Empirical Likelihood Confidence Intervals under the Rao-Hartley-Cochran Sampling Design

Y.G. Berger*

Abstract

The Hartley-Rao-Cochran (RHC) sampling design (Rao *et al.*, 1962) is a popular unequal probability sampling design. We show how empirical likelihood confidence intervals can be derived under this sampling design. Berger and De La Riva Torres (2012) proposed an empirical likelihood approach which can be used for point estimation and to construct confidence intervals under complex sampling designs. We show how this approach can be adjusted for the RHC sampling design. The proposed approach intrinsically incorporates sampling weights and auxiliary information. It may give better coverages than standard methods even when the sampling distribution of the parameters of interest is not normal. The proposed approach is simple to implement and less computer intensive than bootstrap. The proposed approach does not rely on re-sampling, linearisation, variance estimation, or design-effects.

Key Words: Auxiliary information, Design-based approach, Estimating equations, Probability proportional to size sampling design, Regression estimator, Unequal inclusion probabilities.

1. Introduction

Complex estimators, such as quantiles, poverty indicators, M-estimators or parameters of population models are often computed from survey data. The sampling distribution of these estimators may not be normal when the distributions of the underlying variables are skewed. Furthermore, asymptotic linearised variances estimators can be biased in this situation. Therefore, standard confidence intervals based upon normality and variance estimates can have poor coverages. Their bounds can be also out of the range of the parameter space. For example, a lower bound can be negative even when the parameter of interest is positive. Empirical likelihood confidence intervals may have better coverages in this situation, as empirical likelihood confidence intervals are determined by the distribution of the data (e.g. Owen, 2001) and as the range of the parameter space is preserved.

Let U be a finite population of N units; where N denotes the population size. Consider that the population parameter of interest θ_N is the non-random quantities which is the unique solution of the following estimating equation (Godambe, 1960).

$$G(\theta) = 0, \quad \text{with } G(\theta) = \sum_{i \in U} g_i(\theta); \tag{1}$$

where $g_i(\theta)$ is a function of θ and of the values variable of interest and auxiliary variables for the unit i. For example, when $g_i(\theta) = y_i - \theta$, the parameter θ_N is population mean $\mu = N^{-1} \sum_{i \in U} y_i$; where the y_i are the values of a variable of interest. Other examples are ratios, low income measures, regression coefficients, M-estimators (e.g. Qin and Lawless, 1994; Binder and Kovacević, 1995). We consider that $g_i(\theta)$ and θ_N are scalars, although this paper approach can be extended when they are vectors. Note that the $g_i(\theta)$ do not need to be differentiable functions. The aim of this paper is to propose an estimator for θ_N and to derive a confidence interval for θ_N.

Suppose we have a sample s of size n selected with the uni-stage Hartley-Rao-Cochran (RHC) sampling design (Rao *et al.*, 1962) defined in § 2. The parameter θ_N will be estimated from the sampled data. We adopt a design-based approach which considers that the

*University of Southampton, UK, y.g.berger@soton.ac.uk

sampling distribution of the estimator is specified by the RHC sampling design and the values of the variables are fixed (non-random) quantities. Under this approach, the standard likelihood function is flat and cannot be used for inference (Godambe, 1966). Alternatively, empirical likelihood approaches can be used.

Hartley and Rao (1968) introduced the empirical likelihood-based approach. Owen (1988) brought this approach into the mainstream statistics (see also Owen, 2001). The empirical likelihood-based approach cannot be straightforwardly implemented under a design-based approach without some adjustments. Chen and Sitter (1999) proposed a pseudo empirical likelihood approach which can be used to construct confidence intervals (Wu and Rao, 2006). This approach consists in including the first-order inclusion probabilities within the empirical likelihood function and adjusting the empirical log-likelihood ratio function by a design effect which needs to be estimated. Berger and De La Riva Torres (2012) proposed a different empirical likelihood approach which consists in using the design constraints without adjusting the empirical likelihood function. Berger and De La Riva Torres (2012) showed that this approach can be used for point estimation and to construct confidence intervals under a class of high entropy sampling designs. The confidence interval proposed by Berger and De La Riva Torres (2012) does not rely on variance estimates or design effects. This approach cannot be straightforwardly implemented under RHC sampling, because the RHC sampling design does not belong to the of high entropy sampling designs. In this paper, we show how the approach proposed by Berger and De La Riva Torres (2012) can be adjusted to take into account of the RHC sampling design.

We suppose that we have a set of auxiliary variables $\boldsymbol{x}_i$ attached to unit i. We suppose that some population characteristics (denoted $\boldsymbol{\vartheta}_N$) of these variables are known at population level (see § 3.2). For example, these population characteristics can be totals, means, ratios or quantiles. We will show how these characteristics can be used for point estimation, and how it can be taken into account for constructing confidence intervals.

