Neyman Smooth-Type Goodness of Fit Tests in Complex Surveys

Yan Lu [*], Lang Zhou[†], Guoyi Zhang[‡], Ronald Christensen [§]

Abstract

In this research, we extend Neyman smooth-type goodness of fit tests to complex surveys by incorporating consistent estimators under certain survey design, which is accomplished by data-driven nonparametric order selection methods. Asymptotic properties of the proposed estimators are investigated. Simulation results show that the proposed methods improve statistical power while control type I error very well, especially for the cases with slow-varying probabilities. The proposed methods are applied to solve some problems arising from the National Youth Tobacco Survey (NYTS) data.

[*] Yan Lu, Associate Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001, (*luyan@math.unm.edu*)

[†] Lang Zhou, AbbVie Inc., IL, (*z*houlang117@gmail.com)

[‡] Guoyi Zhang, Associate Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001, (*g*zhang123@gmail.com)

[§] Ronald Christensen, Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001, (*fletcher@stat.unm.edu*)

Key Words: Complex Surveys, Goodness of fit, Neyman smooth, Order selection, The first order and second order corrected tests.

1 Introduction

Analyses of categorical data arising from complex surveys is frequently encountered in quantitative sociological and economic research. A number of chi-squared tests were proposed to assess the fit of models for such data, which include but not limited to Wald (1943), Fay (1979, 1985), Rao and Scott (1981, 1984), Bedrick (1983), Rao and Scott (1987), Thomas, Singh, and Roberts (1996) and kwang Kim, Rao, and Wang (2019).

The Neyman smooth-type tests (Neyman, 1937) have been studied for independent and identically distributed (iid) data. Lancaster (1969) discussed the decomposition of the Pearson's chi-squared test statistic. Rayner, Best, and Dodds (1985) examined the similarities and differences between Pearson's chi-squared test and the Neyman smooth test. Rayner and Best (1986) extended Neyman smooth-type tests to location-scale families. A comprehensive overview of the Neyman smooth-type goodness of fit (GOF) tests can be found in Rayner and Best (1989, 1990) and Rayner, Thas, and Best (2009).

The chi-squared test statistics can be decomposed into ordered components with most information contained in the first few ones, so that component deduction is possible. Eubank (1997) introduced Neyman smooth-type GOF tests incorporated with order selection for iid data. A review of order selection can be found from Eubank (1999).

In this research, we extend Eubank's work from iid case to complex surveys. We incorporate design consistent estimators into test statistic, and use data-driven methods to select the optimal orders. This paper is organized as follows. In Section 2, we review a Neyman GOF test and first and second order corrected tests. In Section 3, we propose the Neyman smooth type GOF tests in complex surveys. In section 4, we investigate asymptotic properties of the proposed estimators. In Section 5, simulation studies are used to evaluate the proposed methods and to compare the proposed tests with several tests in literature. Section 6 gives an application example. Finally, we give conclusions and future research work in Section 7.

2 Background

Let $\mathbf{y} = (y_1, y_2 \cdots y_K)'$ follow a multinomial distribution with $\sum_{i=1}^{K} y_i = n$. Suppose that the underlying probability vector is $\mathbf{p} = (p(1), p(2) \cdots p(K-1))'$ with $p_K = 1 - \sum_{k=1}^{K-1} p_k$, and let $\mathbf{p}_0 = (p_0(1), p_0(2) \cdots p_0(K-1))'$ be a known vector. A general hypothesis of interest is:

$$H_0 : \mathbf{p} = \mathbf{p}_0 \text{ versus } H_\alpha : \mathbf{p} \neq \mathbf{p}_0. \tag{1}$$

The Pearson's test statistic $X^2 = \sum_{i=1}^{K} (y_i - e_i)^2/e_i = n \sum_{i=1}^{K} (\tilde{p}_i - p_{0i})^2/p_{0i}$ is used to measure the distance between the observed counts and expected counts e_i's under

H_0, where $\tilde{p}_i = x_i/n$ is the proportion of sample units in category i. With a complex survey data, one natural extension of X^2 is to replace $\tilde{p}_i$ with the design consistent estimator $\hat{p}_i$ (defined as the sum of weights of units in cell i divided by the sum of weights of all units in the sample), where weights are associated with a specified survey design. A weighted-up X^2 is defined as follows:

$$X^2 = n \sum_{i=1}^{K} \frac{(\hat{p}_i - p_{0i})^2}{p_{0i}}. \tag{2}$$

2.1 Neyman smooth-type tests in for iid case

In this section, we review Neyman smooth-type tests for iid case by Eubank (1997). Our interest is to test if $\mathbf{p} = \mathbf{p}_0$ as in Equation (1). Define basis function $\mathbf{x}_i = (x_i(1), x_i(2), \cdots, x_i(K))', i = 1, 2, \cdots, K-1$, which satisfies the following orthogonality conditions:

$$\mathbf{x}_j'\mathbf{x}_i = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i, \end{cases} \quad i, j = 1, \cdots, K-1, \tag{3}$$

and

$$\sum_{k=1}^{K} x_j(k)\sqrt{p_0(k)} = 0, \quad j = 1, \cdots, K-1. \tag{4}$$

Let $\tilde{f}(k) = (\tilde{p}(k) - p_0(k))/\sqrt{p_0(k)}$, for $k = 1, \cdots, K$, with associated (discrete) generalized Fourier coefficients $\tilde{b}_j = \sum_{k=1}^{K} \tilde{f}(k)x_j(k)$, for $j = 1, \cdots, K-1$. It can be shown that $\tilde{f}(k)$ is an unbiased estimator of $f(k) = (p(k) - p_0(k))/\sqrt{p_0(k)}$, for $k = 1, \cdots, K$, with associated Fourier coefficients $\beta_j = \sum_{k=1}^{K} f(k)x_j(k)$, for $j = 1, \cdots, K-$

