2-26-2019

Hypotheses Testing from Complex Survey Data Using Bootstrap Weights: A Unified Approach

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

J. N. K. Rao
Carleton University

Zhonglei Wang
Xiamen University

Hypotheses Testing from Complex Survey Data Using Bootstrap Weights: A Unified Approach

Abstract

Standard statistical methods that do not take proper account of the complexity of survey design can lead to erroneous inferences when applied to survey data due to unequal selection probabilities, clustering, and other design features. In particular, the actual type I error rates of tests of hypotheses based on standard tests can be much bigger than the nominal significance level. Methods that take account of survey design features in testing hypotheses have been proposed, including Wald tests and quasi-score tests that involve the estimated covariance matrices of parameter estimates. Bootstrap methods designed for survey data are often applied to estimate the covariance matrices, using the data file containing columns of bootstrap weights. Standard statistical packages often permit the use of survey weighted test statistics, and it is attractive to approximate their distributions under the null hypothesis by their bootstrap analogues computed from the bootstrap weights supplied in the data file. In this paper, we present a unified approach to the above method by constructing bootstrap approximations to weighted likelihood ratio statistics and weighted quasi-score statistics and establish the asymptotic validity of the proposed bootstrap tests. In addition, we also consider hypothesis testing from categorical data and present a bootstrap procedure for testing simple goodness of fit and independence in a two-way table. In the simulation studies, the type I error rates of the proposed approach are much closer to their nominal level compared with the naive likelihood ratio test and quasi-score test. An application to data from an educational survey under a logistic regression model is also presented.

Keywords

Likelihood ratio test, Quasi-score test, Wald test, Wilk's theorem

Disciplines

Categorical Data Analysis | Design of Experiments and Sample Surveys | Statistical Methodology

Comments

This pre-print is made available through arxiv: https://arxiv.org/abs/1902.08944.

Hypotheses Testing from Complex Survey Data Using Bootstrap Weights: A Unified Approach

Jae-kwang Kim [*] J. N. K. Rao [†] Zhonglei Wang [‡]

February 26, 2019

Abstract

Standard statistical methods that do not take proper account of the complexity of survey design can lead to erroneous inferences when applied to survey data due to unequal selection probabilities, clustering, and other design features. In particular, the actual type I error rates of tests of hypotheses based on standard tests can be much bigger than the nominal significance level. Methods that take account of survey design features in testing hypotheses have been proposed, including Wald tests and quasi-score tests that involve the estimated covariance matrices of parameter estimates. Bootstrap methods designed for survey data are often applied to estimate the covariance matrices, using the data file containing columns of bootstrap weights. Standard statistical packages often permit the use of survey weighted test statistics, and it is attractive to approximate their distributions under the null hypothesis by their bootstrap analogues computed from the bootstrap weights supplied in the data file. In this paper, we present a unified approach to the above method by constructing bootstrap approximations to weighted likelihood ratio statistics and weighted quasi-score statistics and establish the asymptotic validity of the proposed bootstrap tests. In addition, we also consider hypothesis testing from categorical data and present a bootstrap procedure for testing simple goodness of fit and independence in a two-way table. In the simulation studies, the type I error rates of the proposed approach are much closer to their nominal level compared with the naive likelihood ratio test and quasi-score test. An application to data from an educational survey under a logistic regression model is also presented.

Key Words: Likelihood ratio test; Quasi-score test; Wald test; Wilk's theorem.

[*] Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.

[†] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6

[‡] Wang Yanan Institute for Studies in Economics (WISE) and School of Economics, Xiamen University, Xiamen, Fujian, 361005, China

1 Introduction

Testing statistical hypotheses is one of the fundamental problems in statistics. In the parametric model approach, hypothesis testing can be implemented using Wald test, likelihood ratio test, or score test. In each case, a test statistic is computed and then compared with the 100α%-quantile of the reference distribution which is the limiting distribution of the test statistic under the null hypothesis, where α is the significance level. The limiting distribution is often a chi-squared distribution (Shao; 2003).

Statistical inference with survey data involves additional steps incorporating the sampling design features. Korn and Graubard (1999) and Chambers and Skinner (2003) provide comprehensive overviews of the methods for analyzing survey data from complex sampling designs. In hypothesis testing with sample survey data, the limiting distribution of the test statistic is not generally a chi-squared distribution. Rather, it can be expressed as a weighted sum of several independent random variables from a $\chi^2(1)$ distribution, which is a chi-square distribution with one degree of freedom, and the weights depend on unknown parameters (Lumley and Scott; 2014). To handle such problems, one may consider some corrections to the test statistics to obtain a chi-square limiting distribution approximately. Such an approach usually involves computing "design effects" associated with the test statistics. Rao and Scott (1984) and Rao et al. (1998) used this approach to obtain quasi-score tests for survey data.

In this paper, we use a different approach of computing the limiting distribution using bootstrap methods for sample survey data. Beaumont and Bocci (2009) studied the use of bootstrap to compute the limiting distribution of test statistics under complex sampling designs in the context of Wald-type tests for linear regression analysis. We provide extensions of bootstrap tests to likelihood ratio test and score test. We present a unified approach of using the bootstrap method to obtain the limiting distribution of test statistics for generalized regression analysis under complex sampling designs. The theory is developed under Poisson sampling but the proposed method is applicable to more general sampling designs as long as the bootstrap distribution is asymptotically normally distributed, as presented in Lemma 1 of Section 3. The sampling design is allowed to be informative in the sense of Pfeffermann (1993). The proposed method is applicable to Wald test, likelihood ratio test, and score test. The proposed method is also applied to the simple goodness-of-fit and testing independence in a two-way table for categorical survey data. Earlier work (Rao and Scott; 1984) developed corrected test statistics based on design effects, and known as Rao-Scott first order and second order corrections. Rao-Scott corrections have been implemented in several software packages including SAS and Stata (Scott; 2007). The proposed bootstrap method does not require computing the design effect for Rao-Scott corrections.

In Section 2, basic setup is introduced. The proposed bootstrap method is presented in Section 3. In Section 4, the bootstrap likelihood ratio test using survey-weighted log-likelihood ratio is introduced. In Section 5, the proposed bootstrap method is applied to survey-weighted quasi-score test. In Section 6, the proposed bootstrap method is applied to the simple goodness-of-fit test for categorical survey data. In Section 7, test of independence in two-way tables is covered. Results from three simulation studies are presented in Section 8. An application to data from an educational survey under a logistic regression model is presented in Section 9. Concluding remarks are made in Section 10.

2 Basic Setup

Suppose that a finite population U_N of size N is randomly generated from a super-population model with density function $f(y; \theta_0)$ for some $\theta_0 \in \Theta \subset \mathbb{R}^p$, where Θ is the parameter space. From the finite population U_N, a probability sample A of size n is selected with known first-order inclusion probability π_i, and the sampling weights are obtained as $w_i = \pi_i^{-1}$. We are interested in making inference about θ_0.

The pseudo maximum likelihood estimator of θ_0 is given as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} l_w(\theta),$$

where $l_w(\theta) = N^{-1} \sum_{i \in A} w_i \log f(y_i; \theta)$ is the survey weighted log-likelihood or pseudo log-likelihood. Often, the solution $\hat{\theta}$ can be obtained by solving the weighted score equation

$$\hat{S}_w(\theta) = \frac{\partial}{\partial \theta} l_w(\theta) = N^{-1} \sum_{i \in A} w_i S(\theta; y_i) = 0. \tag{1}$$

Under some regularity conditions (Fuller; 2009, Section 1.3), we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow N(0, \Sigma_\theta) \tag{2}$$

in distribution as $n \to \infty$, where $\Sigma_\theta = \mathcal{I}(\theta_0)^{-1}\Sigma_S(\theta_0)\{\mathcal{I}(\theta_0)^{-1}\}^{\mathrm{T}}$, $\mathcal{I}(\theta) = E\{I(\theta; Y)\}$, $I(\theta; y) = -\partial^2 \log f(y;\theta)/\partial\theta\partial\theta^{\mathrm{T}}$, $\Sigma_S(\theta) = \lim_{n\to\infty}[n\mathrm{var}\{\hat{S}_w(\theta)\}]$, and $\mathrm{var}\{\hat{S}_w(\theta)\}$ is the variance of $\hat{S}_w(\theta)$ in (1). The reference distribution in (2) is the joint distribution of the superpopulation model and the sampling mechanism. Since the sampling weights are used in (1), the pseudo maximum likelihood estimator is consistent even when the sampling design is informative.

Using the second-order Taylor expansion, we obtain

$$\begin{aligned} l_w(\theta_0) &= l_w(\hat{\theta}) + \hat{S}_w(\hat{\theta})^{\mathrm{T}}(\theta_0 - \hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^{\mathrm{T}}\hat{I}_w(\hat{\theta})(\theta_0 - \hat{\theta}) + o_p(n^{-1}) \\ &= l_w(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^{\mathrm{T}}\hat{I}_w(\hat{\theta})(\theta_0 - \hat{\theta}) + o_p(n^{-1}), \end{aligned} \tag{3}$$

where

$$\hat{I}_w(\theta) = N^{-1}\sum_{i\in A} w_i I(\theta; y_i).$$

Define

$$W(\theta_0) = -2n\{l_w(\theta_0) - l_w(\hat{\theta})\} \tag{4}$$

to be the pseudo likelihood ratio test statistic. By (3), we obtain

$$W(\theta_0) = n(\hat{\theta} - \theta_0)^{\mathrm{T}}\hat{I}_w(\hat{\theta})(\hat{\theta} - \theta_0) + o_p(1).$$

Thus, using (2), we obtain

$$W(\theta_0) \longrightarrow \mathcal{G} = \sum_{i=1}^{p} c_i Z_i^2 \tag{5}$$

in distribution as $n \to \infty$, where $c_1, \ldots, c_p$ are the eigenvalues of $D = \Sigma_\theta \mathcal{I}(\theta_0)$, and $Z_1, \ldots, Z_p$ are p independent random variables from the standard normal distribution. Result (5) was formally established by Lumley and Scott (2014) and it can be regarded as a version of the Wilks' theorem for survey sampling. Unless the sampling design is simple random sampling and the sampling fraction is negligible, the limiting distribution does not reduce to the standard chi-squared distribution $\chi^2(p)$. If $p = 1$ then we can use $c_1^{-1}W(\theta)$ as the test statistic with $\chi^2(1)$ distribution as the limiting distribution, under the null hypothesis.

3 Bootstrap calibration

We propose using a bootstrap method to approximate the limiting distribution in (5). Such a bootstrap calibration is very attractive because then there is no need to derive the analytic form of the limiting distribution of the test statistic. In this section, we propose a bootstrap calibration method under Poisson sampling, where each unit in the finite population is independently selected using the first-order inclusion probability. That is, $I_i \sim \text{Ber}(\pi_i)$, where I_i is the sampling inclusion indicator function of unit i, and $\text{Ber}(p)$ is a Bernoulli distribution with success probability p. The bootstrap weights for Poisson sampling are constructed as follows:

Step 1 Generate $(N_1^*, \ldots, N_n^*)$ from a multinomial distribution $\text{MN}(N; p)$ with N trials and success probability vector p, where $p = (p_1, \ldots, p_n)$, $p_i \propto w_i$ and $\sum_{i=1}^n p_i = 1$.

Step 2 For each $i \in A$, obtain m_i^* from a binomial distribution $\text{Bin}(N_i^*, \pi_i)$ with N_i^* trials and success probability π_i independently.

Step 3 The bootstrap weight is computed as $w_i^* = w_i m_i^*$.

The above bootstrap procedure is a simplified version of that considered in Wang et al. (2019). Step 1 is used to generate a bootstrap finite population U_N^* containing N_i^* replicates of y_i for $i \in A$. In Step 2, a bootstrap sample is generated from U_N^* under the same Poisson sampling. Based on the generated bootstrap sample, the bootstrap weights $\{w_i^* : i \in A\}$ are obtained in Step 3.

Using the above bootstrap weights, we have

$$l_w^*(\theta) = N^{-1} \sum_{i \in A} w_i^* \log f(y_i; \theta),$$

the pseudo log-likelihood function based on the bootstrap sample. Let $\hat{\theta}^*$ be the maximizer of $l_w^*(\theta)$. The following lemma shows that the conditional distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ given the sample A approximates the sampling distribution of $\sqrt{n}(\hat{\theta} - \theta)$ asymptotically.

