Empirical Likelihood Inference for the Rao-Hartley-Cochran Sampling Design

YVES G. BERGER

Southampton Statistical Sciences Research Institute, University of Southampton

Abstract

The Hartley-Rao-Cochran (RHC) sampling design is an unequal probability sampling design which can be used to select samples from finite populations. We propose to adjust the empirical likelihood approach for the RHC sampling design. The approach proposed intrinsically incorporates sampling weights, auxiliary information and allows for large sampling fractions. It can be used to construct confidence intervals. In a simulation study, we show that the coverage may be better for the empirical likelihood confidence interval than for standard confidence intervals based on variance estimates. The approach proposed is simple to implement and less computer intensive than bootstrap. The confidence interval proposed does not rely on re-sampling, linearisation, variance estimation, design-effects or joint inclusion probabilities.

Key Words: auxiliary information, confidence intervals, design-based approach, estimating equations, regression estimator.

Running Headline: Empirical Likelihood Inference for RHC Design

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 estimators may not be normal when the distributions of the underlying variables are skewed or contain outlying values. Furthermore, asymptotic linearised variances estimators may also be biased for moderate sample sizes. Therefore, standard confidence intervals based upon normality and variance estimates can have poor coverages. The bounds can be also out of the range of the parameter space. For example, the 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 parameters' 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 defined as the unique solution of the following population 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 of a set of 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 & Lawless, 1994; Binder & Kovacević, 1995). We consider that $g_i(\theta)$ and θ_N are scalars, although this paper's approach can be extended when they are vectors. The approach proposed does not require the $g_i(\theta)$ to be differentiable. This is not the case for linearisation (Binder, 1983). 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. We shall extend the approach proposed for two-stage sampling in § 5.3. The parameter θ_N shall be estimated from sample data. We adopt a design-based; that is, the 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 & Rao (1968) introduced the empirical likelihood-based approach. Owen (1988) developed this approach for mainstream statistics (see also Owen, 2001). The empirical likelihood-based approach cannot be straightforwardly implemented under a design-based approach without some adjustments. Chen & Sitter (1999) proposed a pseudoempirical likelihood approach which can be used to construct confidence intervals (Wu & 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 & De La Riva Torres (2016) proposed a different empirical likelihood approach which consists in using the design constraints without adjusting the empirical likelihood function. Berger & De La Riva Torres (2016) showed that this approach can be used for point estimation and to construct confidence intervals under a class of high entropy sampling designs. This approach cannot be straightforwardly implemented under RHC sampling, because the RHC sampling design does not belong to the class of high entropy sampling designs. In this paper, we show how the approach proposed by Berger & De La Riva Torres (2016) can be adjusted to take into account of the RHC sampling design.

In §§ 5.2 and 5.3, we show that standard confidence interval based on linearised variance estimators may produce confidence intervals with poor coverages. In § 5.1, we show that the pseudoempirical likelihood approach, which requires design effects, may give confidence interval with a coverage (and tail error rates) significantly different than the nominal level. The empirical likelihood approach proposed gives better coverage and tail error rates and does not rely on design effects. In other words, even if variance estimates and design effect are available, they do not guarantee that the standard and pseudoempirical likelihood confidence intervals have the correct coverage and/or tail error rates (see Tables 1, 2 and 3 in § 5). Furthermore, design effects are limited to means and totals. For example, in the pseudoempirical likelihood literature, there is no clear definition of the design effect that should be used for quantiles. Chen & Wu (2002) proposed to use a Woodruff's (1952) approach.

Chen & Kim (2014) proposed a population empirical likelihood approach. They showed that the population empirical log-likelihood ratio function follows a χ^2-distribution asymptotically under Poisson sampling when the sampling fraction is negligible. The RHC design is different from the Poisson sampling design. The empirical log-likelihood ratio function that is proposed in this paper follows a χ^2-distribution asymptotically even when the sampling fraction is large.

We suppose that we have a set of auxiliary variables $\boldsymbol{x}_i$ attached to unit i. We suppose that some population characteristics (denoted by the vector $\boldsymbol{\varphi}_N$) of these variables are known at population level (see § 3.2). For example, these population characteristics can be known population totals, means, ratios, proportions or quantiles. We will show how these characteristics can be used for point estimation, and how it can be taken into account when 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. In § 5.3, we show how the approach proposed can be extended for two-stage RHC designs.

2. The RHC sampling design

The RHC design can be used to select units without replacement with probabilities proportional to a size variable. This sampling design has several attractive properties: (i) it is easy to implement; (ii) it is always more efficient than with replacement sampling (Rao, 2005); (iii) the RHC point estimator for a total is unbiased; (iv) the variance can be easily estimated without the need of joint-inclusion probabilities, even when the sampling fraction is large; (v) variance estimates are always positive; (vi) the selection probabilities can be easily updated using the Keyfitz's (1951) method (e.g. Laniel & Mohl, 1994; Statistics Canada, 2008).

The Canadian Labour Force Survey (Statistics Canada, 2008) uses the RHC design to select primary sampling units (e.g. Laniel & Mohl, 1994; Statistics Canada, 2008; Rao, 2005). The RHC design is used in forestry because of its simplicity (Schabenberger & Gregoire, 1994; Rao, 2005). It is also use in audit sampling (Rao, 2005). Chaudhuri *et al.* (2004) implemented an adaptive RHC design to select small scale industries in India.

