Journal of the Royal Statistical Society
SERIES B Statistical Methodology

*J. R. Statist. Soc.* B (2010)
**72**, *Part* 4, *pp.* 533–544

# Bayesian pseudo-empirical-likelihood intervals for complex surveys

J. N. K. Rao

*Carleton University, Canada*

and Changbao Wu

*University of Waterloo, Canada*

[Received June 2009. Revised February 2010]

**Summary.** Bayesian methods for inference on finite population means and other parameters by using sample survey data face hurdles in all three phases of the inferential procedure: the formulation of a likelihood function, the choice of a prior distribution and the validity of posterior inferences under the design-based frequentist framework. In the case of independent and identically distributed observations, the profile empirical likelihood function of the mean and a non-informative prior on the mean can be used as the basis for inference on the mean and the resulting Bayesian empirical likelihood intervals are also asymptotically valid under the frequentist set-up. For complex survey data, we show that a pseudo-empirical-likelihood approach can be used to construct Bayesian pseudo-empirical-likelihood intervals that are asymptotically valid under the design-based set-up. The approach proposed compares favourably with a full Bayesian analysis under simple random sampling without replacement. It is also valid under general single-stage unequal probability sampling designs, unlike a full Bayesian analysis. Moreover, the approach is very flexible in using auxiliary population information and can accommodate two scenarios which are practically important: incorporation of known auxiliary population information for the construction of intervals by using the basic design weights; calculation of intervals by using calibration weights based on known auxiliary population means or totals.

*Keywords*: Auxiliary information; Bayesian interval; Calibration weight; Design effect; Non-informative prior; Pseudo-empirical-likelihood; Unequal probability sampling

## 1. Introduction

The design-based frequentist approach to inference from survey samples is widely used in survey practice because of its non-parametric nature and its ability to handle complex design features. This approach uses estimators of finite population parameters, such as totals and means, that are design consistent for large samples. To achieve increased efficiency, model-assisted design consistent estimators, based on working models, are often used. Further, the estimators and associated design consistent variance estimators are used to construct confidence intervals on the parameters, on the basis of the normal approximation. Such intervals are asymptotically valid 'whatever the unknown properties of the population' (Neyman, 1934).

Attempts have also been made to integrate sample survey theory with mainstream statistical inference via the likelihood function. Godambe (1966a, b) obtained the likelihood function for the full sample data $\{(i, y_i), i \in s\}$ and showed that it provides no information on the non-observed values and hence on the finite population parameters, where $s$ denotes the sample

*Address for correspondence*: Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada.
E-mail: cbwu@uwaterloo.ca

 1369–7412/10/72533

and $i$ denotes labels that are attached to the population units with associated values of interest $y_i$. This feature is due to the label property that treats the $N$ population units as essentially $N$ different post strata. One way out of this difficulty is to take a Bayesian route by assuming an informative (exchangeable) prior on the $N$-dimensional parameter vector $(y_1, \ldots, y_N)'$ and to combine it with the Godambe likelihood (Ericson, 1969). The resulting posterior distribution is informative, but inferences do not depend on how the sample was chosen. Ericson (1969) focused on the posterior mean and the posterior variance of the finite population mean $\bar{Y}$. In stratified sampling, exchangeability is assumed within strata. Meeden and Vardeman (1991) proposed a non-informative Bayesian approach to interval estimation, assuming that 'the unseen are like the seen' (which is equivalent to exchangeability). This approach leads to a Polya urn distribution as a pseudoposterior distribution over the unobserved population values $y_i$. The Polya posterior is based on a stepwise Bayesian argument and it is not a true Bayesian posterior based on a single prior. Hence, it is called a pseudoposterior, but it can be used in the customarily Bayesian manner for point and interval estimation. The Polya posterior is a flexible tool with reasonable design-based frequentist properties under simple random sampling, and it permits Bayesian interval estimation through simulation of finite populations from the Polya posterior. The Polya posterior is equivalent to the finite population Bayesian bootstrap of Lo (1988); see also Rubin (1981). Meeden (1995) extended the Polya posterior approach to utilize auxiliary population information $(x_1, \ldots, x_N)'$ by assuming exchangeability of the ratios $r_i = y_i/x_i$. Lazar *et al.* (2008) developed a constrained Polya posterior to generate simulated populations that are consistent with the known population mean of an auxiliary variable, using Markov chain Monte Carlo methods. A limitation of the Polya posterior approach is that it requires an exchangeability assumption and it is not clear whether the approach can be extended to general single-stage unequal probability sampling designs such that the resulting Bayesian intervals are also asymptotically valid under the design-based framework.

Hartley and Rao (1968) took an alternative design-based route that yields informative non-parametric likelihood functions by ignoring certain aspects of the data depending on the situation at hand. For example, under simple random sampling in the absence of information relating the label $i$ to $y_i$, the sample labels are suppressed and the likelihood function based on $\{y_i, i \in s\}$ is obtained by using a scale load approach that assumes a finite set of possible values for the variable of interest $y$. This likelihood function is identical to the well-known empirical likelihood (EL) of Owen (1988) when the sampling fraction is negligible. Hartley and Rao (1968) also studied Bayesian inference using a non-informative compound multinomial prior (Hoadley, 1969) under simple random sampling, leading to a posterior mean and posterior variance similar to Ericson's (1969). They also studied stratified random sampling by ignoring labels within strata but retaining strata labels because of known strata differences. Under simple random sampling and stratified simple random sampling, Aitkin (2008) applied the compound multinomial prior to the Hartley–Rao likelihood, using the improper Dirichlet–Haldane prior only on the components corresponding to observed non-zero sample scale loads. He showed through simulation that the resulting Bayesian intervals have good frequentist properties. However, Aitkin's full Bayesian analysis and some extensions are all restricted to scenarios where the Hartley–Rao likelihood is available.

