# EMPIRICAL LIKELIHOOD METHODS FOR SAMPLE SURVEY DATA: AN OVERVIEW

J. N. K. Rao[1]

## ABSTRACT

The use of empirical likelihood (EL) in sample surveys dates back to Hartley and Rao (1968). In this paper, an overview on the developments in empirical likelihood methods for sample survey data is presented. Topics covered include EL estimation using auxiliary population information and EL confidence intervals. Issues related to pseudo-EL estimation for general sampling designs are also discussed.

KEY WORDS: Confidence interval; Distribution function; General sampling designs; Population mean.

## RÉSUMÉ

L'utilisation de la vraisemblance empirique (VE) dans le contexte des sondages remonte à Hartley et Rao (1968). Dans cet article, un survol des développements concernant la vraisemblance empirique (VE) dans le contexte des sondages est présenté. Les sujets couverts comprennent l'estimation par VE utilisant de l'information auxiliaire disponible au niveau de la population et les intervalles de confiance par VE. Des problèmes traitant de l'estimation par pseudo-VE pour des plans de sondage quelconques sont également discutés.

MOTS CLÉS : Fonction de répartition; intervalle de confiance; moyenne de population; plans de sondage quelconques;

## 1. INTRODUCTION

### 1.1 Description of the Problem

"Empirical likelihood" is used to denote a non-parametric likelihood. It was first introduced in the survey sampling context by Hartley and Rao (1968) under the name "scale-load" approach. Twenty years later, Owen (1988) introduced it in the main stream statistical inference, under the name "empirical likelihood", developed a unified theory and demonstrated its advantages. In recent years, the empirical likelihood (EL) approach has been revived in survey sampling literature. The main purpose of this paper is to present an overview of some recent developments in applying the EL approach to sample survey data.

### 1.2 Organisation of the Paper

Section 2 of the paper gives a brief account of the original approach of Hartley and Rao (1968). Owen's (1988, 2001) EL theory for the case of independent and identically distributed (IID) variables and the use of supplementary population information are highlighted in Section 3. Some results for stratified random sampling are given in Section 4. Use of pseudo-EL to handle general sampling designs is outlined in Section 5.

## 2. SCALE-LOAD APPROACH

Consider a finite population $U$ consisting of units $i(=1,...,N)$ with associated values $y_i$. A subset $s$ of units is selected from $U$ with probability $p(s)$. The sample data is denoted as $\{(i, y_i), i \in s\}$. Godambe (1966) obtained the non-parametric likelihood for the sample data and showed that it is "flat" in the sense that all possible non-observed values

[1] J. N. K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6, jrao@math.carleton.ca

$y_i, i \in U - s$ lead to the same likelihood. Hence, the likelihood is not informative. This difficulty arises because of the distinct labels $i$, associated with the units in the sample data, that make the sample unique.

Hartley and Rao (1968) suggested the scale-load approach to obtain an informative likelihood. Under this approach, some aspects of the sample data need to be ignored to make the sample non-unique and in turn the likelihood informative. The reduction of sample data is not unique and depends on the situation at hand; data based decisions might be needed (Hartley and Rao, 1971). The basic feature of the Hartley-Rao approach is a specific representation of the finite population, assuming that the variable $y$ is measured on a scale with finite set of known scale points $y_t^*$ ($t = 1,..,T$); the parameter $T$ is only conceptual and inferences do not require the specification of $T$. Let $N_t$ be the number of units in $U$ having the value $y_t^*$ ($\sum N_t = N$) so that the population mean $\bar{Y} = N^{-1}\sum N_t y_t^*$ is completely specified by the "scale loads" $\underset{\sim}{N} = (N_1,..., N_T)^T$.