1. By Parseval's relation (Arfken, 1985, pg. 425), Pearson's chi-squared test statistic can be re-organized as

$$X^2 = n\sum_{k=1}^{K} \frac{(\tilde{p}(k) - p_0(k))^2}{p_0(k)} = n\sum_{k=1}^{K} (\tilde{f}(k))^2 = \sum_{j=1}^{K-1} n\tilde{b}_j^2. \tag{5}$$

It can be shown that hypothesis (1) is equivalent to

$$H_0^* : \beta_1 = \cdots = \beta_{K-1} = 0, \tag{6}$$

and its corresponding alternative becomes

$$H_1^* : \beta_q \neq 0 \text{ and } \beta_{q+1} = \cdots = \beta_{K-1} = 0, \text{ for } q = 1, \cdots, K-1. \tag{7}$$

Under null hypothesis H_0^*, test statistic of order q: $X_q^2 = \sum_{j=1}^{q} n\tilde{b}_j^2$ for $q = 1, \cdots, K-1$ is a χ_q^2 distribution.

The remaining problem is to find an optimal estimate of the order q. Let $\tilde{q}$ be the maximizer of $\tilde{M}(q)$, where

$$\tilde{M}(q) = \frac{n+1}{n-1}\sum_{j=1}^{q} b_j^2 - \frac{2}{n-1}\sum_{j=1}^{q} \tilde{v}_{jj}, \text{ for } q = 1, \cdots, K-1, \tag{8}$$

and $\tilde{v}_{jj} = \sum_{k=1}^{K} x_j(k)^2\tilde{p}(k)/p_0(k)$ for $j = 1, \cdots, K-1$. Eubank (1997) suggested a test statistic of $W = (X_{\tilde{q}}^2 - \tilde{q})/\sqrt{2\tilde{q}}$ for $\tilde{q} \neq 0$. Under H_0^*, the distribution of W is obtained through simulations. The test statistic W is compared to $1-\alpha$ quantile of the distribution of W under H_0^* for testing purposes.

Another estimator suggested by Eubank (1997) is denoted by $\tilde{q}_\alpha$, which is the maximizer of

$$\tilde{M}_\alpha(q) = \frac{n+1}{n-1}\sum_{j=1}^{q} b_j^2 - \frac{a_\alpha}{n-1}\sum_{j=1}^{q} \tilde{v}_{jj}, \text{ for } q = 1, \cdots, K-1, \tag{9}$$

where $\tilde{M}_\alpha(0) = 0$, α is a specified significance level, and a_α is the solution of $1 - \alpha = \exp\{-\sum_{k=1}^{\infty} P(\chi^2_k > k a_\alpha)/k\}$. The null hypothesis will be rejected if $\tilde{q}_\alpha > 0$.

2.2 Corrections to Chi-Squared Test Statistic

Under complex designs, observations are correlated due to clustering. Therefore, independence assumption is violated. Under null hypothesis (1), the weighted-up X^2 in Equation (2) is distributed asymptotically as a linear sum of $\delta_1 W_1^2 + \cdots + \delta_{K-1} W_{k-1}^2$ instead of a $\chi^2(K-1)$ distribution, where W_i's are iid $\chi^2(1)$ distribution. The weights δ_i's are eigenvalues of the design effect matrix $\mathbf{P}^{-1}\mathbf{V}$ under H_0, where $\mathbf{P} = \mathrm{D}(\mathbf{p}) - \mathbf{p}\mathbf{p}'$, $D(\mathbf{p})$ is a $(K-1) \times (K-1)$ matrix with kth diagonal element $p(k)$ and off-diagonal entries 0, and $\mathbf{V}/n$ is the covariance matrix of $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{K-1})'$.

Under null hypothesis (1), let $\delta.$ be the expected value of $\delta_1 W_1^2 + \cdots + \delta_{K-1} W_{K-1}^2$ such that $\delta. = \sum_{i=1}^{K-1} \delta_i/(K-1)$. An approximate first-order corrected test can thus be obtained by comparing the observed value of $X^2/\hat{\delta}.$ to $\chi^2_{K-1}(\alpha)$.

When the full estimated covariance matrix $\hat{\mathbf{V}}$ is known, a better approximation to the asymptotic distribution of X^2 is to match the mean and variance of the test statistic to the mean and variance of a χ^2 distribution. The Rao-Scott (1981) second-order corrected test statistic is $X_S^2 = X^2/[\hat{\delta}.(1 + \hat{a}^2)]$, where

$$\hat{a}^2 = \sum_{i=1}^{K-1} \hat{\delta}_i^2/[(K-1)\hat{\delta}.^2] - 1, \tag{10}$$

This statistic is approximately a chi-squared random variable on $v = (K-1)/(1+\hat{a}^2)$ degrees of freedom. If the design effects of the categories are all similar, the first and second-order corrected tests will behave similarly. Otherwise, the second order corrected test may perform better.

3 Neyman Smooth-Type GOF Tests in Complex Surveys

In this Section, we extend Neyman smooth-type GOF test (Eubank, 1997) to complex surveys. First, we introduce the notation and set up the research problem. Next we propose two Neyman smooth-type GOF tests incorporated with order selection for use in complex surveys.

3.1 Notation and Research Problem

In section 2, we have defined $\mathbf{p}_0$, $\mathbf{p}$, and $\hat{\mathbf{p}}$ as the vector of hypothesized, underlying, and estimated proportions of a categorical data set. The hypothesis of interest stated in Equation (1) is $H_0 : \mathbf{p} = \mathbf{p}_0,$ versus $H_1 : \mathbf{p} \neq \mathbf{p}_0$.