The Hartley–Rao likelihood approach runs into difficulties in the case of complex sampling designs involving design features such as unequal probabilities of selection and clustering. In this paper, we attempt to overcome those difficulties by using the pseudo-EL method for complex designs (Chen and Sitter, 1999; Wu and Rao, 2006). We show that the profile pseudo-EL on the finite population mean can be used as the basis for Bayesian inference on the mean under complex designs by combining it with a non-informative prior on the mean. In particular,

we obtain Bayesian pseudo-EL intervals on the mean that are also asymptotically valid in the design-based framework. Lazar (2003) first proposed the Bayesian EL approach in the case of independent and identically distributed observations. Informative priors on the mean, constructed from prior information, could also be used to obtain more efficient inferences. This prior information is different from the population auxiliary information that can be used in deriving the profile pseudo-EL; see Section 3.1. As a result, the approach proposed can be a useful tool for handling both the auxiliary information and the prior information. We also develop Bayesian pseudo-EL intervals based on the pseudo-EL itself by using a non-informative Dirichlet prior and we show that the intervals are asymptotically valid under the design set-up. Our proposed approach compares favourably with a full Bayesian analysis under simple random sampling without replacement. It is also valid under general single-stage unequal probability sampling designs.

In Section 2, we briefly discuss the Bayesian EL method for independent and identically distributed observations. In Section 3, we present Bayesian pseudo-EL methods for complex survey data and examine their asymptotic properties under the design-based framework. Some related computational aspects are discussed in Section 4. Results of a simulation study are reported in Section 5. We conclude with some remarks in Section 6. Proofs are given in Appendix A.

## 2. Bayesian empirical likelihood

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be independent and identically distributed observations on a random variable $Y$, and let $\theta = E(Y)$ be the population mean. The profile EL of $\theta$ is then obtained by maximizing $\Sigma_{i=1}^n \log(p_i)$ subject to $\Sigma_{i=1}^n p_i = 1$ and $\Sigma_{i=1}^n p_i(y_i - \theta) = 0$. We use the method of Lagrange multipliers by setting the partial derivative of

$$G = \sum_{i=1}^{n} \log(p_i) - n\lambda \sum_{i=1}^{n} p_i(y_i - \theta) - \gamma\left(\sum_{i=1}^{n} p_i - 1\right)$$

with respect to $p_i$ equal to 0, which leads to the profile-EL function given by $L_{\mathrm{EL}}(\theta) = \exp\{l_{\mathrm{EL}}(\theta)\}$, where

$$l_{\mathrm{EL}}(\theta) = -n\log(n) - \sum_{i=1}^{n} \log\{1 + \lambda(\theta)(y_i - \theta)\} \tag{1}$$

and $\lambda = \lambda(\theta)$ solves

$$\sum_{i=1}^{n} (y_i - \theta)/\{1 + \lambda(y_i - \theta)\} = 0.$$

We combine $L_{\mathrm{EL}}(\theta)$ with a specified prior $\pi(\theta)$ on $\theta$ via the Bayes theorem to obtain a pseudo-posterior

$$\pi(\theta|\mathbf{y}) = c(\mathbf{y}) \exp\left[\log\{\pi(\theta)\} - \sum_{i=1}^{n} \log\{1 + \lambda(\theta)(y_i - \theta)\}\right], \tag{2}$$

where $c(\mathbf{y})$ is the normalizing constant such that $\int \pi(\theta|\mathbf{y})\,\mathrm{d}\theta = 1$. The $\alpha$th quantile $t_\alpha$ of $\pi(\theta|\mathbf{y})$ is determined by $H(t_\alpha|\mathbf{y}) = \alpha$ where

$$H(t|\mathbf{y}) = \int_{-\infty}^{t} \pi(\theta|\mathbf{y})\,\mathrm{d}\theta. \tag{3}$$

The Bayesian EL intervals based on equation (2) should be more robust than those based on parametric approaches because EL is a non-parametric likelihood. The validity of posterior inferences based on equation (2) can be assessed through coverage probabilities of intervals $(t_{\alpha_1}, t_{\alpha_2})$ on $\theta$. This can be done through Monte Carlo simulations for selected values of $\alpha_1$ and

$\alpha_2$. Monahan and Boos (1992) proposed that $\pi(\theta|\mathbf{y})$ should be viewed as a valid posterior if $H = H(\theta|\mathbf{y})$ is uniformly distributed over (0,1). Lazar (2003) provided further discussion on the validity of $L_{\mathrm{EL}}(\theta)$ for posterior inferences and on the frequentist properties of the Bayesian EL intervals. Fang and Mukerjee (2006) showed that the profile EL does not admit a probability matching prior $\pi(\theta)$ that ensures frequentist validity of Bayesian EL intervals with margin of error $o(n^{-1})$.

## 3. Bayesian pseudo-empirical-likelihood for complex surveys

Suppose that the finite population consists of $N$ units. Associated with the $i$th unit is the value $y_i$ of the study variable $y$ and the value $\mathbf{x}_i$ of the vector of auxiliary variables $\mathbf{x}$. We assume that the population mean vector $\bar{\mathbf{X}} = N^{-1}\Sigma_{i=1}^{N}\mathbf{x}_i$ is known. Let $\bar{Y} = N^{-1}\Sigma_{i=1}^{N} y_i$ be the unknown population mean of interest. Under the Bayesian framework, we shall use $\theta$ to denote a generic finite population parameter, with particular reference to $\bar{Y}$. Let $\{(i, y_i, \mathbf{x}_i), i \in s\}$ be a survey sample from the finite population, where $s$ is the set of $n$ units included in the sample. Let $\pi_i = P(i \in s)$, $i = 1, \ldots, N$, be the first-order inclusion probabilities.

