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

Jae Kwang Kim, J. N. K. Rao, and Zhonglei Wang

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 calculated and then compared with the 100(1 − α)% quantile of the reference distribution, which is the limiting distribution of the test statistic under the null hypothesis, where α is the nominal significance level. The limiting distribution is often a Chi-squared distribution (Shao 2003).

However, statistical inference with survey data involves additional steps incorporating the sampling design features. Korn and Graubard (1999) and Chambers and Skinner (2003) provided comprehensive overviews of the methods for analyzing survey data from complex sampling designs. In hypothesis testing with survey data, the limiting distribution of the test statistic is generally not a Chi-squared distribution. Rather, it can be expressed as a weighted sum of several independent random variables from \( \chi^2(1) \), which is a Chi-square distribution with one degree of freedom, and the weights depend on unknown parameters. 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 test statistics (Lumley and Scott 2014).

In this article, we use a different approach to computing the limiting distribution using a bootstrap approximation. Beaumont and Bocci (2009) investigated a general weighted bootstrap method for hypothesis testing under complex sampling designs in the context of Wald tests for linear regression analysis. In this article, we generalize the bootstrap testing idea of Beaumont and Bocci (2009) and present a unified bootstrap weight approach to obtain the limiting distribution of test statistics under complex sampling designs, including stratified multi-stage sampling with clusters selected with replacement.

The proposed method is a bootstrap weight method (Mashreghi, Haziza, and Léger 2016) under complex survey sampling. Although Praestgaard and Wellner (1993) and Chatterjee and Bose (2005) rigorously investigate bootstrap weight methods under regularity conditions, both may provide erroneous inferences under complex survey sampling. Praestgaard and Wellner (1993) established central limit theorems for the weighted bootstrap empirical process, but they assumed that the sample is independently and identically distributed (iid). Although it is common to assume that the finite population is iid, elements in a sample may not be due to complex sampling designs; see Pfeffermann (1993) for details. In addition, Praestgaard and Wellner (1993) only focused on mean estimation, but we are interested in hypothesis testing for parameters that solve a certain estimating equation. Chatterjee and Bose (2005) proposed a bootstrap weight method for estimating functions.
and their theoretical results are valid even for correlated random variables. However, the bootstrap weight method of Chatterjee and Bose (2005) is not applicable to informative sampling designs, since survey weights are not taken into consideration; see Pfeffermann (1993) for details.

There exist bootstrap weight methods under complex survey sampling, but the proposed bootstrap method differs from them in the following aspects. Rao, Wu, and Yue (1992) proposed a bootstrap weight method to estimate the variance of a function of population total estimators under stratified random sampling, but did not investigate the theoretical properties of their bootstrap method. Chipperfield and Preston (2007) proposed a without replacement scaled bootstrap to achieve the same goal as Rao, Wu, and Yue (1992) under stratified random sampling, but their method is only applicable when the parameter of interest is a smooth function of population totals. Moreover, neither method is applicable to other complex survey sampling. Bertail and Combris (1997) developed a general bootstrap weight method for parameter solving an estimating equation in general sampling designs, they focused only on variance estimation without a rigorous discussion about hypothesis testing. Antal and Tillé (2014) considered parameters solving an estimating equation without a rigorous discussion about hypothesis testing. Antal and Tillé (2011) proposed a one-one bootstrap method for variance estimation under general sampling designs. Similar to the one-one bootstrap method, Antal and Tillé (2014) developed a doubled half-bootstrap to estimate the variance of the estimator for population means. However, neither Antal and Tillé (2011) nor Antal and Tillé (2014) considered parameters solving an estimating equation. Moreover, existing work did not establish limiting distributions for the corresponding bootstrap estimator, especially for the one solving an estimating equation. Thus, up to our knowledge, there is no theoretical guarantee for existing methods when used for hypothesis testing.

Since the use of bootstrap for hypothesis testing is not fully investigated in the literature, our goal is to fill this important gap under general sampling designs. To do this, we first establish a bootstrap central limit theorem under regularity conditions. The sampling design is allowed to be informative in the sense that the sampling mechanism depends on the response variables being sampled (Pfeffermann 1993). Once the bootstrap central limit theorem is established, the proposed bootstrap method can be applied to the quasi-likelihood-ratio test and the quasi-score test. As long as the bootstrap central limit theorem holds for the sampling design, the bootstrap replicates of the test statistics can be used to approximate their distributions under the null hypothesis. Besides, this article also has a practical contribution to survey sampling, since bootstrap weights are commonly provided in agencies, including Statistics Canada.

The article is organized as follows. In Section 2, the basic setup is introduced. The proposed bootstrap method is presented in Section 3. In Section 4, the quasi-likelihood ratio test and the quasi-score test are introduced. Test for categorical survey data is briefly discussed in Section 5. The results of two simulation studies are presented in Section 5. An application to data from an educational survey under a logistic regression model is presented in Section 6. Concluding remarks are made in Section 7.

2. Basic Setup

Suppose that a finite population \( \mathcal{F}_N = \{(z_i, x_i, y_i) : i = 1, \ldots, N\} \) is a random sample of size \( N \) from a super-population model (Isaki and Fuller 1982; Pfeffermann 1993; Sverchkov and Pfeffermann 2004; Rubin-Bleuer and Schiopu-Kratina 2005), where \( y_i \) is the response of interest, and \( x_i \) is the corresponding auxiliary vector of length \( d \), and \( z_i \) is the vector of design variables such that the first-order inclusion probability for unit \( i \) can be obtained by \( \pi_i = \pi(z_i; \phi_N) \), where \( \phi_N \) is a design parameter determined by \( \{z_1, \ldots, z_N\} \). In other literature, the inclusion probability is expressed as \( \pi_i = \pi(\mathcal{F}_N, i) \); see Sverchkov and Pfeffermann (2004) for details.

We are interested in hypothesis testing for a parameter \( \theta_0 = \arg \max_{\theta \in \Theta} E[l(\theta; X, Y)] \), where \( \Theta \subset \mathbb{R}^p \) is the parameter space, \((X, Y)\) is the random variable associated with elements of \( \mathcal{F}_N \), the expectation is taken with respect to the super-population model, and \( l(\theta; x, y) \) is a pre-defined objective function. For example, if the finite population is a random sample generated from a super-population model with a density function \( f(x, y; \theta_0) \) and if \( \theta_0 \) is the parameter of interest, then \( l(\theta; x, y) = \log f(x, y; \theta) \). For regression problems, \( \theta_0 \) can be the parameter in the conditional expectation of the response of interest given the auxiliary vectors.

From the finite population \( \mathcal{F}_N \), a probability sample \( \{(x_i, y_i) : i \in A\} \) of size \( n \) is selected by a probability sampling design with the first-order inclusion probability \( \pi_i \), where \( A \) is the set of sample indexes. The design variables \( z_i \) and the design parameter \( \phi_N \) are not available at the sample level. From the sample, we can compute a pseudo \( M \)-estimator of \( \theta_0 \) by

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

where

\[
l_w(\theta) = N^{-1} \sum_{i \in A} w_i l(\theta; x_i, y_i)
\]

is the survey-weighted objective function and \( w_i = \pi_i^{-1} \) is the sampling weight for \( i \in A \). Note that \( l_w(\theta) \) is design-unbiased for \( l_N(\theta) = N^{-1} \sum_{i=1}^N l(\theta; x_i, y_i) \) in the sense that \( E[l_w(\theta) \mid \mathcal{F}_N] = l_N(\theta) \), where the conditional expectation conditioning on \( \mathcal{F}_N \) is with respect to the sampling mechanism. To be rigorous, a subscript \( N \) should also be used to index the sample, the sampling weights, the estimator and other quantities, but it is suppressed in the sequel for simplicity.