Lemma 1. *Under Poisson sampling and some regularity conditions in Section S1 of the Supplementary Material, we have*

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \mid A \longrightarrow N(0, \Sigma_\theta) \tag{6}$$

in distribution as $n \to \infty$, where the reference distribution in the left side of (6) is the bootstrap sampling distribution given the observations in sample A.

The proof of Lemma 1 is given in Section S2 of the Supplementary Material. The limiting distribution on the right side of (6) does not involve the sample A since we let the sample size of A increase to ∞; see Section S1 of the Supplementary Material for details. Lemma 1 is a bootstrap version of the central limit theorem for the pseudo maximum likelihood estimator $\hat{\theta}^*$. To establish a bootstrap version of the Wilks' theorem in (5), we use

$$W^*(\hat{\theta}) = -2n\{l_w^*(\hat{\theta}) - l_w^*(\hat{\theta}^*)\}$$

as the bootstrap version of $W(\theta_0)$ in (4). The following theorem shows that the conditional distribution of $W^*(\hat{\theta})$ given the sample converges in distribution to $\mathcal{G}$ as $n \to \infty$, the same asymptotic distribution of $W(\theta_0)$ in (4).

Theorem 1. *Under some regularity conditions in Section S1 of the Supplementary Material,*

$$W^*(\hat{\theta}) \mid A \longrightarrow \mathcal{G}$$

in distribution as $n \to \infty$, where $\mathcal{G}$ is the same stable distribution in (5).

The proof of Theorem 1 is given in Section S3 of the Supplementary Material.

Remark 1. *We propose using a bootstrap method to approximate the distribution of $W(\theta_0)$ under Poisson sampling. In Wang et al. (2019), similar bootstrap methods are discussed for simple random sampling and probability-proportional-to-size sampling with replacement, and Lemma 1 can be proved under certain conditions for both sampling designs; see Fuller (2009, Section 1.2) for details about these two sampling designs. Similar idea can be extended to other complex sampling designs as long as Lemma 1 holds.*

4 Likelihood ratio test

To simplify the notation, we use θ to denote the true parameter θ_0 in the following two sections. Let $\Theta_0 \subset \Theta$ be the parameter space under the null hypothesis. That is, the null hypothesis can be written as $H_0 : \theta \in \Theta_0$. In this section, we consider $H_0 : \theta_2 = \theta_2^{(0)}$ for some known vector $\theta_2^{(0)}$, where $\theta = (\theta_1, \theta_2)$. Thus, we have $\Theta_0 = \{\theta \in \Theta; \theta_2 = \theta_2^{(0)}\}$. Let $\hat{\theta}_1^{(0)}$ be the profile pseudo maximum likelihood estimator of θ_1 under $H_0 : \theta_2 = \theta_2^{(0)}$, which can be obtained by maximizing $l_w(\theta_1, \theta_2^{(0)})$ with respect to θ_1.

The pseudo likelihood ratio test statistic for testing $H_0 : \theta_2 = \theta_2^{(0)}$ is defined as

$$W(\theta_2^{(0)}) = -2n\{l_w(\hat{\theta}^{(0)}) - l_w(\hat{\theta})\}, \tag{7}$$

where $\hat{\theta}^{(0)} = (\hat{\theta}_1^{(0)}, \theta_2^{(0)})$. Under simple random sampling, $W(\theta_2^{(0)})$ in (7) is asymptotically distributed as $\chi^2(q)$ with $q = p - p_0$ and $p_0 = \dim(\Theta_0)$, where $\dim(\Theta)$ is the dimension of Θ. The following theorem, proved by Lumley and Scott (2014), presents the limiting distribution of the pseudo likelihood ratio test statistic in (7) for general sampling designs.

Theorem 2. *Under $H_0 : \theta_2 = \theta_2^{(0)}$ and some regularity conditions in Section S1 of the Supplementary Material,*

$$W(\theta_2^{(0)}) \longrightarrow \mathcal{G}_1 = \sum_{i=1}^{q} c_i Z_i^2 \tag{8}$$

in distribution as $n \to \infty$, where $c_1 \geq c_2 \geq \cdots \geq c_q > 0$ are the eigenvalues of $P = n\mathrm{var}(\hat{\theta}_2)\mathcal{I}_{22\cdot1}(\theta_1, \theta_2^{(0)})$ and $Z_1, \ldots, Z_q$ are q independent random samples from the standard normal distribution, where

$$\mathcal{I}_{22\cdot1}(\theta_1, \theta_2) = \mathcal{I}_{22}(\theta_1, \theta_2) - \mathcal{I}_{21}(\theta_1, \theta_2)\{\mathcal{I}_{11}(\theta_1, \theta_2)\}^{-1}\mathcal{I}_{12}(\theta_1, \theta_2),$$

and $\mathcal{I}_{ij}(\theta_1, \theta_2) = E[-\partial^2 \log f\{y; (\theta_1, \theta_2)\}/(\partial\theta_i \partial\theta_j^{\mathrm{T}})]$ for $i, j = 1, 2$.

Lumley and Scott (2014) proposed to estimate the limiting distribution in (8) using a design-based estimator of $P = n\mathrm{var}(\hat{\theta}_2)\mathcal{I}_{22\cdot1}(\theta_1, \theta_2^{(0)})$, but the computation can be cumbersome.

We consider an alternative test using a novel application of the bootstrap method. To do this, a bootstrap version of the pseudo likelihood ratio test statistic in (7) is obtained as

$$W^*(\hat{\theta}_2) = -2n\{l_w^*(\hat{\theta}_1^{*(0)}, \hat{\theta}_2) - l_w^*(\hat{\theta}^*)\}, \tag{9}$$

where $\hat{\theta}_1^{*(0)} = \arg\max_{\theta_1} l_w^*(\theta_1, \hat{\theta}_2)$ and $\hat{\theta}^* = \arg\max_{\theta} l_w^*(\theta)$. The following theorem shows that the conditional distribution of $W^*(\hat{\theta}_2)$ given the sample A converges in distribution to $\mathcal{G}_1$ as $n \to \infty$, the same asymptotic distribution as $W(\theta_2^{(0)})$ in (7).

Theorem 3. *Under some regularity conditions in Section S1 of the Supplementary Material,*

$$W^*(\hat{\theta}_2) \mid A \longrightarrow \mathcal{G}_1, \tag{10}$$

in distribution as $n \to \infty$, where $\mathcal{G}_1$ is defined in Theorem 2.

The proof of Theorem 3 is given in Section S4 of the Supplementary Material. By Theorem 3, we can use the empirical distribution of $W^*(\hat{\theta}_2)$ in (9) to approximate the sampling distribution of $W(\theta_2^{(0)})$ under $H_0 : \theta_2 = \theta_2^{(0)}$. Thus, the p-value for testing H_0 using $W(\theta_2^{(0)})$ can be obtained by computing the proportion of $W^*(\hat{\theta}_2)$ greater than $W(\theta_2^{(0)})$.

5 Quasi-score test

In this section, we consider another test without assuming a parametric super-population model. In the context of generalized linear regression analysis with survey data, Rao et al. (1998) developed a design-based test procedure using regression model assumptions in the super-population model. Let y be the study variable of interest and x be the vector of auxiliary variables that is used in the regression model. Suppose that the super-population model satisfies

$$E(Y_i \mid x_i) = \mu(x_i; \theta)$$

for known function $\mu(\cdot)$ with unknown parameter θ. Also, we assume a "working" model for the variance

$$\text{var}(Y_i \mid x_i) = V_0(\mu_i)$$

with known $V_0(\cdot)$ and $\mu_i = \mu(x_i; \theta)$. In this setup, a consistent estimator of θ can be obtained by solving

$$\hat{S}_w(\theta) = N^{-1} \sum_{i \in A} w_i u(\theta; y_i) = 0, \tag{11}$$

where $u(\theta; y_i) = (y_i - \mu_i)\{V_0(\mu_i)\}^{-1} (\partial \mu_i / \partial \theta)$ is the quasi-score function of θ. Under some regularity conditions (Binder; 1983), the solution $\hat{\theta}$ to (11) satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \longrightarrow N(0, \Sigma_\theta)$$

in distribution as $n \to \infty$, where $\Sigma_\theta = \mathcal{I}(\theta_0)^{-1} \Sigma_S(\theta) \{\mathcal{I}(\theta)^{-1}\}^{\mathrm{T}}, \mathcal{I}(\theta) = E\{I(\theta; Y)\}$, $I(\theta; y) = -\partial u(\theta; y)/\partial \theta^{\mathrm{T}}$, and $\Sigma_S(\theta) = \lim_{n \to \infty} [n \text{var}\{\hat{S}_w(\theta)\}]$.

We consider a bootstrap method for the quasi-score test of $H_0 : \theta_2 = \theta_2^{(0)}$ under the above setup. Given the partition $\theta = (\theta_1, \theta_2)$, we can write

$$\hat{S}_w(\theta) = \begin{pmatrix} \hat{S}_{w1}(\theta) \\ \hat{S}_{w2}(\theta) \end{pmatrix} = N^{-1} \begin{pmatrix} \sum_{i \in A} w_i u_1(\theta; y_i) \\ \sum_{i \in A} w_i u_2(\theta; y_i) \end{pmatrix} \tag{12}$$

with $u_j(\theta; y_i) = (y_i - \mu_i)\{V_0(\mu_i)\}^{-1} (\partial \mu_i / \partial \theta_j)$ for $j = 1, 2$. Let $\hat{\theta}_1^{(0)}$ be the solution to

$$\hat{S}_{w1}(\theta_1, \theta_2^{(0)}) = 0,$$

where $\hat{S}_{w1}(\theta)$ is defined in (12). The quasi-score test for $H_0 : \theta_2 = \theta_2^{(0)}$ is

$$X^2_{QS}(\theta_2^{(0)}) = \hat{S}_w(\hat{\theta}^{(0)})^{\mathrm{T}} \{\hat{I}_w(\hat{\theta}^{(0)})\}^{-1} \hat{S}_w(\hat{\theta}^{(0)}), \tag{13}$$

where $\hat{\theta}^{(0)} = (\hat{\theta}_1^{(0)}, \theta_2^{(0)})$ and $\hat{I}_w(\theta) = N^{-1} \sum_{i \in A} w_i I(\theta; y_i)$. Now, based on the partition $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, we can write

$$\hat{I}_w(\theta) = \begin{pmatrix} \hat{I}_{w11}(\theta) & \hat{I}_{w12}(\theta) \\ \hat{I}_{w21}(\theta) & \hat{I}_{w22}(\theta) \end{pmatrix}.$$

Since $\hat{\theta}^{(0)}$ satisfies $\hat{S}_{w1}(\hat{\theta}^{(0)}) = 0$, the test statistic in (13) is algebraically equivalent to

$$X^2_{QS}(\theta_2^{(0)}) = \hat{S}_{w2}(\hat{\theta}^{(0)})^{\mathrm{T}} \{\hat{I}_{w22 \cdot 1}(\hat{\theta}^{(0)})\}^{-1} \hat{S}_{w2}(\hat{\theta}^{(0)}), \tag{14}$$

where

$$\hat{I}_{w22 \cdot 1}(\theta) = \hat{I}_{w22}(\theta) - \hat{I}_{w21}(\theta) \{\hat{I}_{w11}(\theta)\}^{-1} \hat{I}_{w12}(\theta). \tag{15}$$

Note that the matrix $\hat{I}_{w22 \cdot 1}$ in (15) is a $q \times q$ matrix, where $q = \dim(\Theta_0)$, while the matrix $\hat{I}_w(\theta)$ in (13) is a $p \times p$ matrix, where $p = \dim(\Theta)$. Thus, if p is much larger than q, the computation for (14) is more efficient than the computation for (13). Now, we have the following theoretical result for $X^2_{QS}(\theta_2^{(0)})$, and its proof is given in Section S5 of the Supplementary Material.

