Empirical likelihood multiplicity adjusted estimator for multiple frame surveys

Ewa Kabzinska and Yves G. Berger

Abstract Multiple frame surveys are commonly used for a variety of reasons, including correcting for frame undercoverage, increasing the precision of estimators of population parameters for groups of interest, targeting rare populations and reducing survey costs. Several approximately design unbiased estimators have been proposed for inference from multiple frame surveys. [15] generalized most of the existing estimators as a class of Generalized Multiplicity-Adjusted Horvitz-Thompson Estimators. We develop an Empirical Likelihood approach to the Multiplicity-adjusted estimator. The proposed estimator allows for several multiplicity adjustments. It can handle auxiliary information and can be applied to a variety of parameters of interest expressed as unique solutions to estimating equations. Under certain sampling designs, Wilks-type confidence intervals [16] can be calculated without variance estimates.

Key words: dual frame surveys, estimating equations, design based inference, multiplicity adjusted estimator

1 Introduction

Using more than one sampling frame may improve the coverage of the target population, increase the precision of estimators or reduce sampling cost, especially when a single frame containing all population units is not available or expensive to sample from. For instance, mobile phone frames are increasingly used together with

Ewa Kabzinska
University of Southampton, University Road, Southampton SO17 1BJ, e-mail: ejk1g12@soton.ac.uk

Yves G. Berger
University of Southampton, University Road, Southampton SO17 1BJ e-mail: Y.G.Berger@soton.ac.uk

landlines in CATI research (e.g. [2]) in order to increase the coverage in surveys. Multiple frames are also used to oversample rare populations [6]. Inference from multiple frame surveys have attracted a lot of attention and several multiple frame estimators are available (see [10, 9, 15, 1] for a review).

[15] showed how most of the existing multiple frame estimators can be expressed in the form of the Generalized Multiplicity Adjusted Horwitz-Thompson (GMHT) estimator [11]. The idea of multiplicity estimation consists of pooling all the units selected from all the frames into one sample and finding adjustment factors which multiply the design weights in order to account for increased sampling probability of units which appear in more than one sampling frame. This approach can handle inference from multiple frame surveys. It can also be applied to other sampling designs. For example, the Generalized Weight Share estimator used to make inference from indirect sampling surveys [7] can be expressed as a GMHT estimator [15].

While there exist various parametric methods of estimation under multiple frame designs, design based non-parametric alternatives are not so common. Among those there is the Pseudo Empirical Likelihood approach [14]. We propose an Empirical Likelihood method which adopts the flexible multiplicity approach and can easily handle additional constrains such as benchmarking. Empirical Likelihood is particularly well suited for estimation of parameters which have a skewed distribution, as no assumptions about normality of the parameter of interest are made. It gives data-driven, range-preserving asymmetric confidence intervals which can be calculated without variance estimates. We follow the design based approach where sampling is the only source of randomness and the parameters are fixed quantities [12].

2 Empirical likelihood approach

Consider T sampling frames $Q_t, t >= 2$ which together cover the entire population U. T samples $(s_1, s_2, ...s_T)$ of sizes $(n_1, n_2, ...n_T)$ respectively, are selected independently, where s_t denotes the sample selected from frame Q_t. We assume that the units are selected with-replacement with unequal probabilities $\rho_{t;i}$, e.g. [5]. Let

$$\pi_{t;i} = n_t \rho_{t;i}. \tag{1}$$

Note that the sampling frames usually overlap. The extent of the overlap may be unknown.

Let s of size $n = \sum_{t=1}^{T} n_t$ be the collection of labels of all the units selected in all the T samples. If a unit is selected k times, its label appears k times in s.

Suppose that the values of two variables $\mathbf{y}$ and $\mathbf{x}$ are collected for every unit in the samples $s_1, s_2, ..., s_T$. The variable $\mathbf{y}$ is the variable of interest. The vector $\mathbf{x}$ contains auxiliary variables for which the population level parameter $\boldsymbol{\vartheta}_U$ is known. The parameter $\boldsymbol{\vartheta}_U$ is defined as the vector of the unique solutions of the population estimating equation:

$$\sum_{i\in U} \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_U) = \mathbf{0}\cdot \tag{2}$$

