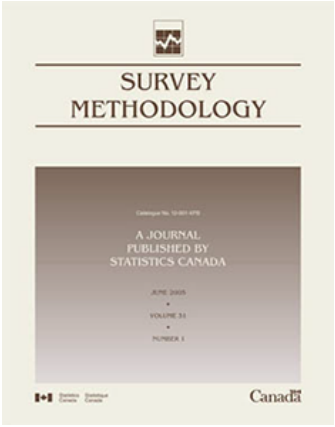
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Sample empirical likelihood approach under complex survey design with scrambled responses



by Sixia Chen, Yichuan Zhao and Yuke Wang

Release date: June 24, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 59-74
Statistics Canada, Catalogue No. 12-001-X

Sample empirical likelihood approach under complex survey design with scrambled responses

Sixia Chen, Yichuan Zhao and Yuke Wang[1]

Abstract

One effective way to conduct statistical disclosure control is to use scrambled responses. Scrambled responses can be generated by using a controlled random device. In this paper, we propose using the sample empirical likelihood approach to conduct statistical inference under complex survey design with scrambled responses. Specifically, we propose using a Wilk-type confidence interval for statistical inference. Our proposed method can be used as a general tool for inference with confidential public use survey data files. Asymptotic properties are derived, and the limited simulation study verifies the validity of theory. We further apply the proposed method to some real applications.

Key Words: Empirical likelihood; Scrambled responses; Statistical disclosure control; Survey data.

1. Introduction

The survey sampling technique has been shown to be one of the most effective ways to collect representative information for the underlying study population of interest; see Kish (1965) and Cochran (1977), among others. This approach has been used frequently in practice to obtain important information related to health, social economics, and public opinions. However, data collection by using a complex sampling design without careful control of statistical disclosure may lead to low response rate and large measurement error (Hundepool, Domingo-Ferrer, Franconi, Giessing, Nordholt, Spicer and Wolf, 2012). Statistical disclosure control (SDC) has been defined as one of few necessary steps to release public use files by agencies such as the US Census Bureau. For instance, Krenzke, Li, Freedman, Judkins, Hubble, Roisman and Larsen (2011) produced transportation data products from the Amercian Community Survey that comply with disclosure rules. Gouweleeuw, Kooiman, Willenborg and Wolf (1998) discussed statistical data protection at Statistics Netherlands.

The idea underlying SDC is to generate some perturbation based on the original raw data file so that the risk of identifying individuals is tiny and the utility of the perturbed data file is high. Currently, there are many SDC approaches including data coarsening, variable suppression, data swapping (Fienberg and McIntyre, 2005), Parametric model-based multivariate sequential replacement (Raghunathan, Lepkowski, van Hoewyk and Solenberger, 2001), and scrambled responses or randomized response methods (Horvitz, Shah and Simmons, 1967; Fox and Tracy, 1986). For more information about those approaches, see Hundepool et al. (2012).

Inference after SDC is an important and challenging problem. Statistical analysis without taking into account SDC leads to a biased variance estimation (Raghunathan, Reiter and Rubin, 2003). Raghunathan

1. Sixia Chen, Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK 73104, U.S.A. E-mail: sixia-chen@ouhsc.edu; Yichuan Zhao, Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, U.S.A.; Yuke Wang, Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, U.S.A.

et al. (2003) proposed using the multiple imputation (MI) procedure to generate perturbed data files and using the Rubin's variance estimator formula for inference. However, most agencies only seek to produce one public use file, instead of many files and the validity of MI depends on the well-known congeniality condition of Meng (1994). This condition may not hold under informative sampling design (Kim and Yang, 2017). Compared with other approaches, the scrambled responses approach is very easy to implement and has good compromise of risk and utility. In addition, valid statistical inference can be developed for most complex sampling designs. Warner (1965) first proposed using a randomization device, such as a deck of cards, to estimate the proportion of sensitive characters, such as induced abortions, drug used, and so on. Tracy and Mangat (1996) contains a comprehensive review of randomized response methods. One effective randomized response method (Scrambled responses technique) is a multiplicative model considered by Eichhorn and Hayre (1983). Bar-Lev, Bobovitch and Boukai (2004) proposed an improved version of their model. Saha (2011) discussed an optional scrambled randomized response technique for practical surveys. More recently, Singh and Kim (2011) proposed using a pseudo empirical likelihood estimator with a simple random sampling without replacement (SRSWOR) design under this model. However, they only considered a point estimation under the SRSWOR design, and their proposed method may not work for other sampling designs, such as probability proportional to size design.

Empirical likelihood approach was proposed by Hartley and Rao (1968) and studied by Owen (1988, 2001) and Qin and Lawless (1994) under traditional statistical settings. Under complex survey settings, Wu and Rao (2006) considered pseudo empirical likelihood approach. Chen and Kim (2014) proposed population and sample empirical likelihood methods which are more efficient than pseudo empirical likelihood method with high entropy designs. Berger and Torres (2016), Berger (2018a, 2018b) extended the sample empirical likelihood approach in Chen and Kim (2014) to a more general setting. In this paper, we only consider single stage sampling designs, which include Poisson sampling and stratified probability proportional to size sampling designs. Our proposed approach can be generalized to multi-stage design by using the method discussed in Berger (2018b). In surveys with multi-stage design, one challenge is that we need to specify the conditions of inclusion probabilities and consider the correlation of observations within the same cluster in different stages. We also consider interval estimation by using the sample empirical likelihood method considered in Chen and Kim (2014). After estimating the scale factor consistently, the adjusted pseudo empirical likelihood ratio converges to a standard Chi-square distribution, which can be used to construct the confidence interval. External aggregated auxiliary information, such as population size by age, gender, and race, can be naturally incorporated into our proposed method to improve the efficiency of the proposed estimators. Our proposed method is practical and can be used in most public-use survey data files, such as those from the National Health and Nutrition Examination Survey (NHANES), National Health Interview Survey (NHIS), and Behavioral Risk Factor Surveillance System (BRFSS).