Remark 1. When discussing a finite population in the preceding paragraph, we implicitly assume that elements of \( \mathcal{F}_N \) are iid. Such an assumption is widely adopted in survey sampling; see Corollary 1.3.2.1 and Section 2.2.1 of Fuller (2009) for details. Besides, under a linear regression model,

\[
Y = X^T \beta_0 + \epsilon
\]

with \( E(\epsilon \mid X) = 0 \), the finite population \( \{(x_i, y_i) : i = 1, \ldots, N\} \) satisfies the iid assumption; see Särndal, Swensson, and Wretman (1989, sec. 5.1), Breidt and Opsomer (2000), Särndal, Swensson, and Wretman (2003, sec. 6.4), Kim et al. (2006, sec. 5), Toth and Eltinge (2011), Wu and Thompson (2020, sec. 5.1) and Han and Wellner (2021) for more general
setups about the super-population model. If we are interested in the regression parameter $\beta_0$ in (1), an objective function is

$$I(\theta; x, y) = -(y - x^T \beta)^2.$$ 

Although we implicitly make an iid assumption for the finite population, elements in the sample $A$ may not be iid due to complex survey designs. That is, the informative sampling no longer preserves the iid structure in the sample; see Pfeffermann (1993) for details. Under certain sampling designs, we can relax the iid assumption, and the super-population model can involve cluster random effects (Rubin-Bleuer and Schiopu-Kratina 2005); refer to Section S4 of supplementary material for an extension to stratified multi-stage sampling with cluster random effects. We focus on element sampling mainly in this article.

Since the parameter of interest is the superpopulation parameter, the final sample can be viewed as a realization of the two-phase sampling design, where the finite population is the first-phase sample and the final sample is the second-phase sample. Chen and Rao (2007) and Rubin-Bleuer and Schiopu-Kratina (2005) present rigorous asymptotic setups for two-phase sampling.

Often, the pseudo $M$-estimator $\hat{\theta}$ can be obtained by solving

$$\hat{S}_w(\theta) = \theta \frac{\partial}{\partial \theta} I_w(\theta) = \frac{1}{N} \sum_{i \in A} w_i S(\theta; x_i, y_i) = 0,$$ 

(2)

where $S(\theta; x, y) = \partial I(\theta; x, y)/\partial \theta$ is the tangent function of $\theta$. To discuss the asymptotic properties of the pseudo $M$-estimator, we assume the following regularity conditions for a sequence of finite populations and samples (Isaki and Fuller 1982).

C1. The parameter space $\Theta$ is an open and convex set containing $\theta_0$ as an interior point. For $\theta \in \Theta$, $I(\theta; x, y)$ is concave and twice-continuously differentiable with respect to $\theta$, $E[I(\theta; X, Y)] < \infty$, and $E[I(\theta; X, Y)]$ is uniquely maximized at $\theta_0$.

C2. For $\theta \in \Theta$, $V[I_w(\theta) | F_N] \rightarrow 0$ in probability with respect to the super-population model as $n \rightarrow \infty$, where $V[I_w(\theta) | F_N]$ is the design variance of $I_w(\theta)$ conditional on the finite population $F_N$. In addition, there exists a nonnegative design variance estimator $\tilde{V}[I_w(\theta) | F_N]$, such that $E[\tilde{V}[I_w(\theta) | F_N] | F_N] = V[I_w(\theta) | F_N]$ almost surely.

C3. There exists a compact set $B \subset \Theta$ containing $\theta_0$ as an interior point, such that $\sup_{\theta \in B} \|nV[\hat{S}_w(\theta) | F_N] - \Sigma_S(\theta)\| \rightarrow 0$ in probability with respect to the super-population model as $n \rightarrow \infty$, where $V[\hat{S}_w(\theta) | F_N]$ is the design variance of $\hat{S}_w(\theta)$ conditional on the finite population, $\|M\| = \sup_{x \in \mathbb{R}^m, \|x\|=1} \|Mx\|$ is the operator norm of an $m \times m$ matrix $M$ associated with the Euclidean vector norm $\|\cdot\|$, and $\Sigma_S(\theta)$ is a positive-definite nonstochastic matrix.

C4. A central limit theorem holds for the survey-weighted score function $\hat{S}_w(\theta)$ in (2). Specifically,

$$n^{1/2} \hat{S}_w(\theta_0) \rightarrow N(0, \Sigma_S(\theta_0))$$

in distribution as $n \rightarrow \infty$ with respect to the super-population model and the sampling mechanism.

C5. The weak convergence of $\hat{I}_w(\theta) = N^{-1} \sum_{i \in A} w_i I(\theta; x_i, y_i)$ is uniform in $\theta \in B$, where $I(\theta; x, y) = -\partial S(\theta; x, y)/\partial \theta^T$. Specifically, $\sup_{\theta \in B} \|\hat{I}_w(\theta) - I(\theta)\| \rightarrow 0$ in probability with respect to the super-population model and the sampling mechanism as $n \rightarrow \infty$, where $I(\theta) = E[I(\theta; X, Y)]$, and $I(\theta_0)$ is invertible.

Conditions C1–C2 are sufficient conditions for the weak consistency of $\hat{\theta}$. Conditions C3–C5 are used to establish the asymptotic normality of $\hat{\theta}$. In Conditions C3–C4, we consider the case where $V[E[\hat{S}_w(\theta) | F_N]]$ is asymptotically negligible compared with $E[V[I_w(\theta) | F_N]]$, and it is often the case when $n/N \rightarrow 0$ as $n \rightarrow \infty$. That is, we can safely ignore the variability for generating the finite population and focus on the design variance of $\hat{S}_w(\theta)$ given the finite population. This assumption is a building block of the proposed bootstrap method in the next section, and a similar assumption is also proposed by Beaumont and Bocci (2009); see Asquith 5 and its discussion of Beaumont and Bocci (2009) for details.

As mentioned by Han and Wellner (2021), "establishing a finite-dimensional CLT ... can be a nontrivial problem for general sampling designs," and it is beyond our scope to prove Condition C4 under general sampling designs. In fact, there is no guarantee that the central limit theorem holds for general sampling designs. Instead, we simply consider the set of sampling designs such that the central limit theorem holds; see Hájek (1964), Rosén (1972a), Rosén (1972b), Berger (1998), Rubin-Bleuer and Schiopu-Kratina (2005), sections 1.3.3–1.3.4 of Fuller (2009), Chauvet (2015), and Han and Wellner (2021) for details. For simplicity, we use the same notation, “$\|\cdot\|$”, to denote the matrix norm and the vector norm. Detailed comments on these regularity conditions are presented in Section S1 of supplementary material. For the proposed conditions, we implicitly assume that $N \rightarrow \infty$ as $n \rightarrow \infty$. The following theorem establishes the asymptotic normality of $\hat{\theta}$. Its proof can be found in Section S2 of supplementary material.

**Theorem 1.** Under Conditions C1–C5, we have

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

(3)

in distribution as $n \rightarrow \infty$ with respect to the joint distribution of the super-population model and the sampling mechanism, where $\Sigma_\theta = \Sigma_\theta(\theta_0)^{-1} \Sigma_S(\theta_0) \Sigma_\theta(\theta_0)^{-1}$.

Now, by the second-order Taylor expansion, we obtain

$$I_w(\theta_0) = I_w(\hat{\theta}) + \nabla^2 I_w(\hat{\theta})(\theta_0 - \hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^T \nabla^2 I_w(\hat{\theta})(\theta_0 - \hat{\theta})$$

$$+ o_p(n^{-1})$$

(4)

where $\tilde{I}_w(\theta) = N^{-1} \sum_{i \in A} w_i I(\theta; x_i, y_i)$. Define

$$W(\theta_0) = -2n[I_w(\theta_0) - I_w(\hat{\theta})].$$

(5)

By (4), we obtain

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

(6)

Thus, by Theorem 1, we have

$$W(\theta_0) \rightarrow G = \sum_{i=1}^p c_i Z_i^2$$

(6)
in distribution as \( n \to \infty \), and the reference distribution is the joint distribution of the super-population model and the sampling mechanism, where \( c_1, \ldots, c_p \) are the eigenvalues of \( D = \Sigma \mathbf{Z}(\theta_0) \), and \( Z_1, \ldots, Z_p \) are \( p \) independent random variables from the standard normal distribution. Result (6) was established by Rao and Scott (1984) and Lumley and Scott (2014), and it can be regarded as a version of the Wilks’ theorem for survey sampling. For \( p = 1 \), we can use \( c_i^{-1} W(\theta_0) \) as the test statistic with \( \chi^2(1) \) distribution as the limiting distribution under the null hypothesis, where \( c_1 = \sigma_1^2(\theta_0)/I(\theta_0) \), and \( \sigma_1^2(\theta) \) and \( I(\theta) \) are the counterparts of \( \Sigma(\theta) \) and \( I(\theta) \), respectively. 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 with \( p \) degrees of freedom.

