# Estimation of confidence intervals for quantiles under unequal probability sampling using a new empirical likelihood approach

Omar DE LA RIVA TORRES*
University of Southampton, Social Statistics
Southampton, SO17 1BJ, United Kingdom
O.De-La-Riva@soton.ac.uk

Yves G.BERGER
University of Southampton, Southampton Statistical Sciences Research Institute
Southampton, SO17 1BJ, United Kingdom

We propose a new empirical likelihood approach which can be used to construct non-parametric (design-based) confidence intervals for quantiles which do not rely on the normality of the point estimator. The proposed approach does not rely on variance estimates, design-effects, re-sampling or linearisation. We show that the proposed approach gives suitable confidence intervals even when the estimator of a quantile is biased. The proposed approach also deals with large sampling fractions. Bootstrap is an alternative approach which can be used to derive non-parametric confidence intervals for quantiles. The proposed approach is less computationally intensive than the bootstrap. We compare our proposed approached with alternative approaches such as linearisation, bootstrap and the Woodruff approach.

Key Words: Design-based approach, Estimating equations, Regression estimator, Unequal inclusion probabilities.

## Introduction

Let $U$ be a finite population of $N$ units; where $N$ is a fixed quantity which is not necessarily known. Suppose that the population parameter of interest $\theta_0$ is the unique solution of the following estimating equation (e.g. Qin and Lawless, 1994).

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

where $g_i(\theta)$ is a function of $\theta$ and of characteristics of the unit $i$. This function does not need to be differentiable. We will show how estimating equation can be used to estimate quantiles.

Suppose that we wish to estimate $\theta_0$ from the data of a sample $s$ of size $n$ selected with a single stage unequal probabilities without replacement sampling design. We consider that the sample size $n$ is a fixed (non-random) quantity. We adopt a non-parametric design-based approach; where the sampling distribution is specified by the sampling design and where the values of the variables of interest are fixed constants.

We propose to use the following empirical likelihood function (e.g. Owen, 2001).

$$(2) \qquad L(m) = \prod_{i=1}^{n} \frac{m_i}{N},$$

where $m_i$ is the unit mass of unit $i$ in the population (e.g. Deville, 1999).

Hartley and Rao (1969) showed that (2) is the correct empirical likelihood function under unequal probability sampling with replacement, as $m_i/N$ is the probability to observe the $i$-th unit. Owen

(2001, Ch. 6) showed that (2) is a suitable empirical likelihood function when the units are selected independently with a Poisson sampling design. Although under fixed size sampling designs, the units are not selected independently, we propose to use the empirical likelihood function (2) under fixed size sampling designs. The aim is to show that this empirical likelihood function can be used for point estimation and to construct confidence intervals (or to derive tests) under fixed size sampling designs.

The maximum likelihood estimators of $m_i$ are the values $\widehat{m}_i$ which maximise the *log-empirical likelihood function*

$$(3) \qquad \ell(m) = \sum_{i=1}^{n} \log\left(m_i\right),$$

subject to the constraints $m_i \geq 0$ and

$$(4) \qquad \sum_{i=1}^{n} m_i \boldsymbol{c}_i \;=\; \boldsymbol{C};$$

where $\sum_{i=1}^{n}$ denotes the sum over the sampled units, $\boldsymbol{c}_i$ is a known $Q \times 1$ vector associated with the $i$-th sampled unit and $\boldsymbol{C}$ is a known $Q \times 1$ vector. We also assume that the constraint (4) is such that the fixed size constraint

$$(5) \qquad \sum_{i=1}^{n} m_i \pi_i = n$$

always holds, where $\pi_i$ denotes the inclusion probability of unit $i$. Under equal probability sampling, we have that $\pi_i = n/N$, and the constraint (5) reduces to $\sum_{i=1}^{n} m_i = N$ which is the constraint adopted under equal probability sampling (e.g. Rao and Wu, 2009). Berger and De La Riva Torres (2012) showed that the solution of this maximisation is given by

$$(6) \qquad \widehat{m}_i = \left(\pi_i + \boldsymbol{\eta}'\boldsymbol{c}_i\right)^{-1},$$

The quantity $\boldsymbol{\eta}$ is such that the constraint (4) holds. This quantity can be computed using an iterative Newton-Raphson procedure (e.g. Berger and De La Riva Torres, 2012; Rao and Wu, 2009).