The RHC sampling design does not belong to the class of high entropy sampling designs. Therefore the empirical likelihood approach proposed by Berger & De La Riva Torres (2016) 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 $A_1, \dots,$ $A_g, \dots, A_n$ of sizes $N_1, \dots, N_g, \dots, N_n$, where $\sum_{g=1}^{n} N_g = N$. The N_g 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 according to the following probabilities:

$$p_i = \frac{M_i}{t_g}; \quad \text{where } t_g = \sum_{j\in A_g} M_j \quad \text{and } i \in A_g. \tag{2}$$

Note that $\sum_{i\in U} p_i = n$. The quantities p_i play the same role as the first-order inclusion probabilities, despite the fact that the p_i are different from the first-order inclusion probabilities.

3. Empirical likelihood point estimator

Consider the following *empirical log-likelihood function* (Berger & De La Riva Torres, 2016).

$$\ell(\boldsymbol{m}) = \sum_{i \in S} \log(m_i), \tag{3}$$

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

$$\sum_{i \in S} m_i \boldsymbol{c}_i = \boldsymbol{C}; \tag{4}$$

where $\boldsymbol{c}_i$ is a $Q \times 1$ vector associated with the i-th sampled unit and $\boldsymbol{C} = \sum_{i \in U} \boldsymbol{c}_i$. The $\widehat{m}_i$ are empirical likelihood weights. The $\boldsymbol{c}_i$, defined in §§ 3.1 and 3.2, are function of the p_i and of the auxiliary information.

We assume that the $\boldsymbol{C}$ is an inner point of the conical hull formed by $\{\boldsymbol{c}_i : i \in s\}$ so that the solution $\{\widehat{m}_i : i \in s\}$ is unique. We assume that $\boldsymbol{c}_i$ and $\boldsymbol{C}$ are such that the regularity conditions (A.1)-(A.6) proposed by Berger & De La Riva Torres (2016) hold. These conditions are given in the Appendix. The p_i are assumed to be incorporated within the $\boldsymbol{c}_i$; that is, we assume that the vectors $\boldsymbol{c}_i$ are such that there exists a non random $Q \times 1$ vector $\boldsymbol{t}$ such that $\boldsymbol{t}^\top \boldsymbol{c}_i = p_i$. This implies that the constraint

$$\sum_{i \in S} m_i \, p_i = n \tag{5}$$

always holds.

The constraint

$$\sum_{i \in S} m_i = N \tag{6}$$

is the leading constraint for the classical empirical likelihood approach based on simple random sampling (Chen & Qin, 1993). The constraint (5) is a generalised leading constraint under unequal probability sampling. Note that the constraint (5) reduces to the constraint (6), under equal probabilities, as $p_i = n/N$ in this case. Note that we do not impose that the constraint (6) always holds (except when $p_i = n/N$). In other words, the constraint (6) may or may not hold. If we wish to use the constraint (6), we need to consider the additional constraint $\sum_{i \in S} m_i(x_i - p_i N n^{-1}) = 0$ with $x_i = 1$, and treat x_i as an auxiliary variable (see (10) in §§ 3.2). This last constraint is equivalent to the constraint (6), because of the constraint (5).

The minimisation of (3) under (4) has a unique solution given by

$$\widehat{m}_i = \left(p_i + \boldsymbol{\eta}^\top \boldsymbol{c}_i\right)^{-1} \tag{7}$$

(see Berger & De La Riva Torres, 2016). The quantity $\boldsymbol{\eta}$ is such that the constraint (4) holds. This quantity can be computed using an iterative modified Newton-Raphson procedure (Polyak, 1987) as in Chen *et al.* (2002).

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

$$\widehat{G}(\theta) = \sum_{i \in S} \widehat{m}_i \, g_i(\theta) = 0; \tag{8}$$

where $\widehat{m}_i$ is defined by (7). Berger & De La Riva Torres (2016) showed that $\widehat{\theta}$ also minimises an empirical log-likelihood ratio function.

3.1 Without auxiliary information

Suppose that we ignore the auxiliary information. In this case, we use $\boldsymbol{c}_i = Nn^{-1}p_i$ and $\boldsymbol{C} = N$. It can be shown that $\widehat{m}_i = p_i^{-1}$ and (8) reduces to

$$\widehat{G}(\theta)_{RHC} = \sum_{i \in S} \frac{g_i(\theta)}{p_i} \cdot \tag{9}$$

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 (9) 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 With auxiliary information

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

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

where $\boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi})$ denotes a vector of known function of $\boldsymbol{x}_i$ and $\boldsymbol{\varphi}$ (e.g. Owen, 1991; Chaudhuri *et al.*, 2008; Lesage, 2011). We suppose that the parameter $\boldsymbol{\varphi}_N$ is a vector of population quantities known without sampling errors. For example, $\boldsymbol{\varphi}_N$ is a vector of known population means when $\boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}) = \boldsymbol{x}_i - \boldsymbol{\varphi}$. The vector $\boldsymbol{\varphi}_N$ may also contain a combination of means, ratios, total and/or quantiles.

