Empirical likelihood confidence intervals in the presence of unit non-response

Yves G. Berger*
University of Southampton, Southampton, United Kingdom - y.g.berger@soton.ac.uk

Abstract

Suppose that the parameter of interest is a complex parameter defined as a solution of an estimating equation (Godambe, 1960). For example, this parameter could be a regression parameter, a quantile, a ratio or a mean. The aim is to estimate, test and construct a confidence interval for this parameter. We assume that the data are based on a stratified sample of units selected with unequal probabilities. Suppose that we have unit non-response according to a uniform response mechanism within cells; that is, we assume that any units can be missing with the same response probability within cells. We consider a reverse approach (Fay, 1991); that is, the response mechanism is the first phase and the second phase is a stratified unequal probability sampling design with negligible sampling fraction. We show how the empirical likelihood approach proposed by Berger and De La Riva Torres (2015) can be used to estimate complex parameters and to construct confidence intervals in the presence of unit non-response. The proposed estimator is based on response rates estimated within cells. The proposed empirical likelihood confidence interval takes into account of (i) the design, (ii) the response mechanism, (iii) the randomness of the estimator of the response rates and (iv) the population level information. This confidence interval does not rely directly on the normality of the point estimator and does not require variance estimation, linearisation, re-sampling and estimation of a design effect.

Keywords: design-based inference, estimating equations, empirical likelihood, stratification, unequal inclusion probabilities.

1. Introduction

Let U be a finite population of N units; where N is a fixed quantity which is not necessarily known. Suppose that the population parameter of interest θ_N is the unique solution of the following estimating equation (Godambe, 1960).

$$G(\theta) = 0, \quad \text{with} \quad G(\theta) = \sum_{i \in U} g_i(\theta);$$

where $g_i(\theta)$ is a function of θ and of the values of variables for unit i.

Let $s \subset U$ be a sample of sample of size n. Consider the following *empirical log-likelihood function*

$$\ell(m) = \log\left(\prod_{i \in s} m_i\right) = \sum_{i \in s} \log(m_i), \tag{1}$$

where $\prod_{i \in s}$ and $\sum_{i \in s}$ denote the product and the sum over the sampled units and the m_i are scale loads (Hartley and Rao, 1968). The m_i can be estimated by the values $\widehat{m}_i$ which maximise $\ell(m)$ subject to the constraints $m_i \geq 0$ and

$$\sum_{i \in s} m_i \boldsymbol{c}_i = \boldsymbol{C}; \tag{2}$$

where $\boldsymbol{c}_i$ is a $Q \times 1$ vector associated with the i-th sampled unit and $\boldsymbol{C}$ is a $Q \times 1$ vector. Berger and De La Riva Torres (2015) showed that the $\widehat{m}_i$ are given by

$$\widehat{m}_i = \left(\pi_i + \boldsymbol{\eta}^\top \boldsymbol{c}_i\right)^{-1}, \tag{3}$$

The quantity $\boldsymbol{\eta}$ is such that the constraint (2) holds.

The vectors $\boldsymbol{c}_i$ and $\boldsymbol{C}$ satisfied the regularity conditions given by Berger and De La Riva Torres (2015). These vectors contains the information about the design (inclusion probabilities and stratification) and the population level information (for more details see Berger and De La Riva Torres, 2015). For example, when $\boldsymbol{c}_i = Nn^{-1}\pi_i$ and $\boldsymbol{C} = N$, the $\widehat{m}_i$ are Horvitz and Thompson (1952) weights: $\widehat{m}_i = \pi_i^{-1}$.

Let the $\widehat{m}_i^\star(\theta)$ be the values which maximise (1) subject to the constraints $m_i \geq 0$ and

$$\sum_{i\in s} m_i \boldsymbol{c}_i^\star = \boldsymbol{C}^\star \tag{4}$$

with $\boldsymbol{c}_i^\star = (\boldsymbol{c}_i^\top, g_i(\theta))^\top$ and $\boldsymbol{C}^\star = (\boldsymbol{C}^\top, 0)^\top$, for a given θ. Let $\ell(\widehat{m}^\star, \theta) = \sum_{i\in s} \log(\widehat{m}_i^\star(\theta))$ be the maximum value of (1) under the constraint (4). The *empirical log-likelihood ratio function* is defined by the following function of θ.

$$\widehat{r}(\theta) \quad = \quad 2\left\{\ell(\widehat{m}) - \ell(\widehat{m}^\star, \theta)\right\}. \tag{5}$$

The *maximum empirical likelihood estimate* $\widehat{\theta}$ of θ_N is defined by the unique value of θ which minimises the function (5) (or maximise the empirical likelihood function $\ell(\widehat{m}^\star, \theta)$). Note that when $\boldsymbol{c}_i = Nn^{-1}\pi_i$, $\boldsymbol{C} = N$ and $g_i(\theta) = y_i - n^{-1}\theta\pi_i$, we have $\widehat{m}_i = \pi_i^{-1}$ and $\widehat{\theta}$ is the Horvitz and Thompson (1952) estimator $\widehat{\theta} = \sum_{i\in s} y_i\pi_i^{-1}$. When $g_i(\theta) = y_i - \theta N^{-1}$, $\widehat{\theta}$ is the Hájek (1971) ratio estimator $\widehat{\theta} = N(\sum_{j\in s}\pi_j^{-1})^{-1}\sum_{i\in s} y_i\pi_i^{-1}$.