### 3.1. Unequal probability sampling with basic design weights

We now consider a general unequal probability sampling design. Let $d_i = 1/\pi_i$, $i \in s$, be the basic design weights and $\tilde{d}_i(s) = d_i/\Sigma_{i\in s} d_i$ be the normalized design weights for the given sample $s$. We also assume that auxiliary population information on $\mathbf{x}$ is available and needs to be incorporated in the inference. Without loss of generality, we assume that $\bar{\mathbf{X}} = \mathbf{0}$ in the following discussion. Otherwise we replace $\mathbf{x}_i$ by $\mathbf{x}_i - \bar{\mathbf{X}}$. Let $n^*$ be the effective sample size defined as $n^* = n/\mathrm{deff}_{\mathrm{GR}}$, where $n$ is the sample size and $\mathrm{deff}_{\mathrm{GR}}$ is the estimated design effect deff based on the generalized regression (GR) estimator. The design effect that is associated with an estimator is defined as the ratio of the variances under the given design and under simple random sampling with replacement. Wu and Rao (2006) have given a formula for $\mathrm{deff}_{\mathrm{GR}}$ and its estimator. It may be noted that the design effect is estimator specific, but under simple random sampling without replacement it is equal to $1 - n/N$, regardless of the estimator.

A pseudo-EL method for complex survey data was first proposed by Chen and Sitter (1999). The profile pseudo-empirical log-likelihood function for the population mean $\theta$, which was defined by Wu and Rao (2006), is given by

$$l_{\mathrm{PEL}}(\theta) = n^* \sum_{i\in s} \tilde{d}_i(s) \log\{\hat{p}_i(\theta)\}, \tag{4}$$

where the $\hat{p}_i(\theta)$ maximize $\Sigma_{i\in s}\tilde{d}_i(s)\log(p_i)$ subject to $p_i > 0$, $\Sigma_{i\in s} p_i = 1$ and

$$\sum_{i\in s} p_i y_i = \theta, \tag{5}$$

$$\sum_{i\in s} p_i \mathbf{x}_i = \mathbf{0}. \tag{6}$$

Note that $\hat{p}_i(\theta)$ depends on $\theta$ through the use of constraint (5). Let $\mathbf{u}_i = (y_i - \theta, \mathbf{x}_i')'$. It can be shown that $\hat{p}_i(\theta) = \tilde{d}_i(s)/\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\}$, where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\theta)$ is the solution to

$$g(\boldsymbol{\lambda}) = \sum_{i\in s} \tilde{d}_i(s)\mathbf{u}_i/(1 + \boldsymbol{\lambda}'\mathbf{u}_i) = \mathbf{0}.$$

It may be noted that the sampling design features are taken into account by the profile pseudo-EL (4) because it depends on the design weights $d_i$, $i \in s$, as well as on the design effect. Under

the non-informative prior $\pi(\theta) \propto 1$, the posterior density function of $\theta$ under the pseudo-EL is given by

$$\pi(\theta|\mathbf{y}, \mathbf{x}) = c(\mathbf{y}, \mathbf{x}) \exp\left[-n^* \sum_{i \in s} \tilde{d}_i(s) \log\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\}\right], \tag{7}$$

where $c(\mathbf{y}, \mathbf{x})$ is the normalizing constant such that $\int \pi(\theta|\mathbf{y}, \mathbf{x})\, \mathrm{d}\theta = 1$. Assuming that $\mathbf{0}$ is an inner point of the convex hull formed by $\{\mathbf{x}_i, i \in s\}$, the support for the pseudoposterior density (7) is $\min_{i \in s}(y_i) < \theta < \max_{i \in s}(y_i)$.

For the special case of simple random sampling without replacement, $\pi(\theta) \propto 1$ and no auxiliary information on $\mathbf{x}$, equation (7) reduces to

$$\pi(\theta|\mathbf{y}) \propto \exp\left[-\left(1 - \frac{n}{N}\right)^{-1} \sum_{i \in s} \log\{1 + \lambda(\theta)(y_i - \theta)\}\right], \tag{8}$$

which is approximately equal to expression (2) when the sampling fraction $n/N$ is small. A limited simulation study on the design-based coverage probability of the 95% pseudoposterior intervals based on expressions (2) and (8) with equal tail probabilities (2.5%) showed that the interval based on equation (2) performs quite well when $n/N = 0.05$, but it exhibits a clear trend of overcoverage as $n$ increases (for a fixed $N$). In contrast, the interval that is based on expression (8) performs well in tracking the nominal coverage probability 0.95 for all sample sizes. The average length of this interval over simulations is also shorter than the corresponding average length of the interval that is based on equation (2). We have also compared the design-based performance of the intervals based on expression (8) with the intervals that are based on a full Bayesian analysis proposed by Aitkin (2008) using the Dirichlet–Haldane prior. Our simulation results indicated that the interval proposed compares favourably with Aitkin's interval in terms of coverage probabilities and average length.

Let $\tilde{p}_i$ be the maximizer of $\Sigma_{i \in s} \tilde{d}_i(s) \log(p_i)$ subject to $p_i > 0$, $\Sigma_{i \in s} p_i = 1$ and the calibration constraint (6). Then $\hat{\theta}_{\mathrm{PEL}} = \Sigma_{i \in s} \tilde{p}_i y_i$ is the maximum pseudo-EL estimator of $\theta$. Let $\sigma^2_{\mathrm{PEL}} = V(\hat{\theta}_{\mathrm{PEL}})$ be the design-based variance of $\hat{\theta}_{\mathrm{PEL}}$.