For $j = 1, \cdots, n$ and $k = 1, \cdots, K$, define

$$y_j(k) = \begin{cases} 1 & \text{if outcome } j \text{ is from category } k, \\ \\ 0 & \text{otherwise,} \end{cases}$$

and let w_j be the sampling weight associated with $y_j(k)$ based on a specified survey design. The estimated proportion of the kth category can be expressed as

$$\hat{p}(k) = \frac{\sum_{j=1}^{n} w_j y_j(k)}{\sum_{j=1}^{n} w_j}, \; k = 1, \cdots, K.$$

Let $\hat{f}(k) = (\hat{p}(k) - p_0(k))/\sqrt{p_0(k)}$, with associated generalized Fourier coefficients $b_j = \sum_{k=1}^{K} \hat{f}(k)x_j(k),\ j = 1, \cdots, K-1$. By the fact that $\mathrm{E}[\hat{p}(k)] = p(k)$, we can show that $\hat{f}(k)$ is an unbiased estimator of $f(k)$ with associated Fourier coefficients $\beta_j = \sum_{k=1}^{K} f(k)x_j(k),\ j = 1, \cdots, K-1$.

Now define basis functions $x_j(k), j = 1, 2, \cdots, K-1$ as in Section 2 that satisfy the orthogonality conditions (3) and (4). Define $\mathbf{x_K} = (\sqrt{p_{01}}, \sqrt{p_{02}}, \cdots, \sqrt{p_{0K}})'$, and let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_K)$, and $\mathbf{X}_{[K]} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{K-1})$. Orthogonality condition (4) can also be written as $\mathbf{x}_j'\mathbf{x}_K = 0,\ j = 1, \cdots, K-1$. To construct the basis functions, we can use the Gram-Schmidt process to orthonormalize the polynomials of degree $K-1$ under the inner product $< w, v >= \sum_{k=1}^{K} w(k)v(k)\sqrt{p_0(k)}$. For example, if $K = 3$ and $\mathbf{p}_0 = (0.5, 0.3, 0.2)'$, we can chose the two basis functions as $x_1 = (-0.6031023, 0.1369881, 0.7858129)$ and $x_2 = (0.3691445, -0.8253692, 0.4271979)$, which satisfy the orthogonality conditions (3) and (4). As another example, for the simple uniform hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0 = 1/K$, the following basis function can be applied:

$$x_j(k) = \sqrt{\frac{2}{K}}\cos\left(\frac{j\pi(k-0.5)}{K}\right),\ \ k = 1, \cdots, K \text{ and } j = 1, \cdots, K-1. \tag{11}$$

Let $\hat{\mathbf{F}} = (\hat{f}(1), \hat{f}(2), \cdots, \hat{f}(K))', \mathbf{F} = (f(1), f(2), \cdots, f(K))', \mathbf{b} = (b(1), b(2), \cdots, b(K-1))'$ and $\boldsymbol{\beta} = (\beta(1), \beta(2), \cdots, \beta(K-1))'$. We can show that $\mathbf{b} = \mathbf{X}_{[K]}'\hat{\mathbf{F}}$, $\mathbf{F}'\mathbf{x}_K = 0$, and $\boldsymbol{\beta} = \mathbf{X}'_{[K]}\mathbf{F}$. The weighted up X^2 defined in Equation (2) can be written as

$$\begin{aligned} X^2 &= n\sum_{k=1}^{K}\frac{(\hat{p}(k)-p_0(k))^2}{p_0(k)} \\ &= n\hat{\mathbf{F}}'\hat{\mathbf{F}} = n\hat{\mathbf{F}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{F}} \\ &= n\mathbf{b}'\mathbf{b} = n\sum_{j=1}^{K} nb_j^2 = n\sum_{j=1}^{K-1} b_j^2, \end{aligned}$$

where the last equality comes from the fact that $b_K = \hat{\mathbf{F}}'\mathbf{x}_K = 0$. Notice that, $f(k)$ is 0 under the null hypothesis $\mathbf{p} = \mathbf{p}_0$. By Lehmann (1986, pg. 495), the hypothesis (1) is equivalent to $H_0^* : \beta_1 = \cdots = \beta_{K-1} = 0$, and its corresponding alternative becomes $H_1^* : \beta_q \neq 0 \text{ and } \beta_{q+1} = \cdots = \beta_{K-1} = 0, \text{ for } q = 1, \cdots, K-1$. That is to say that we want to find an optimal estimate of q, so that the first q components contain most of the information. Let $\mathbf{b}_q = (b(1), b(2), \cdots, b(q))'$, $\boldsymbol{\beta}_q = (\beta(1), \beta(2), \cdots, \beta(q))'$, and let $\mathbf{X}_q = (\mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_q)$. A Neyman smooth-type GOF test statistic is $\sum_{j=1}^{q} nb_j^2,\ q = 1, \cdots, K-1$. If the underlying order is q_0, an optimal estimator of q_0 can be the minimizer of

$$\begin{aligned} &\sum_{k=1}^{K}(f_q(k) - f(k))^2 \text{ where } f_q = \sum_{j=1}^{q} b_j x_j. \qquad (12) \\ = \;& (\mathbf{X}_q\mathbf{X}_q'\hat{\mathbf{F}} - \mathbf{F})'(\mathbf{X}_q\mathbf{X}_q'\hat{\mathbf{F}} - \mathbf{F}) \\ = \;& \mathbf{b}_q'\mathbf{b}_q - 2\mathbf{b}_q'\boldsymbol{\beta}_q + \boldsymbol{\beta}'\boldsymbol{\beta}. \end{aligned}$$

In this research, we propose statistical tests for hypothesis H_0^* versus alternative H_1^* based on criterion (12), and examine properties of the estimators.