In § 2, we define the RHC sampling design. In § 3, we show how the parameter of interest can be estimated using empirical likelihood. In § 4, we introduce a penalised empirical log-likelihood ratio function which can be used under the RHC sampling design. We show how the penalised empirical log-likelihood ratio function can be used for testing and confidence intervals. In § 5, a simulation study supports our findings.

2. The RHC sampling design

The RHC sampling design is an unequal probability sampling design which does not belong to the class of high entropy sampling designs. Therefore the empirical likelihood approach proposed by Berger and De La Riva Torres (2012) cannot be directly implemented without some adjustments.

The RHC sampling design is a probability proportional to size design; that is a unit i is selected with probability proportional to a *measure of size* M_i. We consider that the M_i are standardised such that $\sum_{i\in U} M_i = 1$. Note that this design allows for large sampling fractions.

Suppose that the population is divided randomly into n disjoint groups $A_1, \dots, A_i, \dots, A_n$ of sizes $N_1, \dots, N_i, \dots, N_n$, where $\sum_{i=1}^{n} N_i = N$, where $\sum_{i=1}^{n}$ denote the sum over the sampled units. The N_i are fixed (non-random) quantities which are chosen before sampling. A sample of size n is obtained by selecting one unit independently from each group with the following probabilities:

$$p_i = \frac{M_i}{t_i}; \text{ where } t_i = \sum_{j\in A_i} M_j \cdot \tag{2}$$

Note that $\sum_{i \in U} p_i = n$. As far as weighting is concerned, the p_i play the same role as the first-order inclusion probabilities.

3. Empirical likelihood point estimator

Consider the following *empirical log-likelihood function* (Berger and De La Riva Torres, 2012).

$$\ell(m) = \sum_{i=1}^{n} \log(m_i). \tag{3}$$

The quantity m_i denotes the scale load of unit i (Hartley and Rao, 1968). As the units are selected independently, the empirical log-likelihood function is given by (3). Let $\{\widehat{m}_i : i \in s\}$ be the set of values which maximises $\ell(m)$ subject to the constraints $m_i \geq 0$ and

$$\sum_{i=1}^{n} m_i \boldsymbol{c}_i \quad = \quad \boldsymbol{C}; \tag{4}$$

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. The $\widehat{m}_i$ are empirical likelihood weights. We assume that $\boldsymbol{c}_i$ and $\boldsymbol{C}$ are such that the regularity conditions proposed by Berger and De La Riva Torres (2012) hold. The p_i are assumed to be contained within the $\boldsymbol{c}_i$; that is, we assume that the $\boldsymbol{c}_i$ and $\boldsymbol{C}$ are such that there exists a non random $Q \times 1$ vector $\boldsymbol{t}$ such that $\boldsymbol{t}^\top \boldsymbol{c}_i = p_i$ and $\boldsymbol{t}^\top \boldsymbol{C} = n$. Note that we do not impose that $\sum_{i=1}^{n} m_i = N$ always holds (except when $p_i = n/N$). If we want to impose that constraint, we need to consider an additional constraint $\sum_{i=1}^{n} m_i x_i = N$ with $x_i = 1$, and treat x_i as an auxiliary variable (see §§ 3.2 and 5.1).

Berger and De La Riva Torres (2012) showed that the minimisation of (3) under (4) has a unique solution given by

$$\widehat{m}_i = \Big(p_i + \boldsymbol{\eta}^\top \boldsymbol{c}_i\Big)^{-1}. \tag{5}$$

The quantity $\boldsymbol{\eta}$ is such that the constraint (4) holds. This quantity can be computed using an iterative Newton-Raphson 'type' procedure (Polyak, 1987).

The *maximum empirical likelihood estimate* $\widehat{\theta}$ of θ_N is defined by the unique solution of

$$\widehat{G}(\theta) = \sum_{i=1}^{n} \widehat{m}_i \, g_i(\theta) = 0; \tag{6}$$

where $\widehat{m}_i$ is defined by (5). Berger and De La Riva Torres (2012) showed that $\widehat{\theta}$ is also minimises an empirical log-likelihood ratio function.