The *maximum empirical likelihood estimator* $\widehat{\theta}$ of $\theta_0$ is defined by solution of the following estimating equation.

$$(7) \qquad \widehat{G}(\theta) = 0, \quad \text{with} \quad \widehat{G}(\theta) = \sum_{i=1}^{n} \widehat{m}_i \; g_i(\theta);$$

where $\widehat{m}_i$ is defined by (6). We assume that the $g_i(\theta)$ are such that $\widehat{G}(\theta) = 0$ has a unique solution. The estimator $\widehat{\theta}$ is a maximum empirical likelihood estimator because it also minimises the empirical log-likelihood ratio function (or deviance) defined by (9).

For example, when $\boldsymbol{c}_i = \pi_i$ and $\boldsymbol{C} = n$, we have that $m_i = \pi_i^{-1}$ and

$$(8) \qquad \widehat{G}(\theta) = \widehat{G}_{\pi}(\theta) = \sum_{i=1}^{n} \frac{g_i(\theta)}{\pi_i},$$

is the Horvitz and Thompson (1952) estimator of the function (1). When $g_i(\theta) = y_i - \theta\pi_i/n$, we have that $\widehat{\theta} = \sum_{i=1}^{n} y_i \pi_i^{-1}$ is the Horvitz and Thompson (1952) estimator of the population total $Y = \sum_{i \in U} y_i$.

### Estimation of quantiles

Suppose that the parameter $\theta_0$ of interest is the $q$ quantile $Y_q$ of the population distribution of a variable of interest $y_i$; where $0 < q < 1$. The estimating equation for a quantile cannot be used directly to derive an empirical likelihood estimator for a quantile, because the distribution function is a step function (e.g. Owen, 2001, p. 45). We propose a simple solution which consists in defining the empirical likelihood estimator by the estimating equation (7) where $g_i(\theta) = \varrho(y_{(i)}, \theta) - q$, with

$$\varrho(y_{(i)}, \theta) = \delta\{y_{(i)} \leq \theta\} + \frac{\theta - y_{(i-1)}}{y_{(i)} - y_{(i-1)}} \delta\{y_{(i-1)} \leq \theta\}(1 - \delta\{y_{(i)} \leq \theta\};$$

where the $y_{(i)}$ is the values of the $i$-th sampled units arranged in increasing order, with $y_{(0)} = y_{(1)} - (y_{(2)} - y_{(1)})$. The empirical likelihood estimator of $Y_q$ is the solution of the equation $\widehat{G}(\theta) = 0$ which becomes $\widetilde{F}(\theta) = q$; where $\widetilde{F}(\theta) = (\sum_{i=1}^{n} \widehat{m}_i)^{-1} \sum_{i=1}^{n} \widehat{m}_{(i)} \varrho(y_{(i)}, \theta)$. Note that $\widetilde{F}(\theta) = q$ has always a unique solution because $\widetilde{F}(y)$ is a bijective function given by a piecewise linear interpolation of the step distribution function $\widehat{F}(\theta) = (\sum_{i=1}^{n} \widehat{m}_i)^{-1} \sum_{i=1}^{n} \widehat{m}_i \delta\{y_{(i)} \leq \theta\}$.

## Empirical log-likelihood ratio function

The main advantage of the empirical likelihood approach is its capability of deriving non-parametric confidence intervals which do not depend on variance estimates or on the normality of the point estimator. We propose to use the empirical log-likelihood ratio function defined by (9) to derive empirical likelihood confidence intervals.