Theorem 4. *Under $H_0 : \theta_2 = \theta_2^{(0)}$ and some regularity conditions in Section S1 of the Supplementary Material,*

$$n X^2_{QS}(\theta_2^{(0)}) \longrightarrow \mathcal{G}_1$$

in distribution as $n \to \infty$, where $\mathcal{G}_1$ is defined in Theorem 2.

We now develop a bootstrap method to approximate the limiting distribution of the quasi-score test statistic in (14). Similarly to the likelihood ratio method in Section 4, we first develop a bootstrap version of the quasi-score equation, that is,

$$\hat{S}_w^*(\theta) = N^{-1} \sum_{i \in A} w_i^* u(\theta; y_i) = 0.$$

Let $\hat{\theta}_1^{*(0)}$ be the solution to $S_{w1}^*(\theta_1, \hat{\theta}_2) = 0$. Then, the bootstrap version of the quasi-score test statistic $X_{QS}^2(\theta_2^{(0)})$ is

$$X_{QS}^{2*}(\hat{\theta}_2) = \hat{S}_{w2}^*(\hat{\theta}^{*(0)})^{\mathrm{T}} \{\hat{I}_{w22\cdot 1}^*(\hat{\theta}^{*(0)})\}^{-1} \hat{S}_{w2}^*(\hat{\theta}^{*(0)}), \tag{16}$$

where $\hat{\theta}^{*(0)} = (\hat{\theta}_1^{*(0)}, \hat{\theta}_2)$ and $\hat{I}_{w22\cdot 1}^*$ is computed similarly to (15) using the bootstrap weights. Similarly to Theorem 4, we have the following results.

Theorem 5. *Under some regularity conditions in Section S1 of the Supplementary Material,*

$$nX_{QS}^{2*}(\hat{\theta}_2) \mid A \longrightarrow \mathcal{G}_1$$

in distribution as $n \to \infty$, where $\mathcal{G}_1$ is defined in Theorem 2.

The proof of Theorem 5 is given in Section S6 of the Supplementary Material. Conditional on sample A, we can use $X_{QS}^{2*}(\hat{\theta}_2)$ to approximate the sampling distribution of $X_{QS}^2(\theta_2^{(0)})$ in (14). The bootstrap distribution can be used to control the size of the test based on $X_{QS}^2(\theta_2^{(0)})$ in (14).

Example 1. *For an illustration of the proposed bootstrap methods, we consider a logistic regression model. Specifically, we assume that $Y \in \{0, 1\}$ is a binary random variable, and $\mathrm{logit}\{\mathrm{pr}(Y = 1 \mid X = x)\} = (1, x^{\mathrm{T}})\theta$, where $\mathrm{logit}(x) = \log(p) - \log(1 - p)$ and $\theta = (\theta_1, \theta_2)$. A sample A is obtained by Poisson sampling, and the weighted score equation is*

$$\hat{S}_w(\theta) = N^{-1} \sum_{i \in A} w_i \{y_i - p_i(\theta)\}(1, x_i)^{\mathrm{T}} = 0, \tag{17}$$

where $\mathrm{logit}\{p_i(\theta)\} = \theta_1 + x_i^{\mathrm{T}}\theta_2$. In this case, we obtain $\hat{I}_w(\hat{\theta}) = N^{-1}\sum_{i\in A} w_i\hat{p}_i\{1 - \hat{p}_i\}(1, x_i)(1, x_i)^{\mathrm{T}}$ and $\hat{p}_i = p_i(\hat{\theta})$.

Suppose that we are interested in testing $H_0 : \theta_2 = \theta_2^{(0)}$. The profiled pseudo maximum likelihood estimator of θ under H_0, denoted by $\hat{\theta}^{(0)}$, can be obtained by solving (17) subject to $\theta_2 = \theta_2^{(0)}$. The quasi-score test statistic is

$$X^2_{QS}(\theta_2^{(0)}) = N^{-1}\left\{\sum_{i\in A} w_i(y_i - \hat{p}_i^{(0)})x_i^{\mathrm{T}}\right\}\left\{\sum_{i\in A} w_i\hat{p}_i^{(0)}(1-\hat{p}_i^{(0)})(x_i-\bar{x}_w)^{\otimes 2}\right\}^{-1}\left\{\sum_{i\in A} w_i(y_i-\hat{p}_i^{(0)})x_i\right\},$$

where $B^{\otimes 2} = BB^{\mathrm{T}}$,

$$\bar{x}_w = \frac{\sum_{i\in A} w_i\hat{p}_i^{(0)}(1-\hat{p}_i^{(0)})x_i}{\sum_{i\in A} w_i\hat{p}_i^{(0)}(1-\hat{p}_i^{(0)})},$$

and $\hat{p}_i^{(0)} = p_i(\hat{\theta}^{(0)})$. The bootstrap replicates of $X^2_{QS}(\theta_2^{(0)})$ can be easily constructed by replacing the sampling weights w_i and the profile pseudo maximum likelihood estimator $\hat{\theta}^{(0)}$, respectively, by the bootstrap weights w_i^* and the bootstrap profile pseudo maximum likelihood estimator $\hat{\theta}$ that solves $\hat{S}_w^*(\theta) = N^{-1}\sum_{i\in A} w_i^*\{y_i - p_i(\theta)\}(1, x_i) = 0$ subject to $\theta_2 = \hat{\theta}_2$.

6 Simple goodness-of-fit test for categorical data

Suppose that a finite population U_N is partitioned into K categories with $U_N = U_N^{(1)} \cup \cdots \cup U_N^{(K)}$ being such a partition. Denote $p_k = N_k/N$ to be the population proportion of the k-th category, where N_k is the cardinality of $U_N^{(k)}$. In this section, we are interested in simple goodness-of-fit testing $H_0 : p_k = p_k^{(0)}$ for $k = 1, \ldots, K$, where $(p_1^{(0)}, \ldots, p_K^{(0)})$ is a pre-specified vector satisfying $\sum_{k=1}^K p_k^{(0)} = 1$.

From the sample A, we compute $\hat{p}_k = \hat{N}_k/\hat{N}$ as an estimator of p_k, where $\hat{N}_k = \sum_{i\in A_k} w_i$ is a design-unbiased estimator of N_k, $A_k = A \cap U_N^{(k)}$, and $\hat{N} = \sum_{k=1}^K \hat{N}_k = \sum_{i\in A} w_i$. Then, the Pearson chi-squared goodness-of-fit test statistic for H_0 is

$$X^2(p^{(0)}) = n\sum_{k=1}^K \left(\hat{p}_k - p_k^{(0)}\right)^2/p_k^{(0)},$$

where $p^{(0)} = (p_1^{(0)}, \ldots, p_{K-1}^{(0)})^{\mathrm{T}}$. If we assume a multinomial distribution for the superpopulation model, we can compute the likelihood ratio test statistic as

$$W(p^{(0)}) = 2n\sum_{k=1}^K \hat{p}_k \log\left(\frac{\hat{p}_k}{p_k^{(0)}}\right).$$

Denoting $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_{K-1})^{\mathrm{T}}$, we have

$$\sqrt{n}\left(\hat{p} - p^{(0)}\right) \longrightarrow N(0, \Sigma_p)$$

in distribution as $n \to \infty$ under H_0, where Σ_p is the asymptotic variance of $\sqrt{n}\hat{p}$. Under simple random sampling with replacement, Σ_p is equal to $P_0 = \text{diag}(p^{(0)}) - p^{(0)}(p^{(0)})^{\text{T}}$. For other sampling designs, Σ_p is more complicated. Under some regularity conditions, according to Rao and Scott (1981), we have

$$X^2(p^{(0)}), W(p^{(0)}) \longrightarrow \mathcal{G}_2 = \sum_{k=1}^{K-1} \lambda_k Z_k^2, \tag{18}$$

in distribution as $n \to \infty$ under H_0, where $\lambda_1 \leq \cdots \leq \lambda_{K-1}$ are the eigenvalues of the design effect matrix $D = P_0^{-1}\Sigma_p$ and $Z_1, \ldots, Z_{K-1} \overset{iid}{\sim} N(0,1)$. Under simple random sampling with replacement, the limiting distribution in (18) reduces to a chi-squared distribution with $K-1$ degrees of freedom.

If a chi-squared distribution with $K-1$ degrees of freedom is blindly used as the reference distribution for $X^2(p^{(0)})$, the resulting inference can be misleading. For example, in some two-stage cluster sampling design with $\lambda_i = \lambda(> 1)$ for $i = 1, \ldots, K-1$, the type I error rate of using a chi-squared distribution with $K-1$ degrees of freedom is approximately equal to $\text{pr}\{X^2(p^{(0)}) > \chi^2_{K-1}(\alpha)\} = \text{pr}\{X > \lambda^{-1}\chi^2_{K-1}(\alpha)\}$, where $\chi^2_{K-1}(\alpha)$ is the $(1-\alpha)$ quantile of a chi-squared distribution with $K-1$ degrees of freedom, α is a significance level, and X is chi-squared distributed with $K-1$ degrees of freedom. The resulting type I error rate increases with λ, and this can be arbitrarily large by increasing λ. To overcome this problem, Rao and Scott (1981) proposed a first-order correction which compares $X_C^2(p^{(0)}) = X^2(p^{(0)})/\hat{\lambda}_+$ to $\chi^2_{K-1}(\alpha)$, where

$$\hat{\lambda}_+ = \frac{1}{k-1}\sum_{i=1}^{k} \frac{\hat{p}_i}{p_i^{(0)}}(1-\hat{p}_i)\hat{d}_i$$

and $\hat{d}_i$ is an estimated design effect of $\hat{p}_i$, which depends on the estimated variance of $\hat{p}_i$. The second-order Rao-Scott correction (Rao and Scott; 1981) requires the knowledge of the full estimated covariance matrix of the estimated proportions, but inversion of the covariance matrix is not involved unlike in the case of a Wald statistic. Stata® and other survey softwares use the Rao-Scott corrections as a default option.

We now apply the proposed bootstrap method to approximate the limiting distribution in (18), without having to compute the estimated covariance matrix, $\hat{\Sigma}_p$. To describe the proposed method, let $\hat{p}^*$ be the estimator of the population proportion p based on the bootstrap weights w_i^*. The proposed bootstrap statistics of $X^2(p^{(0)})$ and $W(p^{(0)})$ are

$$X^{2*}(\hat{p}) = n\sum_{i=1}^{K}(\hat{p}_i^* - \hat{p}_i)^2/\hat{p}_i \quad \text{and} \quad W^*(\hat{p}) = 2n\sum_i \hat{p}_i^* \log\left(\hat{p}_i^*/\hat{p}_i\right),$$

respectively. We use $\hat{p}_i$ in place of $p_i^{(0)}$ in the bootstrap test statistics. The following theorem presents the asymptotic properties of the proposed bootstrap test statistics.

Theorem 6. *Under some regularity conditions in Section S1 of the Supplementary Material,*

$$X^{2*}(\hat{p}), W^*(\hat{p}) \mid A \longrightarrow \mathcal{G}_2 \tag{19}$$

in distribution as $n \to \infty$, where $\mathcal{G}_2$ is defined in (18).

The proof of Theorem 6 is given in Section S7 of the Supplementary Material. By Theorem 6, we can use the bootstrap samples to approximate the sampling distribution of the test statistics.

7 Test of Independence

We now discuss a bootstrap test of independence in a two-way table of counts. Let $p_{ij} = N_{ij}/N$ be the population proportion for cell (i,j) with margins p_{i+} and p_{+j} for $i = 1,\ldots,R$ and $j = 1,\ldots,C$, where R and C are the numbers of rows and columns, and $\{N_{ij}; i = 1,\ldots,R, j = 1,\ldots,C\}$ is the set of population counts with margins N_{i+} and N_{+j}. Let $\hat{N}_{ij}$ be a design unbiased estimator of N_{ij} and $\hat{p}_{ij} = \hat{N}_{ij}/\hat{N}$. The chi-squared statistic and the likelihood ratio test statistic by assuming a multinomial distribution for the super-population model for testing independence $H_0 : p_{ij} = p_{i+}p_{+j}$, for all i and j are

$$X_I^2 = n\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}, \tag{20}$$

$$W_I = 2n\sum_{i=1}^{R}\sum_{j=1}^{C}\hat{p}_{ij}\log\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right). \tag{21}$$

Since X_I^2 and W_I are asymptotically equivalent under H_0, Rao and Scott (1981) have shown that

$$X_I^2, W_I \longrightarrow \mathcal{G}_3 = \sum_{l=1}^{d}\delta_l Z_l^2 \tag{22}$$

in distribution as $n \to \infty$ under H_0, where $\delta_1 \leq \ldots \leq \delta_d$ are the d eigenvalues of a design effect matrix discussed in in Section S8 of the Supplementary Material, $d = (R-1)(C-1)$, and $Z_1, \ldots, Z_d \overset{iid}{\sim} N(0,1)$.

