

Estimation of Proportion with Non-probability Survey Samples Using Pseudo-Empirical Likelihood Approach

Yilin Chen, Pengfei Li and Changbao Wu ¹

ABSTRACT

In survey questionnaires, binary response such as, yes/no is one of the most commonly-used format to collect information, and collected binary data are used to estimate the proportion of the population who has a certain characteristics. In this paper, we focus on the estimation of population proportions when survey samples are non-probability based. We propose a pseudo-empirical likelihood (PEL) approach, under which consistent point estimators for population proportions can be obtained. In addition, by utilizing auxiliary information from external data sources, obtained point estimators could acquire double robustness property desirably. Proposed PEL approach is also advantageous to use in the interval estimation. Simulation study shows that when sample size is small and proportion is close to 0 and 1, PEL based confidence intervals have noticeable improvement in the coverage rate and balance of tail error compared to normal approximation based confidence intervals.

KEY WORDS: Confidence interval, double robustness, non-probability survey samples, proportion, pseudo-empirical likelihood.

RÉSUMÉ

Dans les questionnaires d'enquête, les réponses binaires comme oui-non, d'accord-désaccord, pas satisfait - satisfait sont courantes. Les données binaires recueillies sont ensuite utilisées pour estimer les proportions de la population présentant certaines caractéristiques. Dans le cadre de ce projet, nous proposons d'estimer les proportions de la population à partir d'échantillons provenant d'enquête non-probabilistes. Nous proposons une procédure d'inférence de pseudo-vraisemblance empirique et montrons que l'estimateur ponctuel, qui en résulte pour une proportion de la population, a une propriété doublement robuste souhaitable. On propose deux méthodes de construction des intervalles de confiance du rapport de pseudo-vraisemblance empirique pour la proportion de la population: l'une est fondée sur la distribution limite de la statistique du rapport de pseudo-vraisemblance empirique ajusté, et l'autre utilise la pseudo-vraisemblance empirique bootstrap calibrée. Une étude de simulation montre que lorsque la taille de l'échantillon est petite, les intervalles de confiance basés sur le rapport de pseudo-vraisemblance empirique ont un meilleur taux de couverture et un taux d'erreur plus équilibré dans les queues que les intervalles de confiance de type Wald couramment utilisés

MOTS CLÉS : Intervalle de confiance; Double robustesse; Échantillons non-probabilistes; Proportion; Pseudo-vraisemblance empirique

1 INTRODUCTION

With the advancement of information technology, web-based surveys are growing into affordable, accessible and efficient research tools for both academic and commercial use. The rise of web-based surveys has facilitated many small-budget and time-sensitive projects, but inferences made from web-based survey samples are frequently questioned due to the lack of the theoretical grounds.

Web-based survey samples, usually taken by unknown non-probability based strategies, belong to non-probability samples. Contrary to probability samples, non-probability samples are not representative of the population, and require more sophisticated adjustments to make valid inferences about population parameters. There exists a variety of methods to estimate population mean and total with non-probability samples. Lee and Valliant (2009), and Brick (2015) investigate propensity score adjustments, coupled with calibration procedures. Mass imputation methods, such as sample matching and regression prediction are considered by Rivers (2007) and Kim et al. (2018)

¹Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave. W., Waterloo, Ontario N2L 3G1, Canada; E-mails: y992chen@uwaterloo.ca, pengfei.li@uwaterloo.ca and cbwu@uwaterloo.ca. This research is supported by grants from the Natural Sciences and Engineering Research Council of Canada.

respectively. Chen et al. (2018) propose a doubly robust estimation method, which has more robustness against model misspecification.

In this paper, we are interested in population proportion, which is a special type of mean when data is binary. Estimating method proposed by Chen et al. (2018) can naturally be adopted to estimate proportions, and it is more appealing compared to other mentioned approaches due to its double robustness property. However, we found when sample size is small and the true proportion is close to 0 or 1, confidence intervals (CI) derived under Chen et al. (2018) fail to provide satisfactory coverage rates. We propose using pseudo-empirical likelihood (PEL) approach (Chen and Sitter, 1999) for the proportion estimation. Proposed PEL approach can be applied to both discrete and continuous response, and shows particular advantage in the interval estimation for proportions.