*Theorem 1.* Under suitable regularity conditions, the pseudoposterior distribution of $\theta$ under the pseudo-EL and the non-informative prior $\pi(\theta) \propto 1$ is asymptotically normal, with mean $\hat{\theta}_{\mathrm{PEL}}$ and variance $\sigma^2_{\mathrm{PEL}}$.

An important consequence of theorem 1 is that the pseudoposterior intervals based on equation (7) will have asymptotically correct coverage probability under the design-based framework. The equal tail Bayesian pseudo-EL intervals and the highest pseudoposterior density intervals are both asymptotically equivalent to the normal approximation interval that is based on $\hat{\theta}_{\mathrm{PEL}}$ and $\sigma^2_{\mathrm{PEL}}$.

The choice (4) for the pseudo-EL is not necessarily unique or the best. For example, Kim (2009) proposed an alternative pseudo-EL. It would be useful to compare the performance of Bayesian pseudo-EL intervals that are obtained from different pseudo-EL functions.

### 3.2. Unequal probability sampling with calibration weights

It is a routine practice among statistical agencies to release public use survey data with calibration weights $w_i$. An estimator in the form of $\hat{Y}_{\mathrm{c}} = \Sigma_{i \in s} w_i y_i$ would be equivalent to the GR estimator of the population total, $N\bar{Y}$, and it is calibrated in the sense of $\Sigma_{i \in s} w_i \mathbf{x}_i = \mathbf{X}$, where $\mathbf{X} = N\bar{\mathbf{X}}$ is the known population total of the auxiliary variables $\mathbf{x}$. The calibration estimator $\hat{Y}_{\mathrm{c}}$ is typically more efficient than the unbiased estimator $\hat{Y} = \Sigma_{i \in s} d_i y_i$. The question of interest

here is whether the calibration weights can be used to define a version of the pseudo-EL so that Bayesian pseudo-EL intervals are still asymptotically valid under the sampling design.

Let $\tilde{w}_i(s) = w_i / \Sigma_{i \in s} w_i$ be the normalized calibration weights and $\hat{\theta}_c = \Sigma_{i \in s} \tilde{w}_i(s) y_i$ be the calibration estimator of the population mean, $\bar{Y}$. Consider the following pseudoempirical log-likelihood function

$$l_{\mathrm{PEL}}(\theta) = m \sum_{i \in s} \tilde{w}_i(s) \log\{p_i(\theta)\}, \tag{9}$$

where $m$ is a scale factor to be determined, and $p_i(\theta)$ maximize $\Sigma_{i \in s} \tilde{w}_i(s) \log(p_i)$ subject to $p_i > 0$, $\Sigma_{i \in s} p_i = 1$ and constraint (5). Note that the calibration equation (6) is no longer involved here in defining $l_{\mathrm{PEL}}(\theta)$. Now, combining equation (9) with the non-informative prior $\pi(\theta) \propto 1$, the pseudoposterior density is given by

$$\pi(\theta|\mathbf{y}) \propto \exp\left[-m \sum_{i \in s} \tilde{w}_i(s) \log\{1 + \lambda(\theta)(y_i - \theta)\}\right], \tag{10}$$

where $\lambda = \lambda(\theta)$ is the solution to

$$\sum_{i \in s} \tilde{w}_i(s)(y_i - \theta)/\{1 + \lambda(y_i - \theta)\} = 0.$$

Using similar arguments to those in the proof of theorem 1, we can show that

$$-m \sum_{i \in s} \tilde{w}_i(s) \log\{1 + \lambda(\theta)(y_i - \theta)\} = -\frac{m}{2} \left\{\theta - \sum_{i \in s} \tilde{w}_i(s) y_i\right\}^2 \Big/ \sum_{i \in s} \tilde{w}_i(s)(y_i - \theta)^2 + o_p\left(\frac{m}{n}\right).$$

Note that $\hat{\theta}_c = \Sigma_{i \in s} \tilde{w}_i(s) y_i = \hat{\theta}_{\mathrm{PEL}} + o_p(1/\sqrt{n})$ and

$$\sum_{i \in s} \tilde{w}_i(s)(y_i - \theta)^2 = S_y^2 + O_p(1/\sqrt{n})$$

for any $\theta$ such that $\theta = \bar{Y} + O(1/\sqrt{n})$, where $S_y^2 = (N-1)^{-1} \Sigma_{i=1}^N (y_i - \bar{Y})^2$ is the finite population variance for the $y$-variable. It is now clear that the pseudoposterior distribution of $\theta$ is asymptotically normal with mean $\hat{\theta}_c$ and variance $S_y^2/m$. It is also clear that, to have Bayesian intervals that are asymptotically valid under the design-based framework, we need to choose $m$ as a design consistent estimator of $S_y^2 / V(\hat{\theta}_c)$. This is different from the effective sample size $n^*$ that was defined in Section 3.1.

### 3.3. Bayesian pseudo-empirical-likelihood based on $(p_1, \ldots, p_n)$

The methods that were presented in Sections 3.1 and 3.2 are based on the pseudoposterior distribution of $\theta$ under a non-informative prior on $\theta$. Unequal probability sampling designs are handled through the use of design weights $d_i$ or calibration weights $w_i$ as well as the design effect. The starting point is the profile pseudo-empirical log-likelihood function as defined by equation (4). An alternative approach for Bayesian pseudo-EL inference is to treat $(p_1, \ldots, p_n)$ as unknown parameters and to define the pseudo-EL function as $L_{\mathrm{PEL}}(p_1, \ldots, p_n) = \exp\{l_{\mathrm{PEL}}(\theta)\}$. This leads to

$$L_{\mathrm{PEL}}(p_1, \ldots, p_n) = \prod_{i \in s} p_i^{\gamma_i}, \tag{11}$$