3.1 Example without auxiliary information

Suppose that we ignore the auxiliary information. In this case, we use $\boldsymbol{c}_i = p_i$ and $\boldsymbol{C} = n$. It can be shown that $\widehat{m}_i = p_i^{-1}$ and (6) reduces to

$$\widehat{G}(\theta)_{RHC} = \sum_{i=1}^{n} \frac{g_i(\theta)}{p_i}. \tag{7}$$

which is the unbiased Rao *et al.* (1962) estimator of $G(\theta)$ for a given θ. The solution $\widehat{\theta}$ of $\widehat{G}(\theta)_{RHC} = 0$ is the maximum empirical likelihood point estimate for θ_N. When $g_i(\theta) = y_i - n^{-1} p_i \theta$, the solution of (7) is the Rao *et al.* (1962) estimate of a total. When $g_i(\theta) = y_i - \theta$, the solution is the ratio estimate of a mean.

3.2 Example with auxiliary information

Let $\boldsymbol{x}_i$ be a vector of values of auxiliary variables attached to unit i. Let $\boldsymbol{\vartheta}_N$ be some known population characteristics, of the auxiliary variables, which are considered to be the solution of the following estimating equation:

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

where $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta})$ denotes a vector of known function of $\boldsymbol{x}_i$ and $\boldsymbol{\vartheta}$. Suppose that the parameter $\boldsymbol{\vartheta}_N$ is known without sampling error. For example, this the case, when $\boldsymbol{\vartheta}_N$ described some known population quantities. For example, $\boldsymbol{\vartheta}_N$ is a vector of known population means when $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N) = \boldsymbol{x}_i - \boldsymbol{\vartheta}_N$.

The point estimator is the solution of (6) with $\boldsymbol{c}_i = (p_i, \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N)^\top)^\top$ and $\boldsymbol{C} = (n, \mathbf{0}^\top)^\top$. It can be shown that the resulting $\widehat{m}_i$ are such that $\sum_{i=1}^n \widehat{m}_i \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N) = \mathbf{0}$ holds. This implies that the maximum empirical likelihood estimator $\widehat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}_N$ is such that $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}_N$. In other words, the $\widehat{m}_i$ are empirical likelihood weights calibrated with respect to $\boldsymbol{\vartheta}_N$.

4. Penalised empirical log-likelihood ratio function

In § 4.3, we show how confidence intervals can be computed using the penalised empirical log-likelihood ratio function defined by (14). This function is based upon the following *penalised empirical log-likelihood function* (Berger and De La Riva Torres, 2012).

$$\widetilde{\ell}(m) = \log \left(\prod_{i=1}^{n} m_i \exp(1 - p_i m_i) \right) \cdot \tag{8}$$

Let $\{\widetilde{m}_i : i \in s\}$ be the set of values which maximises (8) subject to the constraints $m_i \geq 0$ and

$$\sum_{i=1}^{n} m_i \widetilde{\boldsymbol{c}}_i \;\; = \;\; \widetilde{\boldsymbol{C}}; \tag{9}$$

for some $\widetilde{\boldsymbol{c}}_i$ and $\widetilde{\boldsymbol{C}}$ defined in §§ 4.1 and 4.2. It can be shown that

$$\widetilde{m}_i = \left(p_i + \widetilde{\boldsymbol{\eta}}^\top \widetilde{\boldsymbol{c}}_i \right)^{-1}, \tag{10}$$

where $\widetilde{\boldsymbol{\eta}}$ is such that (9) holds. Note that $\widetilde{\boldsymbol{c}}_i$ and $\widetilde{\boldsymbol{C}}$ are different from $\boldsymbol{c}_i$ and $\boldsymbol{C}$. However, we will see in §§ 4.1 and 4.2 that the choice of $\widetilde{\boldsymbol{c}}_i$ and $\widetilde{\boldsymbol{C}}$ depends on $\boldsymbol{c}_i$ and $\boldsymbol{C}$.

4.1 Without auxiliary information

In § 3.1, we use $\boldsymbol{c}_i = p_i$ and $\boldsymbol{C} = n$ for point estimation. In this case, we consider $\widetilde{\boldsymbol{c}}_i = q_i^\circ p_i$ and $\widetilde{\boldsymbol{C}} = \sum_{i=1}^n q_i^\circ$, where $q_i^\circ = t_i^{1/2}$. Let $\{\widetilde{m}_i : i \in s\}$ be the set of values which maximises (8). In this case, $\widetilde{m}_i = p_i^{-1}$.