The rest of the paper is structured as follows. Section 2 reviews the use of PEL approach in probability samples, and briefly describes the setup of the non-probability sample we consider through the paper. In Section 3, we discuss the estimation of population proportions with non-probability samples. A doubly robust point estimator is derived by maximizing PEL function under the model-calibrated constraint. We also illustrate two PEL based methods of constructing CIs for population proportions. Simulation studies are presented in Section 4, where we compare proposed CIs to several commonly-used CIs. Additional remarks are given in Section 5.

2 PSEUDO-EMPIRICAL LIKELIHOOD APPROACH

2.1 PEL with Probability Survey Samples

Let $\mathcal{F}_N = \{(\mathbf{x}_i, y_i), i = 1, \dots, N\}$ be the data of the finite population \mathcal{U} , where \mathbf{x} is the associated value of auxiliary variables, and y is the associated value of response variable. Response y is univariate, and can be either continuous or discrete. The parameter of interest is the population mean of the response variable, i.e., $\mu_y = N^{-1} \sum_{i=1}^N y_i$, and μ_y is a proportion when y is binary.

Given \mathcal{F}_N , log-empirical likelihood function is given by $l(\mathbf{p}) = \sum_{i=1}^N \log p_i$, where parameter p_i is the density at point (\mathbf{x}_i, y_i) , for $i = 1, \dots, N$. Instead of working with $l(\mathbf{p})$ directly, now we consider a probability sample \mathcal{S} which is drawn from population \mathcal{U} . Suppose $\{(\mathbf{x}_i, y_i, d_i), i \in \mathcal{S}\}$ is the data of sample \mathcal{S} with d_i being the design weights, then pseudo-empirical likelihood in Chen and Sitter (1999) is given by $\hat{l}(\mathbf{p}) = \sum_{i \in \mathcal{S}} d_i \log p_i$. PEL $\hat{l}(\mathbf{p})$ is an approximation of empirical likelihood $l(\mathbf{p})$ since $E_p\{\hat{l}(\mathbf{p})\} = l(\mathbf{p})$, where subscript p indicates the randomization mechanism of the probability sampling. In a simple scenario where no auxiliary information is available, estimator of p_i is obtained by maximizing $\hat{l}(\mathbf{p})$ under normalization constraint $\sum_{i \in \mathcal{S}} p_i = 1$. Let \hat{p}_i be the resulting estimator of p_i , then a pseudo-empirical maximum likelihood estimator (PEMLE) of μ_y is given by $\hat{\mu} = \sum_{i \in \mathcal{S}} \hat{p}_i y_i$.

2.2 Non-probability Sample Setup

Recall population \mathcal{U} , and consider a non-probability sample \mathcal{S}_A of size n_A from \mathcal{U} . Let $\{(\mathbf{x}_i, y_i), i \in \mathcal{S}_A\}$ be the dataset of sample \mathcal{S}_A , and $R_i = I(i \in \mathcal{S}_A)$ be the indicator variable for unit i being included in the sample \mathcal{S}_A . The conditional selection probability for unit i given \mathbf{x}_i and y_i is $\pi_i^A = E_q(R_i | \mathbf{x}_i, y_i) = P_q(R_i = 1 | \mathbf{x}_i, y_i)$, $i = 1, 2, \dots, N$, where the subscript q refers to the selection mechanism for sample \mathcal{S}_A . The value of π_i^A is customarily referred to as propensity score (Rubin, 1976).

Recall PEL function $\hat{l}(\mathbf{p}) = \sum_{i \in \mathcal{S}} d_i \log p_i$ from Section 2.1, and notice weights d_i ensure $\hat{l}(\mathbf{p})$ to be a valid approximation of the population level information $l(\mathbf{p})$. To construct a sample \mathcal{S}_A based PEL function, we need likewise obtain a set of weights for \mathcal{S}_A . The inverse of propensity scores turns out to be a natural choice. To estimate π_i^A , we adopt following *ignorability* condition (Rosenbaum and Rubin, 1983),

$$\pi_i^A = P(R_i = 1 | \mathbf{x}_i, y_i) = P(R_i = 1 | \mathbf{x}_i), \quad i = 1, 2, \dots, N.$$

This condition simplifies the estimation of propensity scores by requiring no measurements from the response variable. More commonly-postulated assumptions, such as strong ignorability assumption, for the selection mechanism of \mathcal{S}_A can be found in Rosenbaum and Rubin (1983).