The point estimator is the solution of (8) with $\boldsymbol{c}_i = (Nn^{-1}p_i, \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)^\top)^\top$ and $\boldsymbol{C} = (N, \boldsymbol{0}^\top)^\top$. The resulting $\widehat{m}_i$ are such that

$$\sum_{i \in S} \widehat{m}_i \, \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N) = \boldsymbol{0} \cdot \tag{10}$$

This implies that the maximum empirical likelihood estimator $\widehat{\boldsymbol{\varphi}}$ of $\boldsymbol{\varphi}_N$ is such that

$\widehat{\boldsymbol{\varphi}} = \boldsymbol{\varphi}_N$. In other words, the $\widehat{m}_i$ are weights calibrated with respect to $\boldsymbol{\varphi}_N$.

4. Penalised empirical log-likelihood ratio function

In § 4.3, we show how confidence intervals can be computed using a penalised empirical log-likelihood ratio function proposed by Berger & De La Riva Torres (2016) and defined by (15). This function is based upon the following *penalised empirical log-likelihood function*.

$$\widetilde{\ell}(\boldsymbol{m}) = \log\left(\prod_{i \in s} m_i \exp(1 - p_i m_i)\right) \cdot \tag{11}$$

Let $\{\widetilde{m}_i : i \in s\}$ be the set of values which maximises (11) subject to the constraints $m_i \geq 0$ and

$$\sum_{i \in S} m_i \widetilde{\boldsymbol{c}}_i \quad = \quad \widetilde{\boldsymbol{C}}; \tag{12}$$

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

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

where $\widetilde{\boldsymbol{\eta}}$ is such that (12) holds. The $\widetilde{\boldsymbol{c}}_i$ and $\widetilde{\boldsymbol{C}}$ have to be chosen to accommodate the RHC sampling design (see §§ 4.1 and 4.2). Note that $\widetilde{\boldsymbol{c}}_i$ and $\widetilde{\boldsymbol{C}}$ are different from $\boldsymbol{c}_i$ and $\boldsymbol{C}$. However, we shall 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 $\boldsymbol{c}_i = N n^{-1} p_i$ and $\boldsymbol{C} = N$ for point estimation. In this case, we use $\widetilde{\boldsymbol{c}}_i = N n^{-1} q_i^\circ p_i$ and $\widetilde{\boldsymbol{C}} = N n^{-1} \sum_{i \in S} q_i^\circ$, where $q_i^\circ = t_i^{1/2}$, where t_i, defined in expression (2), contained some information about the RHC sampling design. Let

$\{\widetilde{m}_i : i \in s\}$ be the set of values which maximises (11). It can be shown that $\widetilde{m}_i = p_i^{-1}$. Note that $\widetilde{\boldsymbol{C}}$ is not the population sum of the $\widetilde{\boldsymbol{c}}_i$.

Let $\{\widetilde{m}_i^{\star}(\theta) : i \in s\}$ be the set of values which maximises the function (11) (for a given θ) subject to the constraints $m_i \geq 0$ and

$$\sum_{i \in S} m_i \widetilde{\boldsymbol{c}}_i^{\star} \;=\; \widetilde{\boldsymbol{C}}^{\star}; \qquad (13)$$

with

$$\begin{aligned} \widetilde{\boldsymbol{c}}_i^{\star} &= \left(\widetilde{\boldsymbol{c}}_i \,,\, q_i^{\bullet} g_i(\theta)\right)^{\top}, \\ \widetilde{\boldsymbol{C}}^{\star} &= \left(\widetilde{\boldsymbol{C}} \,,\, \sum_{i \in S} (q_i^{\bullet} - 1) \frac{g_i(\theta)}{p_i}\right)^{\top} \end{aligned} \qquad (14)$$

and $q_i^{\bullet} = \widehat{\varsigma}^{\,1/2}\, t_i^{-1/2}$. Here, $\widehat{\varsigma} = (\sum_{i \in S} N_i^2 - N)(N^2 - \sum_{i \in S} N_i^2)^{-1}$ is the finite population correction proposed by Rao *et al.* (1962, p. 485) and t_i is defined in expression (2). Note that q_i° and $q_i^{\bullet}$ contained some information about the RHC sampling design. It can be shown that $\widetilde{m}_i^{\star}(\theta) = \left(p_i + \widetilde{\boldsymbol{\eta}}^{\star\top} \widetilde{\boldsymbol{c}}_i^{\star}\right)^{-1}$, where $\widetilde{\boldsymbol{\eta}}^{\star}$ is such that constraint (13) holds.

The *penalised empirical log-likelihood ratio function* is the following function of θ.

$$\widetilde{r}(\theta) \;=\; 2\left\{\widetilde{\ell}(\widetilde{\boldsymbol{m}}) - \widetilde{\ell}(\widetilde{\boldsymbol{m}}^{*}(\theta))\right\}, \qquad (15)$$

where

$$\begin{aligned} \widetilde{\ell}(\widetilde{\boldsymbol{m}}) &= \log\left(\prod_{i \in s} \widetilde{m}_i \exp(1 - p_i \widetilde{m}_i)\right), \\ \widetilde{\ell}(\widetilde{\boldsymbol{m}}^{*}(\theta)) &= \log\left(\prod_{i \in s} \widetilde{m}_i^{\star}(\theta) \exp(1 - p_i \widetilde{m}_i^{\star}(\theta))\right) \end{aligned}$$

are the maximum values of the function (11) respectively under the two different sets of constraints: (12) and (13). Here, $\widetilde{\boldsymbol{m}}$ and $\widetilde{\boldsymbol{m}}^{*}(\theta)$ denote respectively the vectors of $\widetilde{m}_i^{\star}$ and of $\widetilde{m}_i^{\star}(\theta)$