Let $\{\widetilde{m}_i^*(\theta) : i \in s\}$ be the set of values which maximises (8) (for a given θ) subject to the constraints $m_i \geq 0$ and

$$\sum_{i=1}^{n} m_i \widetilde{\boldsymbol{c}}_i^* \;\; = \;\; \widetilde{\boldsymbol{C}}^*; \tag{11}$$

with

$$\widetilde{\boldsymbol{c}}_i^* \;=\; (\widetilde{\boldsymbol{c}}_i \;,\; q_i^\bullet g_i(\theta))^\top \,, \tag{12}$$

$$\widetilde{\boldsymbol{C}}^* \;=\; \left(\widetilde{\boldsymbol{C}} \;,\; \sum_{i=1}^n (q_i^\bullet - 1)\frac{g_i(\theta)}{p_i}\right)^\top \,, \tag{13}$$

where $q_i^\bullet = \widehat{\varsigma}^{\,1/2}\; t_i^{-1/2}$; where $\widehat{\varsigma} = (\sum_{i=1}^n N_i^2 - N)(N^2 - \sum_{i=1}^n N_i^2)^{-1}$ is the finite population correction proposed by Rao *et al.* (1962, p. 485) and t_i is defined in (2). It can be shown that $\widetilde{m}_i^*(\theta) = (p_i + \widetilde{\boldsymbol{\eta}}^{*\top}\widetilde{\boldsymbol{c}}_i^*)^{-1}$, where $\widetilde{\boldsymbol{\eta}}^*$ is such that (11) holds.

Berger and De La Riva Torres (2012) proposed to use the following *penalised empirical log-likelihood ratio function* which is the following function of θ.

$$\widetilde{r}(\theta) \;=\; 2\left\{\widetilde{\ell}(\widetilde{m}) - \widetilde{\ell}(\widetilde{m}^*, \theta)\right\}, \tag{14}$$

where $\widetilde{\ell}(\widetilde{m})$ and $\widetilde{\ell}(\widetilde{m}^*, \theta)$ are given by (8) after substituting m_i by $\widetilde{m}_i$ and $\widetilde{m}_i^*(\theta)$ respectively; that is, $\widetilde{\ell}(\widetilde{m})$ and $\widetilde{\ell}(\widetilde{m}^*, \theta)$ are the maximum values of (8).

The following Theorem gives an asymptotic approximation for $\widetilde{r}(\theta_N)$.

Theorem 1 *We have that*

$$\widetilde{r}(\theta_N) = \frac{\widehat{G}(\theta_N)^2_{RHC}}{\widehat{var}[\widehat{G}(\theta_N)_{RHC}]} + o_p(1); \tag{15}$$

where $\widehat{G}(\theta_N)_{RHC}$ is defined by (7) and

$$\widehat{var}[\widehat{G}(\theta_N)_{RHC}] = \widehat{\varsigma}\left\{\sum_{i=1}^n t_i \frac{g_i(\theta_N)^2}{M_i^2} - \widehat{G}(\theta_N)^2_{RHC}\right\} \tag{16}$$

is the Rao et al. *(1962) variance estimator of $\widehat{G}(\theta_N)_{RHC}$.*

Proof of Theorem 1: As $\widehat{m}_i = p_i^{-1}$, we have that $\widetilde{\ell}(\widetilde{m}) = -\ell(p)$, where $\ell(p) = \sum_{i=1}^n \log(p_i)$. Using Lemma 3 in Berger and De La Riva Torres (2012), we have that

$$-2\{\widetilde{\ell}(\widetilde{m}^*, \theta_N) + \ell(p)\} = (\widetilde{\boldsymbol{C}}_p^* - \boldsymbol{C}^*)^\top \widetilde{\boldsymbol{\Sigma}}^{*-1}(\widetilde{\boldsymbol{C}}_p^* - \boldsymbol{C}^*) + o_p(1), \tag{17}$$

where $\boldsymbol{C}^*$ is defined by (13) and

$$\widetilde{\boldsymbol{C}}_p^* \;=\; \sum_{i=1}^n \frac{\widetilde{\boldsymbol{c}}_i^*}{p_i},$$

$$\widetilde{\boldsymbol{\Sigma}}^* \;=\; = \sum_{i=1}^n \frac{1}{p_i^2}\widetilde{\boldsymbol{c}}_i^* \widetilde{\boldsymbol{c}}_i^{*\top} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{pp} & \widehat{\boldsymbol{\Sigma}}_{pg} \\ \widehat{\boldsymbol{\Sigma}}_{pg}^\top & \widehat{\sigma}_{gg} \end{pmatrix};$$

where

$$\widehat{\boldsymbol{\Sigma}}_{pp} \;=\; \frac{N^2}{n^2}\sum_{i=1}^n q_i^{\circ 2} = \frac{N^2}{n^2}\sum_{i=1}^n t_i = \frac{N^2}{n^2},$$

$$\widehat{\boldsymbol{\Sigma}}_{pg} \;=\; \frac{N}{n}\sum_{i=1}^n q_i^\circ q_i^\bullet g_i(\theta_N) p_i^{-1} = \frac{N}{n}\,\widehat{\varsigma}^{\,1/2}\,\widehat{G}(\theta_N)_{RHC},$$

$$\widehat{\sigma}_{gg} \;=\; \sum_{i=1}^n q_i^{\bullet 2}\frac{g_i(\theta_N)^2}{p_i^2} = \widehat{\varsigma}\sum_{i=1}^n t_i \frac{g_i(\theta_N)^2}{p_i^2}.$$