Consider simple random sampling without replacement of fixed size $n$ and let $n_t$ be the number of units in the sample having the value $y_t^*$, so that $n_t \geq 0$ and $\sum n_t = n$. If we suppress the labels $i$ from the sample data, then the data are given by $\underset{\sim}{n} = (n_1,..., n_T)^T$ and the resulting likelihood is simply given by the hyper-geometric distribution that depends on the parameter $\underset{\sim}{N}$, unlike the flat likelihood based on the full sample data. Loss of information due to ignoring the sample labels may be regarded as negligible if there is no evidence of a relationship between the labels and associated variable values. If the sampling fraction is negligible, the log-likelihood may be approximated by the multinomial log-likelihood $l(\underset{\sim}{p}) = \sum n_t \log(p_t)$, where $p_t = N_t / N$. The resulting maximum likelihood estimator (MLE) of $\bar{Y} = \sum p_t y_t^*$ is the sample mean $\bar{y} = \sum \hat{p}_t y_t^*$, where $\hat{p}_t = n_t / n$.

Hartley and Rao (1968) also showed that the scale-load approach provides a systematic method of finding MLE of the mean $\bar{Y}$ in the presence of known population information on an auxiliary variable $x$, in particular the mean $\bar{X}$. Letting the scale points of $x$ as $x_j^*(j = 1,..., J)$, and the scale loads of $(y_t^*, x_j^*)$ as $N_{tj}$ so that $\bar{Y} = \sum\sum p_{tj} y_t^*$ and $\bar{X} = \sum\sum p_{tj} x_j^*$, where $p_{tj} = N_{tj} / N$. Maximisation of the multinomial log-likelihood subject to the constraint that $\bar{X}$ is known leads to the MLE of $\bar{Y}$ which is asymptotically equal to the customary linear regression estimator of $\bar{Y}$.

Under stratified random sampling, strata are regarded as separate populations, each described by its separate set of parameters, that is, an additional subscript $h$ is used to index the strata, and the strata labels $h$ are regarded as informative because of known strata differences. As a result, the likelihood is a product of multinomial distributions assuming negligible sampling fractions within strata, and the MLE of $\bar{Y}$ is the usual stratified mean in the absence of auxiliary information. Hartley and Rao (1969) studied probability proportional to size (PPS) sampling with replacement, where $y_i$ is approximately proportional to the size $x_i$. Under the latter assumption, it is reasonable to consider the scale points of $r_i = y_i / x_i$, say $r_t^*$, and the resulting MLE of the total $Y$ is equal to the customary unbiased estimator in PPS sampling with replacement.

## 3. EMPIRICAL LIKELIHOOD APPROACH

Owen (1988) considered independent and identically distributed observations $y_1,..., y_n$ from some distribution $F(.)$. A non-parametric (or empirical) likelihood puts masses $p_i = \Pr(y = y_i)$ at the sample points and the log-likelihood is $l(F) = \sum \log(p_i)$. Maximising $l(F)$ under the constraints $p_i > 0$ and $\sum p_i = 1$ leads to the MLE of $p_i$ as $\hat{p}_i = 1/n$

and the MLE of the mean $\mu = E(y)$ as $\hat{\mu} = \sum \hat{p}_i y_i = \bar{y}$, the sample mean. Note that $l(F)$ is equivalent to the multinomial log-likelihood of the scale-load approach.

Chen and Qin (1993) extended Owen's EL approach to the case of known auxiliary information of the form $E\{w(x)\} = 0$, assuming simple random sampling with replacement. They considered parameters of the form $\theta = N^{-1}\sum g(y_i)$ for specified $g(.)$. In the special case of $w(x) = x - \bar{X}$ and $g(y) = y$, their results are equivalent to those of Hartley and Rao (1968) for estimating the mean $\bar{Y}$. By letting $g(y_i)$ be the indicator function $I(y_i \le t)$ for fixed $t$, we get the MLE of the population distribution function as $\tilde{F}(t) = \sum_{i \in s} \tilde{p}_i I(y_i \le t)$, where $\tilde{p}_i$ is the MLE of $p_i$. The estimator $\tilde{F}(t)$ is non-decreasing in $t$ and it can be used to obtain MLE of population quantiles, in particular the population median.