The paper is organized as follows. Basic notations, research questions, and the Hájek estimator are introduced in Section 2. Section 3 discusses the proposed sample empirical likelihood method. One

simulation study is presented in Section 4. We apply the proposed methods to 2015-2016 National Health Nutrition and Examination Survey (NHANES) data in Section 5. In Section 6, we conclude this paper. All technique details are contained in the Appendix.

2. Preliminaries

Suppose the finite population $\mathcal{F}_N = (X_i, Y_i, i = 1, \ldots, N)$ is generated from some unknown super-population model, where Y_i is a study variable and X_i is a covariate. For ease of presentation, given $\mathcal{F}_N$, a random sample A is assumed to be selected from a single stage unstratified sampling design. Let I_i be the sampling indicator for unit i such that $I_i = 1$ if unit i is selected and 0 otherwise. Denote the first-order and second-order inclusion probabilities as $\pi_i = E(I_i)$ and $\pi_{ij} = E(I_i I_j)$ for $i, j = 1, \ldots, N$. Then, the sampling weight can be written as $d_i = \pi_i^{-1}$ and sample size is $n = \sum_{i=1}^{N} I_i$. Suppose the parameter of interest is $\theta_N = N^{-1}\sum_{i=1}^{N} Y_i$. Due to confidentiality, we plan to use scrambled responses Z_i of Y_i such that $Z_i = Y_i S_i$ with probability $1 - p$ and $Z_i = Y_i$ with probability p, where $E(S_i) = a$ and $V(S_i) = b^2$ with p, a, and b^2 known. Bar-lev et al. (2004) and Singh and Kim (2011) considered similar models. Instead of observing Y_i directly, we only observe the scrambled responses Z_i in the data file. Hájek estimator discussed in Hájek (1971) and Fuller (2009) has been used frequently in survey data analysis. Under certain regularity conditions, one can show that the following Hájek (HJ) type estimator is consistent:

$$\hat{\theta}_{\mathrm{HJ}} = \frac{1}{\hat{N}} \sum_{i \in A} d_i Y_i^*, \tag{2.1}$$

where $Y_i^* = Z_i \{(1 - p)\, a + p\}^{-1}$ and $\hat{N} = \sum_{i \in A} d_i$ since $E(\hat{N}) = N$ and

$$E\left(\sum_{i \in A} d_i Y_i^*\right) = \sum_{i=1}^{N} \{E(Y_i^*)\} = \sum_{i=1}^{N} \left[\{E(Y_i S_i)\,(1 - p) + Y_i p\} \{(1 - p)\, a + p\}^{-1}\right] = \sum_{i=1}^{N} Y_i.$$

The asymptotic properties of $\hat{\theta}_{\mathrm{HJ}}$ are described in the following Theorem 1, and the sketched proof is contained in Appendix B.

Theorem 1. *Under the regularity conditions in Appendix A, $\hat{\theta}_{\mathrm{HJ}}$ has the following asymptotic expansion*

$$\hat{\theta}_{\mathrm{HJ}} = \theta_N + \frac{1}{N} \sum_{i \in A} d_i (Y_i^* - \theta_N) + o_p(n^{-1/2}), \tag{2.2}$$

and

$$V_{\mathrm{HJ}}^{-1/2} (\hat{\theta}_{\mathrm{HJ}} - \theta_N) \to^d N(0, 1), \tag{2.3}$$

as $n, N \to \infty$ with

$$V_{\mathrm{HJ}} = V_1 + V_2, \tag{2.4}$$

where

$$V_1 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\left(Y_i-\theta_N\right)\left(Y_j-\theta_N\right),$$

and

$$V_2 = \frac{(1-p)\left\{p(a-1)^2+b^2\right\}}{\left\{(1-p)\,a+p\right\}^2}\,\frac{1}{N^2}\sum_{i=1}^{N}Y_i^2.$$

Note that V_1 is the design variability of Hájek estimator for population mean θ_N by using the true values and V_2 is the additional variability generated by using scrambled responses. According to Theorem 1, the consistent estimator of V_{HJ} can be written as

$$\hat{V}_{\mathrm{HJ}} = \frac{1}{\hat{N}^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,\frac{Y_i^*-\hat{\theta}_{\mathrm{HJ}}}{\pi_i}\,\frac{Y_j^*-\hat{\theta}_{\mathrm{HJ}}}{\pi_j}+\frac{(1-p)\left\{b^2+p(a-1)^2\right\}}{\left(b^2+a^2\right)(1-p)+p}\,\frac{1}{\hat{N}^2}\sum_{i\in A}d_iY_i^{*2}.$$

When $n/N = o(1)$, the second term above can be safely ignored. Therefore, we can use a traditional design consistent estimator with transformed variable Y_i^*. In the next section, we will propose using the pseudo empirical likelihood method to construct both point estimator and confidence interval when we have aggregated auxiliary information.

3. Proposed method

Population-level aggregated information is often available through census or large surveys, such as the American Community Survey (ACS). For instance, we may know the national-level population counts by gender, race, educational level, or income level. Incorporating such information into estimation will often reduce the coverage error and improve the efficiency of the estimators. In this section, we propose using the sample empirical likelihood (SEL) approach proposed by Chen and Kim (2014) to conduct point and interval estimation simultaneously. Suppose a population mean $\bar{X}_N = N^{-1}\sum_{i=1}^{N}X_i$ is known through some external resources. Then, the SEL estimator can be obtained by maximizing the following sample empirical log-likelihood function

$$l_s = \sum_{i\in A}\log(w_i), \tag{3.1}$$

subject to constraints

$$\sum_{i\in A}w_i = 1, \quad \sum_{i\in A}w_i\pi_i^{-1}\left(X_i-\bar{X}_N\right)=0, \quad w_i\geq 0, \tag{3.2}$$

and

$$\sum_{i\in A}w_i\pi_i^{-1}\left(Y_i^*-\theta\right)=0. \tag{3.3}$$

By maximizing objective function (3.1) subject to constraints in (3.2), the SEL weight can be written as

$$\hat{w}_i = \frac{1}{n} \frac{1}{1 + \hat{\lambda}\pi_i^{-1}\left(X_i - \bar{X}_N\right)},$$

with $\hat{\lambda}$ as the Lagrange multiplier, and it can be obtained by solving the second constraint in (3.2). Then, according to (3.3), the SEL estimator of θ_N can be written as

$$\hat{\theta}_{\text{SEL}} = \frac{\sum_{i\in A} \hat{w}_i \pi_i^{-1} Y_i^*}{\sum_{i\in A} \hat{w}_i \pi_i^{-1}}.$$