where $\gamma_i = n^* \tilde{d}_i(s)$. This is not a genuine likelihood but can be used to derive a pseudoposterior distribution on $(p_1, \ldots, p_n)$. Consider the Dirichlet prior $D(\alpha_1, \ldots, \alpha_n)$ on $(p_1, \ldots, p_n)$:

$$\pi(p_1, \ldots, p_n) = c(\alpha_1, \ldots, \alpha_n) \prod_{i \in s} p_i^{\alpha_i - 1},$$

where $c(\alpha_1, \ldots, \alpha_n) = \Gamma(\Sigma_{i \in s} \alpha_i) / \Pi_{i \in s} \Gamma(\alpha_i)$, the posterior distribution of $(p_1, \ldots, p_n)$ given the sample $s$, is Dirichlet $D(\gamma_1 + \alpha_1, \ldots, \gamma_n + \alpha_n)$ and is given by

$$\pi(p_1, \ldots, p_n | s) = c(\gamma_1 + \alpha_1, \ldots, \gamma_n + \alpha_n) \prod_{i \in s} p_i^{\gamma_i + \alpha_i - 1}. \tag{12}$$

A particular choice of the Dirichlet prior is the improper Dirichlet–Haldane prior corresponding to $\alpha_i = 0$ for all $i \in s$; see Aitkin (2008). The posterior distribution of the population mean is the distribution of $\theta = \Sigma_{i \in s} p_i y_i$ with $(p_1, \ldots, p_n)$ following the Dirichlet distribution that is specified by equation (12). The Dirichlet distribution can be approximated by Monte Carlo simulation methods and so are the pseudoposterior inferences that are based on equation (12). Auxiliary information can also be incorporated, with the design weights $d_i$ replaced by the calibration weights $w_i$. We spell out the details for two general cases.

(a) *Unequal probability sampling without auxiliary information*: in this case $\gamma_i = n^* \tilde{d}_i(s)$, $n^* = \hat{n}/\text{deff}_{\text{H}}$ and the estimated design effect $\text{deff}_{\text{H}}$ is associated with the Hájek estimator $\hat{\bar{Y}}_{\text{H}} = \Sigma_{i \in s} \tilde{d}_i(s) y_i$ of $\bar{Y}$.

(b) *Unequal probability sampling with auxiliary information*: when auxiliary information is available, the design weights $d_i$ may be replaced by the calibration weights $w_i$ under appropriate benchmark constraints. Following the results of Section 3.2, we let $\gamma_i = m\{\tilde{w}_i(s)\}$ where $m$ is the estimated value of $S_{\hat{y}}^2 / V(\hat{\theta}_{\text{c}})$. The posterior distribution of $(p_1, \ldots, p_n)$ is again the Dirichlet distribution of equation (12) with the chosen $\gamma_i$.

For each of these two scenarios, the posterior mean and the posterior variance of $\theta = \Sigma_{i \in s} p_i y_i$ under the Dirichlet–Haldane prior match $\hat{\theta}_{\text{PEL}}$ and the design variance of $\hat{\theta}_{\text{PEL}}$ asymptotically. Hence, the Bayesian pseudo-EL intervals based on equation (12) are valid under the design.

*Theorem 2.* Under the improper Dirichlet–Haldane prior, the posterior distribution of $\theta = \Sigma_{i \in s} p_i y_i$ based on equation (12) provides Bayesian intervals on the finite population mean that are valid under the design-based framework.

The three different approaches that were presented in Sections 3.1–3.3 are further examined in Section 5 through a simulation study.

## 4. Computational notes

The computational procedures that are involved in making pseudoposterior inferences are quite different for $\pi(\theta | \mathbf{y}, \mathbf{x})$ given by equation (7) and for $\pi(p_1, \ldots, p_n | s)$ defined by equation (12). The first requires finding the normalizing constant $c(\mathbf{y}, \mathbf{x})$ and computing quantiles $t_\alpha$ of the posterior distribution for specified values of $\alpha$. It also involves routine calculations for finding $\lambda(\theta)$ or $\boldsymbol{\lambda}(\theta)$ for a given set of data. The second requires approximations to the Dirichlet distribution through Monte Carlo simulations.

We first illustrate the computational procedures that are to be employed in Section 5 for the pseudoposterior distribution of $\pi(\theta | \mathbf{y}, \mathbf{x})$ given by equation (7). This posterior density function has a nice unimodal shape just like the normal density, which facilitates the search for the boundary points of Bayesian pseudo-EL intervals.

*Theorem 3.* The pseudoposterior density function $\pi(\theta | \mathbf{y}, \mathbf{x})$ is a concave function of $\theta$ with the maximum value attained at $\theta = \hat{\theta}_{\text{PEL}}$.

The concavity of the posterior density function also holds when the prior distribution of $\theta$ is exponential. The maximum point of $\pi(\theta|\mathbf{y},\mathbf{x})$ occurs at $\theta=\hat{\theta}_{\mathrm{PEL}}$, and $\pi(\theta|\mathbf{y},\mathbf{x})$ diminishes to 0 quickly when $\theta$ moves away from $\hat{\theta}_{\mathrm{PEL}}$. The theoretical support of the pseudoposterior density is $(\min_{i\in s}(y_i),\ \max_{i\in s}(y_i))$. Computation of the pseudoposterior density for $\theta$ near the end points of the support can be problematic, owing to the non-existence of an empirical likelihood solution at the end points. However, a practical support $(c_1,\ c_2)$ can be used, with the end points $c_1$ and $c_2$ satisfying $\min_{i\in s}(y_i)<c_1<\hat{\theta}_{\mathrm{PEL}}<c_2<\max_{i\in s}(y_i)$, $h(c_1)<\varepsilon$ and $h(c_2)<\varepsilon$ for a prespecified tolerance $\varepsilon$, where

$$h(\theta)=\exp\left[-n^*\sum_{i\in s}\tilde{d}_i(s)\log\{1+\boldsymbol{\lambda}'(\theta)\mathbf{u}_i\}\right].$$

In the simulation studies that were reported in Section 5, we used $\varepsilon=0.000\,01$. A grid method can then be used to find $c(\mathbf{y},\mathbf{x})=\{\int_{c_1}^{c_2}h(\theta)\,\mathrm{d}\theta\}^{-1}$. Quantiles of the pseudoposterior distribution can similarly be found by grid approximations.