### 3. Bootstrap Calibration

We propose using a bootstrap weight method to approximate the limiting distribution in (6) and other test statistics under complex survey sampling. Such a bootstrap calibration is very attractive in practice since there is no need to derive the analytic form of the limiting distribution of the test statistic. That is, standard software that takes account of the weights can be used to implement bootstrap hypothesis testing from data files reporting weights and associated sets of bootstrap weights, as done in Statistics Canada and some other agencies. Given the sample \( A \), the proposed bootstrap method is as follows:

**Step 1** Generate a rescaling vector \( r^* \) of size \( n = |A| \) from a bootstrap distribution with \( E_s(r^*) = I_n \), and the covariance of \( r^* \) is determined by the sampling design, where \( |A| \) is the cardinality of set \( A \), \( E_s(\cdot) \) is the expectation with respect to the bootstrap distribution conditional on the sample \( A \), and \( I_n \) is a vector of \( 1 \) with length \( n \). The bootstrap weights are chosen in such a way that \( E_s(S_w^*(\theta)) = \hat{S}_w(\theta) \) and \( V_s(S_w^*(\theta)) = \hat{V}(\hat{S}_w(\theta)) | F_N \), where \( \hat{S}_w(\theta) = \sum_{i \in A} w_i^* \delta(\theta; x_i, y_i) \) and \( w_i^* = r_i^* w_i \) is the bootstrap variance conditional on the sample \( A \), and \( \hat{V}(\hat{S}_w(\theta)) | F_N \) is a nonnegative design variance estimator of \( \hat{S}_w(\theta) \) conditional on the finite population \( F_N \).

**Step 2** Obtain a bootstrap replicate of \( \hat{\theta} \), denoted as \( \hat{\theta}^* \), by solving \( \hat{S}_w(\theta) = 0 \).

Different distributions may be used to generate \( r^* \) under different sampling designs, but the generation of \( r^* \) only depends on the inclusion probabilities and not on the specific form of the estimating function for parameter estimation. Thus, once the bootstrap weights are generated, they can be used for various inference problems based on the sample \( A \). However, the commonly used linearization technique is based on the specific form of the estimating function, so different linearization is required for different hypothesis testing problems. Thus, compared with the linearization technique, the proposed bootstrap method is more appealing to the practitioners.

We now present additional regularity conditions to validate the proposed bootstrap method.

C. For \( \theta \in \Theta \), the bootstrap weights are nonnegative and satisfy \( E_s(\hat{S}_w(\theta)) = \hat{S}_w(\theta) \) and \( V_s(\hat{S}_w(\theta)) = \hat{V}(\hat{S}_w(\theta)) \) conditional on the sample \( A \), where \( \hat{V}(\hat{S}_w(\theta)) \) satisfies

\[
\max_{\theta \in \Theta} \sup_{\delta \in B} \| \hat{V}(\hat{S}_w(\theta)) - V(\hat{S}_w(\theta)) \| \to 0
\]

in probability with respect to the super-population model and the sampling mechanism as \( n \to \infty \).

C7. A central limit theorem holds for the bootstrap weighted score function \( \hat{S}_w(\theta) \) uniformly over \( B \). Specifically,

\[
n^{1/2}(\hat{S}_w(\theta) - \hat{S}_w(\theta)) | A \to N(0, \Sigma(\theta))
\]

in distribution uniformly over \( B \), and the reference distribution is the bootstrap sampling distribution conditional on the sample \( A \).

C8. Suppose \( \sup_{\theta \in B} \| \hat{V}(\hat{S}_w(\theta)) - V(\hat{S}_w(\theta)) \| \to 0 \) in probability with respect to the bootstrap sampling distribution conditional on the sample \( A \) as \( n \to \infty \), where \( \hat{V}(\hat{S}_w(\theta)) = \sum_{i \in A} w_i^* \delta(\theta; x_i, y_i) \).

Condition C6 guarantees that the scaled bootstrap variance of \( \hat{S}_w(\theta) \) converges to \( \Sigma(\theta) \) in Condition C3, and they are also used to show the bootstrap central limit theorem in Theorem C7. Condition C8 is needed to establish the asymptotic normality of the bootstrap estimator \( \hat{\theta}^* \). To check Condition C7, we may use a bootstrap version of the Lyapunov condition. As discussed in Condition C4, we do not claim that Condition C7 holds under general sampling designs, but we can validate both under certain sampling designs. For example, in Section S4 of the supplementary material, we use a Lyapunov-type condition to prove the two central limit theorems under stratified multi-stage sampling. Detailed comments on these regularity conditions are presented in Section S1 of the supplementary material. The high-level conditions are also verifiable under other commonly used sampling designs, including Poisson sampling.

**Theorem 2.** Let the assumptions for Theorem 1 hold. Then, under Conditions C6–C8, we have

\[
\sqrt{n} (\hat{\theta}^* - \theta) | A \to N(0, \Sigma_{\theta})
\]

in distribution as \( n \to \infty \), and the reference distribution on the left side of (8) is the bootstrap sampling distribution conditional on the sample \( A \), where \( \Sigma_{\theta} \) is defined in Theorem 1.

Theorem 2 is proved under general sampling designs; see Section S3 of the supplementary material for details. Theorem 2 establishes the bootstrap central limit theorem for \( \hat{\theta}^* \), and the limiting distribution coincides with the one in (3) for \( \theta \). Thus, the proposed bootstrap method can be used to approximate the sampling distribution of \( \theta \).

For estimating the mean of a finite population, Chao and Lo (1985) first proved a bootstrap central limit theorem under simple random sampling without replacement, and Bickel and Freedman (1984) proved the bootstrap central limit theorem under stratified random sampling without replacement. We now present stratified multi-stage sampling with clusters selected with replacement as an example to validate the bootstrap central limit theorem in Theorem 2. Stratified multi-stage sampling with clusters selected with replacement is extensively used in
socio-economic surveys when the first-stage sampling fractions within strata are negligible, as is often the case with large-scale sample surveys.

Example 1. Under stratified multi-stage sampling, a popular approach is to assume that $n_h (\geq 2)$ clusters are selected with replacement from the $h$th stratum, with the selection probabilities $P_{hi}$ satisfying $\sum_{i=1}^{N_h} P_{hi} = 1$ for $h = 1, \ldots, H$ (Rao and Wu 1988), where $N_h$ is the number of clusters in the $h$th stratum. As mentioned in Section 2, the selection probabilities can be random with respect to the super-population model. Within each sampled cluster, denote $\tilde{S}_{w_i, h}(\theta)$ as a design-unbiased estimator of the cluster mean under sampling at the second and subsequent stages. For simplicity, we still use $\tilde{S}_{w, h}(\theta)$ for the case when the (hi)th cluster is fully observed. Then, the weighted score function is $\tilde{S}_w(\theta) = \sum_{h=1}^{H} W_h \tilde{S}_{w, h}(\theta)$, where $W_h = M_h/M = \sum_{i=1}^{N_h} M_{hi}, M_{hi}$ is the size of the (hi)th cluster, $\tilde{S}_{w, h}(\theta) = n_h^{-1} \sum_{i=1}^{n_h} \tilde{S}_{w_i, h}(\theta), \tilde{S}_{w_i, h}(\theta) = d_{a(h)} p_{a(h)}^{-1} \tilde{S}_{w,a(h)}(\theta), d_{a(h)} = M_{a(h)}/M_h$ and $a(h)$ is the index of the $h$th selected cluster in the $h$th stratum.