In the Appendix, we show that

$$\widetilde{r}(\theta_N) = \frac{\widehat{G}(\theta_N)^2_{RHC}}{\widehat{var}[\widehat{G}(\theta_N)_{RHC}]} + O_p(n^{-\frac{1}{2}}); \tag{16}$$

where $\widehat{G}(\theta_N)_{RHC}$ is defined by (9) and

$$\widehat{var}[\widehat{G}(\theta_N)_{RHC}] = \widehat{\varsigma}\left\{\sum_{i\in S}\frac{t_i}{M_i^2}\, g_i(\theta_N)^2 - \widehat{G}(\theta_N)^2_{RHC}\right\}$$

is the Rao *et al.* (1962) variance estimator of $\widehat{G}(\theta_N)_{RHC}$. The stochastic order $O_p(\cdot)$ denotes a random variable which a convergence in probability with respect to the RHC sampling design, as $n \to \infty$ and $N \to \infty$ (e.g. Isaki & Fuller, 1982).

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 $\widehat{G}(\theta_N)_{RHC}$, the expression (16) implies that $\widetilde{r}(\theta_N)$ follows asymptotically a χ^2-distribution with one degree of freedom, by the Slutsky's lemma.

Note that the $\widetilde{\boldsymbol{c}}_i$ and $\widetilde{\boldsymbol{C}}$ incorporate the adjustment factors q_i° and $q_i^{\bullet}$ which takes into account of the RHC design. The inclusion of these factors in the constraints (12) and (13) implies that the empirical log-likelihood ratio function (11) needs to be adjusted by penalties $\exp(1 - p_i m_i)$ in order for the property (16) to hold (see Appendix A for more details).

4.2 With auxiliary information

For point estimation, we use $\boldsymbol{c}_i = (Nn^{-1}p_i, \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)^{\top})^{\top}$ and $\boldsymbol{C} = (N, \boldsymbol{0}^{\top})^{\top}$, where $\boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)$ is defined in § 3.2). For $\{\widetilde{m}_i : i \in s\}$, we use

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

For $\{\widetilde{m}_i^\star(\theta) : i \in s\}$, we use

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

Using (16) and the Theorem 2 in Berger & De La Riva Torres (2016), it can be shown that $\widetilde{r}(\theta_N)$ defined by (15) still follows asymptotically a χ^2-distribution with one degree of freedom.

4.3 Confidence intervals and hypotheses testing

Empirical likelihood confidence intervals rely on the asymptotic distribution of the pivotal statistics $\widetilde{r}(\theta_N)$. In the previous §, we show that $\widetilde{r}(\theta_N)$ follows asymptotically a χ^2-distribution. Thus, the α level consistent empirical likelihood confidence interval (e.g. Wilks, 1938; Hudson, 1971) for the population parameter θ_N is given by

$$\left\{\theta \,:\, \widehat{r}(\theta) \leq \chi^2_1(\alpha)\right\}; \tag{17}$$

where $\chi^2_1(\alpha)$ is the upper α-quantile of the χ^2-distribution with one degree of freedom. Note that $\widehat{r}(\theta)$ is a convex non-symmetric function with a minimum at 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 (1973, Ch. 4) and Dekker (1969) method. This involves calculating $\widehat{r}(\theta)$ for several values of θ.

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 χ^2-distribution with one degrees of freedom. This p-value is obtained from the statistical table of a χ^2-distribution.

5. Simulation study

In this §, the Monte-Carlo performance of the empirical likelihood 95% confidence interval proposed is compared with linearisation (e.g. Deville, 1999), pseudoempirical likelihood (Wu & Rao, 2006), rescaled bootstrap (Rao *et al.*, 1992) and Woodruff's (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 and 5.3). The Rao *et al.* (1962) variance estimator is used for standard confidence intervals (linearisation) and for the pseudoempirical likelihood approaches. Wu & Rao (2006) proposed two pseudoempirical likelihood approaches denoted PEL1 and PEL2. Both approaches incorporate the constraint on the auxiliary variable. The PEL2 incorporates an additional constraint based on the p_i. The pseudoempirical likelihood approaches are not considered for quantiles (in §§ 5.2 and 5.3), because there is no pseudoempirical likelihood confidence intervals for quantiles in the literature. Chen & Wu (2002) proposed to use a Woodruff's (1952) approach for confidence intervals of pseudoempirical likelihood estimators of quantiles.

In §§ 5.1 and 5.2, the simulation studies are based on $10{,}000$ RHC samples of size $n = 500$ and the quantities N_g are given by $N_g = N/n$. We used the statistical software R (R Development Core Team, 2014). 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 that we have a vector $\boldsymbol{x}_i = (1, x_i)^\top$ of auxiliary variables for each unit i. We suppose that the population means $\boldsymbol{\varphi}_N$ of the $\boldsymbol{x}_i$ is known. In this case, $\boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N) = \boldsymbol{x}_i - \boldsymbol{\varphi}_N$. 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 pseudoempirical likelihood point estimators (PEL1 & PEL2) and the empirical likelihood point estimator are different, because we used auxiliary information.

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 $y_i = 3 + a_i + \beta x_i + \phi\, e_i$, where $\phi = 1.5$. 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 (Ratio Av. 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 (SD Length). The 'Ratio SD Lengths' are the 'SD Lengths' divided by the 'SD Lengths' of the linearisation confidence intervals. The column 'Ratio MSE' gives the relative efficiency (Rel. Eff.) given by the ratio between the mean squared error (MSE) of the point estimator and the regression point estimator.