Berger and De La Riva Torres (2015) showed that $\widehat{r}(\theta_N)$ follows asymptotically a chi-squared distribution with one degree of freedom. Hence a confidence interval for θ_N is given by

$$\left\{\theta \ : \ \widehat{r}(\theta) \leq \chi_1^2(\alpha)\right\}; \tag{6}$$

where $\chi_1^2(\alpha)$ is the upper α-quantile of the chi-squared distribution with one degree of freedom.

2. Unit Non-response

Consider a uniform response mechanism where all the units respond independently with the same response probability p_r. Let r_i be the response indicator: $r_i = 1$ if the unit i is a respondent and $r_i = 0$ otherwise. We propose to substitute $g_i(\theta)$ by $r_i g_i(\theta)$. The maximum empirical likelihood estimate $\widehat{\theta}$ of θ_N is still defined by the value of θ which minimises the function (5).

When the sampling fraction is negligible, it can be shown that the empirical log-likelihood ratio function $\widehat{r}(\theta_N)$ follows asymptotically a chi-squared distribution with one degree of freedom, under a two-phase reverse approach (Fay, 1991). The first phase is the response mechanism and the second phase is the sampling design. Hence, a confidence interval for θ_N is given by (6).

The uniform response assumption is often unrealistic in practice. It is common practice to form a finite number of re-weighting cells and to assume uniform response within cells. We will show how the proposed approach can be extended in this case.

3. Simulation

We give here the results of simulation study which could be found in Berger and De La Riva Torres (2015) for a single re-weighting cell and under a uniform response mechanism. The 1998-1999 British Family Expenditure Survey data were duplicated three times to create an artificial population of $N = 19{,}890$ households. Samples of size $n = 500$ are selected with unequal probabilities proportional to the first-order inclusion probabilities given in the dataset. The parameters of interest are q-quantiles of the equivalent total weekly household expenditure (Department of Social Security, 2001). The population level information is the numbers of individuals within age-sex groups (0-19, 20-39, 40-59, 60+). Non-responding households are generated

according to a uniform response mechanism with the average response rates 60%, 70% and 80%. We consider confidence intervals with the nominal level of 95%. The observed coverages and tail error rates are given in Table 1. Most of the coverages and tail error rates are not significantly different from 95% and 2.5%. With auxiliary variables, the upper tail error rates can be slightly larger than 2.5%.

5. Conclusions

Standard confidence intervals based on the central limit theorem require the normality of the point estimator and variance estimates which often involve linearisation or re-sampling. The proposed method does not rely on variance estimates, linearisation or re-sampling, even if the parameter of interest is not linear. The proposed method does not rely directly on the normality of the point estimator. Empirical likelihood confidence intervals can be easier to compute than standard confidence regions based on variance estimates. It provides a less computationally intensive alternative to bootstrap. The proposed approach naturally includes population level information and the information about the sampling design. Note that the proposed confidence interval takes into account of the effects of the nonresponse, the population level information and the sampling design.

Table 1: Empirical likelihood coverages (%) for quantiles Y_q. Unit nonresponse. FES data. *Source: Berger and De La Riva Torres (2015)*.

		Av. resp. rate = 60%			Av. resp. rate = 70%			Av. resp. rate = 80%		
	q	Overall	Lower	Upper	Overall	Lower	Upper	Overall	Lower	Upper
Without	10%	95.2	2.1*	2.7	95.0	2.6	2.4	94.8	2.4	2.8*
population level	25%	94.9	2.2	2.9*	94.8	2.4	2.8	94.8	2.5	2.6
information	50%	95.1	2.3	2.7	95.0	2.3	2.7	95.3	2.1*	2.5
	75%	94.5*	2.4	3.1*	94.8	2.2	2.9*	95.4	1.9*	2.7
With	10%	94.4*	2.8	2.8*	94.6	2.6	2.7	94.8	2.4	2.8*
population level	25%	94.5*	2.7	2.8	94.9	2.5	2.6	95.0	2.2	2.7
information	50%	94.9	2.2	2.9*	94.8	2.4	2.8*	94.5*	2.3	3.2*
	75%	94.4*	2.2	3.4*	94.6*	2.3	3.1*	94.6*	2.2	3.2*

$*$ Coverages (or tail error rates) significantly different from 95% (or 2.5%). p-value ≤ 0.05.

References

Berger, Y. G. and De La Riva Torres, O. (2015) An empirical likelihood approach for inference under complex sampling design. *To appear in the Journal of the Royal Statistical Society, Series B*, 22pp.

Department of Social Security (2001) Households below average income 1999/00, appendix 2. *London: Department of Social Security.*

Fay, B. E. (1991) A design-based perspective on missing data variance. *Proceeding of the 1191 Annual Research Conference. U.S. Bureau of the Census*, 429–440.

Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, **31**, pp. 1208–1211.

Hájek, J. (1971) Comment on a paper by D. Basu. in Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston.

Hartley, H. O. and Rao, J. N. K. (1968) A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.