Besides sample \mathcal{S}_A , we assume a probability sample \mathcal{S}_B of size n_B is also available. Let $\{(\mathbf{x}_i, d_i^B), i \in \mathcal{S}_B\}$ be observations from \mathcal{S}_B , where $d_i^B = 1/\pi_i^B$ are survey weights and $\pi_i^B = P(i \in \mathcal{S}_B)$ are the inclusion probability of units being in sample \mathcal{S}_B . Sample \mathcal{S}_B is often referred to as reference sample since measurements of response variable y are not available from it.

To estimate π_i^A , we further assume that π_i^A can be modelled parametrically, i.e., $\pi_i^A = P(R_i = 1 | \mathbf{x}_i) = \pi(\mathbf{x}_i, \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the true model parameters. Suppose π_i^A follows a logistic regression model from now on, then by Chen et al. (2018), pseudo-maximum likelihood estimator of $\boldsymbol{\theta}_0$ can be obtained by solving equation

$$\mathbf{0} = \sum_{i \in \mathcal{S}_A} \mathbf{x}_i - \sum_{i \in \mathcal{S}_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\theta}) \mathbf{x}_i. \quad (2.1)$$

Let $\hat{\boldsymbol{\theta}}$ be the solution of (2.1), then the estimated propensity score is given by $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$.

3 ESTIMATION OF PROPORTION WITH PEL APPROACH

3.1 Doubly Robust Estimation through Model Calibration Constraint

Given estimated propensity score $\hat{\pi}_i^A$, non-probability sample based pseudo-empirical likelihood function can be constructed by,

$$\hat{l}^A(\mathbf{p}) = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A \log p_i,$$

with $\hat{d}_i^A = (\hat{N}^A)^{-1} (\hat{\pi}_i^A)^{-1}$, and $\hat{N}^A = \sum_{i \in \mathcal{S}_A} 1/\hat{\pi}_i^A$. Note we use normalized \hat{d}_i^A as weights, instead of $1/\hat{\pi}_i^A$. This modification simplifies the derivation of the theorems we are about to propose.

To maximize $\hat{l}^A(\mathbf{p})$, we start with the simplest case where $\sum_{i \in \mathcal{S}_A} p_i = 1$ is the only constraint. Then trivially, we have $\hat{p}_i = \hat{d}_i^A$, which leads to PEMLE $\hat{\mu}_{IPW} = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A y_i$, where subscript *IPW* indicates this PEL based estimator is equivalent to an inverse probability weighted (IPW) estimator.

Besides normalization constraint, more sophisticated constraints can also be considered for maximization. Similarly to Wu and Sitter (2001), we assume the first moment satisfies $E_{\xi}(y | \mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}_0)$ for the response y given \mathbf{x} , where $\boldsymbol{\beta}_0$ is the true model parameter, and ξ indicates the randomization attributed to the prediction model. Let $\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}_0$, and $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ be the fitted value of the prediction model, then *model-calibrated* constraint is given by $\sum_{i \in \mathcal{S}_A} p_i \hat{m}_i = \hat{m}^B$, where $\hat{m}^B = (\hat{N}^B)^{-1} \sum_{i \in \mathcal{S}_B} d_i^B \hat{m}_i$, and $\hat{N}^B = \sum_{i \in \mathcal{S}_B} d_i^B$. Maximizing $\hat{l}^A(\mathbf{p})$ under normalization and model-calibrated constraint together leads to model-calibrated PEMLE $\hat{\mu}_{MC} = \sum_{i \in \mathcal{S}_A} \hat{p}_i y_i$.