3.2 Proposed Test W

In this section, we first investigate the estimator of q_0 in a simple random sample (SRS). Next, we propose a general form of the estimator for complex surveys. Since $\boldsymbol{\beta}'\boldsymbol{\beta}$ doesn't depend on q, minimizing Equation (12) is equivalent to maximizing $M(q) = -\mathbf{b}_q'\mathbf{b}_q + 2\mathbf{b}_q'\boldsymbol{\beta}_q$. Now we want to find an estimator of $M(q)$ such that $E(\hat{M}(q)) = E(M(q))$. In an SRS, Equation (24) (refer to Appendix) shows that $V(b_j) = \left(\sum_{k=1}^{K} x_j^2(k)p(k)/p_0(k) - \beta_j^2\right)/\tilde{n}$, where $\tilde{n} = n/(1-n/N)$. We propose maximizing the following $\hat{M}_{\tilde{n}}(q)$ for an SRS case,

$$\hat{M}_{\tilde{n}}(q) = \frac{\tilde{n}+1}{\tilde{n}-1}\sum_{j=1}^{q} b_j^2 - \frac{2}{\tilde{n}-1}\sum_{j=1}^{q}\hat{v}_{jj},\ q = 1, \cdots, K-1,$$

where $\hat{M}_{\tilde{n}}(0) = 0$, and $\hat{v}_{jj} = \sum_{k=1}^{K} x_j(k)^2\hat{p}(k)/p_0(k)$, for $j = 1, \cdots, K-1$. It can be shown that $E(\hat{M}_{\tilde{n}}(q)) = E(M(q))$.

Note that in an SRS, $\mathbf{V}/n \approx (1-n/N)\mathbf{P}/n$. Under H_0, $\mathbf{P}_0^{-1}\mathbf{V} = \mathbf{P}_0^{-1}(1-n/N)\mathbf{P}_0 = (1-n/N)$. Therefore, $\delta_.$, the average of eigenvalues of $(\mathbf{P}_0)^{-1}\mathbf{V}_0$ under an SRS is $1-n/N$, and $\tilde{n} = n/\delta_.$. In a complex survey, we can estimate $\delta.$ under a specified survey design, and approximate $\hat{V}(\hat{\mathbf{p}}) = \hat{\delta}.\hat{V}_{SRS}(\hat{\mathbf{p}})$. Define an effective sample size $\hat{n} = n/\hat{\delta}..$ The maximizing criterion in a general complex survey is:

$$\hat{M}(q) = \frac{\hat{n}+1}{\hat{n}-1}\sum_{j=1}^{q} b_j^2 - \frac{2}{\hat{n}-1}\sum_{j=1}^{q}\hat{v}_{jj},\ q = 1, \cdots, K-1. \tag{13}$$

Let $\hat{q}$ be the estimate of q_0 by maximizing criterion (13). $\hat{q}$ may be a natural test statistic and the null hypothesis can be rejected if $\hat{q} > 0$ is obtained through data. However, as shown in Spitzer (1956) and Zhang (1992), the limiting probability of the Type I error $\lim_{K\to\infty}\lim_{n\to\infty} P(\hat{q} > 0|q_0 = 0)$ is 0.29. Follow Eubank (1997), we propose a test statistic W as follows:

$$W = \begin{cases} \dfrac{X_{\hat{q}}^2 - \hat{q}}{\sqrt{2\hat{q}}} & \hat{q} > 0 \\ \\ 0 & \hat{q} = 0, \end{cases} \tag{14}$$

where $X_{\hat{q}}^2 = \sum_{j=1}^{\hat{q}} \hat{n} b_j^2$ for $q = 1, \cdots, K-1$ and $X_{\hat{q}}^2 = 0$ for $\hat{q} = 0$.

The distribution of W under null hypothesis (6), denoted by W_0 is obtained through simulations. For an arbitrary pre-specified level of significance α, the test can be performed by comparing the value of W with the $1-\alpha$ quantile of W_0.

3.3 Proposed Test $\hat{q}_\alpha$

In this section, we propose a test statistic $\hat{q}_\alpha$ that can be used directly to test the hypothesis H_0^*. The criterion in Equation (9) is modified by replacing sample size n with the effective size $\hat{n}$.

$$\hat{M}_\alpha(q) = \frac{\hat{n}+1}{\hat{n}-1}\sum_{j=1}^{q} b_j^2 - \frac{a_\alpha}{\hat{n}-1}\sum_{j=1}^{q} \hat{v}_{jj},\ q = 1, \cdots, K-1 \tag{15}$$

where $\hat{M}_\alpha(0) = 0$, $\hat{v}_{jj} = \sum_{k=1}^{K} x_j(k)^2 \hat{p}(k)/p_0(k)$ for $j = 1, \cdots, K-1$, and a_α will be discussed in next paragraph. The proposed test statistic $\hat{q}_\alpha$ is the maximizer of Equation (15).

Recall that the limiting probability of the Type I error for the proposed estimator $\hat{q}$ in section 3.2 is 0.29. To solve this problem, the proposed test $\hat{q}_\alpha$ replaced the constant 2 in equation (13) with a_α in equation (15) to control the Type I error at a specified level α. According to Eubank and Hart (1992) and Eubank (1997), a_α is the solution of the equation

$$1 - \alpha = \exp\left\{ -\sum_{k=1}^{\infty} \frac{P(\chi_k^2 > k a_\alpha)}{k} \right\} \tag{16}$$

or the solution of the following equation

$$P\left(\max_{1 \le k \le K-1} \left[\frac{1}{k} \sum_{j=1}^{k} Z_j^2 \right] \ge a_\alpha \right) = \alpha, \tag{17}$$

where χ_k^2 is the central chi-squared random variable with k degrees of freedom and Z_j's are independent standard normal random variables. Notice that a large K approximation is needed for equation (16). From simulation studies, a_α converges to the desired value quickly when $K > 10$. A level α test is conducted by rejecting H_0^* if $\hat{q}_\alpha > 0$ is obtained.

4 Properties of the Proposed Estimators

In Section 3, we proposed two Neyman smooth-type GOF tests incorporated with order selection for use in complex surveys. In this section, we first derive distribution of the Fourier coefficients b_j. Next, we examine asymptotic properties of the proposed estimators. Proofs are given in Appendix.