Let $\widehat{m}_i$ be the values which maximise (3) subject to the constraints $m_i \geq 0$ and (4) when $\boldsymbol{c}_i = \pi_i$ and $\boldsymbol{C} = n$. Note that $m_i = \pi_i^{-1}$ in this situation. Hence the empirical likelihood point estimator is the solution of the estimating equation (8). The maximum value of the empirical log-likelihood function is given by $\ell(\widehat{m}) = -\sum_{i=1}^{n} \log(\pi_i)$. Let $\widehat{m}_i^*$ be the values which maximise (3) subject to the constraints $m_i \geq 0$ and (4) with $\boldsymbol{c}_i = \boldsymbol{c}_i^*$ and $\boldsymbol{C} = \boldsymbol{C}^*$, where $\boldsymbol{c}_i^* = (\pi_i, g_i(\theta))'$ and $\boldsymbol{C}^* = (n, 0)'$. Let $\ell(\widehat{m}^*, \theta)$ be the maximum value of of the empirical log-likelihood function. The *empirical log-likelihood ratio function* (or deviance) is defined by the following function of $\theta$.

$$(9) \qquad \widehat{r}(\theta) \;=\; 2\left\{\ell(\widehat{m}) - \ell(\widehat{m}^*, \theta)\right\} \cdot$$

It can be easily shown that $\widehat{r}(\widehat{\theta}) = 0$. Hence $\widehat{\theta}$ is indeed the maximum empirical likelihood estimator of $\theta_0$, because it minimises the empirical log-likelihood ratio function. Berger and De La Riva Torres (2012) showed how the stratification can be taken into account by including the stratification variables within the vectors $\boldsymbol{c}_i$ and $\boldsymbol{c}_i^*$.

Berger and De La Riva Torres (2012) showed that under a set of regularity conditions, $\widehat{r}(\theta_0)$ follows asymptotically a chi-squared distribution with one degree of freedom when the sampling fraction, $n/N$, is negligible. This property relies on the fact that $\widehat{G}_\pi(\theta_0)$ is an Horvitz and Thompson (1952) estimator which follows a normal distribution asymptotically (Berger, 1998; Hájek, 1964; Vísek, 1979).

## Empirical likelihood confidence intervals

As $\widehat{r}(\theta_0)$ follows asymptotically a chi-squared distribution, the $(1 - \alpha)$ level empirical likelihood confidence interval (e.g. Wilks, 1938) for the population parameter $\theta_0$ is given by

$$(10) \qquad \left[\ \min\left\{\theta \middle| \ \widehat{r}(\theta) \leq \chi_1^2(\alpha)\right\};\ \max\left\{\theta \middle| \ \widehat{r}(\theta) \leq \chi_1^2(\alpha)\right\}\ \right];$$

where $\chi_1^2(\alpha)$ is the upper $\alpha$-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 $\theta$ is the maximum empirical likelihood estimator. This interval can be found using a bijection search method. This involves calculating $\widehat{r}(\theta)$ for several values of $\theta$.

With large sampling fractions the empirical log-likelihood ratio function does not necessarily follow a chi-squared distribution. Berger and De La Riva Torres (2012) proposed to adjust the constraint in order to obtain a chi-squared distribution asymptotically. Consider $\boldsymbol{c}_i = \pi_i$ and $\boldsymbol{C} = n$. We propose to use $\boldsymbol{c}_i^* = q_i(\pi_i, g_i(\theta))'$ and $\boldsymbol{C}^* = (\sum_{i=1}^n q_i, \sum_{i=1}^n (q_i - 1)g_i(\theta)\pi_i^{-1})'$, with $q_i = (1-\pi_i)^{1/2}$. Let $\widehat{m}_i^*$ be defined by

$$(11) \qquad \widehat{m}_i^* = \left(\pi_i + \boldsymbol{\eta}^{*'}\boldsymbol{c}_i^*\right)^{-1},$$

where $\boldsymbol{\eta}^*$ is such that $\sum_{i=1}^n \widehat{m}_i^* \boldsymbol{c}_i^* = \boldsymbol{C}^*$ holds. We propose to use the same empirical log-likelihood ratio function (9). The empirical log-likelihood ratio function is still defined by (9) with $\ell(m)$ given by (3). Berger and De La Riva Torres (2012) showed that under a set of regularity conditions, $\widehat{r}(\theta_0)$ follows asymptotically a chi-squared distribution with one degree of freedom for any sampling fractions. Hence empirical likelihood confidence intervals can be constructed using (10). Berger and De La Riva Torres (2012) showed how the stratification can be taken into account by including the stratification variables within the vectors $\boldsymbol{c}_i$ and $\boldsymbol{c}_i^*$.