The following Theorem 2 contains asymptotic properties of the proposed SEL estimator $\hat{\theta}_{\text{SEL}}$. The sketched proof is contained in Appendix C.

Theorem 2. *Under the regularity conditions in Appendix A, $\hat{\theta}_{\text{SEL}}$ has the following asymptotic expansion*

$$\hat{\theta}_{\text{SEL}} = \theta_N + \frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right) - B\,\frac{1}{N}\sum_{i\in A} d_i\left(X_i - \bar{X}_N\right) + o_p\left(n^{-1/2}\right), \tag{3.4}$$

where

$$B = \left\{\frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\left(Y_i - \theta_N\right)\left(X_i - \bar{X}_N\right)\right\}\left\{\frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\left(X_i - \bar{X}_N\right)\left(X_i - \bar{X}_N\right)^{\mathrm{T}}\right\}^{-1}$$

and

$$V_{\text{SEL}}^{-1/2}\left(\hat{\theta}_{\text{SEL}} - \theta_N\right) \to^d N(0, 1),$$

as $n, N \to \infty$ with

$$V_{\text{SEL}} = V_1^* + V_2,$$

where V_2 is defined in Theorem 1 and

$$V_1^* = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\,\eta_i\eta_j,$$

with $\eta_i = Y_i - \theta_N - B\left(X_i - \bar{X}_N\right)$.

Note that V_1^* is the design variability of optimal regression estimator which is less than V_1 defined in Theorem 1. The optimal regression estimator has been discussed by Fuller and Isaki (1981), Montanari (1987), and Rao (1994). According to Theorem 2, the consistent estimator of V_{SEL} can be written as

$$\hat{V}_{\text{SEL}} = \frac{1}{\hat{N}^2}\sum_{i\in A}\sum_{j\in A} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\,\frac{\hat{\eta}_i}{\pi_i}\,\frac{\hat{\eta}_j}{\pi_j} + \frac{(1-p)\left\{b^2 + p(a-1)^2\right\}}{\left(b^2 + a^2\right)(1-p) + p}\,\frac{1}{\hat{N}^2}\sum_{i\in A} d_i Y_i^{*2},$$

where $\hat{\eta}_i = Y_i^* - \hat{\theta}_{\text{SEL}} - \hat{B}\left(X_i - \bar{X}_N\right)$ with

$$\hat{B} = \left\{\sum_{i\in A} d_i^2\left(Y_i^* - \hat{\theta}_{\text{SEL}}\right)\left(X_i - \bar{X}_N\right)\right\}\left\{\sum_{i\in A} d_i^2\left(X_i - \bar{X}_N\right)\left(X_i - \bar{X}_N\right)^{\mathrm{T}}\right\}^{-1}.$$

When $n/N = o(1)$, the second term of $\hat{V}_{\text{SEL}}$ can be ignored. Under the simple random sampling (SRS) design, it can be shown that $\hat{\theta}_{\text{SEL}}$ is asymptotically equivalent to the following well-known regression estimator

$$\hat{\theta}_{\text{REG}} = \frac{1}{\hat{N}} \sum_{i \in A} d_i Y_i^* - \hat{B}_R \left(\frac{1}{\hat{N}} \sum_{i \in A} d_i X_i - \bar{X}_N \right), \tag{3.5}$$

where

$$\hat{B}_R = \left\{ \sum_{i \in A} d_i \left(Y_i^* - \hat{\theta}_{\text{HJ}} \right) \left(X_i - \bar{X}_N \right) \right\} \left\{ \sum_{i \in A} d_i \left(X_i - \bar{X}_N \right) \left(X_i - \bar{X}_N \right)^{\top} \right\}^{-1}.$$

However, for general design, $\hat{\theta}_{\text{SEL}}$ is different from $\hat{\theta}_{\text{REG}}$. Under Poisson sampling design, it can be shown that $\hat{\theta}_{\text{SEL}}$ is more efficient than $\hat{\theta}_{\text{REG}}$. Theorem 1 and Theorem 2 can be used to construct a Wald-type confidence interval for θ_N. The following Theorem 3 can be used to construct a Wilk-type confidence interval. The sketched proof of Theorem 3 is contained in Appendix D.

Theorem 3. *Define* $R_n(\theta_N) = 2\left\{ l_s\left(\hat{\theta}_{\text{SEL}}\right) - l_s(\theta_N) \right\}$, *where* $l_s(\theta)$ *is defined in (3.1) with* w_i *satisfying (3.2) and (3.3). Then under the regularity conditions listed in Appendix A, as* $n, N \to \infty$, $c_1 c_2^{-1} R_n(\theta_N) \to^d \chi_1^2$, *where* $c_1 = N^{-2} \sum_{i=1}^N \pi_i^{-1} \eta_i^{*2}$ *with* $\eta_i^* = Y_i^* - \theta_N - B(X_i - \bar{X}_N)$ *and* $c_2 = V_{\text{SEL}}$.