Let $\kappa_{t;i}$ be the frame inclusion indicator, which is equal to 1 if the frame Q_t contains the i-th unit and to 0 otherwise. We assume that for every sampled unit i, the value of the *multiplicity-adjusted inclusion probability* p_i,

$$p_i = \pi_{t;i}\alpha_{t;i}^{-1} \tag{3}$$

is known, where $\alpha_{t;i}$ are the multiplicity adjustment factors and are such that, for all $i \in s$,[15]

$$\sum_{t=1}^{T} \kappa_{t;i}\alpha_{t;i} = 1\cdot \tag{4}$$

The target parameter $\boldsymbol{\theta}_U$ is the unique solution of the population estimating equation

$$\sum_{i\in U} \mathbf{g}_i(\mathbf{v}_i, \boldsymbol{\theta}_U) = \mathbf{0}, \tag{5}$$

where $\mathbf{g}_i(\mathbf{v}_i, \boldsymbol{\theta}_U)$ is a defined function of $\boldsymbol{\theta}_U$ and $\mathbf{v}_i$, which is a subset of $\{\mathbf{x}, \mathbf{y}\}$.

Consider the following *joint empirical log-likelihood function*:

$$\ell(\mathbf{m}) = \sum_{t=1}^{T}\sum_{i\in s_t} \log(m_i)\cdot \tag{6}$$

Let $\boldsymbol{\theta}$ be a vector in the parameter space of $\boldsymbol{\theta}_U$. Consider the following set of constraints:

1. *Unknown parameter constraint*:

$$\sum_{t=1}^{T}\sum_{i\in s_t} m_i\mathbf{g}_i(\mathbf{v}_i, \boldsymbol{\theta}) = \mathbf{0}. \tag{7}$$

2. *Sample size constraint*:

$$\sum_{i\in s_t} m_i p_i = n_t, t = 1, 2, ..., T. \tag{8}$$

   The sample size constraint can be extended to allow for stratification using the method proposed by [3]. A separate constraint on each of the stratum sample sizes is then used. The sampling frames can be stratified differently.
3. *Known parameter constraint*: [13, 4, 8]

$$\sum_{t=1}^{T}\sum_{i\in s_t} m_i\mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_U) = \mathbf{0}\cdot \tag{9}$$

The above constraint may include domain-specific auxiliary variables. For example, constraints may be created around parameters known for sampling frames, rather than for the population, or parameters known for socio - demographic groups. Note that the constraint (9) may also include domain counts, in particular, frame sizes and the size of the overlapping domain. Alignment-type constraints on the overlapping domains can be seen as special cases of domain-level constraints.

3 Maximum empirical likelihood point estimator

Let $\hat{\mathbf{m}}^* = \{\widehat{m}_i^*(\boldsymbol{\theta}) : i \in s\}$ be the set of values which maximises the expression (6), for a given vector $\boldsymbol{\theta}$, under $m_i > 0$ and constraints (7) - (9). The *maximum empirical likelihood point estimator* of $\boldsymbol{\theta}_U$ is defined as the vector $\widehat{\boldsymbol{\theta}}$ which maximises the following function:

$$\ell(\hat{\mathbf{m}}^*, \boldsymbol{\theta}) = \sum_{i \in s} \log(\widehat{m}_i^*(\boldsymbol{\theta}))\cdot \tag{10}$$

If $\boldsymbol{\theta}_U$ is uniquely defined by the estimating equation (5), the estimator $\widehat{\boldsymbol{\theta}}$ is the unique solution of the sample estimating equation:

$$\hat{\mathbf{G}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{i \in s_t} \hat{m}_i \mathbf{g}_i(\mathbf{v}_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{11}$$

where the set $\{\widehat{m}_i : i \in s\}$ maximises the expression (6) under the constraints (8) - (9).

4 Empirical Likelihood confidence intervals

Let $\hat{\mathbf{m}} = \{\widehat{m}_i : i \in s\}$ be the values which maximise (6) under the constraints (8) - (9).

Consider the following empirical likelihood ratio statistic: [3]

$$r(\boldsymbol{\theta}, \hat{\mathbf{m}}, \hat{\mathbf{m}}^*) = 2\{\ell(\hat{\mathbf{m}}) - \ell(\hat{\mathbf{m}}^*, \boldsymbol{\theta})\}\cdot \tag{12}$$