[Table 1 should be here]

The empirical likelihood approach proposed gives coverages which are not significantly different from the nominal level 95%. Linearisation has also good coverages, but the empirical likelihood approach proposed 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 pseudoempirical likelihood estimators have similar MSE. With small sampling fraction ($N = 25{,}000$), the empirical likelihood approach proposed and the pseudo-EL1 approach give similar coverages, but the empirical likelihood 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 & De La Riva Torres (2016) for quantiles. The standard confidence interval is based on the linearised variable proposed by Deville (1999).

We generated several skewed population data using $y_i = 3 + a_i + \phi\, e_i$ (Wu & Rao, 2006); where the a_i follows an exponential distribution with rate parameters equal to 1 and $e_i \sim \chi^2_1 - 1$. 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 $\phi = 0.5$; $\rho(y, M) = 0.3$ with $\phi = 2.3$. Note that all the approaches give the same point estimate, because auxiliary information is not considered.

[Table 2 should be here]

The results are given in Table 2. 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 rescaled bootstrap is known to give poor coverages. Indeed, the coverages and tail error rates are significantly different from 95% and 2.5% respectively. The bootstrap confidence intervals have more unstable confidence intervals (see the column 'Ratio SD Length') because of re-sampling. Linearisation gives the most stable confidence intervals, but with coverages significantly higher than 95%.

Chen & Wu (2002) proposed to use a Woodruff's (1952) approach for confidence intervals of pseudoempirical likelihood estimators of quantiles. The Woodruff's (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 empirical likelihood approach proposed and the Woodruff's (1952) approach.

5.3 Two-stage design: synthetic EU-SILC data

The *European Union Statistics on Income and Living Conditions* (EU-SILC) survey is an European survey which collects information on income and living conditions (Eurostat, 2012). This surveys is used for measuring poverty within the European Union. Alfons *et al.* (2011) created a synthetic dataset, called AMELIA, based on EU-SILC. AMELIA maintains the association between key variables. A full description of the AMELIA data can be found in Alfons *et al.* (2011). AMELIA is replicated five times to create a population of $18,903,620$ households split into $M = 7,860$ regions, denoted R_i $(i = 1, \ldots, M)$. The regions containing less than 60 households are removed from the population.

We consider a two-stage design. For the first stage, $n = 100$ regions are selected using the RHC design with a measure of size M_i proportional to the number of households within the regions. We used $N_g = M/n$. For the second stage, simple random samples of 20 households are selected within each selected regions. This gives a sample of 2000 households. The target variable is the equalized disposable household income. The Canadian Labour Force Survey is based on a similar design, where the first stage is a RHC design.

Let $g_{ij}(\theta)$ be the estimating function for the household j in the region R_i. Let θ_N be the solution of

$$\sum_{i \in U} g_{i\cdot}(\theta) = 0, \tag{18}$$

where U denote the population of M regions and

$$g_{i\cdot}(\theta) = \sum_{j \in R_i} g_{ij}(\theta)\cdot \tag{19}$$

Let $\widehat{g}_{i\cdot}(\theta)$ be the unbiased estimator of $g_{i\cdot}(\theta)$ for a given value of θ; that is, $\widehat{g}_{i\cdot}(\theta) = 20M_i^{-1}\sum_{j\in s_i} g_{ij}(\theta)$, where s_i is the sample of households selected from R_i. We propose to use an ultimate cluster approach which is described in Oguz-Alper & Berger (2015). That is, $g_i(\theta)$ is substituted by $\widehat{g}_{i\cdot}(\theta)$. The p_i are the region level probabilities given by expression (2). The random variable $\widetilde{r}(\theta_N)$ follows asymptotically a χ^2-distribution, as long as the first-stage sampling fraction n/M is small (e.g. Oguz-Alper & Berger, 2015), because the variance in the quadratic form (16) is now the first-stage Rao *et al.* (1962) variance estimator.

The target parameters are the quantiles of the population distribution of the equalized disposable household income. Thus, the function $g_{ij}(\theta)$ is the same as the estimating function used in § 5.2. The simulation studies are based on 2000 two-stage RHC samples. The standard confidence interval is based on linearisation (Deville, 1999) and on a two-stage RHC variance estimator. The results are given in the Table 3. Note that all the approaches give the same point estimate, because auxiliary information is not considered.

[Table 3 should be here]

For the median $Y_{0.5}$, all the approaches give similar coverages and tail error rates. The differences are more pronounced for the quantiles of the tail of the distribution. With $Y_{0.10}$ and $Y_{0.95}$, the standard approach based on linearisation give poor coverages. This is due to the bias of the linearised variance and lack of normality (see column 'Shapiro-Wilk p-value'). For $Y_{0.25}$ and $Y_{0.75}$, the linearisation approach gives tail error rates significantly different from 2.5%. For $Y_{0.10}$, $Y_{0.25}$, $Y_{0.50}$ and $Y_{0.75}$, the Bootstrap, Woodruff's (1952) and empirical likelihood approaches give coverages which are not significantly different from 95%. For $Y_{0.95}$, the coverage of Woodruff's (1952) and empirical likelihood confidence intervals give better coverages, but they are significantly different from 95%. The bootstrap gives a lower coverage.