The bootstrap replicate of $\tilde{S}_{w, h}(\theta)$ can be obtained in the following way. For the $h$th stratum, generate $n_{h}^{*} = (n_{h1}^{*}, \ldots, n_{hn_{h}}^{*})$ by a multinomial distribution with $n_h - 1$ trials and a success probability vector $n_h^{-1}(1, \ldots, 1)^T$ of length $n_h$. Then, the bootstrap-weighted score function is $\tilde{S}_{w}^{*}(\theta) = \sum_{h=1}^{H} W_h \tilde{S}_{w, h}^{*}(\theta)$, where $\tilde{S}_{w, h}^{*}(\theta) = n_h^{-1} \sum_{i=1}^{n_{h}^{*}} r_{h}^{*} \tilde{S}_{w_i, h}(\theta), r_{h}^{*} = k_h n_{h}^{*1}$ and $k_h = n_h / (n_h - 1)$; Rao, Wu, and Yue (1992) proposed a similar rescaling bootstrap method under stratified multi-stage sampling, but they did not investigate the corresponding theoretical properties with respect to hypothesis testing of super-population model parameters. By Lemma 7 in Section S4 of supplementary material, we conclude that $E_{w} [\tilde{S}_{w}^{*}(\theta)] = \tilde{S}_{w}(\theta)$ and $\text{Var} [\tilde{S}_{w}^{*}(\theta)] = \tilde{V}(\tilde{S}_{w}(\theta)) = \sum_{h=1}^{H} W_h^{2} (n_h - 1)^{-1} \tilde{S}_{w, vec}(\theta)$, where $\tilde{S}_{w, vec}(\theta) = (\tilde{I}_{n_h} - \tilde{P}_{n_h, \theta}) \tilde{S}_{w, vec}(\theta)$, and $\tilde{I}_{n_h}$ is the projection matrix to the linear space spanned by $\tilde{I}_{n_h}$, $\tilde{I}_{n_h} = (1, \ldots, 1)^T$ is a vector consisting of $n_h 1$s, $\tilde{I}_{n_h}$ is an $n_h \times n_h$ identity matrix, and

$$\tilde{S}_{w, vec}(\theta) = (d_{a(h)} p_{a(h)}^{-1} \tilde{S}_{w,a(h)}(\theta), \ldots, d_{a(n_h)} p_{a(n_h)}^{-1} \tilde{S}_{w,a(n_h)}(\theta)).$$

In Section S4 of supplementary material, we show that the bootstrap central limit theorem in Theorem 2 holds with $n = \sum_{h=1}^{H} n_h$.

Under stratified simple random sampling without replacement (Bickel and Freedman 1984), a bootstrap method is proposed in Section S7 of the supplementary material, and it is essentially the same as the one considered by Rao and Wu (1988) when estimating the population mean; see Rao and Wu (1988) and Beaumont and Patak (2012) for details.

To establish a bootstrap version of Wilks' Theorem in (6), we use

$$W^*(\hat{\theta}) = -2n [I_{n}^{\theta} - I_{n}^{*}(\hat{\theta})]$$

as the bootstrap version of $W(\theta)$ in (5). Under the assumptions for Theorem 2, we can show that

$$W^*(\hat{\theta}) | A \rightarrow \mathcal{G} \quad (9)$$

in distribution as $n \rightarrow \infty$, and the reference distribution is the bootstrap sampling distribution conditional on the sample $A$, where $\mathcal{G}$ is the limiting distribution (6) of the quasi-likelihood ratio test statistic $W(\theta_0)$ in (5) under the null hypothesis.

By (9), we conclude that the corresponding test statistic generated by the proposed bootstrap method has the same asymptotic distribution as in (5). Thus, we can use the bootstrap distribution of $W^*(\hat{\theta})$ to approximate the sampling distribution of $W(\theta_0)$.

Remark 2. As mentioned in Section 1, most existing bootstrap methods were proposed for variance estimation under survey sampling. There indeed exist some papers discussing bootstrap confidence intervals under survey sampling, but these methods do not apply to general hypothesis testing problems. For example, Rao, Wu, and Yue (1992) proposed a bootstrap-$t$ confidence interval for the parameter of interest, but their bootstrap confidence interval does not apply when the parameter of interest is multi-dimensional. In addition, their bootstrap method is only valid under the stratified simple random sampling. Beaumont and Patak (2012) and Bertail and Combris (1997) discussed bootstrap confidence intervals in their simulation studies, but it is not clear how the corresponding intervals are constructed. Also, neither Bertail and Combris (1997) nor Beaumont and Patak (2012) investigated their bootstrap confidence intervals theoretically. Theorem 2, on the other hand, is a building block to show that the proposed bootstrap method can be used for hypothesis testing under general sampling designs. Although we adopt almost the same technique to prove Theorems 1–2, we are the first to establish that the bootstrap estimator $\hat{\theta}$ has the same asymptotic distribution as the original $\theta$. Then, we can show that the corresponding bootstrap statistic $W^*(\hat{\theta})$ shares the same asymptotic distribution as the original one $W(\theta_0)$ under the null hypothesis, so the proposed bootstrap method can be used for hypothesis testing without deriving the corresponding variance estimator as in Lumley and Scott (2014). In addition, the proposed bootstrap method can be widely applied in commonly used sampling designs; see Sections S4–S5 of the supplementary material for details.

Remark 3. In many practical situations, the design weights $w_i = \pi_i^{-1}$ are further modified to incorporate population-level constraints (as in calibration weighting) or to handle unit non-response (as in nonresponse weighting adjustment). In this case, the bootstrap design weights can be modified in the same way to obtain the bootstrap final weights. See Bessonneau et al. (2021) for a detailed illustration of constructing bootstrap final weights in household surveys. When the test statistic is computed using the final weights rather than the design weights, the bootstrap final weights can be applied to the same test statistics to obtain the bootstrap distribution of the test statistic.

4. Quasi-Likelihood-Ratio Test

Without loss of generality, we denote $\theta$ as the true parameter $\theta_0$ in this section. Let $\Theta_0 \subset \Theta$ be the parameter space under the null hypothesis, so the null hypothesis can be written as $H_0 : \theta \in \Theta_0$. In this section, we consider $H_0 : \theta_2 = \theta_{20}$ for some
known vector $\theta_2^{(0)}$, where $\theta = (\theta_1^T, \theta_2^T)^T$. Thus, we have $\Theta_0 = \{\theta \in \Theta; \theta_2 = \theta_2^{(0)}\}$. Let $\widehat{\Theta}^{(0)}$ be the profile M-estimator of $\theta_1$ under $H_0 : \theta_2 = \theta_2^{(0)}$, which is obtained by maximizing $l_w(\theta_1, \theta_2^{(0)})$ with respect to $\theta_1$.

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

$$W(\theta_2^{(0)}) = -2n[l_w(\widehat{\theta}^{(0)}) - l_w(\widehat{\theta})],$$

where $\widehat{\theta}^{(0)} = (\widehat{\theta}_1^{(0)^T}, \theta_2^{(0)^T})^T$. Under simple random sampling with replacement, $W(\theta_2^{(0)})$ in (10) is asymptotically distributed as $\chi^2(q)$ with $q = p - p_0$ and $p_0 = \text{dim}(\Theta_0)$, and the reference distribution is the joint distribution of the super-population model and the sampling mechanism, where $\text{dim}(\Theta)$ is the dimension of $\Theta$. The following lemma, proved by Lumley and Scott (2014), presents the limiting distribution of the quasi-likelihood ratio test statistic in (10) for general sampling designs.