In practice, population control totals of auxiliary variables are often known and this information is often taken into account at the estimation stage. Berger and De La Riva Torres (2012) showed how auxiliary variables can be taken into account using an empirical likelihood approach. The resulting weights $\widehat{m}_i$ are asymptotically equal to regression weights. Berger and De La Riva Torres (2012) proposed a restricted empirical likelihood approach to construct confidence intervals in the presence of auxiliary variables.

## Simulation studies

We generated a population data according to the following model proposed by Wu and Rao (2006):

$$(12) \qquad y_i = 3 + a_i + \varphi e_i,$$

where $a_i$ follows an independent exponential distributions with rate parameters equal to one and $e_i \sim \chi_1^2 - 1$. The $\pi_i$ are proportional to $a_i + 2$. The constant 2 is added to $a_i$ to avoid having very small $\pi_i$. Two populations of size of $N = 150$ and $N = 800$ were generated using (12), the values $y_i$ and $a_i$ generated were treated as fixed. The parameter $\varphi$ were used to obtain a weak and a strong correlation between the values $y_i$ and $\hat{y}_i = 3 + a_i$. Let $\rho(y_i, \hat{y}_i)$ denote the correlation between $y_i$ and $\hat{y}_i$. We consider two values for the correlation: 0.30 and 0.80. The parameters of interest $\theta_0$ is the population quantile $Y_q$, where $q = 0.10$.

We used Chao (1982) sampling design to select 1000 samples of size $n = 40$ and 80. We consider 95% confidence intervals. For the proposed approach, we use $\boldsymbol{c}_i = \pi_i$ and $\boldsymbol{C} = n$, therefore $m_i = \pi_i^{-1}$. The point estimator is the solution of (8) with $g_i(\theta) = \varrho(y_{(i)}, \theta) - q$. This estimator has a skewed sampling distribution. The performance of the proposed empirical likelihood confidence intervals are compared with the direct bootstrap approach (Antal and Tillé, 2011) based on the bootstrap variance, the Woodruff (1952) approach, the rescaled bootstrap approach (Rao et al., 1992) based upon the observed confidence intervals of the bootstrap values and the standard approach based on linearisation (Deville, 1999). We used 1000 bootstrap replicates. For the Woodruff approach, the confidence interval was obtained from the inverse of $\widehat{F}(y) = \widehat{N}_\pi^{-1} \sum_{i=1}^n \pi_i^{-1} \delta\{y_i \le y\}$, where $\widehat{N}_\pi = \sum_{i=1}^n \pi_i^{-1}$.

In Table 1, we have the coverage probabilities, the lower and the upper tail error rates, the average lengths and the variances of the lengths of the confidence intervals. The values for large sampling fractions ($N = 150$) are given in brackets. The coverage of the standard approach based on the central limit theorem and linearisation is significantly larger than 95% in all the cases considered. This is due to the fact that the point estimator has a positively skewed sampling distribution. This explains the null upper tail error rate. The linearised variance estimator is also biased. The Woodruff (1952) approach gives confidence intervals which are as good as the empirical likelihood confidence

Table 1: Coverages of 95% confidence intervals for the quantile $Y_{0.10}$. The values not in brackets are for the population of size $N = 800$ (small sampling fractions). The values in brackets are for the population of size $N = 150$ (large sampling fractions).