To summarise, linearisation may be problematic for the quantiles of the tail of the distribution. The Woodruff's (1952) and empirical likelihood confidence intervals seem to perform equally. The bootstrap may give tail error rate significantly different from 2.5%. This was also observed in Table 2 for $Y_{0.05}$. Woodruff‘s (1952) approach is limited to quantiles. Empirical likelihood can be used for a wider class of parameters. Empirical likelihood is easier to implement than bootstrap. For two-stage sampling, the consistency of the bootstrap confidence intervals has only been shown for smooth functions of means with small sampling fraction (Rao & Wu, 1988; Rao *et al.*, 1992). The empirical likelihood confidence interval is consistent for a wider class of parameters.

6. Conclusion and discussion

The main contribution of this article is to propose a new set of constraints (see (12) and (13)) for the empirical likelihood approach proposed by Berger & De La Riva Torres (2016). This set of constraints contains information about the RHC sampling design. We show that the resulting empirical log-likelihood ratio function can be used for testing and constructing confidence intervals. The confidence interval proposed does not rely directly on the normality of the point estimator, variance estimates, linearisation and re-sampling, even when the parameter of interest is not linear. The approach proposed 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 empirical likelihood approach proposed and the calibration developed by Deville & Särndal (1992), as the constraints (4) can be viewed as a calibration constraint. Calibration is based on distance functions between survey weights and calibration weights. These distance functions are disconnected from the mainstream likelihood statistical theory. The empirical likelihood objective function (3) is not a distance function and is related to the concept of likelihood. The advantage of the empirical likelihood approach proposed over standard calibration is the fact that (11) can be used to compute point estimates, construct confidence intervals and test hypotheses. Furthermore, empirical likelihood weights are always calibrated and positive.

Linearisation (Binder, 1983) is restricted to the situation when the $g_i(\theta)$ are differentiable with respect to θ. The empirical likelihood confidence interval proposed can be used even when the $g_i(\theta)$ are not differentiable. The coverage of the confidence interval based on linearisation may have a poor coverage (see Table 3).

Acknowledgement

I wish to thank two anonymous reviewers for helpful comments.

References

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J., Kraft, S. & Münnich, R. (2011) Synthetic Data Generation of SILC Data. Research Project Report WP6 – D6.2, University of Trier. URL `http://ameli.surveystatistics.net`. Report WP6 D6.2, FP7-SSH-2007-217322 AMELI.

Berger, Y. G. & De La Riva Torres, O. (2016) An empirical likelihood approach for inference under complex sampling design. *To appear in the J. R. Stat. Soc. Ser. B. Stat. Methodol. doi: 10.1111/rssb.12115*, 22pp.

Binder, D. A. (1983) On the variance of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.*, **51**, 279–292.

Binder, D. A. & 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*. New-Jersey: Prentice-Hall ISBN 0-13-022335-2.

Chaudhuri, A., Bose, M. & Ghosh, J. K. (2004) An application of adaptive sampling to estimate highly localized population segments. *J. Statist. Plann. Inference*, **121**, 175–189.

Chaudhuri, S., Handcock, M. S. & Rendall, M. S. (2008) Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **70**, 311–328.

Chen, J. & Qin, J. (1993) Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**, 107–116.

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

Chen, J., Sitter, R. R. & Wu, C. (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, **89**, 230–237.

Chen, J. & Wu, C. (2002) estimation of distribution function and quantiles using model-calibrated pseudo empirical likelihood method. *Statist. Sinica*, **12**, 1223–1239.

Chen, S. & Kim, J. K. (2014) Population empirical likelihood for nonparametric inference in survey sampling. *Statist. Sinica*, **24**, 335–355.

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. & Särndal, C. E. (1992) Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376–382.

Eurostat (2012) European union statistics on income and living conditions (EU-SILC). `http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc`.

Godambe, V. (1966) A new approach to sampling from finite population i, ii. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **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. & Rao, J. N. K. (1968) A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Hudson, D. J. (1971) Interval estimation from the likelihood function. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **33**, 256–262.

Isaki, C. T. & Fuller, W. A. (1982) Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, **77**, 89–96.

Keyfitz, N. (1951) Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, **46**, 105–108.

Krewski, D. & Rao, J. N. K. (1981) Inference from stratified sample: properties of linearization jackknife, and balanced repeated replication methods. *Ann. Statist.*, **9**, 1010–1019.

Laniel, N. & Mohl, C. (1994) Analysis of urban cluster size in the canadian labour force survey. *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meeting*, 6pp.

Lesage, E. (2011) The use of estimating equations to perform a calibration on complex parameters. *Survey Methodology*, **37**, 103–108.

Oguz-Alper, M. & Berger, Y. G. (2015) Modelling survey data under complex sampling designs and population level information: an empirical likelihood based approach. Southampton Statistical Sciences Research Institute `http://eprints.soton.ac.uk/376699/`.

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

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

Owen, A. B. (1991) Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725–1747.

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. & Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, pp. 300–325.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria.

Rao, J. N. K. (2005) Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, **31**, 117–138.

Rao, J. N. K., Hartley, H. O. & Cochran, W. G. (1962) On a simple procedure of unequal probability sampling without replacement. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **24**, pp. 482–491.

Rao, J. N. K. & Wu, C. F. J. (1988) Resampling inference with complex survey data. *J. Amer. Statist. Assoc.*, **83**, pp. 231–241.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992) Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209–217.