The estimator of c_1 and c_2 can be written as

$$\hat{c}_1 = \hat{N}^{-2} \sum_{i \in A} \pi_i^{-2} \left\{ Y_i^* - \hat{\theta}_{\text{SEL}} - \hat{B}\left(X_i - \bar{X}_N\right) \right\}^2,$$

and $\hat{c}_2 = \hat{V}_{\text{SEL}}$. Theorem 3 can be used to construct a Wilk-type confidence interval for θ_N.

4. Simulation study

In the simulation study, we consider finite population (X_i, Y_i), $i = 1, 2, \ldots, N$ for $N = 10{,}000$. X_i is uniformly distributed over [0, 1] and $Y_i = m(X_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, 0.01)$. Four functions $m(x)$ are listed below:

(A). $m_1(x) = 2 + 2(x - 0.5)$,

(B). $m_2(x) = 2 + 2(x - 0.5)^2$,

(C). $m_3(x) = 2 + 2(x - 0.5) + \exp\left(-200(x - 0.5)^2\right)$,

(D). $m_4(x) = 2 + 2(x - 0.5)\Delta(x < 0.6) + 0.6\Delta(x \geq 0.6)$, where $\Delta(B)$ is the binary indicator function for condition B such that $\Delta(B) = 1$ if condition B is satisfied and 0 otherwise.

We generated $B = 5{,}000$ Monte Carlo samples from Poisson sampling with inclusion probabilities $\pi_i = nk_i \Big/ \sum_{j=1}^N k_j$, where the size variable $k_j = \max\left(0.5Y_j + 2, 1\right) + u_j$ with $u_j \sim \chi^2(1)$. We considered sample sizes $n = 40$, 50, 100 and 200. For each Monte Carlo sample, the scrambled responses Z_i were generated with $p = 0.6$, and $S_i \sim N(1.5, 0.2/1.5)$. Suppose we only observe X_i and Z_i in the

sample. The performance of the HJ estimator and the proposed SEL estimator were compared with the estimate population mean of Y, which is $\theta_0 = E(Y)$. The results are shown in Table 4.1.

We computed Monte Carlo bias $\text{MCB} = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b - \theta_0$, Monte Carlo standard error $\text{MCSE} = \left\{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2\right\}^{1/2}$ with $\bar{\theta} = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b$ and Monte Carlo mean squared error $\text{MCMSE} = \left\{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \theta_0\right)^2\right\}^{1/2}$. For variance estimation, we calculated coverage rate, average length of interval estimates, and percentage of relative bias of variance estimators $\text{RB} = 100 \times \left[\left(B^{-1}\sum_{b=1}^{B}\hat{V}_b\right)\left\{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2\right\}^{-1} - 1\right]$. Results obtained from the simulation are given in Table 4.1.

Table 4.1
Simulation results of Monte Carlo bias (MCB), Monte Carlo standard error (MCSE), and Monte Carlo mean squared error (MCMSE), coverage rate, average length of 95% confidence intervals, and relative bias (RB) for the Hájek (HJ) estimator and sample empirical likelihood (SEL) estimator

Setting		MCB		MCSE		MCMSE		Coverage Rate		Avg Length		RB	
Model	***n***	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**
m_1	40	0.0035	0.0005	0.123	0.076	0.015	0.006	0.936	0.940	0.470	0.283	-0.027	-0.075
	50	0.0026	0.0006	0.110	0.069	0.012	0.005	0.939	0.941	0.420	0.255	-0.024	-0.078
	100	0.0009	0.0003	0.077	0.048	0.006	0.002	0.946	0.950	0.300	0.183	0.007	-0.000
	200	0.0006	-0.0002	0.054	0.033	0.003	0.001	0.944	0.954	0.211	0.130	-0.010	0.000
m_2	40	0.0006	0.0007	0.083	0.085	0.007	0.007	0.937	0.937	0.319	0.314	-0.020	-0.098
	50	-0.0004	-0.0008	0.074	0.075	0.005	0.006	0.939	0.944	0.286	0.283	-0.014	-0.066
	100	-0.0002	-0.0001	0.053	0.053	0.003	0.003	0.941	0.947	0.203	0.203	-0.036	-0.057
	200	-0.0007	-0.0006	0.037	0.037	0.001	0.001	0.945	0.949	0.144	0.144	0.002	-0.013
m_3	40	0.0022	0.0011	0.138	0.091	0.019	0.008	0.926	0.939	0.512	0.344	-0.081	-0.068
	50	0.0056	0.0028	0.119	0.081	0.014	0.007	0.941	0.942	0.460	0.312	-0.018	-0.045
	100	0.0011	0.0003	0.084	0.058	0.007	0.003	0.945	0.943	0.327	0.222	-0.011	-0.053
	200	-0.0002	-0.0006	0.059	0.041	0.003	0.002	0.950	0.952	0.230	0.157	-0.010	-0.028
m_4	40	0.0040	0.0012	0.119	0.080	0.014	0.006	0.938	0.937	0.460	0.296	-0.007	-0.089
	50	0.0008	0.0002	0.107	0.071	0.012	0.005	0.943	0.943	0.413	0.267	-0.020	-0.069
	100	0.0007	0.0006	0.075	0.049	0.006	0.002	0.942	0.945	0.293	0.190	-0.013	-0.036
	200	-0.0003	-0.0002	0.053	0.034	0.003	0.001	0.946	0.957	0.206	0.135	-0.018	0.029

For model m_1, m_3, and m_4, SEL has a smaller Monte Carlo bias, Monte Carlo standard error, and Monte Carlo mean squared error, especially for small sample sizes (n = 40 or 50). For model m_2, the two methods have comparable performance. For all four models, we found that, for most of the cases (14 of 16) the SEL estimators had a coverage rate higher than or equal to that of the HJ estimator, while the average length of confidence interval was shorter compared with the average length obtained with the HJ estimator. Both methods provided small relative biases of variance estimators. Overall, the proposed SEL outperformed HJ for most cases.