The Rao-Scott first-order correction to X_I^2 can be written as $X_C^2 = X_I^2/\hat{\delta}_+$, which is treated as a chi-squared random variable with $(R-1)(C-1)$ degrees of freedom under H_0, where $(R-1)(C-1)\hat{\delta}_+ = \sum_d \hat{\delta}_d$ requires the terms $\hat{\delta}_d$ which depend only on the cell design effects and the row and column marginal design effects (Rao and Scott; 1984). Two-way tables should report those design effects in addition to estimated cell counts or proportions and their marginals in practice. Rao and Scott (1984) provided a unified theory for log-linear models to cover multi-way tables and other extensions.

We now consider bootstrap tests of H_0 for two-way tables. Let $\hat{p}_{ij}^*$ be the bootstrap cell proportion computed using the bootstrap weights, $\hat{p}_{i+}^* = \sum_{j=1}^{C} \hat{p}_{ij}^*$, and $\hat{p}_{+j}^* = \sum_{i=1}^{R} \hat{p}_{ij}^*$. The proposed bootstrap version of X_I^2 is

$$X_I^{2*} = n \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\{(\hat{p}_{ij}^* - \hat{p}_{i+}^*\hat{p}_{+j}^*) - (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})\}^2}{(\hat{p}_{i+}\hat{p}_{+j})}. \tag{23}$$

Under H_0, terms in the numerator of X_I^2 are identical to $\{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j} - (p_{ij} - p_{i+}p_{+j})\}^2$. That is, the bootstrap test statistic is computed by simply replacing $\{\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{+j}\}$ and $\{p_{ij}, p_{i+}, p_{+j}\}$ with $\{\hat{p}_{ij}^*, \hat{p}_{i+}^*, \hat{p}_{+j}^*\}$ and $\{\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{+j}\}$, respectively, leading to (23).

Let $\Delta_{ij} = p_{ij}/(p_{i+}p_{+j})$. Noting that $\Delta_{ij} = 1$ under H_0, the test statistic W_I may be written as

$$W_I = 2n \sum_{i=1}^{R} \sum_{j=1}^{C} \left[\hat{p}_{ij} \log \left\{ \frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}\Delta_{ij}} \right\} - (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}\Delta_{ij}) \right]. \tag{24}$$

The bootstrap version W_I^* is obtained by replacing $\{\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{+j}\}$ with $\{\hat{p}_{ij}^*, \hat{p}_{i+}^*, \hat{p}_{+j}^*\}$ and Δ_{ij} with $\hat{\Delta}_{ij}$ in (24). That is, the proposed bootstrap version of W_I is

$$W_I^* = 2n \sum_{i} \sum_{j} \left[\hat{p}_{ij}^* \log \left\{ \frac{\hat{p}_{ij}^*}{\hat{p}_{i+}^*\hat{p}_{+j}^*\hat{\Delta}_{ij}} \right\} - (\hat{p}_{ij}^* - \hat{p}_{i+}^*\hat{p}_{+j}^*\hat{\Delta}_{ij}) \right], \tag{25}$$

where $\hat{\Delta}_{ij} = \hat{p}_{ij}/(\hat{p}_{i+}\hat{p}_{+j})$. It can be shown that W_I^* is always nonnegative.

The following theorem presents some asymptotic properties of the proposed bootstrap test statistics.

Theorem 7. *Under some regularity conditions in Section S1 of the Supplementary Material,*

$$X_I^{2*}, W_I^* \mid A \longrightarrow \mathcal{G}_3 \tag{26}$$

in distribution as $n \to \infty$, where $\mathcal{G}_3$ is defined by (22).

The proof of Theorem 7 is given in Section S8 of the Supplementary Material. By Theorem 7, we can use the bootstrap distribution to approximate the sampling distribution of the test statistic X_I^2 or W_I without computing the additional terms to account for the design effects.

8 Simulation Study

8.1 Single-stage sampling

In this section, we use a single-stage sampling design to test the performance of the proposed bootstrap test statistics. A finite population of size $N = 500$ is generated in a way similar to Pfeffermann and Sverchkov (2007):

$$\begin{aligned} y_i \mid x_i &\sim N(\theta_1 + \theta_2 x_i, 1) \quad (i = 1, \ldots, N), \\ x_i &\sim \mathrm{U}(0, 5) \quad (i = 1, \ldots, N), \end{aligned}$$

where $(\theta_1, \theta_2) = (1, 1)$, and $\mathrm{U}(a, b)$ is a uniform distribution on the interval (a, b). To make the sampling design informative, a probability-proportional-to-size sampling design with replacement is used to get a sample of size n with selection probability $p_i \propto 1+|y_i+\epsilon_i|/2$ $(i = 1, \ldots, N)$ and $\sum_{i=1}^N p_i = 1$, where $\epsilon_i \sim N(0, 1)$. Specifically, a sample of size n is randomly selected from the finite population with replacement, and the selection probability of y_i is p_i. We consider two scenarios: $(N, n) = (500, 20)$ and $(N, n) = (1\,500, 50)$.

We are interested in testing $H_0 : \theta_2 = \theta_2^{(0)}$ with α=0·05 significance level and consider three different values for $\theta_2^{(0)} \in \{1, 1.1, 1.2\}$. The following testing methods are compared:

1. Naive likelihood ratio method with $W(\theta_2^{(0)})$ in (7) and $\chi^2(1)$ as the test statistic and the reference distribution, respectively.

2. Naive quasi-score method with $X^2_{QS}(\theta_2^{(0)})$ in (14) and $\chi^2(1)$ as the test statistic and the reference distribution, respectively.

3. Lumley and Scott (2014) method. The test statistic is $W_I(\theta_2^{(0)})/\hat{\delta}$, where $\hat{\delta} = n\hat{V}(\hat{\theta}_2)\hat{I}_{w,22\cdot 1}$, $\hat{V}(\hat{\theta}_2)$ is a design-based variance estimator of $\hat{\theta}_2$, and $\hat{I}_{w,22\cdot 1}$ is in (15). The reference distribution is $F_{1,k}$, where F_{ν_1,ν_2} is an F distribution with parameters ν_1 and ν_2, k is the degrees of freedom of the variance estimator based on the sampling design and k is obtained by subtracting the number of parameters from the effective sample size associated with the sampling design.

4. Bootstrap likelihood ratio method with $W(\theta_2^{(0)})$ in (7) being the test statistic, and the reference distribution is approximated by the empirical distribution of $W^*(\hat{\theta}_2)$ in (9); a brief description of the bootstrap method is given in the Section S9 of the Supplementary Material.

5. Bootstrap quasi-score method with X_{QS}^2 in (14) being the test statistic, and the reference distribution is approximated by the empirical distribution of $X_{QS}^{2*}(\hat{\theta}_2)$ in (16).

For each scenario, we generate 1 000 Monte Carlo samples, and for each sample, 1 000 iterations are used for both bootstrap methods. Table 1 summarizes the simulation results. For both scenarios, the native likelihood ratio method and quasi-score method have a significantly inflated type I error rate when the null hypothesis is true. The type I error rate for the Lumley and Scott (2014) method is larger than the nominal level under H_0 when the sample size is small, but it is approximately the same as those of the two bootstrap methods when the sample size is large. The type I error rates of both the bootstrap likelihood ratio method and the bootstrap quasi-score method are close to 0.05 under H_0. On the other hand, the power of the proposed bootstrap methods are reasonable compared to the Lumley-Scott method. The naive methods have slightly larger power but inflated type I error rates.

8.2 Stratified random sampling

In this section, we use a stratified random sampling design to test the performance of the proposed bootstrap methods. A finite population $U_N = \{(x_{i,j}, y_{i,j}) : i = 1, \ldots, N_I; j = 1, \ldots, M_i\}$ is generated based on the following steps, where N_I is the number of groups, and M_i is the size for the i-th group. That is,

Table 1: Test power for the hypothesis test $H_0 : \theta_2 = \theta_2^{(0)}$ based on 1 000 Monte Carlo simulations, and the significance level is 0.05.

(N, n)	Method	$\theta_2^{(0)}$		
		1.00	1.10	1.20
$(500, 20)$	NLR	0.09	0.14	0.28
	NQS	0.14	0.19	0.36
	LS	0.08	0.14	0.28
	BLR	0.06	0.12	0.24
	BQS	0.06	0.12	0.25
$(1500, 50)$	NLR	0.08	0.22	0.52
	NQS	0.10	0.25	0.56
	LS	0.07	0.19	0.45
	BLR	0.06	0.19	0.45
	BQS	0.06	0.19	0.45

NLR, naive likelihood ratio method; NQS, naive quasi-score method; LS, Lumley and Scott (2014) method; BLR, bootstrap likelihood ratio method; BQS: bootstrap quasi-score method.

Step 1 The finite population consists of $N_I = 5$ groups with size $M_i = N/5$ $(i = 1, \ldots, 5)$.

Step 2 For the i-th group, generate $x_{i,j} \sim N(-1 + 0.5i, 1)$ and $p_{i,j}$ by $\text{logit}(p_{i,j}) = \theta_1 + \theta_2 x_{i,j}$, where $(\theta_1, \theta_2) = (-1, 0.5)$. Generate $y_{i,j} \mid p_{i,j} \sim \text{Ber}(p_{i,j})$.

Step 3 Construct $H = 10$ strata by a cross-classification of groups and y-values. For example, stratum one consists of elements with $y = 1$ within the first group and stratum two consists of elements with $y = 0$ within the first group.

From the finite population, we perform stratified random sampling with sample size n_h in stratum h $(h = 1, \ldots, H)$, and the sampling weights are $w_{h,k} = n_h^{-1} N_h$ for $k = 1, \ldots, N_h$, where N_h is the stratum population size. The sampling design is a special case-control design, and such a sampling design is informative in the sense that we cannot ignore the sampling weights when estimating the parameters in the logistic model. We are interested in testing $H_0 : \theta_2 = \theta_2^{(0)}$ with $\alpha = 0.05$ significance level and consider three values for $\theta_2^{(0)}$: 0.5, 0.4 and 0.3. In this simulation study, we consider two scenarios: $(N, n_h) = (3\,000, 10)$ and $(N, n_h) = (12\,000, 30)$.

We compare the methods discussed in Section 8.1 based on 1 000 Monte Carlo simulations, and for the bootstrap methods, we use 1 000 iterations of the bootstrap methods. Table 2 presents the results of the simulation study. Unlike the inflated type I error rates in the previous section, the type I error rates of the two naive methods are much smaller than 0.05 since the design effect is smaller than 1 under the setup of this section. By the simulation results, we conclude that the Lumley and Scott (2014) method and the two proposed bootstrap test methods perform approximately the same under this setup, and their type I error rates are close to 0.05 for both scenarios. The test powers of the proposed bootstrap methods are much better than those of the two naive methods, and they are approximately the same as those of the Lumley and Scott (2014) method.

Table 2: Power for the hypothesis test $H_0 : \theta_2 = \theta_2^{(0)}$ based on 1 000 Monte Carlo simulations, and the significance level is 0.05.

(N, n_h)	Method	$\theta_2^{(0)}$		
		0.5	0.4	0.3
$(3000, 10)$	NLR	0.02	0.03	0.12
	NQS	0.01	0.02	0.11
	LS	0.06	0.12	0.28
	BLR	0.07	0.13	0.30
	BQS	0.06	0.12	0.28
$(12000, 30)$	NLR	0.01	0.11	0.44
	NQS	0.01	0.09	0.41
	LS	0.05	0.24	0.66
	BLR	0.05	0.24	0.67
	BQS	0.05	0.23	0.66

NLR, naive likelihood ratio method; NQS, naive quasi-score method; LS, Lumley and Scott (2014) method; BLR, bootstrap likelihood ratio method; BQS: bootstrap quasi-score method.