In the following theorem, we explore asymptotic properties of estimator $\hat{\mu}_{MC}$. We assume there is a sequence of finite populations \mathcal{U}_{ν} of size N_{ν} , a sequence of associated non-probability samples $\mathcal{S}_{A,\nu}$ of size $n_{A,\nu}$, and a sequence of associated probability samples $\mathcal{S}_{B,\nu}$ of size $n_{B,\nu}$. Population size N_{ν} , sample size $n_{A,\nu}$ and $n_{B,\nu}$ go to infinity as $\nu \rightarrow \infty$. For the notational convenience, let $m_i = m(\mathbf{x}_i, \boldsymbol{\beta})$, $m_i^* = m(\mathbf{x}_i, \boldsymbol{\beta}^*)$, $\bar{m} = N^{-1} \sum_{i=1}^N m_i$, $\bar{m}^* = N^{-1} \sum_{i=1}^N m_i^*$, and $\pi_i^* = \pi(\mathbf{x}_i, \boldsymbol{\theta}^*)$.

Theorem 3.1. *Suppose that propensity score model is correctly specified, and Conditions C1–C7 specified in the Appendix are satisfied. Then estimator $\hat{\mu}_{MC}$ is doubly robust, and has the following asymptotic expansion,*

$$\hat{\mu}_{MC} = \hat{\mu}_{IPW} + (\hat{m}^B - \hat{m}_{IPW}) \hat{B}_m + o_p(n_A^{-\frac{1}{2}}),$$

where $\hat{B}_m = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A (\hat{m}_i - \hat{m}^B) y_i / \sum_{i \in \mathcal{S}_A} \hat{d}_i^A (\hat{m}_i - \hat{m}^B)^2$, and $\hat{m}_{IPW} = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A \hat{m}_i$.

Moreover, asymptotic variance of $\hat{\mu}_{MC}$ is given by $\text{Var}(\hat{\mu}_{MC}) = V_{MC} + o(n_A^{-1})$, with

$$V_{MC} = \frac{1}{N^2} \sum_{i=1}^N \frac{1 - \pi_i^A}{\pi_i^A} (y_i - m_i^* B_m^* - h_N - \pi_i^A \mathbf{x}_i^T \mathbf{b})^2 + W,$$

where $B_m^* = [\sum_{i=1}^N (m_i^* - \bar{m}^*)^2]^{-1} [\sum_{i=1}^N (m_i^* - \bar{m}^*) y_i]$, $h_N = N^{-1} \sum_{i=1}^N (y_i - m_i^* B_m^*)$,

$\mathbf{b} = \left[\sum_{i=1}^N (1 - \pi_i^A) (y_i - m_i^* B_m^* - h_N) \mathbf{x}_i^T \right] \left\{ \sum_{i=1}^N \pi_i^A (1 - \pi_i^A) \mathbf{x}_i \mathbf{x}_i^T \right\}^{-1}$, and $W = N^{-2} V_p(\sum_{i \in \mathcal{S}_B} d_i^B t_i)$ with $t_i = m_i^* B_m^* + \pi_i^A \mathbf{x}_i^T \mathbf{b} - \bar{m}^* B_m^*$.

The variance formula V_{MC} naturally leads to the plug-in type variance estimator for $\hat{\mu}_{MC}$. Let v_{MC} denote the resulting variance estimator. Then given estimator $\hat{\mu}_{MC}$ and v_{MC} , a normal approximation method based 100(1-a)% CI is given by $[\hat{\mu}_{MC} - z_{a/2} v_{MC}^{1/2}, \hat{\mu}_{MC} + z_{a/2} v_{MC}^{1/2}]$, where $z_{a/2}$ is the (1-a/2)th quantile of the standard normal distribution.

3.2 Adjusted PEL Ratio Confidence Intervals

In addition to normal approximation method mentioned above, we also apply adjusted PEL ratio function method, proposed by Wu and Rao (2006), to the current setting. Consider following two sets of constraint,

$$\sum_{i \in \mathcal{S}_A} p_i = 1, \quad \sum_{i \in \mathcal{S}_A} p_i \hat{m}_i = \hat{m}^B, \quad (3.2)$$

and

$$\sum_{i \in \mathcal{S}_A} p_i = 1, \quad \sum_{i \in \mathcal{S}_A} p_i \hat{m}_i = \hat{m}^B, \quad \sum_{i \in \mathcal{S}_A} p_i y_i = \mu, \quad (3.3)$$

