An R Library to construct empirical likelihood confidence intervals for complex estimators

Yves G. Berger, University of Southampton, UK

Keywords: Calibration, Design-based approach, Estimating equations, Finite population corrections, Hajek estimator, Horvitz-Thompson estimator, Regression estimator, Stratification, Unequal inclusion probabilities.

We developed an R library which can be used to compute empirical likelihood point estimates and confidence intervals. After explaining the empirical likelihood theory, we show how to use this library and an example based on the 2009 EU-SILC survey.

1. INTRODUCTION

Under complex sampling designs, point estimators may not have a normal sampling distributionand linearised variance estimators may be biased. Hence standard confidence intervals based upon the central limit theorem may have poor coverages. We propose an empirical likelihood approach which gives design based confidence intervals. The proposed approach does not rely on the normality of the point estimator, variance estimates, design-effects, re-sampling, joint- inclusion probabilities and linearisation, even when the estimator of interest is not linear. It can be used to construct confidence intervals for a large class of complex sampling designs and complex estimators which are solution of an estimating equation [4]. It can be used for means, regressions coefficients, quantiles, totals or counts even when the population size is unknown. It can be used with large and negligible sampling fractions. It also provides asymptotically optimal point estimators, and naturally includes calibration constraints [2]. The proposed approach is computationally simpler than the pseudo empirical likelihood [9] and the bootstrap approaches [8]. Berger and De La Riva Torres [1] show that the empirical likelihood confidence interval may give better coverages than the approaches based on linearisation [3], bootstrap [8] and pseudo empirical likelihood [9].

2. EMPIRICAL LIKELIHOOD APPROACH

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_0 is the unique solution of the following estimating equation [4].

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

where $g_i(\theta)$ is a function of θ and of the characteristics of the unit i, such as the variables of interests and the auxiliary variables.

We propose to use the following *empirical log-likelihood function* [e.g. 1, 7].

$$\ell(m) = \sum_{i=1}^{n} \log(m_i),$$

where $\sum_{i=1}^{n}$ denotes the sum over the sampled units. The quantities m_i are unknown positive scale loads. The maximum likelihood estimators of m_i are the values $\widehat{m}_i$ which maximise $\ell(m)$ subject to the constraints $m_i \geq 0$ and

$$\sum_{i=1}^{n} m_i \boldsymbol{c}_i = \boldsymbol{C};$$

where $\boldsymbol{c}_i$ is $Q \times 1$ vector associated with the i-th sampled unit and $\boldsymbol{C}$ is $\mathrm{Q}\times 1$ vector (see Section 2.1). The values $\widehat{m}_i$ are survey weights.

2.1. Maximum empirical likelihood estimator

Suppose that the finite population U is stratified into H strata denoted by $U_1, \ldots, U_h, \ldots, U_H$; where $\cup_{h=1}^{H} U_h = U$. Suppose that a sample s_h of fixed size n_h is selected with replacement with unequal probabilities from U_h. Let $\boldsymbol{c}_i = \boldsymbol{z}_i$ and $\boldsymbol{C} = \boldsymbol{n}$; where $\boldsymbol{z}_i$ are the values of the design (or stratification) variables defined by $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})^{\top}$, where $\boldsymbol{n} = (n_1, \ldots, n_H)^{\top}$ denotes the vector of the strata sample sizes, with $z_{ih} = \pi_i$ when $i \in U_h$ and $z_{ih} = 0$ otherwise. It can be shown that $\widehat{m}_i = \pi_i^{-1}$. Let $\ell(\widehat{m}) = \sum_{i=1}^{n} \log(\widehat{m}_i)$ be the maximum value of the empirical log-likelihood function.

Let $\widehat{m}_i^{\star}(\theta)$ be the values which maximise $\ell(m)$ subject to the constraints $m_i \geq 0$ and $\sum_{i=1}^{n} m_i \boldsymbol{c}_i^{\star} = \boldsymbol{C}^{\star}$ 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 $\ell(\widehat{m}^{\star}, \theta) = \sum_{i=1}^{n} \log(\widehat{m}_i^{\star}(\theta))$. The *empirical log-likelihood ratio function* (or deviance) is defined by the following function of θ.