**Lemma 1.** Under $H_0 : \theta_2 = \theta_2^{(0)}$ and the assumptions for Theorem 2,

$$W(\theta_2^{(0)}) \rightarrow G_1 = \frac{q}{\sum_{i=1}^q c_iZ_i^2}$$

(11)

in distribution as $n \rightarrow \infty$, and the reference distribution is the joint distribution of the super-population model and the sampling mechanism, where $c_1 \geq c_2 \geq \cdots \geq c_q > 0$ are the eigenvalues of $P = \sum_{\theta_2}I_{\mathcal{Z}_{22}}(\theta_1, \theta_2^{(0)})$, $Z_1, \ldots, Z_q$ are $q$ independent random variables from the standard normal distribution, $\Sigma_{\theta_2}$ is the asymptotic variance of $n^{1/2}\widehat{\theta}_2$ in (3), $I_{\mathcal{Z}_{22}}(\theta_1, \theta_2) = I_{\mathcal{Z}_{22}}(\theta_1, \theta_2) - I_{\mathcal{Z}_{21}}(\theta_1, \theta_2)I_{\mathcal{I}_{11}}(\theta_1, \theta_2) - I_{\mathcal{I}_{12}}(\theta_1, \theta_2)$, and $I_{\mathcal{I}_{ij}}(\theta_1, \theta_2) = E[-\partial^2 f((\theta_1, \theta_2); X, Y)/(\partial \theta_i \partial \theta_j)]$ for $i, j = 1, 2$.

Lumley and Scott (2014) proposed an alternative to the estimating equation approach discussed in Section 3. To do this, a bootstrap version of the quasi-likelihood ratio test statistic in (10) is obtained as

$$W^*(\widehat{\theta}_2) = -2n[l_w^*(\widehat{\theta}_1^{(0)}, \widehat{\theta}_2) - l_w^*(\widehat{\theta}^*)],$$

(12)

where $\widehat{\theta}_1^{(0)} = \arg \max_{\theta_1} l_w^*(\theta_1, \widehat{\theta}_2)$, and $\widehat{\theta}_2$ is the second component of $\theta = (\widehat{\theta}_1^T, \widehat{\theta}_2^T)^T$. Under the conditions of Theorem 2, we can show that the limiting distribution for $W^*(\theta_2)$ is the same as that in (11). Thus, the bootstrap distribution using $W^*(\theta_2)$ in (12) can be used to obtain the reference distribution of the test statistic $W(\theta_2^{(0)})$ in (10) under $H_0 : \theta_2 = \theta_2^{(0)}$.

**Remark 4.** In addition to the quasi-likelihood ratio test, the proposed bootstrap method is also applicable to quasi-score tests (Rao, Scott, and Skinner 1998). To develop a quasi-score test for $H_0 : \theta_2 = \theta_2^{(0)}$, we first decompose the survey-weighted score function as

$$\widehat{S}_w(\theta) = \left(\frac{\widehat{S}_{w1}(\theta)}{\widehat{S}_{w2}(\theta)}\right) = \left(\frac{N^{-1}\sum_{i \in A} w_i u_{i1}(\theta; x_i, y_i)}{N^{-1}\sum_{i \in A} w_i u_{i2}(\theta; x_i, y_i)}\right)$$

(13)

with $u_i(\theta; x_i, y_i) = (y_i - \mu_i)(V_0(\mu_i))^{-1}(\partial \mu_i/\partial \theta)$ for $j = 1, 2$, where $\mu_i = \mu(x_i; \theta)$, $\mu(x; \theta) = E(Y | x)$ is a known function with an unknown parameter $\theta$, and $V_0(\mu_i) = V(Y | x)$ is a known function of $\mu_i$. Note that quasi-score tests clearly specify the model without requiring a correctly specified density function, so they are more important in practice. Let $\widehat{\theta}_1^{(0)}$ be the solution to $\widehat{S}_{w1}(\theta_1, \theta_2^{(0)}) = 0$, where $\widehat{S}_{w1}(\theta)$ is defined in (13).

The quasi-score test statistic for $H_0 : \theta_2 = \theta_2^{(0)}$ is

$$X^2_{QS}(\theta_2^{(0)}) = \widehat{S}_{w1}^T(\theta_2^{(0)}) \left[\widehat{I}_{w22}^{-1}(\theta_2^{(0)})\right]^{-1}\widehat{S}_{w2}(\theta_2^{(0)}),$$

(14)

where

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

(15)

and

$$\left(\widehat{I}_{w11}(\theta) - \widehat{I}_{w12}(\theta)\right)/\widehat{I}_{w21}(\theta) = -\left(\frac{\widehat{S}_{w1}(\theta)/\partial \theta_1}{\widehat{S}_{w2}(\theta)/\partial \theta_1} \frac{\widehat{S}_{w1}(\theta)/\partial \theta_2}{\widehat{S}_{w2}(\theta)/\partial \theta_2}\right).$$

The bootstrap replicates of $X^2_{QS}(\theta_2^{(0)})$ can be constructed by replacing the sampling weights $w_i$ and $\widehat{\theta}^{(0)}$, respectively, by the bootstrap weights $w_i^*$ and the bootstrap estimator that solves $\widehat{S}_w(\theta) = N^{-1}\sum_{i \in A} w_i^*(y_i - p_i(\theta))(1, x_i^T)^T = 0$ subject to $\theta_2 = \theta_2^{(0)}$. Then, the bootstrap version of the quasi-score test statistic $X^2_{QS}(\theta_2^{(0)})$ is

$$X^2_{QS}(\theta_2^{(0)}) = \widehat{S}_{w1}^*(\theta_2^{(0)}) \left[\widehat{I}_{w22}^{*}(\theta_2^{(0)})\right]^{-1}\widehat{S}_{w2}^{*}(\theta_2^{(0)}),$$

(16)

where $\theta_2^{(0)} = (\theta_1^{(0)^T}, \theta_2^{(0)^T})^T$, and $\widehat{I}_{w22}(\theta)^* = \sum_{i \in A} w_i^* I_{\mathcal{I}_{ij}}(\theta_1, \theta_2^{(0)})$ is computed similarly to (15) using bootstrap weights.

5. Test for Categorical Survey Data

5.1. Goodness-of-Fit Test

Suppose that a finite population $U$ is partitioned into $K$ categories with $U = U(1) \cup \cdots \cup U(K)$. Define $p_k = N_k/N$ to be the population proportion of the $k$th category, where $N_k$ is the size of $U(k)$. Thus, we implicitly assume that the finite population is an iid realization of the super-population model following a multinomial distribution with $K$ categories.

Suppose that we are interested in the simple goodness-of-fit test $H_0 : p_k = p_k^{(0)}$ for $k = 1, \ldots, K$, where $(p_1^{(0)}, \ldots, p_K^{(0)})^T$ is a pre-specified vector satisfying $\sum_{k=1}^K p_k^{(0)} = 1$. From the sample A, we compute $\widehat{p}_k = \widehat{N}_k/N$ as an estimator of $p_k$, where $\widehat{N}_k = \sum_{i \in A} w_i$ is a design-unbiased estimator of $N_k$, $A_k = A \cap U(k)$, and $\widehat{N} = \sum_{k=1}^K \widehat{N}_k = \sum_{i \in A} w_i$. The estimated proportions can be obtained by $\widehat{p} = \arg \max_{p \in \Theta} \sum_{i \in A} w_i l(p; x_i)$, where $l(p; x_i) = \sum_{k=1}^K x_{ik} \log(p_k)$, with $x_{ik} = 1$ if the $k$th element belongs to the $i$th category and 0 otherwise, $p = (p_1, \ldots, p_K)^T$.

Then, the Pearson Chi-squared goodness-of-fit test statistic for $H_0$ is

$$X^2(p^{(0)}) = n \sum_{k=1}^K \left(\left(p_k - \widehat{p}_k\right)^2 / p_k^{(0)}\right),$$
where $\mathbf{p}^{(0)} = (p_1^{(0)}, \ldots, p_{K-1}^{(0)})^T$. Under regularity conditions, according to Rao and Scott (1981), we have

$$X^2(\mathbf{p}^{(0)}) \rightarrow \mathcal{G}_2 = \sum_{k=1}^{K-1} \lambda_k Z_k^2 \quad (17)$$

in distribution as $n \rightarrow \infty$ under $H_0$, and the reference distribution is the joint distribution of the super-population model and the sampling mechanism, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{K-1} > 0$ are the eigenvalues of the design effect matrix $\mathbf{D} = \mathbf{P}_0^{-1} \Sigma_p$, $\mathbf{P}_0 = \text{diag}(\mathbf{p}^{(0)}) - \mathbf{p}^{(0)}(\mathbf{p}^{(0)})^T$, $\Sigma_p$ is the asymptotic variance of $\sqrt{np}$, and $Z_1, \ldots, Z_{K-1}$ are $K-1$ independent random variables from the standard normal distribution. Under simple random sampling with replacement, the limiting distribution in (17) is a Chi-squared distribution with $K-1$ degrees of freedom since $\mathbf{P}_0 = \Sigma_p$.

