# Empirical likelihood ratio confidence intervals for unequal probability sampling

Berger, Yves G
*University of Southampton, Southampton Statistical Sciences Research Institute*
*Southampton SO17 1BJ United Kingdom*
*E-mail: Y.G.Berger@soton.ac.uk*

De La Riva Torres, Omar
*University of Southampton, Faculty of Human and Social Sciences*
*Southampton SO17 1BJ United Kingdom*
*E-mail: O.De-La-Riva@soton.ac.uk*

Empirical likelihood (EL) (Owen, 1988) has first been introduced by Hartley and Rao (1968) under the name scale load approach. Since Chen and Qin (1993) suggested its first application in survey sampling, there have been many recent developments of (EL) based methods in survey sampling (e.g. Rao & Wu, 2009) and adaptive sampling (Salehi, *et al.* 2008). Standard confidence intervals based upon a normal distribution can perform poorly when the sampling distribution is not normal. On the other hand, EL confidence intervals may be better in this situation, as EL confidence intervals are determined by the distribution of the data (Rao & Wu 2009). The range of the parameter space is also preserved. This may not be the case for standard confidence intervals based upon a normal distribution, as standard confidence intervals can have negative lower bounds for a positive point estimator. Chen and Sitter (1999) proposed a pseudo EL approach which can be used to construct confidence intervals for the Hájek (1971) ratio estimator (Wu & Rao, 2006). The pseudo EL approach is not entirely appealing from a theoretical point of view (Rao & Wu 2009) as it is not a genuine EL approach, and it is not applicable to the Horvitz-Thompson (1952) estimator. We propose a true EL approach for unequal probability sampling without replacement based on Kim (2009) EL approach. We derive the asymptotic distribution of the profile empirical likelihood and support our results with a simulation study.

## Empirical likelihood for total

Let $U = \{1, \ldots, N\}$ denote a finite population of size $N$, and $s$ denote a sample selected with unequal probabilities from $U$. A parameter of interest is denoted by $\theta(M)$ where $M$ is the population mass which gives a unit mass of 1 for all units $i \in U$; that is, $M_i = 1$ for all $i \in U$. For example, a total of a variable $y$ is given by $\theta(M) = \int y dM = \sum_{i \in U} y_i = Y$. The mass $M$ is estimated by $\hat{m}$ which given the mass $\hat{m}_i$ for all units $i \in s$. Deville (1999) suggests using $\hat{m}_i = 1/\pi_i$ which gives the substitution estimator $\theta(\hat{m}) = \int y d\hat{M} = \sum_{i \in s} y_i/\pi_i = \widehat{Y}_{\text{HT}}$ which is the well known Horvitz-Thompson (1952) estimator. We propose to use an EL method to estimate $M$.

Consider a conditional Poisson sampling design (Hájek, 1981). Let $p_i$ denote the first-order inclusion probabilities of the unconditional Poisson sampling. The first-order inclusion probabilities of the conditional Poisson sampling are denoted by $\pi_i$ (Hájek, 1964).

Let $y_1, ..., y_N$ be the vector of realised values of the finite population with the cumulative distribution function $F(y) = N^{-1} \sum_{i=1}^{N} I_{(y_i \leqslant y)}$ where $I(\cdot)$ is the indicator function, i.e., $I(y_i \leqslant y)$ takes the value one if $y_i \leqslant y$ and takes the value zero otherwise. As $F(y)$ belongs to a family of distributions with support on $\{y_1, \ldots, y_n\}$ we have that

$$F_s(y) = \frac{1}{\hat{N}_\pi} \sum_{i \in s} m_i I_{(y_i \leqslant y)} \tag{1}$$

for some $m_i$'s which are such that $\sum_{i\in s} m_i = \hat{N}_\pi = \sum_{i\in s} \pi_i^{-1}$ and $m_i \geqslant 0$. Following Kim's (2009) approach we found that under conditional Poisson sampling, the distribution in (1) implies the following EL

$$L(m) = \prod_{i\in s} \left( \frac{p_i m_i}{\sum_{i\in s} p_j m_j} \right). \tag{2}$$

The estimator $\hat{m}$ of $M$ with $\hat{m} = \{\hat{m}_1, \ldots, \hat{m}_n\}$ maximises the EL (2) under the following constraints

$$\sum_{i\in s} m_i = \hat{N}_\pi \text{ and } \sum_{i\in s} m_i x_i = X \tag{3}$$

with $x_i = \pi_i$ and $X = \sum_{i\in U} \pi_i = n$. The first constraint guarantees that (1) is a distribution function. The second constraint is the fixed sample size constraint. We can use an iterative procedure (Newton-Raphson) to calculate $m_i$. The point estimator of the total is $\widehat{Y}_{\text{EL}} = \sum_{i\in s} \hat{m}_i y_i$ where $\hat{m}_i$ maximizes (2) under the constraints (3). Note that as $p_i \backsimeq \pi_i$, we have that $\hat{m}_i \backsimeq 1/\pi_i$ and $\widehat{Y}_{\text{EL}} \backsimeq \widehat{Y}_{\text{HT}} = \sum_{i\in s} y_i/\pi_i$.