We also have that

$$\widetilde{\boldsymbol{C}}_p^* - \boldsymbol{C}^* = \left(0, \widehat{G}(\theta_N)_{RHC}\right)^\top \cdot$$

Using $\widetilde{\ell}(\widetilde{m}) = -\ell(p)$, (14) and (17), we have

$$\begin{aligned}\widetilde{r}(\theta_N) &= \left(0, \widehat{G}(\theta_N)_{RHC}\right)\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{pp} & \widehat{\boldsymbol{\Sigma}}_{pg}\\ \widehat{\boldsymbol{\Sigma}}_{pg}^\top & \widehat{\sigma}_{gg}\end{pmatrix}^{-1}\begin{pmatrix}0\\ \widehat{G}(\theta_N)_{RHC}\end{pmatrix} + o_p(1),\\ &= \frac{\widehat{G}(\theta_N)^2_{RHC}}{\widehat{\sigma}_{gg} - \widehat{\boldsymbol{\Sigma}}_{pg}^\top \widehat{\boldsymbol{\Sigma}}_{pp}^{-1}\widehat{\boldsymbol{\Sigma}}_{pg}} + o_p(1)\cdot \qquad (18)\end{aligned}$$

It can be shown that $\widehat{\sigma}_{gg} - \widehat{\boldsymbol{\Sigma}}_{pg}^\top \widehat{\boldsymbol{\Sigma}}_{pp}^{-1}\widehat{\boldsymbol{\Sigma}}_{pg} = \widehat{var}[\widehat{G}(\theta_N)_{RHC}]$. Thus, (18) implies (15). The theorem follows.

□

Ohlsson (1986) proposed regularity conditions under which the Rao *et al.* (1962) estimator $\widehat{G}(\theta_N)_{RHC}$ is asymptotically normal. Assuming that these conditions holds for $\widehat{G}(\theta_N)_{RHC}$, Theorem 1 implies that $\widetilde{r}(\theta_N)$ follows asymptotically a chi-squared distribution with one degree of freedom, by the Slutsky's lemma.

4.2 With auxiliary information

For point estimation, we use $\boldsymbol{c}_i = (\boldsymbol{z}_i^\top, \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N)^\top)^\top$ and $\boldsymbol{C} = (\boldsymbol{n}^\top, \mathbf{0}^\top)^\top$ (see § 3.2). In this case, for $\widetilde{\ell}(\widetilde{m})$, we use

$$\begin{aligned}\widetilde{\boldsymbol{c}}_i &= \left(q_i^\circ p_i\ ,\ q_i^\bullet \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N)^\top\right)^\top,\\ \widetilde{\boldsymbol{C}} &= \left(\sum_{i=1}^n q_i^\circ p_i\ ,\ \sum_{i=1}^n (q_i^\bullet - 1)\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N)^\top p_i^{-1}\right)^\top,\end{aligned}$$

For $\widetilde{\ell}(\widetilde{m}^*, \theta)$, we use

$$\begin{aligned}\widetilde{\boldsymbol{c}}_i^* &= \left(\widetilde{\boldsymbol{c}}_i^\top, q_i^\bullet g_i(\theta)\right)^\top,\\ \widetilde{\boldsymbol{C}}^* &= \left(\widetilde{\boldsymbol{C}}^\top, \sum_{i=1}^n (q_i^\bullet - 1)\breve{g}_i(\theta)\right)^\top \cdot\end{aligned}$$

It can be shown that $\widetilde{r}(\theta_N)$ defined by (14) still follows asymptotically a chi-squared distribution with one degree of freedom.

4.3 Confidence intervals and hypotheses testing

Empirical likelihood confidence intervals rely on the asymptotic distribution of the pivot $\widetilde{r}(\theta_N)$. In the previous §, we show that $\widetilde{r}(\theta_N)$ follows asymptotically a chi-squared distribution. Thus, the α level empirical likelihood confidence interval (e.g. Wilks, 1938; Hudson, 1971) for the population parameter θ_N is given by

$$\left\{\theta\ :\ \widehat{r}(\theta) \le \chi_1^2(\alpha)\right\}; \qquad (19)$$

where $\chi_1^2(\alpha)$ is the upper α-quantile of the chi-squared distribution with one degree of freedom. Note that $\widehat{r}(\theta)$ is a convex non-symmetric function with a minimum when θ is the maximum empirical likelihood estimate $\widehat{\theta}$. This interval can be found using any root search method. In the simulation study, we used the Brent-Dekker method (Dekker, 1969, Brent, 1973, Ch. 4). This involves calculating $\widehat{r}(\theta)$ for several values of θ. Note that (19) will give a confidence intervals with the right coverage asymptotically even when $\widehat{\theta}$ is biased. This confidence interval will take into account of the auxiliary variable when the penalised empirical log-likelihood ratio function is computed as described in § 4.2.