where μ is some constant. Let $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_{n_A})$ and $\hat{\mathbf{p}}(\mu) = (\hat{p}_1(\mu), \dots, \hat{p}_{n_A}(\mu))$ be the maximizers of $\hat{l}^A(\mathbf{p})$ subject to the constraints in (3.2) and (3.3) respectively. Then $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}(\mu)$ associated PEL ratio function is given by $\Lambda_2(\mu) = -2\{\hat{l}^A(\hat{\mathbf{p}}(\mu)) - \hat{l}^A(\hat{\mathbf{p}})\}$. Moreover, we construct an adjustment factor $s_2 = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A \hat{r}_i^2 / v_{MC}$, where $\hat{r}_i = y_i - \hat{\mu}_{MC} - (\hat{m}_i - \hat{m}^B) \hat{B}_m$.

Theorem 3.2. *Suppose that propensity score model is correctly specified, and Conditions C1–C8 are satisfied. Then adjusted PEL ratio $s_2 \Lambda_2(\mu_y)$ is asymptotically χ_1^2 distributed, where χ_1^2 denotes chi-squared distribution with one degree of freedom.*

Under Theorem 3.2, with the presence of the auxiliary information, an approximate $100(1 - a)\%$ CI for μ_y is given by $PEL_{2,adj} = \{\mu \mid s_2 \Lambda_2(\mu) \leq \chi_1^2(a)\}$, where $\chi_1^2(a)$ is the $(1 - a)$ th quantile of χ_1^2 distribution. When auxiliary information or model-calibrated constraint is unavailable, we can likewise construct PEL ratio function and adjustment factor, denoted by $\Lambda_1(\mu)$ and s_1 respectively, and obtain corresponding PEL ratio CI, i.e., $PEL_{1,adj} = \{\mu \mid s_1 \Lambda_1(\mu) \leq \chi_1^2(a)\}$.

3.3 Bootstrap-calibrated PEL Ratio Confidence Intervals

There are two major hurdles in obtaining adjusted PEL ratio CIs. One is the complexity in computing adjustment factors, and the other is the failure of Theorem 3.2 when the propensity score model is misspecified.

To bypass the complications, we consider a bootstrap calibrated PEL procedure, which was investigated by Wu and Rao (2010). We take unadjusted ratio function $\Lambda_2(\mu)$ for illustration. Let $d(a)$ be the $(1 - a)$ th quantile of the distribution of $\Lambda_2(\mu_y)$. If $d(a)$ is known, then a $100(1 - a)\%$ CI for μ_y is given by $\{\mu \mid \Lambda_2(\mu) \leq d(a)\}$. However, since $d(a)$ is unknown, we apply the following bootstrap procedure to approximate $d(a)$.

- (1) Draw a bootstrap sample $\mathcal{S}_A^{(j)}$ of size n_A from $\{\mathbf{x}_i, y_i, i \in \mathcal{S}_A\}$ with simple random sampling with replacement method, and draw a bootstrap sample $\mathcal{S}_B^{(j)}$ from dataset $\{\mathbf{x}_i, d_i^B, i \in \mathcal{S}_B\}$.
- (2) Replace sample \mathcal{S}_A and \mathcal{S}_B by $\mathcal{S}_A^{(j)}$ and $\mathcal{S}_B^{(j)}$ respectively, and apply the proposed procedure in Section 3.2 to obtain PEL $\hat{l}^{A,(j)}(\mathbf{p})$ and PEL ratio statistic $\Lambda_2^{(j)}(\hat{\mu}_{MC}) = -2\{\hat{l}^{A,(j)}(\hat{\mathbf{p}}(\hat{\mu}_{MC})) - \hat{l}^{A,(j)}(\hat{\mathbf{p}})\}$.
- (3) Repeat step (1)–(2) for J times to obtain $\Lambda_2^{(1)}(\hat{\mu}_{MC}), \dots, \Lambda_2^{(J)}(\hat{\mu}_{MC})$. Then $d(a)$ can be approximated by $\tilde{d}(a)$, which is the $(1 - a)$ th quantile of $\{\Lambda_2^{(1)}(\hat{\mu}_{MC}), \dots, \Lambda_2^{(J)}(\hat{\mu}_{MC})\}$. Finally, we obtain the bootstrap-calibrated interval $PEL_{2,bts} = \{\mu \mid \Lambda_2(\mu) \leq \tilde{d}(a)\}$.