8.3 Two-stage cluster sampling

In this section, we consider a two-stage cluster sampling design to test the performance of the proposed bootstrap methods. A finite population $U_N = \{(x_{i,j}, y_{i,j}) : i = 1, \ldots, N_I; j = 1, \ldots, M_i\}$ is generated based on the following steps, where N_I is the number of clusters, and M_i is the size of the i-th cluster. Specifically,

$$\begin{aligned} a_i &\sim N(0,1) \quad (i = 1, \ldots, N_I), \\ M_i \mid a_i &\sim \text{Po}(25|a_i|) + C_0 \quad (i = 1, \ldots, N_I), \\ x_{i,j} &\sim N(0,4) \quad (j = 1, \ldots, M_i), \\ y_{i,j} \mid (a_i, x_{i,j}) &\sim N(\mu_{i,j}, 1) \quad (j = 1, \ldots, M_i), \end{aligned}$$

where $\text{Po}(\lambda)$ is a Poisson distribution with parameter λ, C_0 is the minimum cluster size, $\mu_{i,j} = \theta_1 + \theta_2 x_{i,j} + a_i/2$, and $(\theta_1, \theta_2) = (1, 1)$. Based on a finite population, we use a two-stage cluster sampling design to obtain a sample. The first-stage sampling design is probability-proportional-to-size sampling with replacement, where the selection probability is proportional to the cluster size, and the second-stage is based on simple random sampling. The sample size for the first stage is n_1, and that for the second stage is n_2. We consider two scenarios: $(G, C, n_1, n_2) = (30, 30, 5, 5)$ and $(G, C_0, n_1, n_2) = (60, 40, 10, 5)$.

We are interested in testing $H_0 : \theta_2 = \theta_2^{(0)}$ with $\alpha = 0.05$ significance level and consider two different values for $\theta_2^{(0)}$: 1 and 1.5. Since the likelihood function involves intractable integral forms, we do not consider likelihood ratio test and only compare the bootstrap quasi-score method with the naive quasi-score method; a brief description of the bootstrap method under two-stage cluster sampling is given in the Section S9 of the Supplementary Material.

For each scenario, we generate 1 000 Monte Carlo samples, and 1 000 iterations for each sample are used for both bootstrap methods. Table 3 summarizes the simulation results. For both scenarios, the native quasi-score method has a higher type I error rate under H_0 for both scenarios. In contrast, the type I error rate of the bootstrap quasi-score method is close to 0.05 when the null hypothesis is true for both scenarios. On the other hand, the power of the proposed bootstrap methods is reasonable compared with the naive quasi-score method.

8.4 Test of independence

In this section, we consider test of independence in a 3 × 3 table of counts to check the performance of the proposed bootstrap methods. A finite population $U_N = \{y_i : i =$

Table 3: Test power for the hypothesis test $H_0 : \theta_2 = \theta_2^{(0)}$ based on 1 000 Monte Carlo simulations, and the significance level is 0.05.

(G, C_0, n_1, n_2)	Method	$\theta_2^{(0)}$	
		1.00	1.50
$(30, 30, 5, 5)$	NQS	0.10	0.98
	BQS	0.04	0.93
$(60, 40, 10, 5)$	NQS	0.10	1.00
	BQS	0.04	1.00

NQS, naive quasi-score method; BQS: bootstrap quasi-score method.

$1, \dots, N\}$ is generated by $y_i \sim \mathrm{MN}(1; p)$, where $\mathrm{MN}(1; p)$ is a multinomial distribution with one trial and a success probability vector p, where $p = (p_{11}, \dots, p_{ij}, \dots, p_{33})^{\mathrm{T}}$, and p_{ij} is the success probability for the cell in the i-th row and j-th column for $i, j = 1, \dots, 3$. For simplicity, we assume that y_i is a dummy variable consisting of eight 0's and one 1. For the success probability vector p, we consider three cases:

$$\text{Case I} : p_{11} = 1/4, p_{12} = p_{13} = p_{21} = p_{31} = 1/8, p_{22} = p_{23} = p_{32} = p_{33} = 1/16.$$

$$\text{Case II} : p_{11} = 1/4, \, p_{12} = p_{13} = (1.4)/8, \, p_{21} = p_{31} = (0.6)/8, \, p_{22} = p_{33} = 1/16, \, p_{23} = (1.4)/16, \, p_{32} = (0.6)/16.$$

$$\text{Case III} : p_{11} = p_{2,3} = p_{3,2} = 1/6, \, p_{12} = p_{13} = p_{21} = p_{31} = p_{22} = p_{33} = 1/12.$$

Case I satisfies independence for the two-way table of counts, but Cases II–III do not. The level of non-independence can be expressed using a non-centrality parameter γ, where

$$\gamma = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(p_{ij} - p_{i+} p_{+j})^2}{p_{i+} p_{+j}}.$$

The values of γ are 0, 0.017 and 0.125 for Cases I–III, respectively.

For each y_i, we generate an auxiliary variable $x_i = \beta^{\mathrm{T}} y_i$, where $\beta = (\beta_1, \dots, \beta_9)$, $\beta_j = 1 + e_j$ for $j = 1, \dots, 9$, $e_j \sim \mathrm{Ex}(1)$, and $\mathrm{Ex}(\lambda)$ is an exponential distribution with rate λ. A probability-proportional-to-size sampling design with replacement is used to generate a sample of size n with selection probability proportional to x_i. We consider two scenarios for the population and sample sizes: $(N, n) = (2\,000, 100)$ and $(N, n) = (10\,000, 500)$.

We are interested in testing independence in the two-way table with $\alpha = 0.05$ significance level. For each sample, we consider the following five test methods:

1. Naive Pearson method based on X_I^2 in (20) with $\chi^2(4)$ being the reference distribution.
2. Naive likelihood ratio method using W_I in (21) with $\chi^2(4)$ being the reference distribution.
3. The Rao-Scott method using $X_C^2 = X_I^2/\hat{\delta}_+$ as the test statistic with $\chi^2(4)$ being the reference distribution.
4. Bootstrap Pearson method using X_I^2, and its distribution is approximated by that of X_I^{2*} in (23).
5. Bootstrap likelihood ratio method based on W_I, and its distribution is approximated by that of W_I^* in (25).

For each scenario, we generate 1 000 Monte Carlo samples, and 1 000 iterations for each sample are used for both bootstrap methods. Table 4 summarizes the simulation results. For Case I when the null hypothesis is true, the type I error rates of the two naive methods are much larger than 0.05 for different sample sizes, and the type I errors of the Rao-Scott method is lower than 0.05 for both scenarios. However, the type I error rates for the two bootstrap methods are approximately equal to 0.05 The test power of the proposed bootstrap methods increases with γ.

9 Application

We present an analysis of the 2011 Private Education Expenditure Survey (PEES) in South Korea using the proposed bootstrap methods. This dataset has been studied by Kim et al. (2017). The purpose of this survey is to study the relationship between private education expenditure and the academic performance of students before entering college.

A stratified two-stage cluster sampling design was used for the 2011 PEES, and strata consist of 16 first-tier administrative divisions, including most provinces and metropolitan cities of South Korea. For each stratum, the probability-proportional-to-size sampling design

Table 4: Power of the test procedures for independence based on 1 000 Monte Carlo simulation samples, and the significance level is 0.05.

(N, n)	Method	Case I $(\gamma = 0)$	Case II $(\gamma = 0.017)$	Case III $(\gamma = 0.125)$
$(2\,000, 100)$	NP	0.11	0.17	0.79
	NLR	0.09	0.15	0.79
	RS	0.02	0.04	0.70
	BP	0.05	0.10	0.71
	BLR	0.05	0.09	0.70
$(10\,000, 500)$	NP	0.11	0.51	1.00
	NLR	0.10	0.52	1.00
	RS	0.02	0.23	1.00
	BP	0.05	0.38	1.00
	BLR	0.05	0.37	1.00

NP, naive Pearson method; NLR: naive likelihood ratio method; RS, Rao-Scott method; BP, bootstrap Pearson method; BNR: bootstrap likelihood ratio method.

without replacement was conducted in the first stage, and the primary sampling unit was the school. Students are randomly selected in the second stage. There are about 1 000 sample schools and 45 000 students involved in this survey.

For student i in the sample A, let y_i be the academic performance assessed by the teacher, and it takes a value from 1 through 3 corresponding to low, middle and high academic performance, respectively. Associated with y_i, let x_i be the covariates of interest. As discussed by Kim et al. (2017), we consider the following covariates: after-school education, hours taking lessons provided by the school after regular classes in a month; private education, hours taking private lessons in a month; gender, 1 for female and 0 for male; household income per month; father's education, 1 for college or higher and 0 otherwise; mother's education, 1 for college or higher and 0 otherwise.

In this section, we study the academic performance of students in middle school and high school separately, and we are interested in estimating the conditional probability of achieving high academic performance. Specifically, consider the following logistic model,

$$\text{logit}\{\text{pr}(Y = 1 \mid x)\} = (1, x^{\mathrm{T}})\theta, \tag{27}$$

where $Y = 1$ if high academic performance is achieved and 0 otherwise, x is a vector of six covariates, $\theta = (\theta_0, \theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $p = 6$. We are interested in testing the null hypotheses $H_{0,i} : \theta_i = 0$ for $i = 1, \ldots, p$ with $\alpha = 0.05$ significance level. Since the Wald method is widely used in practice, the naive likelihood ratio method, naive quasi-score method, bootstrap likelihood ratio method, bootstrap quasi-score method with 1 000 iterations and a two-sided Wald test are compared. The p-values for the two naive methods and the Wald test are obtained using reference distributions for simple random sampling. Specifically, the reference distribution for the two naive methods is $\chi^2(1)$, and that of the Wald test is a normal distribution with estimated variance (Fuller; 2009, Section 1.2.8) for $\hat{\theta}$ by the sandwich formula. The p-values for the proposed bootstrap methods are obtained by bootstrap empirical distributions of the corresponding test statistics.

Estimation results are summarized in Table 5. The two bootstrap testing methods and the Wald test perform approximately the same in terms of the p-values. The p-values of two naive methods are approximately the same, but they differ from those of the bootstrap methods, especially for "after school education" and "gender" covariates in the middle school level and "gender" and "father's education" covariates in the high school level, as the naive methods may not properly reflect the intra-cluster correlation in the cluster sampling.

Based on the two bootstrap testing methods in Table 5, we have the following conclusions under 0.05 significance level. Controlling other covariates, the probability of female students achieving high academic performance is significantly higher than that of male students in middle school, but the gender effect is not significant in the high school. The hours spent on private education and after-school education can increase the probability of achieving high academic performance significantly in both middle school and high school. The household income and mother's education level have a significant positive influence on their child's academic performance. However, father's education only has a significant influence during the middle school period. The estimated coefficients and testing results are approximately the same as Kim et al. (2017), who used a random effect model to analyze this dataset.

10 Concluding Remarks

Many statistical agencies provide microdata files to analysts containing survey weights and several sets of associated replication weights, in particular bootstrap weights. Standard

Table 5: Estimates (Est) and the p-values (Unit: 10^{-2}) for testing $H_{0,i} : \theta_i = 0$, where $i = 1, \ldots, 6$, in (27) for the middle school and high school.