We now apply the proposed bootstrap method to approximate the limiting distribution in (17), without computing the estimated covariance matrix $\hat{\Sigma}_p$. To describe the proposed method, let $\hat{\mathbf{p}}$ be the estimator of the population proportion $\mathbf{p}$ based on the bootstrap weights $w_i^*$. The proposed bootstrap statistic of $X^2(\mathbf{p}^{(0)})$ is

$$X^{2*}(\hat{\mathbf{p}}) = n \sum_{i=1}^{R} (\hat{p}_i^* - \hat{p}_i)^2 / \hat{p}_i.$$

We use $\hat{p}_i$ in place of $p_i^{(0)}$ in the bootstrap test statistics. Under the conditions of Theorem 2, we can show that the limiting distribution for $X^{2*}(\hat{\mathbf{p}})$ is the same as that in (17).

### 5.2. Test of Independence in a Two-Way Table

The proposed bootstrap can also be used to test independence in a two-way count table. 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 $\hat{p}_{ij} = \hat{N}_{ij}/\hat{N}$. By assuming a multinomial distribution for the super-population model, the Chi-squared statistic for testing independence $H_0 : p_{ij} = p_{i+}p_{j+}$ for all $i$ and $j$ is

$$X^2_I = n \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(\hat{y}_{ij} - \hat{p}_{i+}\hat{p}_{j+})^2}{\hat{p}_{i+}\hat{p}_{j+}}. \quad (18)$$

The limiting distribution of $X^2_I$ is a mixture of $\chi^2(1)$ distributions (Rao and Scott 1981).

To develop bootstrap replicates of the test statistics under $H_0$, 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 $X^2_I$ is

$$X^{2*}_I = n \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(\hat{y}_{ij} - \hat{p}_{i+}\hat{p}_{j+})^2}{\hat{p}_{i+}\hat{p}_{j+}}. \quad (19)$$

Under $H_0$, the terms in the numerator of $X^2_I$ are identical to $\{(\hat{y}_{ij} - \hat{p}_{i+}\hat{p}_{j+}) - (\hat{y}_{ij} - \hat{p}_{i+}\hat{p}_{j+})\}^2$. That is, the bootstrap test statistic is calculated by simply replacing $\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{j+}$, and $\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{j+}$ with $\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{j+}$ and $\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{j+}$, respectively. It can be shown that the limiting distribution of $X^{2*}_I$ conditional on the sample is the same as that of $X^2_I$.

### 6. Simulation Study

#### 6.1. Single-Stage Sampling

In this section, we test the performance of the proposed bootstrap method under the PPS (probability proportional to size) sampling with replacement. A finite population of size $N$ is generated as follows. For $i = 1, \ldots, N$, $x_i \sim U(-5,5)$ and $y_i = \theta_1 + \theta_2 x_i + \epsilon_i$, where $U(a,b)$ is a uniform distribution on the interval $(a,b)$, $(\theta_1, \theta_2) = (1,1)$, and $\epsilon_i \sim N(0,2^2)$. To make the sampling design informative, the PPS sampling with replacement is carried out to generate a sample of size $n$ with selection probabilities $p_i \propto (\ell_i + 1)$ and $\sum_{i=1}^{N} p_i = 1$, where $\ell_i = 1$ if $x_i \epsilon_i > 0$ and 0 otherwise; see Section S6 of the supplementary material for details about the probability proportional to size sampling and the bootstrap weights.

The sampling design is informative (Pfeffermann 1993), since the selection probabilities are partially determined by the residuals $\{\epsilon_i : i = 1, \ldots, N\}$. A similar model setup was also used by Verret, Rao, and Hidiroglou (2015) in the context of small area estimation. We consider two scenarios: $(N, n) = (2,000,200)$ and $(N,n) = (15,000,500)$.

We are interested in testing $H_0 : \theta_1 = \theta_1^{(0)}$ with the nominal significance level of $\alpha = 0.05$ 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 (10) and $\chi^2(1)$ as test statistic and reference distribution, respectively.
2. Naive quasi-score method with test statistic (14) and $\chi^2(1)$ as the reference distribution.
3. Lumley and Scott (2014) method. The test statistic is $W(\theta_2^{(0)}) / \hat{\sigma}$, where $\hat{\sigma} = \sqrt{\hat{\theta}_2 \hat{\ell}_{22,1} - \hat{\ell}_{22,1}}$ is defined in (15), and $\hat{V}(\theta_2)$ is obtained from $\hat{I}^{-1}_w(\theta) \hat{V}(\hat{\theta})(\hat{I}^{-1}_w(\theta))^T$. The reference distribution is $F_{1,k}$, where $F_{v_1,v_2}$ is an $F$ distribution with parameters $v_1$ and $v_2$, $k$ 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 quasi-likelihood-ratio method with $W(\theta_2^{(0)})$ in (10) being the test statistic, and the reference distribution is approximated by the empirical distribution of $W^*(\hat{\theta}_2)$ in (12).
5. Bootstrap quasi-score method with $X^{2*}(\theta_2^{(0)})$ in (14) being the test statistic, and the reference distribution is approximated by the empirical distribution of the bootstrap test statistics.

When calculating the likelihood-ratio test statistics, we assume a normal distribution with a constant but unknown variance for the conditional distribution $y_i$ given $x_i$. For each scenario, we generate 1000 Monte Carlo samples, and we consider $M = 500$ and $M = 1000$ iterations for both bootstrap methods.

Table 1 summarizes the simulation results. For both scenarios, the naive likelihood-ratio method and the naive quasi-score
method have a significantly inflated Type I error rate when the null hypothesis is true. The corresponding design effects (Wu and Thompson 2020) are more than 2 for estimating \( \theta_2 \) under both scenarios; see Section S8.1 of the supplementary material for details. Thus, the distributions of \( W(\hat{\theta}_2^{(0)}) \) and the quasi-score test statistic (14) are more “dispersed” than \( \chi^2(1) \) under the null hypothesis, and due to the lack of design-effect corrections, the two naive methods are not theoretically valid. The Type I error rates of the Lumley and Scott (2014) method and the two proposed bootstrap testing methods are similar under \( H_0 \), regardless of the population and sample sizes. The Type I error rates of the three methods come closer to the nominal truth as the population and sample sizes increase. Furthermore, the powers of the two bootstrap testing methods are reasonable compared to the Lumley-Scott method. The naive methods have a slightly larger power at the expense of elevated Type I error rates. In addition, we get similar test results regardless of the number of bootstrap repetitions.

Besides, Section S9 of the supplementary material reports results of an additional simulation study for a stratified two-stage sampling design.

### 6.2. Test of Independence

In this section, we consider test of independence in a \( 3 \times 3 \) table of counts to check the performance of the proposed bootstrap testing method. For a finite population \( U = \{ y_i : i = 1, \ldots, N \} \), \( y_i \) is generated by a multinomial distribution with one trial and a success probability vector \( p \), where \( p = (p_{11}, \ldots, p_{ij}, \ldots, p_{33})^T \), and \( p_{ij} \) is the success probability for the cell in the \( i \)th row and \( j \)th column for \( i,j = 1, \ldots, 3 \). That is, \( y_i \) is a dummy variable consisting of eight 0’s and a 1. For the success probability vector \( p \), we consider three cases:

- Case I: \( p_{11} = 1/4, p_{12} = p_{13} = p_{21} = p_{31} = 1/8, p_{22} = p_{23} = p_{32} = p_{33} = 1/16 \).
- Case II: \( p_{11} = 1/4, p_{12} = p_{13} = (1.4)/8, p_{21} = p_{31} = (0.6)/8, p_{22} = p_{23} = 1/16, p_{32} = (1.4)/16, p_{33} = (0.6)/16 \).
- Case III: \( p_{11} = p_{23} = p_{32} = 1/6, p_{12} = p_{13} = p_{21} = p_{31} = 1/12 \).