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

Schabenberger, O. & Gregoire, T. G. (1994) Competitors to genuine πps sampling designs. *Survey Methodology*, **20**, 185–192.

Statistics Canada (2008) *Methodology of the Canadian Labour Force Survey*. Authority of the Minister responsible for Statistics Canada.

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. *J. Amer. Statist. Assoc.*, **47**, 635–646.

Wu, C. & Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canad. J. Statist.*, **34**, 359–375.

Address for correspondence:

Yves G. Berger, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.

E-mail address: *y.g.berger@soton.ac.uk*

Appendix

Consider that the RHC sampling design is such that the following regularity conditions hold, when $\theta = \theta_N$.

$$
\begin{aligned}
nN^{-1}\max_{i\in s}\{p_i^{-1}\} &= O_p(1), &\text{(A.1)}\\
N^{-1}\|\widehat{\boldsymbol{C}}_p^\star - \widetilde{\boldsymbol{C}}^\star\| &= O_p(n^{-\frac{1}{2}}), &\text{(A.2)}\\
\max_{i\in s}\|\widetilde{\boldsymbol{c}}_i^\star\| &= o_p(n^{\frac{1}{2}}), &\text{(A.3)}\\
\|\widehat{\boldsymbol{S}}^\star\| &= O_p(1), &\text{(A.4)}\\
\|\widehat{\boldsymbol{S}}^{\star-1}\| &= O_p(1), &\text{(A.5)}\\
\frac{1}{nN^\tau}\sum_{i\in S}\frac{\|\widetilde{\boldsymbol{c}}_i^\star\|^\tau}{p_i^\tau} &= O_p(n^{-\tau}) \quad (\tau = 2,\ 3,\ 4), &\text{(A.6)}
\end{aligned}
$$

with

$$
\widehat{\boldsymbol{S}}^\star = -\frac{n}{N^2}\sum_{i\in S}\frac{1}{p_i^2}\widetilde{\boldsymbol{c}}_i^\star\,\widetilde{\boldsymbol{c}}_i^{\star\top}, \quad \text{and} \quad \widehat{\boldsymbol{C}}_p^\star = \sum_{i\in S}\frac{\widetilde{\boldsymbol{c}}_i^\star}{p_i},
$$

where $\|\cdot\|$ denotes the Euclidean (Frobenius) norm.

Note that theses conditions only need to hold for $\theta = \theta_N$. They do not need to hold for $\widehat{\theta}$ or for any θ. The condition (A.1) is the key condition. It ensures that the p_i are not disproportionately small compared to the sampling fraction (Krewski & Rao, 1981, p. 1014). The condition (A.2) assumes that the law of large numbers holds for $\widehat{\boldsymbol{C}}_p^\star$ (Isaki & Fuller, 1982; Krewski & Rao, 1981). The condition (A.3) ensures that the maximum of $\|\widetilde{\boldsymbol{c}}_i^\star\|$ does not converge to infinity with a rate larger than $n^{\frac{1}{2}}$ (e.g. Chen & Sitter, 1999, Appendix 2). It can be shown that the conditions

(A.4) and (A.5) hold when $-\widehat{\boldsymbol{S}}^{\star}$ is positive definite and when there exists a positive definite matrix $-\boldsymbol{S}$ such that $\|\widehat{\boldsymbol{S}}^{\star}-\boldsymbol{S}\|=o_p(1)$ and $\|\boldsymbol{S}\|=O(1)$. The condition (A.6) is a Lyapunov-type condition for the existence of moments (e.g. Krewski & Rao, 1981, p. 1014, Deville & Särndal, 1992, p. 381).

Proof of expression (16)

As $\widehat{m}_i=p_i^{-1}$, we have that $\widetilde{\ell}(\widetilde{\boldsymbol{m}})=-\ell(p)$, where $\ell(p)=\sum_{i\in S}\log(p_i)$. Using Lemma 3 in Berger & De La Riva Torres (2016), we have that under the conditions (A.1)-(A.6)

$$-2\{\widetilde{\ell}(\widetilde{m}^{\star},\theta_N)+\ell(p)\}=(\widetilde{\boldsymbol{C}}_p^{\star}-\boldsymbol{C}^{\star})^{\top}\widetilde{\boldsymbol{\Sigma}}^{\star-1}(\widetilde{\boldsymbol{C}}_p^{\star}-\boldsymbol{C}^{\star})+O_p(n^{-\frac{1}{2}}),\qquad\text{(A.7)}$$

where $\boldsymbol{C}^{\star}$ is defined by (14) and

$$\begin{aligned}\widetilde{\boldsymbol{C}}_p^{\star}&=\sum_{i\in S}\frac{\widetilde{\boldsymbol{c}}_i^{\star}}{p_i},\\ \widetilde{\boldsymbol{\Sigma}}^{\star}&==\sum_{i\in S}\frac{1}{p_i^2}\widetilde{\boldsymbol{c}}_i^{\star}\widetilde{\boldsymbol{c}}_i^{\star\top}=\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{pp}&\widehat{\boldsymbol{\Sigma}}_{pg}\\ \widehat{\boldsymbol{\Sigma}}_{pg}^{\top}&\widehat{\sigma}_{gg}\end{pmatrix};\end{aligned}$$