School level	Cov	Est	p-value (Unit: 10^{-2})				
			NLR	NQS	BLR	BQS	Wald
Middle School	After-school Edu	0.03	0.0	0.0	2.6	2.1	2.6
	Private Edu	0.05	0.0	0.0	0.0	0.0	0.0
	Gender	0.19	0.3	0.3	2.2	2.2	2.6
	Income	0.1	0.0	0.0	0.0	0.0	0.0
	Father's Edu	0.45	0.0	0.0	0.0	0.0	0.0
	Mother's Edu	0.36	0.0	0.0	0.1	0.1	0.1
High School	After-school Edu	0.03	0.0	0.0	0.0	0.0	0.0
	Private Edu	0.04	0.0	0.0	0.0	0.0	0.0
	Gender	0.09	6.3	6.3	21.4	21.4	20.3
	Income	0.05	0.0	0.0	2.3	2.3	2.3
	Father's Edu	0.14	2.3	2.3	14.9	14.9	13.8
	Mother's Edu	0.24	0.0	0.0	1.2	1.2	0.9

NLR, naive likelihood ratio method; NQS, naive quasi-score method; BLR, bootstrap likelihood ratio method; BQS: bootstrap quasi-score method; Wald, two-sided Wald test.

statistical packages often permit the use of survey weighted test statistics, and we have shown how to approximate their distributions under the null hypothesis by their bootstrap analogues computed from the bootstrap weights supplied in the data file. We studied weighted likelihood ratio tests and weighted score tests based on weighted score or estimating equations. It would be useful to extend our results to the case of imputation for missing item responses.

We also studied the case of categorical data by developing bootstrap procedures for testing simple goodness of fit and independence in a two-way table. We plan to extend our bootstrap method for categorical data to testing hypotheses from multi-way tables of weighted counts or proportions, using a log-linear model approach proposed by Rao and Scott (1984).

Our theory depends on establishing bootstrap central limit theorems under the specified sampling design. We have established such theorems for simple random sampling without replacement, probability proportional to size sampling with replacement and Poisson sampling. We plan to establish similar central limit theorems for other sampling designs including stratified multi-stage sampling.

Supplementary Material

The supplementary material contains regularity conditions and proofs for Lemma 1, Theorem 1 and Theorems 3–7.

S1 Regularity conditions

To discuss the asymptotic properties of the proposed test statistics, we need the following regularity conditions:

Condition 1 The density function $f(y;\theta)$ in Sections 2–4 and the conditional mean function $\mu(x;\theta)$ in Section 5 have continuous second-order derivatives with respect to θ.

Condition 2 For any $\theta \in \mathcal{B}$, $E\{\|S(\theta;Y)\|^6\}$ is bounded away from infinity, where $\|\cdot\|_2$ is the l_2 norm, and $\mathcal{B}$ is a close interval with $\theta_0 \in \mathcal{B}$.

Condition 3 $n\text{var}\{\hat{S}_w(\theta)\}$ converges to a positive definitive matrix $\Sigma_S(\theta)$ for $\theta \in \mathcal{B}$.

Condition 4 For $\theta \in \mathcal{B}$, $\hat{I}_w(\theta) \to \mathcal{I}(\theta)$ in probability.

Condition 5 A central limit theorem holds for the weighted score function $\hat{S}_w(\theta_0)$. Specifically,

$$\text{var}\{\hat{S}_w(\theta_0)\}^{-1/2}\hat{S}_w(\theta_0) \to N(0, I)$$

in distribution as $n \to \infty$, where I is the identity matrix.

Condition 6 For Poisson sampling, $n_0 \to \infty$, $n_0/N \to 0$, and the inclusion probability π_i satisfies $C_1 \le N{n_0}^{-1}\pi_i \le C_2,$, where $n_0 = \sum_{i=1}^N \pi_i$, and C_1 and C_2 are two positive constants.

Condition 1 is a commonly used to study the pseudo maximum likelihood estimator, and it guarantees the existence of the score equation and information matrix. Condition 2 is used to show the central limit theorem for the bootstrap estimator in Lemma 1. The convergence results in Conditions 3–4 are used to derive the asymptotic distributions, and Condition 3 is changed to $n_0\text{var}(\hat{S}_w(\theta))$ converges to a positive definitive matrix $\Sigma_S(\theta)$ for $\theta \in \mathcal{B}$ under Poisson sampling since the realized sample size n is a random variable. Condition 5 is widely assume to get the central limit theorem in survey sampling (Fuller; 2009, Section 1.3.2).

Condition 6 is specific to Poisson sampling, and different conditions should be assumed for other complex sampling designs. These conditions hold for the likelihood-based methods discussed in Sections 2–4 and the semi-parametric method in Section 5.

S2 Proof of Lemma 1

Let

$$\hat{V}\{\hat{S}_w(\theta)\} = \frac{1}{N^2}\sum_{i\in A}\frac{1-\pi_i}{\pi_i^2}S(\theta;y_i)^{\otimes 2} \tag{S.1}$$

be a design-unbiased variance estimator of $\hat{S}_w(\theta)$ under Poisson sampling. First, we show that

$$n_0 a^{\mathrm{T}}\hat{V}\{\hat{S}_w(\theta)\}a \to a^{\mathrm{T}}\Sigma_S(\theta)a \tag{S.2}$$

in probability for $a \in \mathbb{R}^p$ such that $\|a\|_2 = 1$. For $\theta \in \mathcal{B}$, consider

$$\begin{aligned}\mathrm{var}\{a^{\mathrm{T}}\hat{S}_w(\theta)\} &= E[\mathrm{var}\{a^{\mathrm{T}}\hat{S}_w(\theta)\mid U_N\}] + \mathrm{var}[E\{a^{\mathrm{T}}\hat{S}_w(\theta)\mid U_N\}] \\ &= E\left[\frac{1}{N^2}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\left\{a^{\mathrm{T}}S(\theta;y_i)\right\}^2\right] + \mathrm{var}\left\{\frac{1}{N}\sum_{i=1}^{N}a^{\mathrm{T}}S(\theta;y_i)\right\}. \end{aligned}\tag{S.3}$$

Under Condition 2, we can show that the second term of (S.3) is $O(N^{-1})$, and the first term of (S.3) is $O(n_0^{-1})$ under Condition 2 and Condition 6. Denoting $V_N = N^{-2}\sum_{i=1}^{N}\pi_i^{-1}(1-\pi_i)\left\{a^{\mathrm{T}}S(\theta;y_i)\right\}^2$, it can be shown that $V_N - E(V_N) = o_p(n_0^{-1})$. To sum up, under Condition 2 and Condition 6, we have

$$\mathrm{var}\{a^{\mathrm{T}}\hat{S}_w(\theta)\} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\left\{a^{\mathrm{T}}S(\theta;y_i)\right\}^2 + o_p(n_0^{-1}). \tag{S.4}$$

Consider

$$E[a^{\mathrm{T}}\hat{V}\{\hat{S}_w(\theta)\}a \mid U_N] = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\left\{a^{\mathrm{T}}S(\theta;y_i)\right\}^2 \tag{S.5}$$

$$\begin{aligned}\mathrm{var}[a^{\mathrm{T}}\hat{V}\{\hat{S}_w(\theta)\}a \mid U_N] &= \frac{1}{N^4}\sum_{i=1}^{N}\frac{(1-\pi_i)^3}{\pi_i^3}\left\{a^{\mathrm{T}}S(\theta;y_i)\right\}^4 \\ &\le \frac{1}{(C_1 n_0)^3}\frac{1}{N}\sum_{i=1}^{N}\left\{a^{\mathrm{T}}S(\theta;y_i)\right\}^4 \\ &= o_p(n_0^{-2}), \end{aligned}\tag{S.6}$$

where the inequality of (S.6) holds by Condition 6, and the last equality of (S.5) holds by Condition 2 and the Etemadi's law of large numbers (Athreya and Lahiri; 2006, Theorem 8.2.7). By (S.4)–(S.6) and Condition 3, we have proved (S.2). In a similar manner, we can show that

$$n_0 a^{\mathrm{T}} \hat{V}\{\hat{S}_w(\theta)\} b \to a^{\mathrm{T}} \Sigma_S(\theta) b \tag{S.7}$$

for vectors $a \in \mathbb{R}^k$ and $b \in \mathbb{R}^k$ such that $\|a\| = \|b\| = 1$.

Let $\hat{S}_w^*(\theta) = \partial l_w^*(\theta)/\partial\theta$ be the bootstrap score function, and recall that $l_w^*(\theta) = N^{-1}\sum_{i\in A} w_i^* \log f(y_i;\theta)$. Next, we show, conditional on U_N^*,

$$\mathrm{var}_*\{\hat{S}_w^*(\theta) \mid U_N^*\}^{-1/2}\{\hat{S}_w^*(\theta) - \hat{S}_w(\theta)\} \to N(0, I) \tag{S.8}$$

in distribution as $n \to \infty$, and it implies

$$\mathrm{var}_*\{\hat{S}_w^*(\hat{\theta}) \mid U_N^*\}^{-1/2}\hat{S}_w^*(\hat{\theta}) \to N(0, I) \tag{S.9}$$

in distribution conditional on U_N^* as $n \to \infty$ since $\hat{S}_w(\hat{\theta}) = 0$, where $\mathrm{var}_*(\cdot)$ is the variance operator with respect to the bootstrap procedure.

Under Condition 6, we can show $N^{-1}\sum_{i\in A}\pi_i^{-1} = 1 + O_p(n_0^{-1/2})$. Based on Step 1 of the bootstrap method, we have

$$\begin{aligned} E_*(N_i^*\pi_i) &= \pi_i N \frac{\pi_i^{-1}}{\sum_{j\in A}\pi_j^{-1}} = 1 + O_p(n_0^{-1/2}), &\text{(S.10)}\\ \mathrm{var}_*(N_i^*\pi_i) &\leq \pi_i^2 N \frac{\pi_i^{-1}}{\sum_{j\in A}\pi_j^{-1}} = O(n_0/N). &\text{(S.11)}\end{aligned}$$

By (S.10), (S.11) and Condition 6, we have shown that

$$N_i^*\pi_i = 1 + o_p(1). \tag{S.12}$$

Recall that $\hat{S}_w^*(\hat{\theta}) = N^{-1}\sum_{i\in A}\pi_i^{-1} m_i^* S(\hat{\theta}; y_i)$, where $m_i^* \sim \mathrm{Bin}(N_i^*, \pi_i)$. Denote $\mu_i^* = E_*(m_i^* \mid N_i^*) = N_i^*\pi_i$, and we have

$$E_*\{\hat{S}_w^*(\theta) \mid U_N^*\} = \frac{1}{N}\sum_{i\in A}\frac{\mu_i^*}{\pi_i} S(\theta; y_i) = \hat{S}_w(\theta)\{1 + o_p(1)\},$$

where $E_*(\cdot)$ is the expectation operator with respect to the bootstrap procedure, and the last equality holds by (S.12), Condition 2, Condition 6 and the Cauchy-Schwarz inequality; see (0.6.3) of Horn and Johnson (2017).

Now, consider

$$\begin{aligned}
\mathrm{var}_*\{\hat{S}_w^*(\theta) \mid U_N^*\} &= \frac{1}{N^2}\sum_{i\in A}\frac{\mathrm{var}_*(m_i^* \mid N_i^*)}{\pi_i^2}S(\theta;y_i)^{\otimes 2} \\
&= \frac{1}{N^2}\sum_{i\in A}\frac{N_i^*\pi_i(1-\pi_i)}{\pi_i^2}S(\theta;y_i)^{\otimes 2} \\
&= \frac{1+o_p(1)}{N^2}\sum_{i\in A}\frac{1-\pi_i}{\pi_i^2}S(\theta;y_i)^{\otimes 2} \\
&= \{1+o_p(1)\}\hat{V}(\hat{S}_w(\theta)), \qquad \text{(S.13)}
\end{aligned}$$

where the third equation of (S.13) follows by (S.12), Condition 2 and the Cauchy-Schwarz inequality.

For $\theta \in \mathcal{B}$ and $a \in \mathbb{R}^p$ with $\|a\|_2 = 1$, by (S.13), we have

$$\mathrm{var}_*\{a^{\mathrm{T}}\hat{S}_w^*(\theta) \mid U_N^*\} = \frac{1+o_p(1)}{N^2}\sum_{i\in A}\frac{1-\pi_i}{\pi_i^2}\{a^{\mathrm{T}}S(\theta;y_i)\}^2 = O_p(n_0^{-1}), \qquad \text{(S.14)}$$

where the last equality holds by (S.2) and (S.13). Next, consider

$$\begin{aligned}
E_*(|m_i^* - \mu_i^*|^3 | \, U_N^*) &= E_*\{(m_i^* - \mu_i^*)^3 \mid U_N^*\} + 2P(m_i^* = 0 \mid N_i^*) + o_p(1) \\
&= N_i^*\pi_i(1-\pi_i)(1-2\pi_i) + 2P(m_i^* = 0 \mid N_i^*) + o_p(1) \\
&= O_p(1), \qquad \text{(S.15)}
\end{aligned}$$

where the first and last equalities holds by Condition 6 and (S.12).