4.1 Limiting Distribution of The Fourier Coefficients b_j's

Theorem 1. *Assume that there is a sequence of superpopulations* $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \cdots \subset \mathcal{U}_t \subset \cdots$ *as defined in Isaki and Fuller (1982). Let* $\pi_{it} = p(\textit{psu } i \textit{ is in the sample from } \mathcal{U}_t)$

and $\pi_{ijt} = p(psu\ i \text{ and } psu\ j \text{ are both in the sample from population } \mathcal{U}_t)$ *be the inclusion and joint inclusion probabilities for the samples from population* $\mathcal{U}_t$*. Assume there are constants* c_1 *and* c_2 *such that* $0 < c_2 < \pi_{it} < c_1 < 1$ *for all* i *and any superpopulation in the sequence. Also assume there exists an* α_t *with* $\alpha_t = o(1)$ *such that* $\pi_{it}\pi_{jt} - \pi_{ijt} \leq \alpha_t \pi_{it}\pi_{jt}$*. The Fourier coefficient* $\mathbf{b}$ *is approximately* $(k-1)$*-variate normal with mean vector* 0 *and covariance matrix* $V/\hat{n}$ *for sufficiently large* $\hat{n}$*, where* $\hat{n}$ *is the effective sample size.*

4.2 Asymptotic Properties of $\hat{q}$

In this section, we state the asymptotic properties of $\hat{q}$, which is the maximizer of criterion (13) and is used to construct the proposed test W in Section 3.2.

Theorem 2. *Following Eubank (1999, pg. 51), let*

$$c_r = \sum_r^* \left\{ \prod_{k=1}^r \frac{1}{N_k!} \left(\frac{P(\chi_k^2 > 2k)}{k} \right)^{N_k} \right\},$$

and

$$d_r = \sum_r^* \left\{ \prod_{k=1}^r \frac{1}{N_k!} \left(\frac{P(\chi_k^2 < 2k)}{k} \right)^{N_k} \right\},$$

where $c_0 = d_0 = 1$*,* χ_k^2 *denotes a central chi-squared random variable with* k *degrees of freedom, and* $\sum_r^*$ *denotes the sum extending over all* r*-tuples of integers* $(N_1, \cdots, N_r)$*, such that* $N_1 + 2N_2 + \cdots + rN_r = r$*. Under the null hypothesis (6),*

$$\lim_{\hat{n}\to\infty} P(\hat{q} = q) = c_q d_{K-1-q}, \ \text{for } q = 0, \cdots, K-1.$$

In addition, under alternative hypothesis (7),

$$\lim_{\hat{n}\to\infty} P(\hat{q} < q_0) = 0, \ \textit{and} \ \lim_{\hat{n}\to\infty} P(\hat{q} = q_0 + r) = P(r^* = r), \ r = 0, \cdots, K - q_0 - 1,$$

where r^* is the maximizer of the criterion,

$$R(r) = \sum_{j=1}^{r} v_{(j+q_0)(j+q_0)}(Z_j^2 - 2), \ \textit{for } r = 1, \cdots, K - q_0 - 1, \tag{18}$$

with $R(0) = 0$, and $(Z_1, \cdots, Z_{K-q_0-1})'$ a vector of normal random variables with mean $\mathbf{0}$ and covariance $\operatorname{cov}(Z_i, Z_j) = v_{(i+q_0)(j+q_0)} \big/ \sqrt{v_{(i+q_0)(i+q_0)} v_{(j+q_0)(j+q_0)}}$.

There are several useful conclusions from Theorem 2. First, the limiting probability that $\hat{q}$ is underselected goes to 0, under both null (6) and alternative (7) hypotheses. Second, the limiting probability that $\hat{q}$ is overselected is not negligible under both null and alternative hypotheses. If the maximizing criterion with $a = 2$ is taken in Theorem 2, it is known that the limiting probability of the Type I error is 0.29 as $K \to \infty$ and $\hat{n} \to \infty$. As a result of Theorem 2, the following corollary can be derived.

Corollary 1. *Under both null (6) and alternative (7) hypotheses,*

$$X_{\hat{q}}^2 - X_{q_0}^2 \xrightarrow{d} W_{r^*}, \ \textit{where } W_r = \sum_{j=1}^{r} v_{(j+q_0)(j+q_0)} Z_j^2,$$

and r^* is the maximizer of the criterion (18), in which the vector of normal random variables $(Z_1, \cdots, Z_{K-q_0-1})'$ is the same as defined in Theorem 2. In addition, for any fixed finite constant C, $\lim_{\hat{n}\to\infty} P\left(X_{\hat{q}}^2 \geq C | q_0 \neq 0\right) = 1$.

Recall criterion (13), $\hat{M}(q) = \frac{\hat{n}+1}{\hat{n}-1}\sum_{j=1}^{q} b_j^2 - \frac{2}{\hat{n}-1}\sum_{j=1}^{q}\hat{v}_{jj}$. The limiting probability of the Type I error is about 0.29 for this case. Now let's consider another maximizing criterion

$$\hat{M}_2(q) = \frac{\hat{n}+1}{\hat{n}-1}\sum_{j=1}^{q} b_j^2 - \frac{a_{\hat{n}}}{\hat{n}-1}\sum_{j=1}^{q}\hat{v}_{jj},$$

where $a_{\hat{n}}$ is allowed to grow with the effective sample size $\hat{n}$ at an appropriate rate. In next theorem, we prove that the estimator $\hat{q}_{a_{\hat{n}}}$ (maximizer of $\hat{M}_2(q)$) is consistent with q_0 if $a_{\hat{n}}$ is large enough.

Theorem 3. *If $a_{\hat{n}} = o(\sqrt{\hat{n}})$ and $a_{\hat{n}} > 2\ln(\ln(\hat{n}))$, we have*

$$\hat{q}_{a_{\hat{n}}} \xrightarrow{P} q_0, \textit{ for } q_0 \geq 0.$$

Theorem 3 says that the limiting probability of Type I error goes to 0, when the effective sample size $\hat{n}$ is large enough and the penalty term $a_{\hat{n}}$ grows with sample size $\hat{n}$ at an appropriate rate.