To test the sensitivity of the proposed approach, under current simulation study setups, we added noise, W_i, to the simulation. Then, $Y_i = m(\alpha X_i + (1-\alpha)W_i) + \varepsilon_i$ with α = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 1, $X_i \sim \text{Uniform}(0, 1)$, $W_i \sim N(0, 1)$, and $\varepsilon_i \sim N(0, 0.01)$. Suppose we only observe X_i and Z_i (the scrambled response of Y_i) in the sample, the HJ estimator and SEL estimator were again compared. The results are shown in Tables 4.2 and 4.3. We found that as α decreases, the coverage rates of the SEL

estimator are smaller than those of the HJ estimator, and the average length of CI for SEL estimator is not shorter than that of the HJ estimator. Therefore, the SEL estimator has better performance than the HJ estimator, provided that most of the information is contained in the current covariate.

Table 4.2
Simulation results of the Hájek (HJ) estimator and sample empirical likelihood (SEL) estimator after adding noise

Setting		$\alpha = 0$				$\alpha = 0.1$				$\alpha = 0.3$			
		Coverage Rate		Avg Length		Coverage Rate		Avg Length		Coverage Rate		Avg Length	
Model	n	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL
m_1	40	0.924	0.903	1.419	1.368	0.926	0.911	1.289	1.251	0.938	0.928	1.045	1.022
	50	0.926	0.915	1.292	1.256	0.928	0.920	1.146	1.125	0.937	0.930	0.958	0.938
	100	0.940	0.935	0.927	0.927	0.941	0.935	0.839	0.838	0.948	0.943	0.679	0.668
	200	0.949	0.941	0.651	0.657	0.942	0.943	0.589	0.593	0.948	0.948	0.478	0.469
m_2	40	0.942	0.943	1.872	1.909	0.929	0.930	1.328	1.358	0.933	0.933	1.455	1.458
	50	0.935	0.937	1.704	1.732	0.933	0.937	1.181	1.206	0.931	0.935	1.327	1.325
	100	0.941	0.947	1.191	1.202	0.942	0.949	0.843	0.854	0.945	0.948	0.931	0.927
	200	0.949	0.952	0.841	0.845	0.949	0.955	0.593	0.597	0.948	0.948	0.645	0.640
m_3	40	0.917	0.899	1.438	1.382	0.925	0.906	1.313	1.273	0.933	0.922	1.044	1.020
	50	0.922	0.908	1.297	1.264	0.928	0.916	1.154	1.131	0.939	0.935	0.927	0.911
	100	0.937	0.928	0.960	0.958	0.941	0.935	0.838	0.838	0.940	0.938	0.660	0.654
	200	0.940	0.940	0.674	0.679	0.945	0.944	0.615	0.619	0.945	0.941	0.474	0.467
m_4	40	0.903	0.885	1.226	1.167	0.912	0.894	0.994	0.947	0.927	0.909	0.518	0.511
	50	0.921	0.912	1.093	1.057	0.917	0.912	0.902	0.870	0.928	0.918	0.460	0.457
	100	0.931	0.925	0.805	0.802	0.936	0.935	0.646	0.644	0.935	0.931	0.337	0.338
	200	0.941	0.939	0.581	0.585	0.936	0.939	0.460	0.462	0.945	0.946	0.236	0.237

Table 4.3
Simulation results of the Hájek (HJ) estimator and sample empirical likelihood (SEL) estimator after adding noise

Setting		$\alpha = 0.5$				$\alpha = 0.7$				$\alpha = 0.9$			
		Coverage Rate		Avg Length		Coverage Rate		Avg Length		Coverage Rate		Avg Length	
Model	n	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL
m_1	40	0.934	0.933	1.002	0.934	0.933	0.935	1.091	0.959	0.937	0.940	1.292	1.096
	50	0.935	0.936	0.902	0.841	0.939	0.935	0.979	0.862	0.936	0.938	1.156	0.986
	100	0.947	0.948	0.635	0.596	0.944	0.949	0.697	0.616	0.946	0.948	0.820	0.705
	200	0.951	0.949	0.451	0.421	0.947	0.945	0.493	0.437	0.951	0.951	0.579	0.500
m_2	40	0.933	0.936	2.371	2.139	0.938	0.934	3.418	2.469	0.933	0.942	5.095	2.980
	50	0.940	0.941	2.148	1.938	0.940	0.937	3.057	2.210	0.945	0.944	4.583	2.687
	100	0.939	0.941	1.493	1.345	0.948	0.946	2.196	1.588	0.948	0.951	3.223	1.916
	200	0.942	0.942	1.054	0.938	0.944	0.947	1.545	1.113	0.949	0.947	2.264	1.356
m_3	40	0.939	0.935	1.004	0.940	0.935	0.937	1.101	0.970	0.939	0.947	1.288	1.093
	50	0.937	0.935	0.890	0.832	0.938	0.942	0.978	0.864	0.936	0.940	1.152	0.982
	100	0.946	0.945	0.635	0.595	0.951	0.952	0.698	0.616	0.948	0.952	0.821	0.706
	200	0.949	0.950	0.450	0.420	0.943	0.948	0.493	0.437	0.952	0.952	0.579	0.500
m_4	40	0.937	0.942	0.365	0.358	0.936	0.941	0.362	0.354	0.932	0.938	0.362	0.354
	50	0.935	0.939	0.326	0.322	0.939	0.943	0.325	0.320	0.938	0.947	0.324	0.320
	100	0.941	0.948	0.232	0.230	0.948	0.953	0.230	0.229	0.941	0.946	0.231	0.229
	200	0.947	0.948	0.165	0.164	0.942	0.944	0.163	0.163	0.949	0.951	0.163	0.163

5. Real application

In this section, we applied the proposed approach to 2015-2016 National Health and Nutrition Examination Survey (NHANES) to evaluate its practical performance. NHANES provides timely health- and nutrition-related information for the noninstitutionalized civilian resident population of the United States. It uses a complex, multistage probability design based on in-person survey to collect information. (see https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2015 for more information). The sample size for the 2015-2016 NHANES is about 9,000. We treated the original NHANES sample as a finite population and selected one sample by using a simple random sampling design with sample sizes (n) as 30, 40, 50, 100, and 200, respectively. Suppose our parameters of interest include population means of systolic blood pressure, diastolic blood pressure, HDL cholesterol, and total cholesterol. We created scramble responses for these parameters by using $p = 0.6$, $a = 1.5$, and $b^2 = 0.2/1.5$. In addition, body mass index (BMI) was selected as a covariate in the estimation process, since BMI is correlated with those study variables.

