values of γ should provide insight into the sensitivity of the MLE for θ to the extent of "informativeness" in the sample design.

Another relatively straightforward extension of array sampling is where the distribution of X varies from one row of the array to the next. Of course, if one is interested in a parameter θ which characterises the "overall" distribution of X, then a link between the "row specific" distributions of X and its overall distribution has to be established before likelihood inference about θ is possible. An example is where the rows are defined in terms of another variable Z which is correlated with X. This seems a reasonable way of modelling ordered systematic sampling, for instance, where the ordering is in terms of the values of Z.

The other major extension of the preceding theory is to situations where sample selection is not draw by draw, as in array sampling, but where the entire sample is selected without replacement at one draw. This situation can be modelled by a generalisation of (12). As usual we assume a population of N independent and identically distributed realisations of a non-negative random variable X, with a fixed sample size n drawn without replacement from this population. Let $\mathcal{S}$ denote the set of all $C_{N,n} = N![(N-n)!n!]^{-1}$ possible realisations of the sample label set s. We suppose that for each s in $\mathcal{S}$ we know a non-negative value ξ(s) such that the probability p(s) of selection of the sample label set s given the population values of X can be modelled as

$$p(s) = \frac{\xi(s) + \gamma T(s)}{\sum_{t \in S} \left\{ \xi(t) + \gamma T(t) \right\}} \qquad (16)$$

where T(s) denotes the total of the X values for the population units in s. When $\gamma > 0$ (16) corresponds to single draw informative sampling. Let K denote the total over $\mathcal{S}$ of ξ(s) and let T denote the population total of X. Then (16) can equivalently be written

$$p(s) = \frac{\xi(s) + \gamma T(s)}{K + \gamma T C_{N-1,n-1}}.$$

Let **x** denote a point in n-dimensional Euclidean space. Using the same argument as that leading to (2) we can show that the joint density of the sample X-values at **x** under the sampling scheme defined by (16) is

$$\begin{aligned} f_s(\mathbf{x}) &= \prod_{i=1}^{n} f(x_i) \sum_{s \in \mathcal{S}} E\left[\frac{\xi(s) + \gamma\Sigma(\mathbf{x})}{K + \gamma[\Sigma(\mathbf{x}) + U]C_{N-1,n-1}}\right] \\ &= \prod_{i=1}^{n} f(x_i)\left[K + \gamma\Sigma(\mathbf{x})C_{N-1,n-1}\right] E\left[\frac{1}{K + \gamma[\Sigma(\mathbf{x}) + U]C_{N-1,n-1}}\right] \\ &= \prod_{i=1}^{n} f(x_i) N\left[\overline{K} + \gamma\Sigma(\mathbf{x})\right] E\left[\frac{1}{\overline{K}N + \gamma[\Sigma(\mathbf{x}) + U]n}\right]. \end{aligned} \tag{17}$$

Here $\overline{K}$ is the average value of $\xi(s)$ over all $C_{N,n}$ samples in $\mathcal{S}$, $\Sigma(\mathbf{x})$ denotes the summation of the values in **x** and U is the sum of N – n independent and identically distributed realisations of X.

Maximum likelihood inference for a parameter θ of the marginal distribution of X follows directly from (17). Note that it is straightforward to combine this one draw sampling procedure with the array sampling concept to provide a very general model for informative sampling. For example, stratified informative sampling occurs where each row in the array is a stratum, and within each row a single draw selection procedure like (16) above is implemented.

Acknowledgements

This research was supported in part by the Texas Adanced Research Program (010366-0286-1999), the National Cancer Institute (CA 57030) and the Texas A&M Center of Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences (P30-E509106).

References

Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M. and Welsh, A. H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review*, **62**, 349-363.

Chambers, R. L., Dorfman, A. H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society* **B**, 397 - 411.

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

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and application. *Proc. 6th Berkeley Symp. Math. Statist.*, **1**, 697 - 715.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society B*, **24**, 482-491.