4.3 Asymptotic Properties of $\hat{q}_\alpha$

In this section, we examine properties of the direct test statistic $\hat{q}_\alpha$ (the maximizer of criterion (15) in section 3.3). For a pre-specified level of significance α, it is reasonable that there exists a value a_α such that $P(\hat{q}_\alpha \neq 0 | q_0 = 0) \to \alpha,$ as $K \to \infty$ and $\hat{n} \to \infty$, where the estimator $\hat{q}_\alpha$ is determined by a_α. Theorem 4 gives the asymptotic behaviors of $\hat{q}_\alpha$.

Theorem 4. *Let $\hat{q}_\alpha$ be maximizer of the criterion (15),* $\hat{M}_\alpha(q) = \frac{\hat{n}+1}{\hat{n}-1}\sum_{j=1}^{q} b_j^2 - \frac{a_\alpha}{\hat{n}-1}\sum_{j=1}^{q}\hat{v}_{jj}$*, for* $q = 1, \cdots, K-1$*, where* $\hat{M}_\alpha(0) = 0$*,* $\hat{v}_{jj} = \sum_{k=1}^{K} x_j^2(k)\hat{p}(k)/p_0(k)$*, for* $j = 1, \cdots, K-1$*, and* a_α *is the solution of Equation (16) or (17). As* $\hat{n} \to \infty$*, we have*

$$P(\hat{q}_\alpha > 0 | q_0 = 0) \to \alpha \textit{ and } P(\hat{q}_\alpha > 0 | q_0 \neq 0) \to 1.$$

5 Simulation Studies

In this section, we proceed with simulation studies to evaluate the proposed methods. For all the settings considered, the proposed tests control Type I error at the prespecified level very well; and the proposed tests also have higher or similar empirical statistical power compared to some existing methods.

5.1 Simulation Set Up

Consider a data with $K = 10$ categories from a complex design, the hypothesis of interest in the simulation studies is:

$$H_0 : p(1) = \cdots = p(10) = 0.1, \tag{19}$$

which is equivalent to $H_0^* : \beta_1 = \cdots = \beta_9 = 0$.

The simulation studies are performed with factors: (a) level of significance $\alpha = 0.05$; (b) basis function in Equation (11) is applied; (c) 50 clusters (psus) with 15 individuals (ssus) sampled from each cluster, which makes a total of 750 units; (d)

Intraclass Correlation Coefficients (ICC) (Lohr, 2010, pg. 174-176, ICC reflects the dependency of the data) considered: 0.1, 0.3, and 0.6 to illustrate low, medium, and high levels of correlation respectively. Notice that if ICC $= 0$, all the observations are uncorrelated. If ICC $= 1$, the ssus within the the same cluster are perfectly correlated to each other, i.e., the individuals within the cluster will give exactly the same answers to the survey questionnaire related to the response of interest; (e) following Eubank (1997), three alternatives of (19) are considered. They are

$$p(k) = \frac{1}{10} + \beta(k - 5.5)/10, \text{ for } k = 1, \cdots, 10, \tag{20}$$

$$p(k) = \frac{1}{10} + \beta\text{cos}\left(\frac{j\pi(k - 0.5)}{10}\right), \text{ for } k = 1, \cdots, 10, \tag{21}$$

and

$$p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right], \text{ for } k = 1, \cdots, 10, \tag{22}$$

where $\Phi(\cdot)$ and $\Phi(\cdot)^{-1}$ are cumulative distribution function and inverse cumulative distribution function of the standard normal random variable respectively.

For alternatives (20) and (21), magnitude of the parameter β controls the distance of the alternative's departure from the null model, which is recovered when $\beta = 0$. Since the influence of β is symmetric about $\beta = 0$, we only report the cases of $\beta > 0$. For alternative (22), null model is recovered when $\beta = 1$. Values of β that are larger or smaller than 1 produce more or less probability mass for the centrally numbered categories.

5.2 Simulation Steps

To generate a sample of clustered multinomial responses, we first generate correlated multivariate normal random vectors, next we use probit functions to convert the continuous responses to categorical responses.

For the proposed test W, $\hat{q}$ is obtained via maximizing equation (13). In order to estimate $\delta_{.}$, $100,000$ complex multinomial data under the null hypothesis (19) are generated. The covariance matrix $\mathbf{V}$ is estimated by these 100,000 samples. $\hat{\delta}_{.}$ is obtained by averaging the eigenvalues of the matrix $\mathbf{P}_0^{-1}\hat{\mathbf{V}}$. Test statistic is calculated by (14). The empirical distribution of W_0 (W under the null hypothesis (19)) is obtained through the $100,000$ iterations, and is used to find the critical value at significance level 0.05.

The proposed test $\hat{q}_\alpha$ is the maximizer of equation (15). Instead of calculating $a_{0.05}$ for each estimated $\hat{\delta}_{.}$, we approximate $a_{0.05}$ by a product of $\hat{\delta}_{.}$ and the values of $a_{0.05}$ estimated under an SRS, which is 4.18 (Eubank & Hart, 1992). By using this method, we only need to estimate $\delta_{.}$ for each setting, which is much faster than calculating the corresponding $a_{0.05}$ for every simulated data. Notice that the solution of Equation (16) requires large K approximations. From simulation study, solution converges to the same value as long as $K > 10$.

We compare Type I error and statistical power of the proposed tests with those from Pearson's chi-squared test, the first order and second order corrected tests. Three alternatives (20), (21) and (22) are used to illustrate possible influence on Type I error and statistical power from different data patterns. Below are the simulation steps for an arbitrary alternative.