| $\rho(y_i, \hat{y}_i)$ | $n$ | Approaches | Coverage Prob (%) | Lower Tail Error (%) | Upper Tail Error (%) | Average Length | Variance Length |
|---|---|---|---|---|---|---|---|
| 0.3 | 40 | Proposed $\mathbf{c}_i = \pi_i$ | 93.3 (92.8) | 4.1 (1.9) | 2.6 (5.3) | 0.73 (0.67) | 0.065 (0.046) |
| | | Direct bootstrap | 92.1 (91.7) | 4.3 (5.4) | 3.6 (2.9) | 0.79 (0.72) | 0.080 (0.058) |
| | | Woodruff | 91.3 (93.5) | 4.5 (2.5) | 4.2 (4.0) | 0.66 (0.66) | 0.056 (0.043) |
| | | Rescaled bootstrap | 93.9 (95.6) | 4.7 (3.0) | 1.4 (1.4) | 0.77 (0.76) | 0.062 (0.045) |
| | | Linearisation | 98.9 (99.6) | 1.1 (0.4) | 0.0 (0.0) | 1.17 (1.34) | 0.055 (0.049) |
| 0.3 | 80 | Proposed $\mathbf{c}_i = \pi_i$ | 96.5 (93.6) | 0.8 (1.0) | 2.7 (5.4) | 0.57 (0.43) | 0.024 (0.013) |
| | | Direct bootstrap | 92.2 (92.3) | 4.5 (4.8) | 3.3 (2.9) | 0.56 (0.44) | 0.028 (0.016) |
| | | Woodruff | 95.7 (95.1) | 1.3 (1.8) | 3.0 (3.1) | 0.55 (0.44) | 0.023 (0.014) |
| | | Rescaled bootstrap | 95.4 (98.3) | 3.3 (1.3) | 1.3 (0.4) | 0.57 (0.56) | 0.024 (0.018) |
| | | Linearisation | 99.4 (99.9) | 0.6 (0.1) | 0.0 (0.0) | 0.86 (0.85) | 0.014 (0.007) |
| 0.8 | 40 | Proposed $\mathbf{c}_i = \pi_i$ | 92.9 (91.8) | 4.0 (2.3) | 3.1 (5.9) | 0.54 (0.37) | 0.039 (0.018) |
| | | Direct bootstrap | 92.6 (91.3) | 4.1 (6.0) | 3.3 (2.7) | 0.59 (0.39) | 0.044 (0.018) |
| | | Woodruff | 90.7 (92.7) | 4.8 (3.6) | 4.5 (3.7) | 0.48 (0.36) | 0.033 (0.015) |
| | | Rescaled bootstrap | 93.9 (94.6) | 4.8 (3.8) | 1.3 (1.6) | 0.57 (0.42) | 0.034 (0.020) |
| | | Linearisation | 96.5 (99.5) | 3.5 (0.5) | 0.0 (0.0) | 0.73 (0.62) | 0.028 (0.012) |
| 0.8 | 80 | Proposed $\mathbf{c}_i = \pi_i$ | 95.9 (94.2) | 1.6 (0.6) | 2.5 (5.2) | 0.41 (0.23) | 0.015 (0.005) |
| | | Direct bootstrap | 93.3 (89.7) | 3.3 (7.7) | 3.4 (2.6) | 0.42 (0.23) | 0.018 (0.005) |
| | | Woodruff | 94.3 (94.3) | 2.0 (2.4) | 3.7 (3.3) | 0.39 (0.22) | 0.015 (0.004) |
| | | Rescaled bootstrap | 96.7 (98.7) | 2.7 (0.8) | 0.6 (0.5) | 0.42 (0.29) | 0.015 (0.005) |
| | | Linearisation | 97.8 (99.8) | 2.2 (0.2) | 0.0 (0.0) | 0.55 (0.41) | 0.007 (0.002) |