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

The *maximum empirical likelihood estimate* $\widehat{\theta}$ of θ_0 is defined by the value of θ which minimises the function $\widehat{r}(\theta)$. As the minimum value of $\widehat{r}(\theta)$ is zero, $\widehat{\theta}$ is the solution of $\widehat{r}(\theta) = 0$. It can be easily shown that this implies that $\widehat{\theta}$ is the solution of the following estimating equation.

$$\widehat{G}(\theta) = 0, \quad \text{with} \quad \widehat{G}(\theta) = \sum_{i=1}^{n} \widehat{m}_i \; g_i(\theta);$$

when $g_i(\theta) = y_i - n^{-1}\theta\pi_i$, we have $\widehat{G}(\theta) = \sum_{i=1}^{n} g_i(\theta)\pi_i^{-1}$ and $\widehat{\theta}$ is the Horvitz-Thompson estimator [6] given by $\widehat{Y}_{\pi} = \sum_{i=1}^{n} y_i \pi_i^{-1}$. When $gi = y_i - \theta N^{-1}$, $\widehat{\theta}$ is the Hajek [5] ratio estimator $\widehat{Y}_H = N\widehat{N}_{\pi}^{-1}\widehat{Y}_{\pi}$, where $\widehat{N}_{\pi} = \sum_{i=1}^{n} \pi_i^{-1}$.

2.2. Empirical likelihood confidence intervals

Berger and De La Riva Torres [1] show that the random variable $\widehat{r}(\theta_0)$ follows asymptotically a chi-squared distribution with one degree of freedom. Thus, the α level empirical likelihood confidence interval for the population parameter θ_0 is given by

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

Note that $\widehat{r}(\theta)$ is a convex non-symmetric function with a minimum when θ is the maximum empirical likelihood estimator. This interval can be found using a bisection search method. This involves calculating $\widehat{r}(\theta)$ for several values of θ. Berger and De La Riva Torres [1] showed how this approach can be used to accommodate large sampling fractions and non-response.

It is also possible to calibrate towards parameters more complex than totals. For example, we may want to calibrate with respect to population means, quantiles or variances. In this case, the calibration constraint is specified by the estimating equations $\sum_{i=1}^{n} m_i \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_0) = \mathbf{0}$; where $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_0)$ is a vector function of the auxiliary variables and of a known parameter $\boldsymbol{\vartheta}_0$ which is the solution of the following estimating equation

$$\sum_{i \in U} \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_0) = \mathbf{0}.$$

Calibration constraints are taken into account by including the auxiliary variables within the $\boldsymbol{c}_i$. In this case, we use $\boldsymbol{c}_i = (\boldsymbol{z}_i^\top, \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_0)^\top)^\top$ and $\boldsymbol{C} = (\boldsymbol{n}^\top, \mathbf{0}^\top)^\top$. For example, if we want to calibrate towards known population means $\boldsymbol{\vartheta}_0 = \boldsymbol{X} N^{-1}$, we need to use $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_0) = \boldsymbol{x}_i - \boldsymbol{\vartheta}_0$; where $\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i$ is a vector of known population totals. Simultaneous calibration on totals, means, proportion or any known parameter is also feasible.

3. AN R LIBRARY

In order to implement the approach describes in Section 2, we developed a library in R called *emplikfpop*. First, we need to specify the design and calibration variables. Secondly, we need to specify the parameter of interest (i.e. the definition of $g_i(\theta)$) and the finite population correction. This information is needed for point estimation and for confidence intervals.

i. *Design and auxiliary information*: This information is contained in the matrix `MatDA` and the vector `Var.Labels`. `MatDA` is a $n \times p$ matrix containing the stratification labels, the π_i and $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_0)$ (depending on $\boldsymbol{\vartheta}_0$). The vector `Var.Labels` gives the columns (in `MatDA`) of stratification and π_i.

ii. *Definition of the parameter of interest*: A function object `FunctG` defining the function $g_i(\theta)$. This function depends on a matrix `Data` and a vector `Vect.Const` which specify the data needed in the definition of $g_i(\theta)$. This function also depends on `Theta`, a given value for θ. This function has the following format:

$$\texttt{FunctG} = \texttt{function}(\texttt{Data}, \texttt{Vect.Const}, \texttt{Theta})\{\ldots\}$$

iii. *Finite population corrections*: The $n \times 1$ vector `fpc` contains the $1 - \pi_i$ or $1 - nN^{-1}$ or 1 (if the finite population correction is ignored).