Case I satisfies independence for the two-way table of counts, but Cases II–III do not. The level of nonindependence can be expressed by \( \gamma = \sum_{i=1}^{3} \sum_{j=1}^{3} (p_{ij} - p_{i+}p_{+j})^2/(p_{i+}p_{+j}) \), and \( \gamma = 0 \) corresponds to independence. The values of \( \gamma \) are 0, 0.17, and 0.125 for Cases I–III, respectively.

For each \( y_i \), we generate an auxiliary variable \( x_i = \beta^T y_i \), where \( \beta = (\beta_1, \ldots, \beta_9)^T \), \( \beta_j = 0.5 + e_j \) for \( j = 1, \ldots, 9 \), \( e_j \sim \text{Exp}(1) \). Probability proportional to size sampling 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: \( (N, n) = (2000, 75) \) and \( (N, n) = (10000, 150) \).

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

1. Naive Pearson method based on \( X^2_{I} \) in (18) with \( \chi^2(4) \) being the reference distribution.
2. The second-order Rao-Scott correction method using the Satterthwaite approximation (Thomas and Rao 1987). This method is widely used in analyzing two-way contingency table and is implemented by the command \texttt{svychisq} in the \texttt{R} package \texttt{survey} (Lumley 2020).
3. Bootstrap Pearson method using \( X^2_{I} \), and its distribution is approximated by that of \( X^2_{I} \) in (19).

For each scenario, we generate 1000 Monte Carlo samples, and we consider \( M = 500 \) and \( M = 1000 \) iterations for both bootstrap methods. Table 2 summarizes the simulation results. For Case I when the null hypothesis is true, the Type I error rate of the naive Pearson method is much larger than the nominal significance level 0.05 for different sample sizes, since the asymptotic distribution of the test statistic \( X^2_{I} \) in (18) is more “dispersed” than a \( \chi^2(4) \) distribution under the null hypothesis; see Section S8.2 of the supplementary material for details. The performance of the Rao-Scott method with second-order correction is similar to the bootstrap testing method in terms of Type I error and power. The power of the bootstrap testing method increases with the value of \( \gamma \). Similar to the previous simulation, the power of the naive method is larger at the expense of elevated Type I error rates. In addition, we get similar test results regardless of the number of bootstrap repetitions.

### 7. Application

We present an analysis of the 2011 Private Education Expenditure Survey (PEES) in South Korea using the proposed bootstrap
testing methods. The purpose of this survey is to study the relationship between private education expenditure and academic performance of students before entering college.

A stratified two-stage sampling design was used for the 2011 PEES, and the strata consist of 16 first-tier administrative divisions, including most provinces and metropolitan cities in South Korea. For each stratum, PPS sampling with replacement was conducted in the first stage, and the primary sampling unit was the school. The students were randomly selected from the sample school in the second stage. There are about 1000 sample schools and 45,000 students involved in this survey.

For the student 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 vector of covariates of interest. As discussed by Kim, Park, and Lee (2017), we consider the following covariates: after-school education, hours taking lessons provided by the school after regular classes in a month; father’s education, 1 for college or higher and 0 otherwise; gender, 1 for female and 0 for male; household income per month; mother’s education, 1 for college or higher and 0 otherwise; private education, hours taking private lessons in a month.

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

\[
\logit(\Pr(Y = 1 \mid x)) = (1, x^T)\theta, \tag{20}
\]

where \( \logit(x) = \log(x) - \log(1 - x) \) for \( x \in (0, 1) \), \( Y = 1 \) if high academic performance is achieved and 0 otherwise, \( x \) is a vector of six covariates, and \( \theta = (\theta_0, \theta_1, \ldots, \theta_6)^T \). We are interested in testing the null hypotheses \( H_{0,i} : \theta_i = 0 \) for \( i = 1, \ldots, 6 \) with \( \alpha = 0.05 \) nominal significance level. Since the Wald method is widely used in practice, the naive likelihood ratio method, naive quasi-score method, bootstrap quasi-likelihood ratio method, bootstrap quasi-score method with 1000 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, sec. 1.2.8) for \( \theta \) using the sandwich formula. The \( p \)-values for the proposed bootstrap testing methods are obtained by bootstrap empirical distributions of the corresponding test statistics.

The results of our analysis are summarized in Table 3. The two bootstrap testing methods and the Wald test perform approximately the same in terms of \( p \)-values. The \( p \)-values of the two naive methods are approximately the same, but they differ from those of the bootstrap testing methods, especially for “after school education” at the middle school level and most covariates at the high school level, as the naive methods may not properly reflect the intra-cluster correlation due to cluster sampling.

Based on the two bootstrap testing methods in Table 3, we have the following conclusions under \( \alpha = 0.05 \) nominal significance level. Controlling other covariates, the probability that female students achieve high academic performance is significantly higher than that of male students in middle school, but the gender effect is not significant in high school. Hours spent on private education and after-school education can significantly increase the probability of achieving high academic performance in both middle school and high school. The household income has a significantly positive influence on their child’s academic performance. However, the level of education of the father and mother has a significant influence during the middle school period only.

### 8. Concluding Remarks

Many statistical agencies provide microdata files containing survey weights and several sets of associated replication weights, in particular bootstrap weights. Standard 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. Using bootstrap weights, we can easily apply hypothesis testing without using specialized software designed for complex survey data, as noted by Beaumont and Bocci (2009). The proposed bootstrap method is applied to quasi-likelihood ratio tests and quasi-score tests.

Our theories depend on establishing bootstrap central limit theorems under specified sampling designs, including stratified multi-stage sampling with clusters selected with replacement. Under some sampling designs, the central limiting theorem does not necessarily hold and the proposed method is not applicable. In the case of sampling designs that allow for the central limit theorem, the bootstrap central limit theorems can be established under some additional assumptions. Under the assumption that the bootstrap central limit theorem holds in the sampling design, the proposed hypothesis testing methods are applicable. The hypotheses we have discussed are special cases of the one considered in Section 4.2 of Jiang (2019), so the theoretical properties of the proposed bootstrap method can also be investigated using the general method developed in Section 4.2 of Jiang (2019). The proposed method can be applied to the case of categorical data by developing bootstrap procedures for testing simple goodness

### Table 3. Estimates (Est) and the \( p \)-values (Unit: \( 10^{-2} \)) for testing \( H_{0,i} : \theta_i = 0 \), where \( i = 1, \ldots, 6 \), in (20) 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 | 16.9 | 16.6 | 39.4 | 39.0 | 35.6 |
| Father’s Edu | 0.42 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Gender | 0.3 | 0.3  | 0.3  | 0.7  | 0.6  | 0.6  |
| Income | 0.14 | 0.0  | 0.0  | 0.2  | 0.2  | 0.4  |
| Mother’s Edu | 0.37 | 0.2  | 0.2  | 0.0  | 0.0  | 0.0  |
| Private Edu | 0.05 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| High School |     |      |        |        |      |
| After-school Edu | 0.03 | 2.0  | 2.0  | 0.3  | 0.3  | 0.1  |
| Father’s Edu | 0.19 | 10.4 | 10.3 | 14.1 | 14.0 | 13.8 |
| Gender | 0.21 | 2.5  | 2.5  | 17.0 | 16.8 | 15.7 |
| Income | 0.1 | 0.0  | 0.0  | 0.1  | 0.1  | 0.1  |
| Mother’s Edu | 0.15 | 20.4 | 20.4 | 41.2 | 41.1 | 40.3 |
| Private Edu | 0.04 | 0.0  | 0.0  | 3.5  | 3.0  | 3.2  |

NOTE: Cov, covariate; Est, estimate; NLR, naive likelihood-ratio method; NQS, naive quasi-score method; BLR, bootstrap quasi-likelihood-ratio method; BQS: bootstrap quasi-score method; Wald, two-sided Wald test.
of fit and independence in a two-way table. We plan to extend our bootstrap method for categorical data to test hypotheses from multi-way tables of weighted counts or proportions, using a log-linear model approach proposed by Rao and Scott (1984). Moreover, the bootstrap method can also be applied to nonresponse and weighting adjustments for variance estimation under regularity conditions, and extending the proposed bootstrap method for weight adjustments for hypothesis testing is also a promising project.