intervals in term of coverage and stability of the confidence intervals, but with lower coverages with small sampling fraction. The rescaled bootstrap confidence intervals may have slightly higher coverage probabilities compared to the other approaches. For small sampling fractions, the performance of the proposed empirical likelihood approach is similar to the rescaled bootstrap. However, with large sampling fractions, the rescaled bootstrap confidence intervals may have a large coverage, because this approach does not includes finite population correction factors. With large sampling fractions, the direct bootstrap has better coverage because it includes finite population corrections. However, the coverage of the direct bootstrap tends to be smaller than 95%. The direct bootstrap has a low coverage of 89.7% when $\rho(y_i, \hat{y}_i) = 0.8$ and $n = 80$. The proposed approach gives a coverage of 94.2% in this situation. Note that the direct bootstrap larger variances for the lengths. This means that the direct bootstrap confidence intervals are more volatile than empirical likelihood confidence intervals.

## Conclusions

Standard confidence intervals based upon the central limit theorem can perform poorly when the sampling distribution is not normal. For example, the lower bounds can be negative even when the parameter of interest is positive. The coverage and the tail errors can be also different from their intended levels. On the other hand, empirical likelihood confidence intervals may be better in this situation, as empirical likelihood confidence intervals are determined by the distribution of the data (Rao and Wu, 2009) and the range of the parameter space is preserved. Note that the distribution of a point estimator of is not necessarily normal, and the proposed empirical likelihood approach does not rely on the normality of the point estimator.

Standard confidence intervals based on the central limit theorem require normality and variance estimates which often involve linearisation or re-sampling. The proposed method does not rely on normality, variance estimates, linearisation or re-sampling, even if the parameter of interest is not linear. Empirical likelihood confidence intervals can be easier to compute than standard confidence intervals based on variance estimates. It provides an alternative to bootstrap, when linearisation cannot be used. The proposed approach has some advantages over the bootstrap approach. It is less computationally intensive than the bootstrap. Our simulations study also shows that bootstrap confidence intervals may not have the right coverage and may be more unstable. When the sample size is large, the Woodruff (1952) approach gives confidence intervals which are as good as the empirical likelihood confidence intervals in term of coverage and stability of the confidence intervals. However, the Woodruff (1952) approach relies on variance estimates, joint inclusion probabilities. Furthermore, this approach is only designed for quantiles. The empirical likelihood approach can be used for a wider class of point estimators.

## REFERENCES

Antal, E., and Tillé, Y. (2011), "A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population," *Journal of the American Statistical Association*, 106, 534–543.

Berger, Y. G. (1998), "Rate of Convergence to Normal Distribution for the Horvitz-Thompson Estimator," *Journal of Statistical Planning and Inference*, 67, 209–226.

Berger, Y. G., and De La Riva Torres, O. (2012), "A unified theory of empirical likelihood ratio confidence intervals for survey data with unequal probabilities and non negligible sampling fractions," *Southampton Statistical Sciences Research Institute*, `http://eprints.soton.ac.uk/337688`.

Chao, M. T. (1982), "A General Purpose Unequal Probability Sampling Plan," *Biometrika*, 69, 653–656.

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

Hájek, J. (1964), "Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population," *The Annals of Mathematical Statistics*, 35(4), 1491–1523.

Hartley, H. O., and Rao, J. N. K. (1969), *A new estimation theory for sample surveys, II*, A Symposium on the Foundations of Survey Sampling held at the University of North Carolina, Chapel Hill, North Carolina: Wiley-Interscience, New York.

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(260), 663–685.

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

Qin, J., and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22(1), pp. 300–325.

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.

Rao, J. N. K., and Wu, W. (2009), "Empirical Likelihood Methods," *Handbook of statistics: Sample Surveys: Inference and Analysis, D. Pfeffermann and C. R. Rao eds. The Netherlands (North-Holland)*, 29B, 189–207.

Vísek, J. (1979), "Asymptotic distribution of simple estimate for rejectif, sampford and successive sampling," *Contribution to Statistics, Jaroslav Hajek Memorial Volume. Academia of Prague, Czech Republic*, pp. 71–78.

Wilks, S. S. (1938), "Shortest Average Confidence Intervals from Large Samples," *The Annals of Mathematical Statistics*, 9(3), 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(3), 359–375.