The p-value of the test $H_0 : \theta_N = \theta_0$ is given by p-value $= \int_{\widehat{r}(\theta_0)}^{\infty} f(x)dx$, where $f(x)$ is the density of the chi-squared distribution with r degrees of freedom. This p-value is obtained from the statistical table of a chi-squared distribution.

5. Simulation study

In this §, we compare the Monte-Carlo performance of the proposed empirical likelihood 95% confidence interval with the linearisation (e.g. Deville, 1999), the pseudo empirical likelihood (Wu and Rao, 2006), the rescaled bootstrap (Rao *et al.*, 1992) and the Woodruff (1952) confidence intervals (in § 5.2). The bootstrap confidence intervals are based upon the quantiles of the set of 1000 bootstrap values (the histogram approach). The parameters of interest considered are population means (in § 5.1) and population quantiles (in § 5.2). The Rao *et al.* (1962) variance estimator is used for standard confidence intervals (linearisation) and for the pseudo empirical likelihood approaches. All the simulation studies are based on $10,000$ RHC samples of size $n = 500$ and the quantities N_i are given by $N_i = N/n$. We used the statistical software R (R Development Core Team, 2012). The algorithms were coded in C.

5.1 Estimation of means with auxiliary variables

Consider that the parameter of interest θ_N is the population mean; that is, $g_i(\theta) = y_i - \theta$. Suppose we have a vector $\boldsymbol{x}_i = (1, x_i)^\top$ of auxiliary variables for each unit i. Let μ_x be the population mean of the variable x_i. We suppose that the population means $\boldsymbol{\vartheta}_N = (1, \mu_x)^\top$ of these variables are known. In this case, $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\vartheta}_N) = \boldsymbol{x}_i - \boldsymbol{\vartheta}_N$. The standard confidence interval is based on the standard regression estimator defined by (6.4.2) in Särndal *et al.* (1992), with the p_i playing the role of first-order inclusion probabilities. The linearisation variance is used for the regression estimator. Note that the regression estimator, the pseudo empirical likelihood point estimators (pseudo-EL1 & pseudo-EL2) and the empirical likelihood point estimator are different.

We generate 80% of the values of y_i from a normal distribution with mean 8 and variance 1. The remaining 20% are outlying values generated from $y_i = 3 + a_i + \beta x_i + \varphi\, e_i$, where $\varphi = 1.5$. The variable a_i and x_i ($i \in U$) are generated from independent exponential distributions with rate parameters equal to 0.5. The M_i are proportional to $a_i + 2$. The values y_i, x_i and a_i generated are treated as fixed. Populations of size $N = 2000$ and $N = 25,000$ are generated.

The simulation results are given in Table 1. The values not within brackets are for the populations of size $N = 2000$ (large sampling fractions). The values within brackets are for the populations of size $N = 25,000$ (small sampling fractions). The ratio of average length is the average length of the confidence intervals divided by the average length of the confidence intervals based on linearisation. We measure the stability of the confidence intervals using the standard deviation of the lengths. The standard deviations are divided by the standard deviation of lengths of the linearisation confidence intervals. The column "Ratio MSE" gives the relative efficiency given by the ratio between the mean squared error (MSE) of the point estimator and the regression point estimator.

Table 1: Coverages of the 95% confidence intervals for the mean. $n = 500$. The values not within brackets for $N = 2000$ (large sampling fractions). The values within brackets for $N = 25,000$ (small sampling fractions). The symbol $*$ or $**$ indicate that the coverages (or tail error rates) significantly different from 95% (or 2.5%): $* \rightarrow 0.01 <$ p-value ≤ 0.05, $** \rightarrow$ p-value ≤ 0.01.