Supplementary Materials

Supplementary Material contains the following: comments on the regularity conditions (S1), proof of Theorem 1 (S2), proof of Theorem 2 (S3), verification of the proposed bootstrap method under stratified multi-stage sampling (S4), bootstrap method for Poisson sampling (S5) bootstrap method for probability proportional to size sampling with replacement (S6), stratified simple random sampling (S7), design effect (S8) and an additional simulation study: stratified two-stage sampling (S9).

Acknowledgments

We are grateful to the associate editor and three anonymous referees for their constructive comments, which have greatly improved the article.

Funding

Kim is partially supported by NSF (OAC-1931380). Rao is supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Wang is partially supported by NSFC (11901487, 71988101, 72033002, 12231011) and National Key R&D Program of China (2022YFA10038002).

References

Antal, E., and Tillé, Y. (2011), “A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population,” Journal of the American Statistical Association, 106, 534–543. [1230]
—— (2014), “A New Resampling Method for Sampling Designs Without Replacement: The Doubled Half Bootstrap,” Computational Statistics, 29, 1345–1363. [1230]
Beaumont, J.-F., and Bocci, C. (2009), “A Practical Bootstrap Method for Testing Hypotheses from Survey Data,” Survey Methodology, 35, 25–35. [1229,1231,1237]
Beaumont, J.-F., and Patak, Z. (2012), “On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling,” International Statistical Review, 80, 127–148. [1230,1233]
Berger, Y. G. (1998), “Rate of convergence to normal distribution for the Horvitz-Thompson estimator,” Journal of Statistical Planning and Inference, 67, 209–226. [1231]
Bertail, P., and Combris, P. (1997), “Bootstrap Généralisé d’un Sondage,” Annales d’Economie et de Statistique, 46, 49–83. [1230,1233]
Bessonnette, P., Brillhart, G., Chauvet, G., and Garcia, C. (2021), “With- Replacement Bootstrap Variance Estimation for Household Surveys: Principles, Examples and Implementation,” Survey Methodology, 47, 313–347. [1233]
Bickel, P. J., and Freedman, D. A. (1984), “Asymptotic Normality and the Bootstrap in Stratified Sampling,” Annals of Statistics, 12, 470–482. [1232,1233]
Breidt, F. J., and Opsomer, J. D. (2000), “Local Polynomial Regression Estimators in Survey Sampling,” Annals of Statistics, 28, 1026–1053. [1230]
Chambers, R., and Skinner, C. J. (2003), Analysis of Survey Data, Chichester: Wiley. [1229]
Chao, M.-T., and Lo, S.-H. (1985), “A Bootstrap Method for Finite Population,” Sankhya: Series A, 47, 399–405. [1232]
Chatterjee, S., and Bose, A. (2005), “Generalized Bootstrap for Estimating Equations,” Annals of Statistics, 33, 414–436. [1229,1230]
Chauvet, G. (2015), “Coupling Methods for Multistage Sampling,” Annals of Statistics, 43, 2484–2506. [1231]
Chen, J., and Rao, J. (2007), “Asymptotic Normality of Two-Phase Sampling Designs,” Statistica Sinica, 17, 1047–1064. [1231]
Chipperfield, J., and Preston, J. (2007), “Efficient Bootstrap for Business Surveys,” Survey Methodology, 33, 167–172. [1230]
Fuller, W. A. (2009), Sampling Statistics, New Jersey: Wiley. [1230,1231,1237]
Hájek, J. (1964), “Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population,” Annals of Mathematical Statistics, 35, 1491–1523. [1231]
Han, Q., and Wellner, J. A. (2021), “Complex Sampling Designs: Uniform Limit Theorems and Applications,” Annals of Statistics, 49, 459–485. [1230,1231]
Isaki, C. T., and Fuller, W. A. (1982), “Survey Design under the Regression Superpopulation Model,” Journal of the American Statistical Association, 77, 89–96. [1230,1231]
Jiang, J. (2019), Robust Mixed Model Analysis, Singapore: World Scientific. [1237]
Kim, J. K., Michael Brick, J., Fuller, W. A., and Kalton, G. (2006), “On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling,” Journal of the Royal Statistical Society, Series B, 68, 509–521. [1230]
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, 479–497. [1237]
Korn, E. L., and Graubard, B. I. (1999), Analysis of Health Surveys, New York: Wiley. [1229]
Lumley, T. (2020), Analysis of Complex Survey Samples, R package version 3.37. [1236]
Lumley, T., and Scott, A. J. (2014), “Tests for Regression Models Fitted to Survey Data,” Australasian & New Zealand Journal of Statistics, 56, 1–14. [1229,1232,1233,1234,1235,1236]
Mashreghi, Z., Haziza, D., and Léger, C. (2016), “A Survey of Bootstrap Methods for Finite Population Sampling,” Statistics Surveys, 10, 1–52. [1229]
Pfeffermann, D. (1993), “The Role of Sampling Weights when Modeling Survey Data,” International Statistical Review, 61, 317–337. [1229,1230,1231,1235]
Praestgaard, J., and Wellner, J. A. (1993), “Exchangeably Weighted Bootstraps of the General Empirical Process,” Annals of Probability, 21, 2053–2086. [1229]
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. [1235]
—— (1984), “On Chi-Squared Tests for Multitray Contingency Tables with Cell Proportions Estimated from Survey Data,” Annals of Statistics, 12, 46–60. [1232,1238]
Rao, J. N. K., Scott, A. J., and Skinner, C. J. (1998), “Quasi-Score Tests with Survey Data,” Statistica Sinica, 8, 1059–1070. [1234]
Rao, J. N. K., and Wu, C. F. J. (1988), “Resampling Inference with Complex Survey Data,” Journal of the American Statistical Association, 83, 231–241. [1233]
Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), “Some Recent Work on Resampling Methods for Complex Surveys,” Survey Methodology, 18, 209–217. [1230,1233]
Rosen, B. (1972a), “Asymptotic Theory for Successive Sampling with Varying Probabilities without Replacement, I,” Annals of Mathematical Statistics, 43, 373–397. [1231]
—— (1972b), “Asymptotic Theory for Successive Sampling with Varying Probabilities without Replacement, II,” The Annals of Mathematical Statistics, 43, 748–776. [1231]
Rubin-Bleuler, S., and Schiopu-Kratina, I. (2005), “On the Two-Phase Framework for Joint Model and Design-Based Inference,” Annals of Statistics, 33, 2789–2810. [1230,1231]
Särndal, C.-E., Swensson, B., and Wretman, J. (2003), *Model Assisted Survey Sampling*, New York: Springer. [1230]

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989), “The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total,” *Biometrika*, 76, 527–537. [1230]

Shao, J. (2003), *Mathematical Statistics*, New York: Springer. [1229]

Sverchkov, M., and Pfeffermann, D. (2004), “Prediction of Finite Population Totals based on the Sample Distribution,” *Survey Methodology*, 30, 79–92. [1230]

Thomas, D. R., and Rao, J. N. K. (1987), “Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics under Cluster Sampling,” *Journal of the American Statistical Association*, 82, 630–636. [1236]

Toth, D., and Eltinge, J. L. (2011), “Building Consistent Regression Trees From Complex Sample Data,” *Journal of the American Statistical Association*, 106, 1626–1636. [1230]

Verret, F., Rao, J. N. K., and Hidiroglou, M. A. (2015), “Model-based Small Area Estimation Under Informative Sampling,” *Survey Methodology*, 41, 333–347. [1235]

Wu, C., and Thompson, M. E. (2020), *Sampling Theory and Practice*, Gewerbestrasse: Springer. [1230,1236]