1. Generate $100,000$ samples of clustered multinomial under null hypothesis (19). For each generated sample, we calculate the estimated proportion $\hat{\mathbf{p}} = (\hat{p}(1), \cdots, \hat{p}(K-1))$. The estimated mean of $\hat{\mathbf{p}}$, say $\bar{\mathbf{p}}$, is calculated by averaging the $100,000$ $\hat{\mathbf{p}}$'s. The estimated covariance matrix $\hat{\mathbf{V}}/\sqrt{n}$ is obtained by $(\hat{\mathbf{p}} - \bar{\mathbf{p}})(\hat{\mathbf{p}} - \bar{\mathbf{p}})'/(100,000 - 1)$. The eigenvalues of the the matrix $\mathbf{P}_0^{-1}\hat{\mathbf{V}}$ are calculated using the `eigen()` function in `R`. $\hat{a}^2$ is calculated by (10).

2. Under null hypothesis (19), we search for $\hat{q}$: the maximizer of criterion (13), and obtain a value of W_0 by (14). This procedure is repeated $100,000$ times to create the empirical distribution of W_0, and to find the 95% quantile of W_0.

3. Under the given alternative, Pearson's chi-squared test statistic, the first order and second order corrected test statistics are calculated. Next, by searching all $q = 1, \cdots, K-1$, $\hat{q}$ is obtained by maximizing criterion (13). W is calculated by equation (14). We then search $\hat{q}_{0.05}$ among $q = 1, \cdots, K-1$ to maximize equation (15), where $a_{0.05} = 4.18 * \hat{\delta}_{\cdot}$.

4. We compare the test statistics in Step 3 with their corresponding rejection criteria. The Pearson's chi-square test statistic, the first order and the second order corrected test statistics are compared with the 95% quantile of the central chi-squared distribution with 9 degrees of freedom. W is compared with the

95% critical value of W_0 obtained in Step 2. $\hat{q}_{0.05}$ is compared with 0. If a method rejects the alternative, the count of the rejection of this method is 1, otherwise is 0.

5. Step 3 and 4 are repeated for $10,000$ times. The number of rejection of each method, divided by $10,000$ is the empirical power of each method for the given alternative, or is the empirical Type I error of each method when the alternative is set to be the null model.

6. Steps 1-5 are repeated if other alternatives are given.

5.3 Simulation Results

This section presents the simulation results. Type I error and empirical power comparisons for the different tests are reported under three alternatives (20), (21), and (22). Alternative (20) generates slow varying probabilities, alternative (21) generates both slow varying and non-slow varying probabilities, and alternative (22) focuses on certain data patterns.

5.3.1 Simulation Results by Alternative (20)

In this section, we examine the simulation results with alternative (20), i.e., $p(k) = \frac{1}{10} + \beta(k - 5.5)/10,$ for $k = 1, \cdots, 10$. The null hypothesis is recovered when $\beta = 0$. For β ranged from 0 to 0.14 with an increment of 0.01, 15 sets of probability vectors (including the null hypothesis) are generated by (20). Two sets of probabilities for

$\beta = 0.01$ and $\beta = 0.14$ are plotted in Figure 1. These are the probabilities that are used to generate the multinomial data. The probabilities for different categories are very similar when $\beta = 0.01$, and become moderately slow varying as β increases.

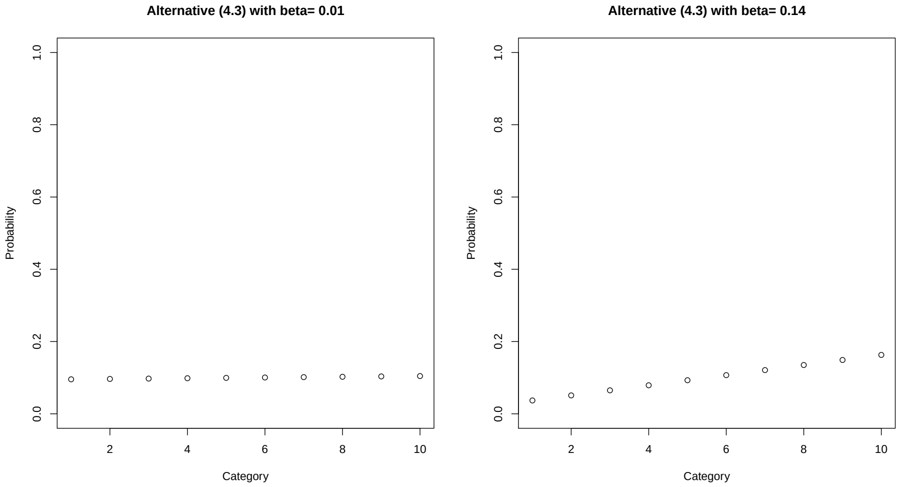
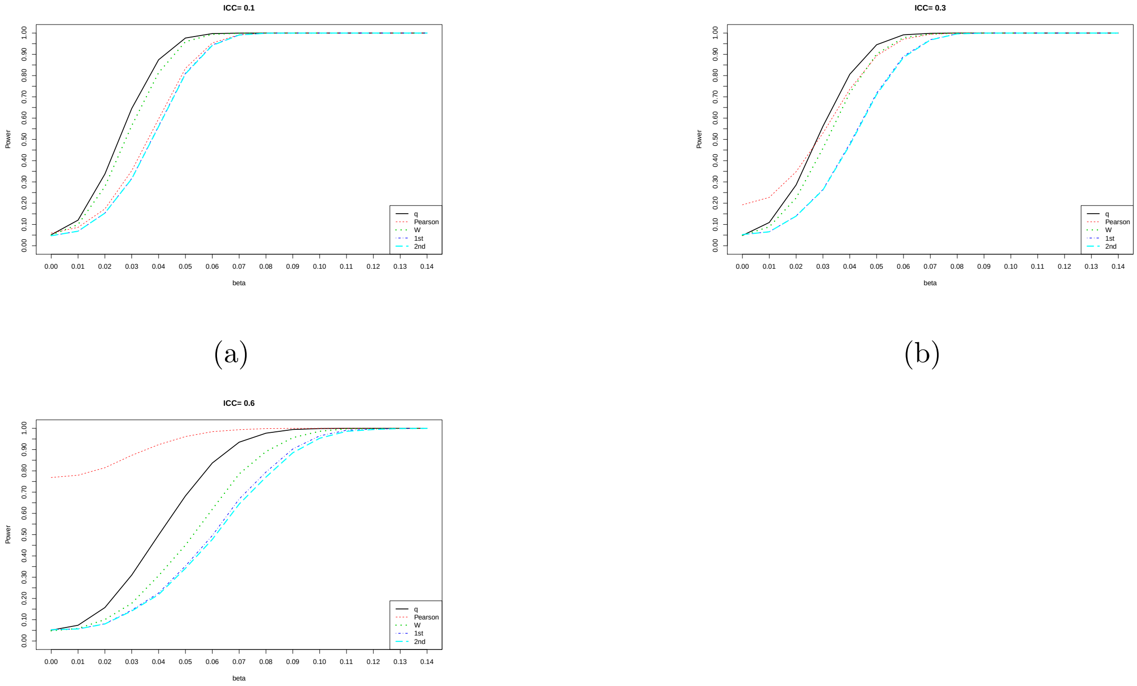