Appendix: Weighted Distribution Estimator Bias in the Poisson Case

Under array sampling, the score function for the Poisson case can be written

$$\mathrm{sc}_s(\theta) = \frac{nM}{\theta}\left[\frac{1}{n}\sum_{i=1}^{n}\left(E\left\{\frac{1}{a_i + B}\right\}\right)^{-1} - \alpha\right],$$

where $\alpha = \xi + \theta$, $a_i = \alpha + (X_{is} - \theta)/M$, $B = M^{-1}[U-(M-1)\theta]$, and the expectation operator conditions on the given value of X_{is}. We shall assume $\alpha > \theta > 1$. Noting that

$$\begin{aligned} E\left(\frac{1}{a_i + B}\right) &= E\left[\frac{1}{a_i}\left(1 - \frac{B}{a_i} + \frac{B^2}{a_i^2} - \frac{B^3}{a_i^3} + \frac{B^4}{a_i^4}\right)\right] + O_p\left(\frac{1}{M^3}\right) \\ &= \frac{1}{a_i}\left[1 + \frac{(M-1)\theta}{a_i^2 M^2} - \frac{\theta}{a_i^3 M^2} + \frac{3\theta^2}{a_i^4 M^2}\right] + O_p\left(\frac{1}{M^3}\right) \end{aligned}$$

we then have

$$\begin{aligned} \mathrm{sc}_s(\theta) &= \frac{nM}{\theta}\left[\frac{1}{n}\sum_{i=1}^{n} a_i\left(1 - \frac{(M-1)\theta}{a_i^2 M^2} + \frac{\theta}{a_i^3 M^2} - \frac{3\theta^2}{a_i^4 M^2} + \frac{\theta^2}{a_i^4 M^2}\right) + O_p\left(\frac{1}{M^3}\right) - \alpha\right] \\ &= \frac{nM}{\theta}\left[\frac{\overline{X}_s - \theta}{M} - \frac{(M-1)\theta}{M^2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}\right) + \left(\frac{\theta}{M^2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i^2}\right) - \frac{2\theta^2}{M^2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i^3}\right)\right) + O_p\left(\frac{1}{M^3}\right)\right] \\ &= \frac{n}{\theta}\left[\overline{X}_s - \theta - \frac{\theta}{\theta + \xi} + \theta\left(\frac{1}{\alpha} - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}\right) + \left(\frac{\theta}{n}\sum_{i=1}^{n}\frac{1}{a_i} + \frac{\theta}{n}\sum_{i=1}^{n}\frac{1}{a_i^2} - \frac{2\theta^2}{n}\sum_{i=1}^{n}\frac{1}{a_i^3}\right)\frac{1}{M} + O_p\left(\frac{1}{M^2}\right)\right] \\ &\approx \mathrm{sc}_{wd}(\theta) + \frac{n}{M\theta}\left[\frac{\theta(\overline{X}_s - \theta)}{\alpha^2} + \left(\frac{\theta}{\alpha} + \frac{\theta}{\alpha^2} - \frac{2\theta^2}{\alpha^3}\right)\right] \\ &> \mathrm{sc}_{wd}(\theta) + \frac{n}{M}\frac{(\overline{X}_s - \theta)}{\alpha^2} \end{aligned}$$

for $\alpha > 1$, since then $\theta/\alpha > \theta^2/\alpha^3$ and $\theta/\alpha^2 > \theta^2/\alpha^3$. Moreover, under array sampling $E(\overline{X}_s - \theta) > 0$. Thus, asymptotically,

$$E(\mathrm{sc}_{wd}(\theta)) < E(\mathrm{sc}_s(\theta)) - n\frac{E(\overline{X}_s) - \theta}{M\alpha^2} = -n\frac{E(\overline{X}_s) - \theta}{M\alpha^2} < 0.$$

By direct calculation one readily sees that $n^{-1}\partial_\theta[\mathrm{sc}_{wd}(\theta)] \to d < 0$ for some constant d. Hence we obtain $E(\tilde{\theta}_{wd}) - \theta < 0$ asymptotically and so the weighted distribution estimator is negatively biased up to order $O(M^{-1})$.