By (S.15), we have

$$\frac{1}{N^3}\sum_{i\in A}\frac{\{a^{\mathrm{T}}S(\theta;y_i)\}^3}{\pi_i^3}E_*(|m_i^* - \mu_i^*|^3 | \, U_N^*) \;=\; \frac{O_p(1)}{N^3}\sum_{i\in A}\frac{\{a^{\mathrm{T}}S(\theta;y_i)\}^3}{\pi_i^3}. \qquad \text{(S.16)}$$

Following the limits of (S.5) and (S.6), we can show that the order of (S.16) is $O_p(n^{-2})$ under Condition 2 and Condition 6. By (S.14) and (S.16), we can apply Lyapunov's central limit theorem (Athreya and Lahiri; 2006, Corollary 11.1.4) and the Cramér-Wold device (Athreya and Lahiri; 2006, Theorem 10.4.5) to prove (S.8).

Using the second-order Taylor expansion, we obtain

$$\begin{aligned}
0 = \hat{S}_w^*(\hat{\theta}^*) &= \hat{S}_w^*(\hat{\theta}) - \{\hat{I}_w^*(\hat{\theta})\}(\hat{\theta}^* - \hat{\theta}) \\
&= \hat{S}_w^*(\hat{\theta}) - \{\hat{I}_w(\hat{\theta}) + o_p(1)\}(\hat{\theta}^* - \hat{\theta})
\end{aligned}$$

where $\hat{I}_w^*(\theta) = -\partial \hat{S}_w^*(\theta)/\partial\theta$ and $\hat{I}_w(\theta)$ is defined in (2). Therefore, we have

$$\hat{\theta}^* - \hat{\theta} = \left\{\hat{I}_w(\hat{\theta})\right\}^{-1} \hat{S}_w^*(\hat{\theta}) + o_p(n^{-1/2}). \tag{S.17}$$

Combining (S.9) with (S.2), (S.7), (S.13) and (S.17), we can establish the following result (S.18) using Condition 1 and Condition 4, noting that $\hat{\theta} \to \theta_0$ in probability and $n/n_0 = 1 + o_p(1)$. That is,

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \mid U_N^* \to N(0, \Sigma_\theta). \tag{S.18}$$

By (S.18), we have $\text{pr}\{\sqrt{n(a^{\mathrm{T}}\Sigma_\theta a)^{-1}} a^{\mathrm{T}}(\hat{\theta}^* - \hat{\theta}) < x \mid U_N^*\} \to \Phi(x)$ for $x \in \mathbb{R}$ and $a \in \mathbb{R}^p$ with $\|a\|_2 = 1$, where $\Phi(x)$ is the distribution function of a standard normal distribution. That is, for $x \in \mathbb{R}$ and $\epsilon \in (0, \infty)$, there exists $N_{x,\epsilon}$ such that $|\text{pr}\{\sqrt{n(a^{\mathrm{T}}\Sigma_\theta a)^{-1}} a^{\mathrm{T}}(\hat{\theta}^* - \hat{\theta}) < x \mid U_N^*\} - \Phi(x)| < \epsilon$ for $N > N_{x,\epsilon}$. Since $\text{pr}\{\sqrt{n(a^{\mathrm{T}}\Sigma_\theta a)^{-1}} a^{\mathrm{T}}(\hat{\theta}^* - \hat{\theta}) < x \mid A\} = E_*[\text{pr}\{\sqrt{n(a^{\mathrm{T}}\Sigma_\theta a)^{-1}} a^{\mathrm{T}}(\hat{\theta}^* - \hat{\theta}) < x \mid U_N^*\}]$, we have

$$\Phi(x) - \epsilon < \text{pr}\{\sqrt{n(a^{\mathrm{T}}\Sigma_\theta a)^{-1}} a^{\mathrm{T}}(\hat{\theta}^* - \hat{\theta}) < x \mid A\} < \Phi(x) + \epsilon$$

for $N > N_{x,\epsilon}$. That is, $\text{pr}\{\sqrt{n(a^{\mathrm{T}}\Sigma_\theta a)^{-1}} a^{\mathrm{T}}(\hat{\theta}^* - \hat{\theta}) < x \mid A\} \to \Phi(x)$ for $x \in \mathbb{R}$. Thus, we have proved (6).

S3 Proof of Theorem 1

Using (6) and by the second-order Taylor expansion, we have

$$l_w^*(\hat{\theta}) = l_w^*(\hat{\theta}^*) + \hat{S}_w^*(\hat{\theta}^*)^{\mathrm{T}}(\hat{\theta} - \hat{\theta}^*) - (\hat{\theta}^* - \hat{\theta})^{\mathrm{T}} \hat{I}_w^*(\hat{\theta})(\hat{\theta}^* - \hat{\theta}) + o_p(n^{-1}). \tag{S.19}$$

Since $\hat{S}_w^*(\hat{\theta}^*) = 0$ and $\hat{I}_w^*(\hat{\theta}) = \hat{I}_w(\hat{\theta}) + o_p(1)$, we have

$$\begin{aligned} -2n\left\{l_w^*(\hat{\theta}) - l_w^*(\hat{\theta}^*)\right\} &= n(\hat{\theta}^* - \hat{\theta})^{\mathrm{T}} \hat{I}_w(\hat{\theta})(\hat{\theta}^* - \hat{\theta}) + o_p(1) \\ &= n S_w^*(\hat{\theta})^{\mathrm{T}}\{\hat{I}_w(\hat{\theta})\}^{-1} S_w^*(\hat{\theta}) + o_p(1), \end{aligned} \tag{S.20}$$

where the second equality follows from (S.17). By (S.2), (S.9) and (S.13), we can show that, conditional on U_N^*, $W^*(\hat{\theta})$ converges in distribution to the weighted sum of p independent $\chi^2(1)$ variables as $n \to \infty$, where the weights are the eigenvalues of $n\{\hat{I}_w(\hat{\theta})\}^{-1}\hat{V}\{\hat{S}_w\}$ which converges in probability to $\mathcal{I}(\theta_0)^{-1}\Sigma_S(\theta_0)$. By the fact that $\mathcal{I}(\theta_0)^{-1}\Sigma_S(\theta_0) = \Sigma_\theta \mathcal{I}(\theta_0)^{\mathrm{T}}$ and $\mathcal{I}_\theta$ is symmetric under Condition 1, Theorem 1 is established.

S4 Proof of Theorem 3

Since $\hat{\theta}^{*(0)} = (\hat{\theta}_1^{*(0)}, \hat{\theta}_2)$ where $\hat{\theta}_1^{*(0)}$ is the maximizer of $l_w^*(\theta_1, \hat{\theta}_2)$, we can obtain, similarly to (S.19),

$$l_w^*(\hat{\theta}) \;=\; l_w^*(\hat{\theta}^*) + \hat{S}_{w1}^*(\hat{\theta}^{*(0)})^{\mathrm{T}}(\hat{\theta}_1 - \hat{\theta}_1^{*(0)}) - (\hat{\theta}_1^{*(0)} - \hat{\theta}_1)^{\mathrm{T}}\hat{I}_{w11}^*(\hat{\theta})(\hat{\theta}_1^{*(0)} - \hat{\theta}_1) + o_p(n^{-1}).$$

where $\hat{S}_{w1}^*(\theta) = \partial l_w^*(\theta)/\partial\theta_1$ and $\hat{I}_{w11}^*(\theta) = \partial^2 l_w^*(\theta)/(\partial\theta_1\partial\theta_1^{\mathrm{T}})$. By the definition of $\hat{\theta}^{*(0)}$, we have $\hat{S}_{w1}^*(\hat{\theta}^{*(0)}) = 0$. Thus, using $\hat{I}_{w11}^*(\hat{\theta}) = \hat{I}_{w11}(\hat{\theta}) + o_p(1)$, we have

$$\begin{aligned} -2n\{l_w^*(\hat{\theta}) - l_w^*(\hat{\theta}^{*(0)})\} &= n(\hat{\theta}_1^{*(0)} - \hat{\theta}_1)^{\mathrm{T}}\hat{I}_{w11}(\hat{\theta})(\hat{\theta}_1^{*(0)} - \hat{\theta}_1) + o_p(1) \\ &= nS_{w1}^*(\hat{\theta})^{\mathrm{T}}\{\hat{I}_{w11}(\hat{\theta})\}^{-1}S_{w1}^*(\hat{\theta}) + o_p(1), \qquad \text{(S.21)} \end{aligned}$$

where the last equality follows from $\hat{\theta}_1^{*(0)} - \hat{\theta}_1 = \{\hat{I}_{w11}(\hat{\theta})\}^{-1}\hat{S}_{w1}^*(\hat{\theta}) + o_p(n^{-1/2})$. Thus, combining (S.20) with (S.21), we have

$$\begin{aligned} W^*(\hat{\theta}_2) &= -2n\{l_w^*(\hat{\theta}^{*(0)}) - l_w^*(\hat{\theta}^*)\} \\ &= -2n\{l_w^*(\hat{\theta}) - l_w^*(\hat{\theta}^*)\} + 2n\{l_w^*(\hat{\theta}) - l_w^*(\hat{\theta}^{*(0)})\} \\ &= n\begin{pmatrix} S_{w1}^*(\hat{\theta}) \\ S_{w2}^*(\hat{\theta}) \end{pmatrix}^{\mathrm{T}} \begin{bmatrix} \hat{I}_{w11}(\hat{\theta}) & \hat{I}_{w12}(\hat{\theta}) \\ \hat{I}_{w21}(\hat{\theta}) & \hat{I}_{w22}(\hat{\theta}) \end{bmatrix}^{-1} \begin{pmatrix} S_{w1}^*(\hat{\theta}) \\ S_{w2}^*(\hat{\theta}) \end{pmatrix} \\ &\quad -nS_{w1}^*(\hat{\theta})^{\mathrm{T}}\{\hat{I}_{w11}(\hat{\theta})\}^{-1}S_{w1}^*(\hat{\theta}) + o_p(1) \\ &= n\{S_{w2}^*(\hat{\theta}) - \hat{B}_{21}S_{w1}^*(\hat{\theta})\}^{\mathrm{T}}\{\hat{I}_{w22\cdot1}(\hat{\theta})\}^{-1}\{S_{w2}^*(\hat{\theta}) - \hat{B}_{21}S_{w1}^*(\hat{\theta})\} \end{aligned}$$

where

$$\hat{B}_{21} = \hat{I}_{w21}(\hat{\theta})\{\hat{I}_{w11}(\hat{\theta})\}^{-1}$$

and

$$\hat{I}_{w22\cdot1}(\hat{\theta}) = \hat{I}_{w22}(\hat{\theta}) - \hat{I}_{w21}(\hat{\theta})\{\hat{I}_{w11}(\hat{\theta})\}^{-1}\hat{I}_{w12}(\hat{\theta}).$$

Therefore, by (S.9), using the same argument for proving Theorem 1, we can show that $W^*(\hat{\theta}_2)$ converges in distribution to the weighted sum of q independent $\chi^2(1)$ variables as $n \to \infty$, where the weights are the eigenvalues of $n\mathrm{var}\{\hat{S}_{w,2\cdot1}(\theta)\}\mathcal{I}_{2\cdot1}(\theta_1, \theta_2^{(0)})^{-1}$, and $\hat{S}_{w,2\cdot1}(\theta) = \hat{S}_{w2}(\theta) - \hat{B}_{21}\hat{S}_{w1}(\theta)$. By a similar argument used in the proof of Lemma 1 and some basic algebra, we can show that the eigenvalues of $n\mathrm{var}\{\hat{S}_{w,2\cdot1}(\theta)\}\mathcal{I}_{2\cdot1}(\theta_1, \theta_2^{(0)})^{-1}$ are the same as those of $\Sigma_{\theta,2}\mathcal{I}_{2\cdot1}(\theta_1, \theta_2^{(0)})$.