Under some mild regularity conditions,

$$r(\boldsymbol{\theta}, \hat{\mathbf{m}}, \hat{\mathbf{m}}^*) = \hat{\mathbf{G}}(\boldsymbol{\theta})_{RG}^{\top} \hat{\mathbf{V}}_p[\hat{\mathbf{G}}(\boldsymbol{\theta})_{RG}]^{-1} \hat{\mathbf{G}}(\boldsymbol{\theta})_{RG} + o_p(1), \tag{13}$$

where $\hat{\mathbf{V}}_p[\hat{\mathbf{G}}(\boldsymbol{\theta})_{RG}]$ is a consistent estimator of the variance-covariance matrix of $\hat{\mathbf{G}}(\boldsymbol{\theta})_{RG}$ under high entropy sampling designs. Therefore, (12) follows a χ^2_d distribution asymptotically, with d being the dimension of $\boldsymbol{\theta}$. The empirical likelihood

ratio statistic can be used to construct confidence regions or confidence intervals for the parameter $\boldsymbol{\theta}$. A $1-\alpha$ confidence region for $\boldsymbol{\theta}$ is defined by the values $\boldsymbol{\theta}$ such that $r(\boldsymbol{\theta}, \hat{\mathbf{m}}, \hat{\mathbf{m}}^*) < \chi^2_{d;\alpha}$, where $\chi^2_{d;\alpha}$ is the upper α-quantile of the χ^2 distribution with d degrees of freedom.

5 Conclusion

We propose an Empirical Likelihood approach for finite population parameters in the multiple frame context. The estimator is based on the multiplicity adjustment principle [15, 14], and can accommodate various multiplicity adjustment factors. Additional benchmark constraints constructed around known population level parameters may be incorporated easily. In particular, constraints on the frame size and size of the overlapping domain can be included. Alignment type constraints can also be defined, i.s., the requirement that both frames produce the same point estimates for parameters of the overlapping domain can be made.

A wide class of parameters, expressed as solutions to population estimating equations, can be estimated through the proposed estimator. A single weight, which can be used for estimation of various parameters, is obtained for every unit. The weights are positive by definition.

Empirical likelihood confidence intervals for finite population parameters can be constructed based on the empirical likelihood ratio statistic (12).

References

1. Arcos, A., Molina, D., Rueda, M., Ranalli, M.: Frames2: A package for estimation in dual frame surveys. The R Journal. To be printed (2015)
2. Barr, M., van Ritten, J.J., Steel, D., Thackway, S.V.: Inclusion of mobile phone numbers into an ongoing population health survey in australia using an overlapping dual frame: description of methods, call outcomes and acceptance by staff and respondents (2012)
3. Berger, Y., Torres, O.D.L.R.: Empirical likelihood confidence intervals for complex sampling designs. Journal of the Royal Statistical Society: Series B (Statistical Methodology) pp. 1–23 (2016)
4. Chaudhuri, S., Handcock, M.S., Rendall, M.S.: Generalized linear models incorporating population level information: An empirical-likelihood-based approach. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **70**(2), pp. 311–328 (2008)
5. Hansen, M.H., Hurwitz, W.N.: On the theory of sampling from finite populations. The Annals of Mathematical Statistics **14**(4), pp. 333–362 (1943)
6. Kalton, G.: Methods for oversampling rare subpopulations in social surveys. Survey Methodology **35**(2), 125–141 (2009)
7. Lavallée, P.: Indirect sampling, vol. 7397. Springer Science & Business Media (2009)
8. Lesage, E.: The use of estimating equations to perform a calibration on complex parameters. Survey Methodology **37**(1), 103–108 (2011)
9. Lohr, S.: Recent developments in multiple frame surveys. cell **46**(42.2), 6 (2007)
10. Lohr, S.L., Rao, J.: Inference from dual frame surveys. Journal of the American Statistical Association **95**(449), 271–280 (2000)

11. Mecatti, F.: A single frame multiplicity estimator for multiple frame surveys. Survey Methodology **33**, 151–158 (2007)
12. Neyman, J.: On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society **97**(4), 558–625 (1938)
13. Owen, A.B.: Empirical likelihood for linear models. The Annals of Statistics **19**(4), 1725–1747 (1991)
14. Rao, J., Wu, C.: Pseudo–empirical likelihood inference for multiple frame surveys. Journal of the American Statistical Association **105**(492), 1494–1503 (2010)
15. Singh, A.C., Mecatti, F.: Generalized multiplicity-adjusted horvitz-thompson estimation as a unified approach to multiple frame surveys. Journal of official statistics **27**(4), 633 (2011)
16. Wilks, S.S.: Shortest average confidence intervals from large samples. The Annals of Mathematical Statistics **9**(3), 166–175 (1938)