We compared the performances of two approaches, HJ and SEL, in terms of point estimates and interval estimates (Table 5.1). Point estimates obtained by using both methods were similar, and they were close to finite population parameters (120.47, 66.17, 54.43, and 180.25 for systolic blood pressure, diastolic blood pressure, HDL cholesterol, and total cholesterol), especially for larger sample sizes (Table 5.1). For systolic blood pressure, diastolic blood pressure, and total cholesterol, intervals produced by SEL shifted slightly to the right compared with the results produced by HJ for small sample sizes. However, when sample sizes increased, the results from the two approaches were similar. For HDL cholesterol, the results are comparable. The results from this application verified the validity of the proposed SEL approach.

Table 5.1
Point estimates and 95% CI for estimating means of different outcomes using scrambled response outcome and BMI from the NHANES data

	Systolic Blood Pressure in mm Hg		Diastolic Blood Pressure in mm Hg		HDL Cholesterol in mg/dL		Total Cholesterol in mg/dL	
n	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**
30	124.5	124.5	67.7	69.4	57.9	57.6	187.0	188.3
	(112.3, 136.8)	(113.5, 139.6)	(61.5, 73.8)	(63.9, 75.2)	(50.3, 65.5)	(50.8, 65.9)	(160.0, 214.0)	(166.6, 225.5)
40	125.6	125.5	70.2	70.2	52.0	51.2	178.7	178.1
	(115.4, 135.8)	(116.5, 136.1)	(64.6, 75.8)	(64.9, 76.1)	(48.0, 56.0)	(47.3, 55.8)	(160.6, 196.8)	(162.1, 199.0)
50	118.3	116.9	67.1	67.1	57.1	56.8	173.7	173.3
	(110.2, 126.4)	(109.0, 126.1)	(60.9, 73.3)	(61.4, 73.8)	(50.8, 63.4)	(51.3, 63.2)	(160.2, 187.1)	(161.2, 187.8)
100	120.8	120.5	70.0	69.7	52.3	52.4	173.1	172.8
	(115.1, 126.5)	(115.1, 126.3)	(65.9, 74.0)	(65.9, 73.6)	(48.9, 55.7)	(49.2, 55.9)	(163.5, 182.7)	(164.0, 183.2)
200	124.1	123.9	67.6	67.5	54.0	53.8	181.4	181.5
	(119.4, 128.9)	(119.4, 128.8)	(64.9, 70.3)	(64.8, 70.3)	(51.1, 56.8)	(51.3, 56.5)	(172.7, 190.1)	(173.3, 190.9)

6. Conclusions

In this paper, we proposed a sample empirical likelihood (SEL)-based approach using scrambled responses to protect the confidentiality of complex survey data. The proposed SEL approach is easy to implement in practice and can be used as a general tool for statistical disclosure control. The idea of our proposed approach is to replace the true values by some scrambled values through random device, then the existing sample empirical likelihood approach can be applied with scrambled values to obtain the point estimation. However, the variance estimation and confidence interval estimation are different from that by treating the scrambled values as true values since we need to incorporate the randomness due to random device in the statistical inference. Such theoretical properties have been investigated and verified through simulation study and real data application. The SEL outperforms traditional approaches, such as HJ, by improving coverage rates and reducing the coverage lengths of confidence intervals. Chen and Kim (2014) has compared Wald-type CI and Wilk-type CI in the simulation studies by using sample empirical likelihood method. In general, the Wilk-type confidence intervals show better coverage properties than the Wald-type confidence intervals in terms of coverage rates. We would expect similar results by using our proposed approaches here. In future research, we will extend the proposed approach to estimate more general parameters, such as population quantiles and distribution functions. The corresponding statistical computational tools, such as R package, will also be developed.

Acknowledgements

Dr. Sixia Chen was partially supported by the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The research of Yichuan Zhao was supported by the National Security Agency (NSA) Grant (H98230-12-1-0209) and the National Science Foundation Grant (DMS-1613176).

Appendix

A. Regularity conditions

We present the regularity conditions needed for proving Theorem 1 to Theorem 3 as following:

(C1). $c_1 < \pi_i N n^{-1} < c_2$ for $i = 1, 2, \ldots, N$ with $0 < c_1 < c_2$.

(C2). $n^{1/2}\left(N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} Y_i - N^{-1}\sum_{i=1}^{N} Y_i\right) \to^d N(0, V_1)$ as $n \to \infty$ and $N \to \infty$, where $V_1 = nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j) d_i Y_i d_j Y_j$.

(C3). $n^{1/2}\left(N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right) \to^d N(0, V_2)$ as $n \to \infty$ and $N \to \infty$, where $V_2 = nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j) d_i X_i d_j X_j^{\top}$.

(C4). $N^{-1}\sum_{i=1}^{N} |Y_i|^4$ and $N^{-1}\sum_{i=1}^{N} \| X_i \|^4$ are bounded.

(C5). $\max_{i \in A} |Y_i| = o_p(n^{1/2})$ and $\max_{i \in A} \| X_i \| = o_p(n^{1/2})$.