The pseudoposterior distribution of $(p_1,\ldots,p_n)$ is a Dirichlet distribution, which can be simulated through independent gamma random variables. Let $X_i$ be a gamma random variable with density function

$$f_i(x)=\Gamma(a_i)^{-1}x^{a_i-1}\exp(-x)$$

for $x>0$, and $X_1,\ldots,X_n$ are independent. Let $p_i=X_i/\Sigma_{i=1}^n X_i$, $i=1,\ldots,n$. Then $(p_1,\ldots,p_n)\sim D(a_1,\ldots,a_n)$. The distribution of $\theta=\Sigma_{i\in s}\,p_i y_i$ can then be approximated by simulated values $\theta^{[j]}=\Sigma_{i\in s}\,p_i^{[j]}y_i$, $j=1,\ldots,M$ for a large $M$, where $(p_1^{[j]},\ldots,p_n^{[j]})$ is generated from the Dirichlet distribution (Aitkin, 2008).

## 5. Simulation study

We consider unequal probability sampling designs with inclusion probabilities that are proportional to a size measure and compare the design-based coverage of the proposed Bayesian pseudo-EL intervals and the design-based pseudo-EL confidence intervals based on a $\chi^2$-approximation (Wu and Rao, 2006), which are denoted here as PEL. We report simulation results under model I, $y_i=\beta_0+\beta_1 z_i+\sigma\varepsilon_i$, that was used in Wu and Rao (2006). We refer the reader to Wu and Rao (2006) for details on the simulation models. Simulation results under other models can be found in an unpublished technical report (Rao and Wu, 2009). The design variable is $z$ and the Rao–Sampford unequal probability sampling method (Rao, 1965; Sampford, 1967) with $\pi_i\propto z_i$ is used. All the results that are reported here are based on $B=1000$ simulation runs. Our simulation was programmed in R and the code is available from

`http://www.blackwellpublishing.com/rss`

For the Bayesian pseudo-EL method, the interval by using the basic design weights $d_i$ (Section 3.1) is denoted by $\mathrm{BEL}_d$, the interval by using the calibration weights $w_i$ (Section 3.2) is denoted by $\mathrm{BEL}_w$ and the interval based on the posterior distribution of $(p_1,\ldots,p_n)$ (Section 3.3) is denoted by $\mathrm{BEL}_h$, where $h$ indicates the Haldane prior on $(p_1,\ldots,p_n)$. The conventional design-based normal approximation confidence intervals are not included here, since Wu and Rao (2006) have shown that the PEL intervals perform better in terms of design-based coverage properties.

Table 1 summarizes the results on design-based coverage probabilities, lower and upper tail error rates and average length of 95% intervals for the population mean. The first column gives the correlation coefficient $\rho$ between $y$ and $z$ and the third column shows whether the

**Table 1.** 95% Bayesian pseudo-EL intervals for the population mean

| $\rho$ | $n$ | *Auxiliary variable* | *Method* | *Coverage probability (%)* | *Lower tail error rate* | *Upper tail error rate* | *Average length* |
|---|---|---|---|---|---|---|---|
| 0.30 | 40 | — | PEL | 92.7 | 2.3 | 5.0 | 1.97 |
| | | | $\text{BEL}_d$ | 93.2 | 3.8 | 3.0 | 2.00 |
| | | | $\text{BEL}_h$ | 92.8 | 1.9 | 5.3 | 1.86 |
| | | $z$ | PEL | 91.4 | 2.8 | 5.8 | 1.86 |
| | | | $\text{BEL}_d$ | 92.6 | 3.4 | 4.0 | 1.88 |
| | | | $\text{BEL}_w$ | 91.9 | 3.1 | 5.0 | 1.88 |
| | | | $\text{BEL}_h$ | 91.1 | 1.9 | 7.0 | 1.76 |
| | 80 | — | PEL | 93.4 | 2.5 | 4.1 | 1.38 |
| | | | $\text{BEL}_d$ | 94.8 | 2.8 | 2.4 | 1.42 |
| | | | $\text{BEL}_h$ | 93.5 | 2.9 | 3.6 | 1.33 |
| | | $z$ | PEL | 93.7 | 2.5 | 3.8 | 1.32 |
| | | | $\text{BEL}_d$ | 94.5 | 2.5 | 3.0 | 1.35 |
| | | | $\text{BEL}_w$ | 94.5 | 2.3 | 3.2 | 1.35 |
| | | | $\text{BEL}_h$ | 92.4 | 2.3 | 5.3 | 1.27 |
| 0.80 | 40 | — | PEL | 94.7 | 2.3 | 3.0 | 0.75 |
| | | | $\text{BEL}_d$ | 93.3 | 3.1 | 3.6 | 0.75 |
| | | | $\text{BEL}_h$ | 93.5 | 2.4 | 4.1 | 0.72 |
| | | $z$ | PEL | 92.3 | 2.7 | 5.0 | 0.47 |
| | | | $\text{BEL}_d$ | 92.7 | 3.3 | 4.0 | 0.48 |
| | | | $\text{BEL}_w$ | 90.9 | 2.3 | 6.8 | 0.46 |
| | | | $\text{BEL}_h$ | 92.8 | 1.3 | 5.9 | 0.46 |
| | 80 | — | PEL | 94.6 | 1.8 | 3.6 | 0.51 |
| | | | $\text{BEL}_d$ | 95.4 | 2.7 | 1.9 | 0.52 |
| | | | $\text{BEL}_h$ | 92.6 | 3.0 | 4.4 | 0.50 |
| | | $z$ | PEL | 93.8 | 2.5 | 3.7 | 0.33 |
| | | | $\text{BEL}_d$ | 94.5 | 2.5 | 3.0 | 0.34 |
| | | | $\text{BEL}_w$ | 94.4 | 1.6 | 4.0 | 0.33 |
| | | | $\text{BEL}_h$ | 91.7 | 1.2 | 7.1 | 0.32 |