S5 Proof of Theorem 4

Since $\hat{\theta}$ solves $\hat{S}_w(\theta) = 0$, we have

$$0 = \hat{S}_w(\hat{\theta}) = \hat{S}_w(\theta) - \hat{I}_w(\theta)(\hat{\theta} - \theta) + o_p(n^{-1/2}), \tag{S.22}$$

where the Taylor expansion holds by Condition 1, and the remaining term is guaranteed by Condition 1 and Condition 5. Thus, by (S.22), we have

$$(\hat{\theta} - \theta) = \hat{I}_w(\theta)^{-1}\hat{S}_w(\theta) + o_p(n^{-1/2})$$

using Condition 4. Since $\hat{S}_{w1}(\hat{\theta}^{(0)}) = 0$, we have $\hat{S}_{w,2\cdot1}(\hat{\theta}^{(0)}) = \hat{S}_{w2}(\hat{\theta}^{(0)})$ and

$$\hat{\theta}_2 - \theta_2^{(0)} = \{\hat{I}_{w,22\cdot1}(\hat{\theta}^{(0)})\}^{-1}\hat{S}_{w2}(\hat{\theta}^{(0)}) + o_p(n^{-1/2}). \tag{S.23}$$

By (14) and (S.23), we have

$$\begin{aligned} X^2_{QS}(\theta_2^{(0)}) &= \hat{S}_{w2}(\hat{\theta}^{(0)})^{\mathrm{T}}\{\hat{I}_{w,22\cdot1}(\hat{\theta}^{(0)})\}^{-1}\hat{S}_{w2}(\hat{\theta}^{(0)}) \\ &= (\hat{\theta}_2 - \theta_2^{(0)})^{\mathrm{T}}\hat{I}_{w,22\cdot1}(\hat{\theta}^{(0)})(\hat{\theta}_2 - \theta_2^{(0)}) + o_p(n^{-1}). \end{aligned} \tag{S.24}$$

By the asymptotic result in (2), Theorem 2 and (S.24), we have proved Theorem 4.

S6 Proof of Theorem 5

The proof of Theorem 5 is essentially similar to that of Theorem 4. Instead of using the asymptotic normality result in (2), we use the results in Lemma 1 to establish Theorem 5, and we omit the proof for simplicity.

S7 Proof of Theorem 6

We can express

$$X^{2*}(\hat{p}) = (\hat{p}^* - \hat{p})^{\mathrm{T}}\hat{P}_0^{-1}(\hat{p}^* - \hat{p}) \tag{S.25}$$

where $\hat{P}_0 = \mathrm{diag}(\hat{p}) - \hat{p}\hat{p}^{\mathrm{T}}$. Using a similar argument as in Lemma 1, we can show that the proposed bootstrap method satisfies

$$\sqrt{n}(\hat{p}^* - \hat{p}) \mid A \longrightarrow N(0, \Sigma_p) \tag{S.26}$$

in distribution as $n \to \infty$. It now follows from (S.25) and (S.26) that (4) holds since we can show that $\hat{p} \to p^{(0)}$ in probability under H_0. Results (19) for $W^*(\hat{p})$ also holds noting that $X^{2*}(\hat{p})$ and $W^*(\hat{p})$ are asymptotically equivalent with respect to the bootstrap distribution.

S8 Proof of Theorem 7

We present a brief justification of the proposed bootstrap method for testing independence in a two-way table of cell proportions or counts. Using the notation of Rao and Scott (1981), let $h(p)$ be the $d = (R-1)(C-1)$ dimensional vector with elements $h_{ij}(p) = p_{ij} - p_{i+}p_{+j}$, $i = 1, \ldots, R-1; j = 1, \ldots, C-1$, where $p = (p_{11}, p_{12}, \ldots, p_{RC-1})^{\mathrm{T}}$. Then the chi-squared statistic X_I^2, under H_0, may be expressed in a matrix form as

$$X_I^2 = n\{h(\hat{p}) - h(p)\}^{\mathrm{T}}(\hat{P}_{R+}^{-1} \otimes \hat{P}_{+C}^{-1})\{h(\hat{p}) - h(p)\},$$

where $\hat{P}_{R+} = \mathrm{diag}(\hat{p}_{R+}) - \hat{p}_{R+}\hat{p}_{R+}^{\mathrm{T}}$ and $\hat{P}_{+C} = \mathrm{diag}(\hat{p}_{+C}) - \hat{p}_{+C}\hat{p}_{+C}^{\mathrm{T}}$ with $\hat{p}_{R+} = (\hat{p}_{1+}, \ldots, \hat{p}_{R-1,+})^{\mathrm{T}}$ and $\hat{p}_{+C} = (\hat{p}_{+1}, \ldots, \hat{p}_{+,C-1})^{\mathrm{T}}$ and $\otimes$ denotes the direct product. Now, noting that $\sqrt{n}(\hat{p} - p) \longrightarrow N(0, \Sigma_p)$ in distribution as $n \to \infty$, it follows that

$$\sqrt{n}\{h(\hat{p}) - h(p)\} \longrightarrow N(0, H\Sigma_p H^{\mathrm{T}})$$

in distribution as $n \to \infty$, where $H = \partial h(p)/\partial p^{\mathrm{T}}$ is the $d \times (RC-1)$ matrix of partial derivatives of $h(p)$. Using the above result, we get (22) where the δ_l $(l = 1, \ldots, d)$ are the eigenvalues of the design effect matrix $D_h = \left(P_{R+}^{-1} \otimes P_{+C}^{-1}\right)(H\Sigma_p H^{\mathrm{T}})$.

Turning to the proposed bootstrap method, we can express the bootstrap version of X_I^2 in a matrix form as

$$X_I^{2*} = n\{h(\hat{p}^*) - h(\hat{p})\}^{\mathrm{T}}(\hat{P}_{R+}^{*-1} \otimes \hat{P}_{+C}^{*-1})\{h(\hat{p}^*) - h(\hat{p})\}.$$

Similar to the proof of Lemma 1, we have

$$\sqrt{n}\,\{h(\hat{p}^*) - h(\hat{p})\} \mid A \longrightarrow N(0, H\Sigma_p H^{\mathrm{T}})$$

in distribution as $n \to \infty$, so the result (26) for X_I^{2*} holds.

Since X_I^2 and W_I are asymptotically equivalent under H_0, we can show the results for W_I^* in a similar way by assuming a multinomial distribution for the super-population model.

S9 A brief description for the bootstrap method

In the simulation and application parts, there are two kinds of sampling designs. One is single-stage probability-proportional-to-size sampling with replacement (PPSWR), and the other one is a two-stage cluster sampling design with PPSWR being the first-stage sampling design and simple random sampling being the second-stage sampling design. In this section, we briefly describe the bootstrap methods under those two kinds of sampling designs.

S9.1 Single-stage sampling design

Denote the finite population to be $U_N = \{y_1, \ldots, y_N\}$, and we assume that N is known. We present the bootstrap method under PPSWR with selection probabilities $\{p_1, \ldots, p_N\}$ satisfying $\sum_{i=1}^N p_i = 1$. Suppose that the original sample can be written as $\{y_{a,i} : i = 1, \ldots, n\}$, where $y_{a,i} = y_k$ if y_k is selected for the i-th draw. The bootstrap method under PPSWR is described as follows.

Step 1. Obtain $(N^*_{a,1}, \ldots, N^*_{a,n})$ from a multinomial distribution $\mathrm{MN}(N; \rho)$, where $\rho = (\rho_1, \ldots, \rho_n)$ and $\rho_i = p_{a,i}^{-1}(\sum_{j=1}^n p_{a,j}^{-1})^{-1}$ for $i = 1, \ldots, n$. Then, the bootstrap population $U^*_N = \{y^*_1, \ldots, y^*_N\}$ consists of $N^*_{a,i}$ replicates of $y_{a,i}$, and the bootstrap selection probabilities are $\{(C^*_N)^{-1}p^*_1, \ldots, (C^*_N)^{-1}p^*_N\}$, where $C^*_N = \sum_{i=1}^N p^*_i = \sum_{i=1}^n N^*_{a,i}p_{a,i}$, and $\{p^*_1, \ldots, p^*_N\}$ contains $N^*_{a,i}$ copies of $p_{a,i}$ for $i = 1, \ldots, n$.

Step 2. From the bootstrap population U^*_N, generate a bootstrap sample of size n by PPSWR using $\{(C^*_N)^{-1}p^*_1, \ldots, (C^*_N)^{-1}p^*_N\}$ as selection probabilities. Based on the bootstrap sample $\{y^*_{b,i} : i = 1, \ldots, n\}$, the corresponding bootstrap weights are computed by $w^*_i = (np^*_{b,i})^{-1}C^*_N$, where $y^*_{b,i} = y^*_k$ and $p^*_{b,i} = p^*_k$ if y^*_k is selected in the i-th draw. Then, we can obtain the bootstrap test statistics.

Step 3. Repeat the two steps above independently B times, where B is a large number.

Then, we can use the empirical distribution of the bootstrap test statistics to approximate that of the corresponding test statistic obtained from the original sample.

S9.2 Two-stage cluster sampling design

From a finite population $U_N = \{y_{i,j} : i = 1, \ldots, N_I; j = 1, \ldots, M_i\}$, we can use two-stage cluster sampling to generate a sample, where N_I is the number of clusters and M_i is the size of the i-th cluster; assume that N_I and M_i are known for $i = 1, \ldots, N_I$. Suppose that the number of clusters for the first-stage sampling is n_1, and the sample size within the i-th cluster is m_i with respect to the second-stage sampling. The bootstrap method for two-stage cluster sampling is an extension of the bootstrap mentioned above.

Step 1. Generate a bootstrap population $U_N^* = \{y_{i,j}^* : i = 1, \ldots, N_I; j = 1, \ldots, M_i^*\}$ by the following two steps, where M_i^* is the size of the i-th cluster in the bootstrap population.

Step 1-a. Based on the n_1 clusters of the original sample, generate a bootstrap population of clusters using the same step (Step 1) for PPSWR under single-stage sampling.

Step 1-b. Within the i-th bootstrap cluster, use the original sample within this cluster to generate $\{y_{i,1}^*, \ldots, y_{i,M_i^*}^*\}$ by the bootstrap method for simple random sampling.

Step 2. From the bootstrap population U_N^*, generate a bootstrap sample by the same two-stage cluster sampling. Then, we can obtain the bootstrap test statistics.

Step 3. Repeat the two steps above independently B times.

References

Athreya, K. B. and Lahiri, S. N. (2006). *Measure theory and probability theory*, Springer Science & Business Media.

Beaumont, J.-F. and Bocci, C. (2009). A practical bootstrap method for testing hypotheses from survey data, *Surv. Methodol.* **35**(1): 25–35.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys, *Int. Stat. Rev.* **51**(3): 279–292.

Chambers, R. and Skinner, C. J. E. (2003). *Analysis of Survey Data*, John Wiley & Sons, England.

Fuller, W. A. (2009). *Sampling Statistics*, John Wiley & Sons, New Jersey.

Horn, R. A. and Johnson, C. R. (2017). *Matrix Analysis*, Cambridge, New York.

Kim, J. K., Park, S. and Lee, Y. (2017). Statistical inference using generalized linear mixed models under informative cluster sampling, *Canadian Journal of Statistics* **45**(4): 479–497.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*, John Wiley & Sons, New York.

Lumley, T. and Scott, A. J. (2014). Tests for regression models fitted to survey data, *Australian & New Zealand Journal of Statistics* **56**: 1–14.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data, *Int. Stat. Rev.* **61**(2): 317–337.

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas, *J. Amer. Statist. Assoc.* **102**: 1427–1439.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables, *Journal of the American Statistical Association* **76**: 221–230.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *The Annals of statistics* **12**: 46–60.

Rao, J. N. K., Scott, A. J. and Skinner, C. J. (1998). Quasi-score tests with survey data, *Statistica Sinica* pp. 1059–1070.

Scott, A. J. (2007). Rao-scott corrections and their impact, *Proceedings of the 2007 joint statistical meetings*, Salt Lake City.

Shao, J. (2003). *Mathematical statistics*, Springer, New York.

Wang, Z., Kim, J. K. and Peng, L. (2019). Bootstrap inference for the finite population total under complex sampling designs, *arXiv:1901.01645* .