B. Sketched proof of Theorem 1

$\hat{\theta}_{\mathrm{HJ}}$ can be written as the solution of estimating equation $\hat{U}_{\mathrm{HJ}}(\theta) = 0,$ where

$$\hat{U}_{\mathrm{HJ}}(\theta) = \frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta\right).$$

Under the assumptions that $\hat{U}_{\mathrm{HJ}}(\theta)$ converges to $U_{\mathrm{HJ}}(\theta) = N^{-1}\sum_{i=1}^{N}(Y_i - \theta)$ uniformly, $E\left(Y^2\right) < \infty,$ and because of $U_{\mathrm{HJ}}(\theta_N) = 0,$ it can be shown that $\hat{\theta}_{\mathrm{HJ}} \to^p \theta_N.$ By using a Taylor expansion,

$$0 = \hat{U}_{\mathrm{HJ}}\left(\hat{\theta}_{\mathrm{HJ}}\right) = \hat{U}_{\mathrm{HJ}}(\theta_N) + \frac{\partial \hat{U}_{\mathrm{HJ}}(\theta_N)}{\partial\theta}\left(\hat{\theta}_{\mathrm{HJ}} - \theta_N\right) + o_p\left(n^{-1/2}\right).$$

After some algebra, it can be shown that

$$\hat{\theta}_{\mathrm{HJ}} = \theta_N + \frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right) + o_p\left(n^{-1/2}\right).$$

Because

$$E\left(Y_i^*\right) = Y_i, \quad V\left(Y_i^*\right) = \frac{(1-p)\left\{b^2 + p(a-1)^2\right\}}{\{(1-p)a + p\}^2} Y_i^2, \tag{B.1}$$

$$\begin{aligned} V\left\{\frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right)\right\} &= E\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\left(Y_i^* - \theta_N\right)\left(Y_j^* - \theta_N\right)\right\} \\ &\quad + V\left\{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^* - \theta_N\right)\right\}. \end{aligned} \tag{B.2}$$

According to (B.1), (B.2), and after some algebra, we can show that

$$V\left\{\frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right)\right\} = V_{\mathrm{HJ}},$$

where V_{HJ} is defined in equation (2.4). Under the regularity conditions in Fuller and Isaki (1981), the asymptotic normality can be derived.

C. Sketched proof of Theorem 2

Define

$$\hat{U}_1(\lambda) = \frac{1}{N}\sum_{i\in A}\frac{\pi_i^{-1}\left(X_i - \bar{X}_N\right)}{1 + \lambda\pi_i^{-1}\left(X_i - \bar{X}_N\right)}$$

and

$$\hat{U}_2(\lambda, \theta) = \frac{1}{N} \sum_{i \in A} \frac{\pi_i^{-1}(Y_i^* - \theta)}{1 + \lambda \pi_i^{-1}(X_i - \bar{X}_N)}.$$

Then, $\hat{\theta}_{\text{SEL}}$ and $\hat{\lambda}$ are the solutions of $\hat{U}_1(\lambda) = \hat{U}_2(\lambda, \theta) = 0$. By using techniques similar to those of Chen and Kim (2014), it can be shown that $\hat{\lambda} = O_p(n^{-1/2})$ and $\hat{\theta}_{\text{SEL}} \to^p \theta_N$. Then, by using Taylor expansion, we have

$$0 = \hat{U}_1(\hat{\lambda}) = \hat{U}_1(0) + \frac{\partial \hat{U}_1(0)}{\partial \lambda} \hat{\lambda} + o_p(n^{-1/2}), \tag{C.1}$$

and

$$0 = \hat{U}_2(\hat{\lambda}, \hat{\theta}_{\text{SEL}}) = \hat{U}_2(0, \theta_N) + \frac{\partial \hat{U}_2(0, \theta_N)}{\partial \theta}(\hat{\theta}_{\text{SEL}} - \theta_N) + \frac{\partial \hat{U}_2(0, \theta_N)}{\partial \lambda} \hat{\lambda} + o_p(n^{-1/2}). \tag{C.2}$$

According to (C.1), (C.2), and after some algebra, it can be shown that

$$\hat{\lambda} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \pi_i^{-1}(X_i - \bar{X}_N)(X_i - \bar{X}_N)^{\top} \right\}^{-1} \frac{1}{N} \sum_{i \in A} d_i (X_i - \bar{X}_N) + o_p(n^{-1/2}) \tag{C.3}$$

and

$$\hat{\theta}_{\text{SEL}} - \theta_N = \frac{1}{N} \sum_{i \in A} d_i (Y_i^* - \theta_N) - B \frac{1}{N} \sum_{i \in A} d_i (X_i - \bar{X}_N) + o_p(n^{-1/2}),$$

where B is defined in Theorem 2. Because

$$\begin{aligned} V(\hat{\theta}_{\text{SEL}}) = V\left(\frac{1}{N} \sum_{i \in A} d_i \eta_i^* \right) + o(n^{-1}) &= V\left(\frac{1}{N} \sum_{i \in A} d_i \eta_i \right) + E\left\{ V\left(\frac{1}{N} \sum_{i \in A} d_i \eta_i^* \,\middle|\, A \right) \right\} \\ &= V_2 + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \eta_i \eta_j + o(n^{-1}), \end{aligned}$$

where V_2 is defined in Theorem 1, η_i is defined in Theorem 2 and $\eta_i^* = Y_i^* - \theta_N - B(X_i - \bar{X}_N)$. After some algebra, we can show that

$$\hat{V}_{\text{SEL}} = V_{\text{SEL}} + o(n^{-1}),$$

with V_{SEL} defined in Theorem 2. Furthermore, under the regularity conditions in Fuller and Isaki (1981), we obtain the asymptotic normality.