iv. *Survey weights*: The survey weight $\widehat{m}_i$ are computed using the function
$> \texttt{Mi} = \texttt{Vect.Mi}(\texttt{MatDA}, \texttt{Var.Labels})$

v. *Point estimate*: The point estimate is computed using the following function
$> \texttt{Theta.Hat} = \texttt{SolEE}(\texttt{FunctG}, \texttt{Data}, \texttt{Vect.Const}, \texttt{Mi}, \texttt{Min}, \texttt{Max})$
where `Min` and `Max` specify the range of $\widehat{\theta}$.

vi. *Confidence interval*: The confidence interval is computed using the following function
$> \texttt{Bounds} = \texttt{ELBound}(\texttt{MatDA}, \texttt{Var.Labels}, \texttt{FunctG},$
$\texttt{Level}, \texttt{Data}, \texttt{Vect.Const}, \texttt{Theta.Hat}, \texttt{fpc})$
where `Level` is the level of the confidence interval. For example, for the 95% confidence interval, we use $\texttt{Level} = 0.95$. The object `Bounds` contains the bounds of the confidence interval. We have also predefined function for means, totals and quantiles: `ELBoundMean()`, `ELBoundTotal()` and `ELBoundQuantile()`. For the Rao-Hartley-Cochran sampling design, we have the function `ELBound.RHC()`.

4. AN APPLICATION TO THE EU-SILC HOUSEHOLD SURVEY

We use the 2009 EU-SILC user database to estimate the *persistent at-risk-of-poverty rate*. we adopted an ultimate cluster approach, where the units are the primary sampling units. In the table below, we have the point estimate and several confidence intervals for a couple of countries: the empirical likelihood confidence intervals, the standard confidence intervals based on variance estimates and the rescaled bootstrap confidences intervals [8]. Note that the bounds of the standard intervals are negative for Ireland, Austria, Malta and Denmark. The bootstrap bounds and the empirical likelihood bounds are larger than the bounds of the standard intervals. These differences are more pronounced for Austria, Malta, Denmark, the Netherlands, Estonia, Latvia and Greece. This is due to the skewness of the sampling distribution.

Table 1: Persistent at-risk-of-poverty rate & confidence intervals. 2009 EU-SILC.

	Rate	Emp. Likelihood		Standard		Rescaled Bootstrap	
Country	(%)	Lower	Upper	Lower	Upper	Lower	Upper
Ireland	0.53	0.08	1.76	-0.26	1.31	0.00	1.58
Austria	2.14	0.53	6.50	-0.52	4.80	0.14	5.26
Malta	2.90	0.97	7.75	-0.10	5.89	0.62	6.09
Denmark	3.46	1.09	8.95	-0.06	6.98	0.67	7.76
France	4.50	3.33	5.99	3.21	5.8	3.23	6.04
UK	5.18	2.56	9.90	1.78	8.57	2.15	8.85
Netherlands	5.22	1.88	11.66	0.69	9.75	1.31	10.25
Estonia	7.45	4.07	14.69	2.87	12.03	3.47	13.11
Poland	8.58	5.89	12.49	6.32	10.85	5.32	12.13
Latvia	10.34	6.09	17.36	5.05	15.63	5.36	15.27
Greece	11.34	7.51	18.32	6.72	15.96	7.03	16.95

REFERENCES

[1] Berger, Y. G. and De La Riva Torres, O. (2012) Empirical likelihood confidence intervals for complex sampling designs. S3RI Working paper, http://eprints.soton.ac.uk/337688/

[2] Deville, J. C. and Sarndal (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

[3] Deville, J. C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology, 25} 193-203.

[4] Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics, 31, 1208-1211.

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

[6] 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.

[7] Owen, A. B. (2001) Empirical Likelihood. New York: Chapman & Hall.

[8] Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992) Some recent work on resampling methods for complex surveys. Survey Methodology, 18, 209--217.

[9] Wu, C. and Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. The Canadian Journal of Statistics, 34, 359-375.