How to draw bootstrap sample $\mathcal{S}_B^{(j)}$ depends on the original sampling design of \mathcal{S}_B . One can refer to Rao and Wu (1988) for bootstrap procedures in complex survey data. Through a similar procedure, we can also construct a bootstrap-calibrated CI based on the unadjusted ratio $\Lambda_1(\mu)$. We denote the resulting interval by $PEL_{1,bts}$.

4 SIMULATION STUDIES

We consider a finite population of size $N = 10,000$, with binary response y and auxiliary variable x_1, x_2 , and x_3 . Each y_i is generated from a Bernoulli distribution with mean u_i , which follows logistic regression model ξ , $\log\{u_i/(1 - u_i)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$, where $x_{1i} = z_{1i}$, $x_{2i} = z_{2i} + 0.1x_{1i}$, $x_{3i} = z_{3i} + 0.1x_{2i}$, with $z_{1i} \sim \text{Bernoulli}(0.5)$, $z_{2i} \sim \text{Uniform}(0, 1)$, and $z_{3i} \sim \text{Exponential}(mean = 0.5)$. Values of model parameters

$(\beta_0, \beta_1, \beta_2, \beta_3)$ are set as $(-4.1, 1.0, 1.0, 1.0)$, $(-0.8, 0.5, 0.5, 0.5)$ and $(4.1, -1.0, -1.0, -1.0)$ such that true proportion approximately equals to 0.1, 0.5 and 0.9 respectively.

True propensity scores π_i^A follow the logistic regression model q , $\log\{\pi_i^A/(1-\pi_i^A)\} = \theta_0 + x_{1i} + x_{2i} + x_{3i}$, where θ_0 is chosen such that $\sum_{i=1}^N \pi_i^A = n_A$, with n_A being the target sample size. The non-probability sample \mathcal{S}_A is selected by the Poisson sampling method with the inclusion probabilities specified by π_i^A . The probability sample \mathcal{S}_B , with the target size n_B , is taken by Rao-Sampford sampling method with the inclusion probabilities π_i^B proportional to $z_i = c + x_{3i} + 0.03y_i$. The value of c is chosen to control the variation of the survey weights such that $\max z_i / \min z_i = 20$.

We consider three scenarios for the model specification. (i) Both models are correctly specified, denoted by ‘‘TT’’. (ii) The prediction model ξ is misspecified, and the propensity score model q is correctly specified, denoted by ‘‘FT’’; The working model for ξ is chosen as $\log\{u_i/(1-u_i)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. (iii) The prediction model ξ is correctly specified, and the propensity score model q is misspecified, denoted by ‘‘TF’’; The working model for q is chosen as $\log\{\pi_i^A/(1-\pi_i^A)\} = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$.

We examine proposed PEL ratio based CIs, as well as normal approximation based CIs,

$$NA_1 : [\hat{\mu}_{IPW} - z_{a/2} v_{IPW}^{1/2}, \hat{\mu}_{IPW} + z_{a/2} v_{IPW}^{1/2}], \quad NA_2 : [\hat{\mu}_{DR} - z_{a/2} v_{BOOT}^{1/2}, \hat{\mu}_{DR} + z_{a/2} v_{BOOT}^{1/2}],$$

where v_{IPW} is a plug-in type variance estimator, $\hat{\mu}_{DR}$ is the doubly robust estimator considered in Chen et al. (2018), and v_{BOOT} is a bootstrap variance estimator for $\hat{\mu}_{DR}$. Performance of CIs are evaluated through simulated coverage probability (%CP), lower tail error rate (%L) and upper tail error rate (%U) based on 2,000 samples. Results for sample size $n_A = 100$ and $n_B = 100$ are reported in Table 1. Key observations are reported below. (1) When q model is correctly specified, bootstrap-calibrated PEL ratio CIs have better coverage rates than other CIs reported. This superiority is especially pronounced when the true proportion is 0.1 and 0.9. (2) No matter which approaches are taken, performance of CIs deteriorates when the true proportion moves from 0.5 to more extreme values. (3) When q model is misspecified, $PEL_{2,bts}$ and NA_2 have acceptable performance while others collapse. (4) PEL approach in general tends to provide more balanced tail error rates for resulting CIs, compared to normal approximation based approach.