**Asymptotic distribution of the profile likelihood**

Under a set of regularity conditions, we show that

$$r^*(Y) = r(Y) \times \frac{\widetilde{var}(\widehat{Y}_{\text{EL}}, Y)}{\widehat{var}(\widehat{Y}_{\text{HT}})}$$

is asymptotically distributed as $\chi^2_1$ under $\pi$ps sampling; where $r(Y) = -2[l(m(Y)) - l(m)]$, $l(m) = \log L(m)$ is the maximum value of the EL log likelihood function defined in (2) subject to constraints (3) and $l(m(Y)) = \log L(m)$ is the maximum value of EL log likelihood subject to these latter constraints (3) and the additional constraint $\sum_{i\in s} m_i y_i = Y$; $\widetilde{var}(\widehat{Y}_{\text{EL}}, Y) = \hat{\sigma}_{yy} - \hat{\sigma}_{yx}\hat{\sigma}_{xx}^{-1}\hat{\sigma}_{xy}$ where $\hat{\sigma}_{yy}$, $\hat{\sigma}_{yx}$, $\hat{\sigma}_{xx}$ and $\hat{\sigma}_{xy}$ are components of the following matrix

$$\begin{pmatrix} \hat{\sigma}_{xx} & \hat{\sigma}_{xy} \\ \hat{\sigma}_{yx} & \hat{\sigma}_{yy} \end{pmatrix} \equiv \left\{ \sum_{i\in s} \frac{1}{p_i^2} \left( x_{ir} - \frac{X_r}{\hat{N}_\pi} \right) \left( x_{iq} - \frac{X_q}{\hat{N}_\pi} \right) \right\}_{r,q\in\{1,2\}}$$

where $x_{i1} = x_i = \pi_i$, $x_{i2} = y_i$, $X_r = \sum_{i\in U} x_{ir}$, $X_q = \sum_{i\in U} x_{iq}$ and $\widehat{var}(\widehat{Y}_{\text{HT}})$ is a consistent estimator for the variance of the Horvitz-Thompson $\widehat{Y}_{\text{HT}}$ estimator of $Y$. Note that $\widetilde{var}(\widehat{Y}_{\text{EL}}, Y)$ is a biased estimator for the variance of $\widehat{Y}_{\text{EL}}$.

**Estimation of confidence intervals using profile likelihood**

As

$$r^*(Y) \to \chi^2_1$$

an approximate $100(1-\alpha)\%$ confidence interval for $Y$ is given by the minimum and maximum values of the following set

$$\{Y : r^*(Y) \leqslant \chi^2_{1-\alpha,1}\}$$

where $\chi^2_{1-\alpha,1}$ is the $(1-\alpha)$th quantile of the $\chi^2_1$ distribution.

**A numerical example**

Consider $N = 2000$ values given by $\tilde{y}_i = 1 + \sqrt{(0.5)}(z_i - 3) + e_i$ (Kim, 2009) where $z_i = (1/N)^\gamma - 1/\gamma$ for $\gamma = 1, 2, 3$ and $e_i \sim N(0, 1)$. We consider $\pi_i = nz_i/\sum_{i\in s} z_i$. The variable of interest is defined by $y_i = (y_{(i)} - \min[y_{(i)}])/N$ and 1000 $\pi$ps samples of size $n = 100$ were selected with unequal

probability using Rao-Sampford sampling design (Rao 1965, Sampford 1967). Coverages of confidence interval are given in Table 1. The Monte-Carlo coverages are close to the target value of 95%.

***Table 1. Monte-carlo coverage of confidence intervals***

| $skewness(\pi_i)$ | Coverage | Average length |
|---|---|---|
| 0.00 | 94.9% | 0.431 |
| 0.63 | 93.7% | 0.546 |
| 1.06 | 93.7% | 0.683 |

## REFERENCES

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. Statistica Sinica 9, 385-406.

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, 1491-1523.

Hájek, J. (1971). Foundations of Statistical Inference. Toronto, Canada: Holt, Rinehart, Winston. Chap. Discussion of an essay on the logical foundations of survey sampling, part on by D. Basu.

Hájek, J. (1981). Sampling from a finite population. New York: Marcel Dekker.

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

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

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. Statistica Sinica 19, 145-157.

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

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. Journal of the Indian Statistical Association 3, 173-180.

Rao, J.N.K. & Wu, C. (2009). Empirical Likelihood Methods. In : D. Pfeffermann and C.R. Rao. (editors). Elsevier.

Salehi, M., Mohammadi, M., Rao, J.N.K., Berger, Y.G (2008). Empirical likelihood confidence intervals for adaptive cluster sampling. Environmental and Ecological Statistics 17, 111-123.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. Biometrika 54, 499-513.

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

## RÉSUMÉ

*Nous proposons une approche du type vraisemblance empirique pour estimer l'intervalle de confiance de l'estimateur de Horvitz-Thompson (1952) pour un plan de sondage a probabilité inégales sans remise. Nous montrons que sous des conditions de régularité, le profile de la vraisemblance empirique transformé a une distribution chi carré. Nous présenterons également quelques résultats de simulation.*