A major advantage of the EL approach is that it provides non-parametric confidence intervals on parameters of interest, similar to the parametric likelihood ratio intervals. For the mean $\mu$, we obtain the profile empirical likelihood ratio function $R(\mu)$ by maximising $\prod(np_i)$ under the constraints $\sum p_i = 1$ and $\sum p_i y_i = \mu$. Noting that $r(\mu) = -2\log R(\mu)$ is asymptotically $\chi^2$ with one degree of freedom, the $1-\alpha$ level EL interval is then given by $\{\mu \mid r(\mu) \le \chi_1^2(\alpha)\}$, where $\chi_1^2(\alpha)$ is the upper $\alpha-$point of $\chi^2$ distribution with one degree of freedom. The shape and orientation of the EL intervals are determined entirely by the data, and the intervals are range preserving and transformation respecting. Unlike normal theory confidence intervals, EL intervals do no require the evaluation of standard errors of estimators and are particularly useful if balanced tail error rates are needed. Chen, Chen and Rao (2003) obtained EL intervals on the population mean for populations containing many zero values. Such populations are encountered in audit sampling, where $y$ denotes the amount of money owed to the government and $\bar{Y}$ is the average amount of excessive claims. Previous work on audit sampling used parametric likelihood ratio intervals based on parametric mixture distributions for the variable $y$. Such intervals perform better than the standard normal theory intervals in terms of coverage, but EL intervals perform better under deviations from the assumed mixture model, by providing non-coverage rate below lower bound closer to the nominal value and also larger lower bound.

## 4. STRATIFIED RANDOM SAMPLING

Zhong and Rao (1996, 2000) studied EL inference under stratified random sampling with separate index for each stratum $h$. In this case, the log-likelihood $l(\underset{\sim}{p}) = l(\underset{\sim 1}{p},...,\underset{\sim L}{p}) = \sum_h \sum_i \log p_{hi}$, $h = 1,...,L$ and $i = 1,...,n_h$ assuming negligible sampling fractions within strata, where $n_h$ is the sample size in stratum $h$, $\underset{\sim h}{p} = (p_{h1},...,p_{hn_h})^T$ and $\sum_i p_{hi} = 1$ for each $h$. Now suppose only the overall mean $\bar{X}$ of the auxiliary variable $x$ is known. Then the MLE of the population mean $\bar{Y}$ is given by $\sum_h \sum_i \tilde{p}_{hi} y_{hi}$, where the $\tilde{p}_{hi}$ are obtained by maximising the log-likelihood under the constraints $p_{hi} > 0$ and $\sum_h \sum_i p_{hi} x_{hi} = \bar{X}$. Zhong and Rao have shown that the MLE is asymptotically equivalent to an optimal linear regression estimator that is known to have good conditional design –based properties (Rao, 1994). The MLE of the distribution function is given by $\sum_h \sum_i \tilde{p}_{hi} I(y_{hi} \le t)$. It is non-decreasing in $t$ and it can be used to obtain the MLE of quantiles. Wu (2004a) has given an algorithm for computing the estimators $\tilde{p}_{hi}$.

Zhong and Rao also studied EL intervals on the population mean. The empirical log-likelihood ratio statistic is given by $r(\bar{Y}) = -2\{l(\underset{\sim}{\breve{p}}) - l(\underset{\sim}{\tilde{p}})\}$, where $\underset{\sim}{\breve{p}}$ is the MLE of $\underset{\sim}{p}$ under the previous constraints and the additional constraint

$\sum_h \sum_i p_{hi} y_{hi} = \bar{Y}$. Zhong and Rao adjusted the empirical log-likelihood ratio statistic to account for within strata sampling fractions and showed that the adjusted statistic is asymptotically $\chi^2$ with one degree of freedom. The adjustment factor reduces to $1 - n/N$ in the special case of proportional sample allocation to the strata.