D. Sketched proof of Theorem 3

Because $\hat{\lambda} = O_p(n^{-1/2})$ and by using a Taylor expansion of $\log(1 + x)$ at $x = \hat{\lambda} \pi_i^{-1}(X_i - \bar{X}_N)$ and (C.3), we have

$$
\begin{aligned}
l_s\left(\hat{\theta}_{\mathrm{SEL}}\right) &= \sum_{i\in A}\log\left\{\frac{1}{n}\,\frac{1}{1+\hat{\lambda}\pi_i^{-1}\left(X_i-\bar{X}_N\right)}\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\log\left\{1+\hat{\lambda}\pi_i^{-1}\left(X_i-\bar{X}_N\right)\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\left\{\hat{\lambda}^{\top}\pi_i^{-1}\left(X_i-\bar{X}_N\right)-\frac{1}{2}\hat{\lambda}^{\top}\pi_i^{-2}\left(X_i-\bar{X}_N\right)^{\otimes 2}\hat{\lambda}\right\}+o_p(1)\\
&= n\log\left(\frac{1}{n}\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\otimes 2}\right\}^{-1}\\
&\quad\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)+o_p(1), \qquad \text{(D.1)}
\end{aligned}
$$

with $a^{\otimes 2}=aa^{\top}$. We now consider to maximize $l_s=\sum_{i\in A}\log(w_i)$, subject to the following constraints

$$
\sum_{i\in A}w_i=1,\qquad \sum_{i\in A}w_i\pi_i^{-1}\left(X_i-\bar{X}_N\right)=0, \qquad \text{(D.2)}
$$

and

$$
\sum_{i\in A}w_i\pi_i^{-1}\eta_i^{*}=0, \qquad \text{(D.3)}
$$

where $\eta_i^{*}=Y_i^{*}-\theta_N-B\left(X_i-\bar{X}_N\right)$. The above constraints are equivalent with the original constraints (3.2) and (3.3). Define $u_i=\left(X_i^{\top}-\bar{X}_N^{\top},\,\eta_i^{*}\right)^{\top}$. Therefore, by using a similar argument, we have

$$
\begin{aligned}
l_s(\theta_N) &= \sum_{i\in A}\log\left\{\frac{1}{n}\,\frac{1}{1+\hat{\lambda}(\theta_N)\pi_i^{-1}u_i}\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\log\left\{1+\hat{\lambda}(\theta_N)\pi_i^{-1}u_i\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\left\{\hat{\lambda}^{\top}(\theta_N)\pi_i^{-1}u_i-\frac{1}{2}\hat{\lambda}^{\top}(\theta_N)\pi_i^{-2}u_i^{\otimes 2}\hat{\lambda}(\theta_N)\right\}+o_p(1)\\
&= n\log\left(\frac{1}{n}\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}u_i^{\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}u_i^{\otimes 2}\right\}^{-1}\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}u_i+o_p(1)\\
&= n\log\left(\frac{1}{n}\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\otimes 2}\right\}^{-1}\\
&\quad\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\eta_i^{*\otimes 2}\right\}^{-1}\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}+o_p(1) \qquad \text{(D.4)}
\end{aligned}
$$

provided $\sum_{i=1}^{N}\pi_i^{-1}\left(X_i-\bar{X}_N\right)\eta_i=0$. According to (D.1), (D.4), and after some algebra, we have

$$
\begin{aligned}
\frac{c_1}{c_2}R_n(\theta_N) &= \frac{2c_1}{c_2}\left\{l_s\left(\hat{\theta}_{\mathrm{SEL}}\right)-l_s(\theta_N)\right\}=\frac{c_1}{c_2}\left\{\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}\right\}^2\left(\frac{1}{N^2}\sum_{i=1}^{N}\pi_i^{-1}\eta_i^{*2}\right)^{-1}\\
&= \left\{V\left(\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}\right)^{-1/2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}\right\}^2\ \to^{d}\ \chi_1^2. \qquad \text{(D.5)}
\end{aligned}
$$

Therefore, Theorem 3 is proven.

References

Bar-Lev, S.K., Bobovitch, E. and Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60, 255-260.

Berger, Y.G. (2018a). Empirical likelihood approaches in survey sampling. *The Survey Statistician*, 78, 22-31.

Berger, Y.G. (2018b). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics*, doi:10.1007/s10463-018-0681-x.

Berger, Y.G., and Torres, O.D.L.R. (2016). An empirical likelihood approach for inference under complex sampling design. *Journal of the Royal Statistical Society, Series B*, 78(2), 319-341.

Chen, S., and Kim, J.K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24, 335-355.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.

Eichhorn, B.H., and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

Fienberg, S.E., and McIntyre, J. (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, 21, 309-323.

Fox, J.A., and Tracy, P.E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Fuller, W.A., and Isaki, C.T. (1981). Survey design under superpopulation models. In *Current Topics in Survey Sampling*, (Eds., D. Krewski, J.N.K. Rao, and R. Platek). New York: Academic Press, 199-226.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R. and Wolf, P. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, Part one". In *The Foundations of Survey Sampling*, (Eds., V.P. Godambe and D.A. Sprott), Holt, Rinehart, and Winston, 236.

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

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. In *Proceedings of the Social Statistics Section,* American Statistical Association, 65-72.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. and Wolf, P. (2012). *Statistical Disclosure Control*, Wiley Series In Survey Methodology.

Kim, J.K., and Yang, S. (2017). A note on multiple imputation under informative sampling. *Biometrika*, 104, 221-228.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R. and Larsen, M. (2011). Producing transportation data products from the American Community Survey that comply with disclosure rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.

Montanari, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.

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

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

Qin, J., and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 1, 85-95. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5857-eng.pdf.

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

Saha, A. (2011). An optional scrambled randomized response technique for practical surveys. *Metrika*, 73, 139-149.

Singh, S., and Kim, J.M. (2011). A pseudo-empirical log-likelihood estimator using scrambled responses. *Statistics and Probability Letters*, 81, 345-351.

Tracy, D.S., and Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4, 147-158.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

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