Approaches	Overall Cov. %	Lower tail err. rates %	Upper tail err. rates %	Ratio Av. Length	Ratio SD Length	Ratio MSE (Rel. Eff.)
Linear. (Reg. Est.)	95.1 (94.6)	2.6 (2.8)	2.3 (2.6)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
Bootstrap	94.8 (93.9**)	0.8** (1.1**)	4.4** (5.0**)	1.05 (1.01)	0.93 (1.01)	1.00 (1.00)
Pseudo-EL1	94.6 (95.4)	2.4 (2.7)	3.0** (1.9**)	0.51 (0.52)	0.45 (0.41)	0.50 (0.47)
Pseudo-EL2	93.1** (93.2**)	3.3** (4.0**)	3.5** (2.8*)	0.49 (0.47)	0.40 (0.37)	0.49 (0.47)
Emp. Lik.	94.8 (94.7)	2.4 (2.8*)	2.8 (2.5)	0.50 (0.49)	0.37 (0.37)	0.49 (0.47)

Table 2: Coverages for quantiles Y_q ($q = 5\%$ and 25%). $n = 500$. The values not within brackets for $N = 2000$ (large sampling fractions). The values within brackets for $N = 25,000$ (small sampling fractions). The symbol $*$ or $**$ indicate that the coverages (or tail error rates) significantly different from 95% (or 2.5%): $* \rightarrow 0.01 <$ p-value ≤ 0.05, $** \rightarrow$ p-value ≤ 0.01.

	$\rho(y, \pi)$	Approaches	Overall Cov. %	Lower tail err. rates %	Upper tail err. rates %	Ratio Av. Length	Ratio SD Length
$Y_{0.05}$	0.8	Linear.	99.3** (98.0**)	0.7** (1.8**)	0.0** (0.2**)	1.0 (1.0)	1.0 (1.0)
		Bootstrap	97.0** (95.1)	1.5** (2.3)	1.5** (2.6)	0.8 (0.8)	3.0 (2.2)
		Woodruff	95.1 (95.0)	2.1* (2.0**)	2.8 (3.0**)	0.7 (0.8)	2.8 (2.2)
		Emp. Lik.	94.5* (94.7)	2.0** (2.1*)	3.6** (3.2**)	0.7 (0.8)	2.8 (2.2)
	0.3	Linear.	98.9** (98.8**)	1.1** (1.1**)	0.0** (0.0**)	1.0 (1.0)	1.0 (1.0)
		Bootstrap	97.1** (95.3)	1.5** (2.2*)	1.5** (2.5)	0.7 (0.7)	2.6 (2.2)
		Woodruff	95.3 (95.4)	2.0** (1.7**)	2.8 (2.9**)	0.6 (0.7)	2.6 (2.2)
		Emp. Lik.	94.9 (94.8)	1.8** (2.0**)	3.2** (3.1**)	0.6 (0.7)	2.5 (2.2)
$Y_{0.25}$	0.8	Linear.	94.2** (95.1)	2.4 (2.1*)	3.5** (2.7)	1.0 (1.0)	1.0 (1.0)
		Bootstrap	97.1** (95.0)	1.4** (2.2)	1.4** (2.7)	1.1 (1.0)	3.6 (2.3)
		Woodruff	95.1 (94.9)	2.6 (2.5)	2.3 (2.6)	1.0 (1.0)	3.4 (2.2)
		Emp. Lik.	95.1 (95.0)	2.3 (2.2)	2.6 (2.8)	1.0 (1.0)	3.4 (2.2)
	0.3	Linear.	97.4** (97.2**)	1.8** (1.4**)	0.8** (1.4**)	1.0 (1.0)	1.0 (1.0)
		Bootstrap	97.2** (95.4)	1.2** (2.3)	1.5** (2.4)	1.0 (0.9)	3.3 (2.5)
		Woodruff	95.1 (95.3)	2.3 (2.5)	2.6 (2.2*)	0.9 (0.9)	3.1 (2.5)
		Emp. Lik.	94.9 (95.3)	2.0** (2.3)	3.1** (2.5)	0.9 (0.9)	3.1 (2.4)

The proposed empirical likelihood approach gives coverages which are not significantly different from 95%. Linearisation has also good coverages, but the proposed empirical likelihood approach gives shorter and more stable confidence intervals. From the last column, we notice that the MSE of the empirical likelihood point estimator is about 50% lower than the MSE of the regression estimator. The pseudo empirical likelihood estimators have similar MSE. With small sampling fraction ($N = 25,000$), the proposed empirical likelihood approach and the pseudo-EL1 approach give similar coverages, but the proposed confidence intervals are slightly shorter and more stable. The bootstrap and the pseudo-EL2 approaches give coverages and tail error rates which may be significantly different from 95% and 2.5%.

5.2 Estimation of quantiles

We consider the 5% and 25% quantiles: $Y_{0.05}$ and $Y_{0.25}$. We use the $g_i(\theta)$ proposed by Berger and De La Riva Torres (2012). The standard confidence interval is based on the linearised variance proposed by Deville (1999).