auxiliary variable $z$ is also used as a calibration variable. Major observations from Table 1 are as follows.

(a) The $\text{BEL}_d$ method intervals perform better than those from method PEL, with coverage probabilities closer to the nominal value and more balanced tail error rates.
(b) The $\text{BEL}_w$ method intervals have performance that is similar to that of method PEL when $n = 40$ but are better when $n = 80$.
(c) The three intervals PEL, $\text{BEL}_d$ and $\text{BEL}_w$ have comparable average length.
(d) The $\text{BEL}_h$ method intervals seem to have an undercoverage problem for some cases, perhaps because the average lengths are smaller for those cases. When we increased the sample size to $n = 100$ (for brevity, the results are not reported here), the performance of $\text{BEL}_h$ improved dramatically. The coverage probabilities become very close to the nominal value and average lengths are smaller than the corresponding average lengths of $\text{BEL}_d$ and $\text{BEL}_w$.

## 6. Additional remarks

In this paper we have examined Bayesian pseudo-EL inferences for complex surveys under the design-based framework. The Bayesian pseudo-EL approach proposed uses the first-order

inclusion probabilities either explicitly through the basic design weights or implicitly through the calibration weights and requires the second-order inclusion probabilities through the design effect. When the sampling fractions are negligible, the design effect may be calculated by using variance formulae under sampling with replacement and hence bypassing the need for second-order inclusion probabilities.

The non-informative prior on the mean $\theta$ that was used here is attractive from a practical point of view. However, one can also use informative priors on $\theta$ if desired, and for large samples the resulting Bayesian intervals remain asymptotically valid under the design set-up. It may be noted that the specification of an informative prior on $\theta$ may be easier than the specification of an informative Dirichlet prior based on the approach that is presented in Section 3.3, which can be viewed as a generalized scale–load method. As noted by Lazar (2003), 'putting a prior on $\theta$ is the most sensible approach and could be done in the usual way'.

Unlike other Bayesian methods that have been proposed in the survey literature, the Bayesian pseudo-EL approach can handle general unequal probability without replacement sampling designs and is extremely flexible in using known auxiliary population information. Moreover, the approach can be extended to cover other parameters such as the finite population distribution function and quantiles or even parameters that are defined through estimating equations such as regression coefficients. Further details on calculating the related design effect and the validity of the inference under the design-based frequentist framework are currently under investigation. Extension of this work to multistage sampling designs will also be studied.

## Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada. We are grateful to the referees and the Joint Editor for many valuable suggestions.

## Appendix A: Proofs

### A.1. Proof of theorem 1

The pseudoposterior density has maximum value attained at $\theta = \hat{\theta}_{\mathrm{PEL}}$. Considering $\theta$ such that $\theta - \hat{\theta}_{\mathrm{PEL}} = O_p(1/\sqrt{n})$ and following along the lines of the proof of theorem 2 in Wu and Rao (2006), we can show that

$$-n^* \sum_{i \in s} \tilde{d}_i(s) \log\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\} = -\tfrac{1}{2}(\theta - \hat{\theta}_{\mathrm{PEL}})^2/\sigma^2_{\mathrm{PEL}} + o_p(1).$$

The required regularity conditions are conditions C1–C5 as outlined in Wu and Rao (2006).

### A.2. Proof of theorem 2

Suppose that $(p_1, \ldots, p_n)$ follows a Dirichlet distribution $D(\alpha_1, \ldots, \alpha_n)$. It follows that $E(p_i) = \alpha_i/k$, $\mathrm{var}(p_i) = (\alpha_i/k^2)\{1 - (\alpha_i + 1)/(k+1)\}$ and $\mathrm{cov}(p_i, p_j) = -\alpha_i\alpha_j/k^2(k+1)$ for $i \neq j$, where $k = \Sigma_{i=1}^n \alpha_i$. Under the Dirichlet–Haldane prior, the pseudoposterior distribution $(p_1, \ldots, p_n)|s \sim D(\gamma_1, \ldots, \gamma_n)$, where $\gamma_i = n^* \tilde{d}_i(s)$ for case (a) and $\gamma_i = m\{\tilde{w}_i(s)\}$ for case (b). Noting that $k = \Sigma_{i \in s} \gamma_i = n^*$ for case (a), it immediately follows that $E(\theta|s) = \Sigma_{i \in s}\, y_i E(p_i|s) = \Sigma_{i \in s}\, \tilde{d}_i(s) y_i = \hat{\bar{Y}}_{\mathrm{H}}$, the Hájek estimator of the mean $\bar{Y}$. In addition,

$$\begin{aligned}
\mathrm{var}(\theta|s) &= \sum_{i \in s} \frac{\tilde{d}_i(s)\{\tilde{d}_i(s) + 1\}}{n^* + 1} y_i^2 - \sum_{i \in s} \sum_{j \neq i} \frac{1}{n^* + 1} \tilde{d}_i(s)\, \tilde{d}_j(s) y_i y_j \\
&= \frac{1}{n^* + 1} \left\{ \sum_{i \in s} \tilde{d}_i(s)(y_i - \hat{\bar{Y}}_{\mathrm{H}})^2 + 2 \sum_{i \in s} \tilde{d}_i(s)^2 y_i^2 \right\} \\
&= \frac{S_y^2}{n^*} + o\left(\frac{1}{n^*}\right) = V(\hat{\bar{Y}}_{\mathrm{H}}) + o\left(\frac{1}{n^*}\right).
\end{aligned}$$