where

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{pp}&=\frac{N^2}{n^2}\sum_{i\in S}q_i^{\circ 2}=\frac{N^2}{n^2}\sum_{i\in S}t_i=\frac{N^2}{n^2},\\ \widehat{\boldsymbol{\Sigma}}_{pg}&=\frac{N}{n}\sum_{i\in S}q_i^{\circ}q_i^{\bullet}g_i(\theta_N)p_i^{-1}=\frac{N}{n}\,\widehat{\varsigma}^{\,1/2}\,\widehat{G}(\theta_N)_{RHC},\\ \widehat{\sigma}_{gg}&=\sum_{i\in S}q_i^{\bullet 2}\frac{g_i(\theta_N)^2}{p_i^2}=\widehat{\varsigma}\sum_{i\in S}\frac{g_i(\theta_N)^2}{t_i\,p_i^2}.\end{aligned}$$

We also have that

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

Using $\widetilde{\ell}(\widetilde{\boldsymbol{m}}) = -\ell(p)$, (15) and (A.7), we have

$$\begin{aligned}\widetilde{r}(\theta_N) &= \left(0, \widehat{G}(\theta_N)_{\scriptscriptstyle RHC}\right) \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{pp} & \widehat{\boldsymbol{\Sigma}}_{pg} \\ \widehat{\boldsymbol{\Sigma}}_{pg}^{\top} & \widehat{\sigma}_{gg} \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \widehat{G}(\theta_N)_{\scriptscriptstyle RHC} \end{pmatrix} + O_p(n^{-\frac{1}{2}}), \\ &= \frac{\widehat{G}(\theta_N)^2_{\scriptscriptstyle RHC}}{\widehat{\sigma}_{gg} - \widehat{\boldsymbol{\Sigma}}_{pg}^{\top}\widehat{\boldsymbol{\Sigma}}_{pp}^{-1}\widehat{\boldsymbol{\Sigma}}_{pg}} + O_p(n^{-\frac{1}{2}})\cdot \end{aligned} \quad (A.8)$$

It can be shown that $\widehat{\sigma}_{gg} - \widehat{\boldsymbol{\Sigma}}_{pg}^{\top}\widehat{\boldsymbol{\Sigma}}_{pp}^{-1}\widehat{\boldsymbol{\Sigma}}_{pg} = \widehat{var}[\widehat{G}(\theta_N)_{\scriptscriptstyle RHC}]$. Thus, (A.8) implies (16).

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 $*$ indicates that the coverages (or tail error rates) significantly different from 95% (or 2.5%): p-value ≤ 0.05.

Approaches	Overall	Lower tail	Upper tail	Ratio	Ratio SD	Ratio MSE
	Cov.%	err. rates%	err. rates%	Av. Length	Length	(Rel. Eff.)
Lin. Reg.	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)
PEL1	94.6 (95.4)	2.4 (2.7)	3.0* (1.9*)	0.51 (0.52)	0.45 (0.41)	0.50 (0.47)
PEL2	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 ($q = 0.5$ and 0.25). $n = 500$. The values not within brackets for $N = 2000$ (large sampling fractions). The values within brackets for $N = 25,000$ (small sampling fractions). The symbol $*$ indicate that the coverages (or tail error rates) significantly different from 95% (or 2.5%): p-value ≤ 0.05.

	$\rho(y,p)$	Approaches	Overall	Lower tail	Upper tail	Ratio Av.	Ratio SD
			Cov. %	err. rates %	err. rates %	Length	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)

Table 3: Coverages for quantiles Y_q. Two-stage design. Synthetic EU-SILCdata: AMELIA. The symbol $*$ indicate that the coverages (or tail error rates) significantly different from 95% (or 2.5%): p-value $\leq$ 0.05. The p-values of the Shapiro-Wilk normality test are reported in the last column.

Quantile	Approaches	Overall	Lower tail	Upper tail	Ratio Av.	Shapiro-Wilk
		Cov. %	err. rates%	err. rates %	Length	p-value
$Y_{0.10}$	Linear.	89.0*	5.0*	6.0*	1.00	$<$ 0.001
	Bootstrap	94.3	3.4*	2.3	1.18	
	Woodruff	94.8	2.9	2.4	1.18	
	Emp. Lik.	94.3	2.9	2.8	1.17	
$Y_{0.25}$	Linear.	94.5	3.5*	2.0	1.00	0.00043
	Bootstrap	94.2	3.3*	2.5	0.99	
	Woodruff	94.4	3.1	2.5	1.00	
	Emp. Lik.	94.3	3.1	2.6	0.99	
$Y_{0.50}$	Linear.	94.6	2.6	2.8	1.00	0.1057
	Bootstrap	94.8	2.5	2.6	1.00	
	Woodruff	94.6	2.7	2.7	1.00	
	Emp. Lik.	94.6	2.7	2.7	1.00	
$Y_{0.75}$	Linear.	94.8	2.1	3.6*	1.00	0.2543
	Bootstrap	94.5	2.6	2.8	1.00	
	Woodruff	94.7	3.0	2.4	1.01	
	Emp. Lik.	94.7	2.7	2.7	1.00	
$Y_{0.95}$	Linear.	86.3*	2.9	10.8*	1.00	$<$ 2.2e-16
	Bootstrap	92.5*	2.5	5.1*	1.10	
	Woodruff	93.6*	3.0	3.4*	1.15	
	Emp. Lik.	93.5*	2.4	4.2*	1.12	