We generated several skewed population data using $y_i = 3 + a_i + \varphi\, e_i$ (Wu and Rao, 2006); where the a_i follows an exponential distribution with rate parameters equal to 1 and $e_i \sim \chi_1^2 - 1$. The M_i are proportional to $a_i + 2$. Populations of size $N = 2000$ and $N = 25000$ are generated. The parameter φ is used to specify the correlation $\rho(y, M)$ between the values y_i and M_i: $\rho(y, M) = 0.8$ with $\varphi = 0.5$; $\rho(y, M) = 0.3$ with $\varphi = 2.3$.

The coverages and tail error rates of the linearised confidence intervals are significantly different from 95% and 2.5% respectively, except with $Y_{0.25}$, $N = 25,000$ and a correlation of 0.8. The rescaled bootstrap gives acceptable coverages for small sampling fractions. However, for large sampling fraction, the coverages and tail error rates are significantly different from 95% and 2.5% respectively. This is not surprising, as rescaled bootstrap is design for small sampling fractions. The bootstrap confidence intervals have more unstable confidence intervals (see last column of Table 2) because of re-sampling. Linearisation gives the most stable confidence intervals, but with coverages significantly higher than 95%.

The Woodruff (1952) confidence intervals gives good coverages and tail error rates in most situations. We notice that the tail error rates of $Y_{0.05}$ are significantly different from 2.5%. We observe similar coverages and average lengths with the proposed empirical likelihood approach and the Woodruff (1952) approach.

6. Conclusion and discussion

Standard confidence intervals based on the central limit theorem and pseudo empirical likelihood confidence intervals require variance estimates which may involve linearisation or re-sampling. Even if the parameter of interest is not linear, the proposed confidence interval does not rely on normality of the point estimator, variance estimates, linearisation and re-sampling. Our simulation study shows that the coverage of standard confidence intervals can be poor with skewed variables.

The proposed approach is simpler to implement and less computationally intensive than bootstrap, especially with calibration weights. Our simulations study also shows that bootstrap confidence intervals may not have the right coverage and may be more unstable.

There is an analogy between the proposed empirical likelihood approach and calibration (e.g. Huang and Fuller, 1978; Deville and Särndal, 1992), as the function (3) can be viewed as a calibration objective function. The objective functions used for calibration are disconnected from mainstream likelihood statistical theory. However, the proposed objective function (3) is related to the concept of likelihood. The advantage of the proposed empirical likelihood approach over standard calibration is the fact that (3) can be used to make inference and construct confidence intervals. Furthermore, empirical likelihood weights are always calibrated and positive.

References

Berger, Y. G. and De La Riva Torres, O. (2012) *An empirical likelihood approach for inference under complex sampling design*. Southampton: Southampton Statistical Sciences Research Institute `http://eprints.soton.ac.uk/337688`.

Binder, D. A. and Kovacević, M. S. (1995) Estimating some measure of income inequality from survey data: an application of the estimating equation approach. *Survey Methodology*, **21**, 137–145.

Brent, R. P. (1973) *Algorithms for Minimization without Derivatives*. Prentice-Hall ISBN 0-13-022335-2.

Chen, J. and Sitter, R. R. (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, **9**, 385–406.

Dekker, T. J. (1969) Finding a zero by means of successive linear interpolation. *Constructive Aspects of the Fundamental Theorem of Algebra: Dejon, B.; Henrici, P.(editors). ondon: Wiley-Interscience*.

Deville, J. C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203.

Deville, J. C. and Särndal, C. E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Godambe, V. (1966) A new approach to sampling from finite population i, ii. *Journal of the Royal Statistical Society, series B*, **28**, 310–328.

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

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

Huang, E. T. and Fuller, W. A. (1978) Nonnegative regression estimation for survey data. *Proceedings Social Statistics Section American Statistical Association*, 300–303.

Hudson, D. J. (1971) Interval estimation from the likelihood function. *Journal of the Royal Statistical Society*, **33**, 256–262.

Ohlsson, E. (1986) Normality of the Rao, Hartley, Cochran estimator: An application of the martingale CLT. *Scandinavian Journal of Statistics*, **13**, 17–28.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

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

Polyak, B. T. (1987) *Introduction to Optimization*. New York: Optimization Software, Inc., Publications Division.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, pp. 300–325.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. `http://www.R-project.org`, Vienna, Austria. ISBN 3-900051-07-0.

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. Series B (Methodological)*, **24**, pp. 482–491.

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.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Wilks, S. S. (1938) Shortest average confidence intervals from large samples. *The Annals of Mathematical Statistics*, **9**, 166–175.

Woodruff, R. S. (1952) Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, **47**, 635–646.

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.