## 5. PSEUDO EMPIRICAL LIKELIHOOD APPROACH

It is difficult to obtain an informative empirical likelihood under general sampling designs. Because of this difficulty, Chen and Sitter (1999) proposed an alternative approach based on a pseudo empirical likelihood function. The finite population is assumed to be a random sample from an infinite super-population, leading to the "census" empirical log-likelihood $\sum_{i \in U} \log(p_i)$. The Horvitz-Thompson (HT) estimator $\hat{l}(\underset{\sim}{p}) = \sum_{i \in s} d_i \log(p_i)$ of the census empirical log-likelihood is then used as a pseudo empirical log-likelihood, where $d_i = \pi_i^{-1}$ and $\pi_i$ is the inclusion probability for the unit $i$. Maximising the pseudo empirical log-likelihood subject to $p_i \geq 0$ and $\sum_{i \in s} p_i = 1$ leads to the psuedo MLE of the total $Y$ as $\sum_{i \in s} \hat{p}_i y_i$. It is equal to the Hajek estimator $\hat{Y}_H = N(\sum_{i \in s} d_i)^{-1}(\sum_{i \in s} d_i y_i)$ and it is significantly less efficient than the HT estimator $\hat{Y}_{HT} = \sum_{i \in s} d_i y_i$ under PPS sampling without replacement with $\pi_i$ proportional to the size $x_i$, when $y_i$ is approximately proportional to $x_i$. Note that the Harltey-Rao approach for PPS sampling with replacement based on the scale loads of the ratios $y_i / x_i$ led to the customary PPS estimator of $Y$. It would be useful to develop a similar approach under the empirical likelihood set-up.

When the population mean $\bar{X}$ of an auxiliary variable $x$ is known, the pseudo-MLE is obtained by minimising the pseudo empirical log-likelihood subject to the previous constraints on the $p_i$'s and the additional constraint $\sum_{i \in s} p_i x_i = \bar{X}$. Chen and Sitter (1999) have shown that the pseudo-MLE of the mean $\bar{Y}$ is asymptotically equal to a generalised regression (GREG) estimator of the mean based on the Hajek estimators of the means $\bar{Y}$ and $\bar{X}$. But the associated weights $\tilde{p}_i$ in the pseudo-MLE $\sum_{i \in s} \tilde{p}_i y_i$ are always positive unlike the weights associated with the GREG estimator. This property enables us to get pseudo-MLE of the distribution function and quantiles. Calibration to $\bar{X}$ leads to an efficient pseudo-MLE or GREG estimator of $\bar{Y}$ under an implicit linear regression model with mean $x_i'\beta$, but the same estimator will be inefficient if the regression relationship is non-linear, as in the case of a binary variable $y$ in which case a logistic model is more relevant. Let the model expectation of $y_i$ be $\mu_i = h(x_i'\beta)$, then it is more efficient to use the constraint $\sum_{i \in s} p_i \hat{\mu}_i = N^{-1} \sum_{i \in U} \hat{\mu}_i$, where $\hat{\mu}_i = h(x_i'\hat{\beta})$ is the predicted value of $y_i$ under the model and $\hat{\beta}$ is an estimator of the model parameter $\beta$. Wu and Sitter (2001) obtained the pseudo-MLE of $\bar{Y}$ under the above constraint, and named it model-calibrated pseudo-MLE. Note that the constraint requires the knowledge of the individual population values $x_1, ..., x_N$ unless $h(a) = a$ which gives the previous calibration constraint $\sum_{i \in s} p_i x_i = \bar{X}$. Chen and Sitter (1999) and Chen and Wu (2002) studied pseudo-MLE of the distribution function $F_y(t)$ and quantiles and obtained model-calibrated pseudo-MLE, using the predicted values of the indicator variables $I(y_i \leq t)$ under the model for $y_i$.