5 ADDITIONAL REMARKS

This paper deals with the estimation of population proportion when samples are non-probability based. A pseudo-empirical likelihood approach is proposed, which is comparable to classic doubly robust estimation approach to some extent. By utilizing auxiliary information, the point estimator derived under PEL approach is also doubly robust, and has the same efficiency as the estimator in Chen et al. (2018) when both propensity score model and prediction model are correctly specified. This PEL framework also provides an alternative way to construct confidence intervals. We found wald-type CIs have poor coverage probabilities for boundary proportions, which is possibly due to the deficiency of sample size for normal approximation. Proposed PEL based CIs, especially $PEL_{2,bst}$, have some improvement over wald-type CIs in terms of coverage probabilities and balance of tail errors in simulation considered.

APPENDIX

Regularity Conditions

- C1** The population size N , and sample sizes n_A and n_B satisfy $\lim_{N \rightarrow \infty} n_A/N = f_A \in (0, 1)$ and $\lim_{N \rightarrow \infty} n_B/N = f_B \in (0, 1)$.
- C2** There exist c_1 and c_2 such that $0 < c_1 \leq N\pi_i^A/n_A \leq c_2$ and $0 < c_1 \leq N\pi_i^B/n_B \leq c_2$ for all units i .
- C3** The finite population and the sampling design for \mathcal{S}_B satisfy $N^{-1} \sum_{i \in \mathcal{S}_B} d_i^B \mathbf{u}_i - N^{-1} \sum_{i=1}^N \mathbf{u}_i = O_p(n_B^{-1/2})$ for $\mathbf{u}_i = 1, \mathbf{x}_i, y_i, \dot{m}(\mathbf{x}_i, \boldsymbol{\beta}^*), m(\mathbf{x}_i, \boldsymbol{\beta}^*)$.
- C4** The finite population satisfies $N^{-1} \sum_{i=1}^N \|\mathbf{x}_i\|^3 = O(1)$, $N^{-1} \sum_{i=1}^N y_i^2 = O(1)$, $N^{-1} \sum_{i=1}^N \{m(\mathbf{x}_i, \boldsymbol{\beta}^*)\}^2 = O(1)$, and $N^{-1} \sum_{i=1}^N \pi_i^A (1 - \pi_i^A) \mathbf{x}_i \mathbf{x}_i^\top$ is a positive definite matrix.

Table 1: Performance of CIs for μ_y ($n_A = 100$, $n_B = 100$)

μ_y	Model		$PEL_{1,adj}$	$PEL_{1,bts}$	$PEL_{2,adj}$	$PEL_{2,bts}$	NA_1	NA_2
0.1	TT	%CP	90.85	92.50	91.80	92.55	88.35	90.25
		%L	1.55	1.00	1.15	0.80	0.50	0.20
		%U	7.60	6.50	7.05	6.65	11.15	9.55
	FT	%CP	91.85	92.95	92.00	92.75	89.65	92.65
		%L	1.20	0.95	1.45	1.05	0.50	0.25
		%U	6.95	6.10	6.55	6.20	9.85	7.10
	TF	%CP	74.10	76.75	89.60	93.70	82.35	91.35
		%L	25.75	23.10	2.50	1.35	17.20	0.60
		%U	0.15	0.15	7.90	4.95	0.45	8.05
0.5	TT	%CP	94.05	95.40	92.65	95.10	93.20	93.85
		%L	3.50	2.65	3.20	2.00	3.90	2.60
		%U	2.45	1.95	4.15	2.90	2.90	3.55
	FT	%CP	94.05	95.40	93.10	94.55	93.20	94.10
		%L	3.50	2.65	3.40	2.60	3.90	2.95
		%U	2.45	1.95	3.50	2.85	2.90	2.95
	TF	%CP	87.45	89.80	91.15	95.35	86.75	94.75
		%L	12.20	9.85	3.55	1.85	12.80	1.85
		%U	0.35	0.35	5.30	2.80	0.45	3.40
0.9	TT	%CP	92.45	93.15	91.80	93.50	89.60	90.15
		%L	6.05	5.90	6.95	5.65	9.95	9.55
		%U	1.50	0.95	1.25	0.85	0.45	0.30
	FT	%CP	91.20	92.30	91.50	92.45	88.90	91.55
		%L	6.95	6.55	6.85	6.50	10.60	8.00
		%U	1.85	1.15	1.65	1.05	0.50	0.45
	TF	%CP	73.05	76.05	90.10	94.85	83.00	92.15
		%L	0.15	0.15	7.70	4.20	0.25	7.50
		%U	26.80	23.80	2.20	0.95	16.75	0.35