The last step follows from $n^* = S_y^2 / V(\hat{\bar{Y}}_{\mathrm{H}})$, where $S_y^2$ is the finite population variance, and $V(\hat{\bar{Y}}_{\mathrm{H}})$ is the design-based variance of $\hat{\bar{Y}}_{\mathrm{H}}$. The second-last step is based on the assumption that $\tilde{d}_i(s) = O_p(n^{-1})$ and $\Sigma_{i\in s} \tilde{d}_i(s)^2 y_i^2 = O_p(n^{-1})$. It is sufficient that this last term is of order $o_p(1)$. For case (b), it can be shown that $E(\theta|s) = \hat{\theta}_c$ and $\mathrm{var}(\theta|s) = V(\hat{\theta}_c) + o(n^{-1})$, where $\hat{\theta}_c = \Sigma_{i\in s} \tilde{w}_i(s) y_i$ is the calibration estimator of $\bar{Y}$, as defined in Section 3.2.

### *A.3. Proof of theorem 3*

Let

$$K(\theta) = -\sum_{i\in s} \tilde{d}_i(s) \log\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\}$$

for $\theta \in (\min_{i\in s}(y_i), \max_{i\in s}(y_i))$. We need to show that $\mathrm{d}^2 K(\theta)/\mathrm{d}\theta^2 < 0$. Let $\lambda_1(\theta)$ be the first component of $\boldsymbol{\lambda}(\theta)$. Noting that $g(\boldsymbol{\lambda}) = \mathbf{0}$ and $\Sigma_{i\in s} \tilde{d}_i(s)/\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\} = 1$, we can show that $\mathrm{d}K(\theta)/\mathrm{d}\theta = \lambda_1(\theta)$. Further, it can be shown that $\mathrm{d}g(\boldsymbol{\lambda})/\mathrm{d}\theta = \mathbf{0}$ reduces to

$$\sum_{i\in s} \frac{\tilde{d}_i(s)\mathbf{u}_i\mathbf{u}_i'}{\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\}^2} \frac{\mathrm{d}\boldsymbol{\lambda}(\theta)}{\mathrm{d}\theta} = -\mathbf{e}_1,$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$. It follows that $\mathrm{d}\lambda_1(\theta)/\mathrm{d}\theta = -h^{11} < 0$, where $h^{11}$ is the (1, 1) element (upper left-hand corner) of $H^{-1}$ and

$$H = \sum_{i\in s} \frac{\tilde{d}_i(s)\,\mathbf{u}_i\mathbf{u}_i'}{\{1 + \boldsymbol{\lambda}'(\theta)\mathbf{u}_i\}^2},$$

which is positive definite. This shows that $\mathrm{d}^2 K(\theta)/\mathrm{d}\theta^2 < 0$.

## References


Aitkin, M. (2008) Applications of the Bayesian bootstrap in finite populations inference. *J. Off. Statist.*, **24**, 21–51.

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

Ericson, W. A. (1969) Subjective Bayesian models in sampling finite populations (with discussion). *J. R. Statist. Soc.* B, **31**, 195–233.

Fang, K.-T. and Mukerjee, R. (2006) Empirical-type likelihoods allowing posterior credible sets with frequentist validity: higher-order asymptotics. *Biometrika*, **93**, 723–733.

Godambe, V. P. (1966a) A new approach to sampling from finite populations: I, sufficiency and linear estimation. *J. R. Statist. Soc.* B, **28**, 310–319.

Godambe, V. P. (1966b) A new approach to sampling from finite populations: II, distribution-free sufficiency. *J. R. Statist. Soc.* B, **28**, 320–328.

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

Hoadley, B. (1969) The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *J. Am. Statist. Ass.*, **64**, 216–229.

Kim, J. K. (2009) Calibration estimation using empirical likelihood in survey sampling. *Statist. Sin.*, **19**, 145–157.

Lazar, N. A. (2003) Bayesian empirical likelihood. *Biometrika*, **90**, 319–326.

Lazar, R., Meeden, G. and Nelson, D. (2008) A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Surv. Methodol.*, **34**, 51–64.

Lo, A. Y. (1988) Bayesian bootstrap for a finite population. *Ann. Statist.*, **16**, 1684–1695.

Meeden, G. (1995) Median estimation using auxiliary information. *Surv. Methodol.*, **21**, 71–77.

Meeden, G. and Vardeman, S. (1991) A non-informative Bayesian approach to interval estimation in finite population sampling. *J. Am. Statist. Ass.*, **86**, 972–980.

Monahan, J. F. and Boos, D. D. (1992) Proper likelihoods for Bayesian analysis. *Biometrika*, **79**, 271–279.

Neyman, J. (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Statist. Soc.*, **97**, 558–606.

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. *J. Ind. Statist. Ass.*, **3**, 173–180.

Rao, J. N. K. and Wu, C. (2009) Bayesian pseudo empirical likelihood intervals for complex surveys. *Working Paper 2009-03*. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo.

Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Sampford, M. R. (1967) On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499–513.
Wu, C. and Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Can. J. Statist.*, **34**, 359–375.