Wu and Rao (2004a) proposed an alternative pseudo empirical log-likelihood function given by $\tilde{l}(p) = n^* \sum_{i \in s} \tilde{d}_i \log(p_i)$, where $\tilde{d}_i = d_i / \sum_{i \in s} d_i$ are the normalised design weights and $n^*$ is the "effective sample size" taken as n/(estimated design effect). For simple random sampling, $\tilde{l}(p)$ reduces to the usual empirical likelihood function $\sum_{i \in s} \log(p_i)$. The pseudo-MLE under the alternative formulation remains the same as the pseudo-MLE of Chen and Sitter (1999), but the resulting pseudo empirical log-likelihood ratio statistic for getting confidence intervals on the

population mean $\overline{Y}$ is asymptotically $\chi^2$ with one degree of freedom , unlike the pseudo empirical log-likelihood ratio statistic based on $\hat{l}(p)$. The latter requires an adjustment to make it asymptotically $\chi^2$ with one degree of freedom. Wu (2004b) has given R/S-PLUS codes for implementing the pseudo-EL methods under PPS sampling without replacement.

Wu (2004c) has developed a pseudo-EL approach that attempts to combine information from two independent surveys from the same population with some common variables of interest. This method ensures consistency between the surveys over the common variables in the sense that the estimators from the two surveys are identical. Wu and Rao (2004b) have studied more efficient methods of combining information from two surveys using the pseudo-EL approach. Singh and Wu (2003) proposed an extension of generalised regression estimator for integrating information from two independent surveys from the same population as well as from dual frame surveys.

## REFERENCES


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

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.

Chen, J. and Wu, C. (2002). "Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method". *Statistica Sinica,* **12**, 1223-1239.

Chen, J., Chen, S.Y. and Rao, J.N.K. (2003). "Empirical likelihood confidence intervals for the mean of a population containing many zero values". *The Canadian Journal of Statistics*, **31**, 53-68.

Godambe, V.P (1996). "A unified theory of sampling from finite populations". *Journal of the Royal Statistical Society, Series B,* **28**, 310-328.

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

Hartley, H.O. and Rao, J.N.K. (1969). "A new estimation theory for sample surveys II". In *New Developments in Survey Sampling,* eds. N.L. Johnson and H.Smith, New York: Wiley Inter-Science, 147-169.

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

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

Rao, J.N.K. (1994). "Estimating totals and distribution functions using auxiliary information at the estimation stage". *Journal of Official Statistics*, **10**, 153-165.

Rao, J.N.K. and Wu, C. (2004b). "Pseudo empirical likelihood inference for multiple surveys and dual frame surveys". Paper under preparation.

Singh, A.C. and Wu, S. (2003). "An extension of generalized regression estimator to dual frame surveys". *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 3911-3918.

Wu, C. (2004a). "Some algorithmic aspects of the empirical likelihood method in survey sampling". *Statistica Sinica*, **14**, 1057-1067.

Wu, C. (2004b). "R/S-PLUS implementation of pseudo empirical likelihood methods under unequal probability sampling". Working Paper 2004-07, Department of Statistics and Actuarial Science, University of Waterloo.

Wu, C. (2004c). “Combining information from multiple surveys through empirical likelihood method”. *The Canadian Journal of Statistics*, **34**, 15-26.

Wu, C. and Sitter, R.R. (2001). “A model-calibration approach to using complete auxiliary information from survey data”. *Journal of the American Statistical Association*, **96**, 185-193.

Wu, C. and Rao, J.N.K. (2004a). “Empirical likelihood ratio confidence intervals for complex surveys”. Submitted for publication.

Wu, C. and Rao, J.N.K. (2004b). “Pseudo empirical likelihood inference for multiple surveys and dual frame surveys”. Paper under preparation.

Zhong, C.X. Bob and Rao, J.N.K. (1996). “Empirical likelihood inference under stratified random sampling using auxiliary information”. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp.798-803.

Zhong, Bob and Rao, J.N.K. (2001). “Empirical likelihood inference under stratified sampling using auxiliary population informaton”. *Biometrika*, **87**, 929-938.