- C5** For each \mathbf{x} , $\partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is continuous in $\boldsymbol{\beta}$ and $|\partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}| \leq h(\mathbf{x}, \boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}^*$, and $N^{-1} \sum_{i=1}^N h(\mathbf{x}_i, \boldsymbol{\beta}^*) = O(1)$.
- C6** For each \mathbf{x} , $\partial^2 m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top$ is continuous in $\boldsymbol{\beta}$ and $\max_{j,l} |\partial^2 m(\mathbf{x}, \boldsymbol{\beta})/\partial \beta_j \partial \beta_l| \leq k(\mathbf{x}, \boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}^*$, and $N^{-1} \sum_{i=1}^N k(\mathbf{x}_i, \boldsymbol{\beta}^*) = O(1)$.
- C7** $\max \{|m_i^* - \bar{m}^*| : i \in \mathcal{S}_A\} = o_p(n^{\frac{1}{2}})$, and $\max \{|\partial(m_i - \bar{m})/\partial \boldsymbol{\beta}^\top|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n} : i \in \mathcal{S}_A\} = o_p(n)$, where $\boldsymbol{\beta}_n \in (\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}})$ or $\boldsymbol{\beta}_n \in (\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$.
- C8** Estimator $N^{-1} \sum_{i \in \mathcal{S}_B} d_i^B \mathbf{u}_i$ is normally distributed, where \mathbf{u}_i is defined in **C3**, and estimator $N^{-1} \sum_{i=1}^N R_i/\pi_i^A \mathbf{v}_i$ is normally distributed, for $\mathbf{v}_i = 1$, y_i , $\pi_i^A \mathbf{x}_i$, $m(\mathbf{x}_i, \boldsymbol{\beta}^*)$.

REFERENCES

- Brick, J. M. (2015). ‘‘Compositional model inference’’. *Proceedings of the Survey Research Methods Section*. Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 299-307.
- Chen, J. and Sitter, R. R. (1999). ‘‘Empirical likelihood estimation for finite populations and the effective usage of auxiliary information’’. *Biometrika*, **80**, 107-116.
- Chen, Y., Li, P. and Wu, C. (2018). ‘‘Doubly robust inference with non-probability survey samples’’. *arXiv preprint arXiv: 1805.06432*.
- Kim, J. K., Park, S., Chen, Y. and Wu, C (2018), ‘‘Combining non-probability and probability survey samples through mass imputation’’. *arXiv preprint arXiv: 1812.10694*

- Lee, S. and Valliant, R. (2009). "Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment". *Sociological Methods & Research*, **37**, 319-343.
- Rao, J. N. K. and Wu, C. F. J. (1988). "Resampling inference with complex survey data". *Journal of the American Statistical Association*, **83**, 231-241.
- Rivers, D. (2007), "Sampling for web surveys". *Proceedings of the Survey Research Methods Section*. Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 1-26.
- Rosenbaum, P.R. and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects". *Biometrika*, **70**, 41-55.
- Rubin, D. B. (1976). "Inference and missing data". *Biometrika*, **63**, 581-592.
- Wu, C. and Rao, J. N. K. (2010). "Bootstrap procedures for the pseudo empirical likelihood method in sample surveys". *Statistics and Probability Letters*, **80**, 1472-1478.
- Wu, C. and Rao, J. N. K. (2006). "Pseudo-empirical likelihood ratio confidence intervals for complex surveys". *The Canadian Journal of Statistics*, **34**, 359-375.
- 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.