Figure 1: Probabilities generated by alternative (20) for $\beta = 0.01$ (left) and $\beta = 0.14$ (right). Probabilities vary slowly when $\beta = 0.01$, but vary greater when $\beta = 0.14$.

Figure 2 plots the empirical powers of the five tests, the proposed tests $\hat{q}_{0.05}$ and W, Pearson's chi-squared GOF test, and the first order and the second order corrected tests, versus β in alternative (20) under ICC $= 0.1, 0.3$, and 0.6 respectively. For ICC varies from 0.1, 0.3 to 0.6, observations within the same cluster are more correlated.

We first look at how the tests control the Type I error. This is equivalent to check the power of the tests under alternative (20) when $\beta = 0$. When ICC $= 0.1$, Pearson's test controls the probability of the Type I error around 0.05. However, when ICC $= 0.3$, probability of the Type I error of Pearson's test is around 0.18 and is around 0.76 when ICC $= 0.6$. The larger the ICC is, the more off the Type I error is



(a)

(b)

(c)

Figure 2: The power curves of selected methods for simulated complex survey data with respect to the alternative (20) $p(k) = \frac{1}{10} + \beta(k - 5.5)/10$, for $k = 1, \cdots, 10$. From left to right, the plots are with ICC=0.1, 0.3 and 0.6 respectively

for Pearson's chi-squared test. All the other four tests control the type I error pretty well. This is evident that Pearson's chi-squared test should not be directly applied to multinomial data from complex surveys.

We now examine the empirical powers of the proposed tests, and the first order and second order corrected tests. For ICC $= 0.1, 0.3$, and 0.6, both proposed tests W and $\hat{q}_{0.05}$ have higher power than the first order and second order corrected tests when the underlying probabilities are varying slowly ($\beta \leq 0.07$). On the other hand, when the probabilities vary greatly ($\beta > 0.07$), the proposed tests perform similarly as the first order and second order corrected tests in regards to empirical statistical powers. In particular, the test $\hat{q}_{0.05}$ has the best empirical statistical power, and test W has the second best empirical statistical power for alternative (20). In summary, the proposed tests have improved statistical powers compared with the first order and second order corrected tests, when the underlying probabilities vary slowly for multinomial data from complex surveys.

5.3.2 Simulation Results by Alternative (21)

For alternative (21), $p(k) = \dfrac{1}{10} + \beta\text{cos}\left(\dfrac{j\pi(k-0.5)}{10}\right)$, for $k = 1, \cdots, 10$, both β and j are parameters that control variability among underlying probabilities. The probability vector vary greatly when j increased. We choose $j = 2$ (medium varying underlying probabilities) and $j = 4$ (high varying underlying probabilities) to examine the performance of the proposed tests. For each selected j, 11 values of β from 0 to

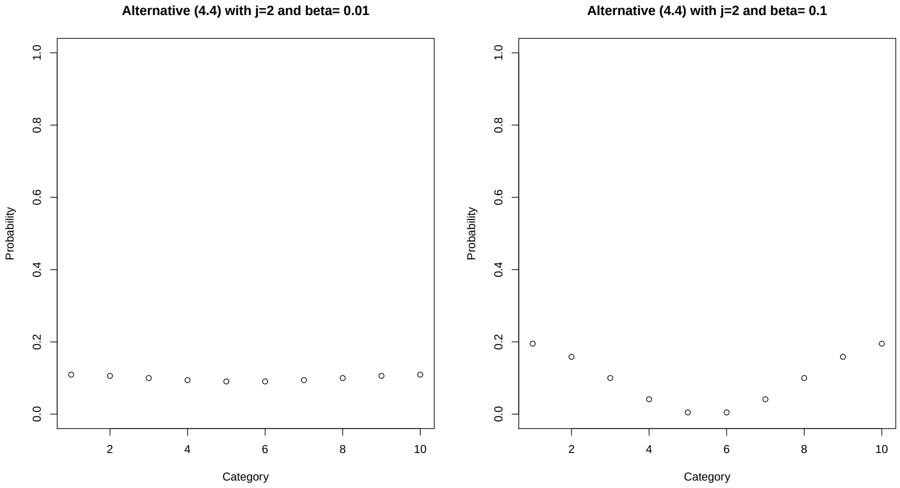


Figure 3: Probabilities in simulation studies generated by alternative (21) with $j = 2$ for $\beta = 0.01$ (left) and $\beta = 0.1$ (right). Probabilities vary slowly when $\beta = 0.01$, but vary greatly when $\beta = 0.1$.

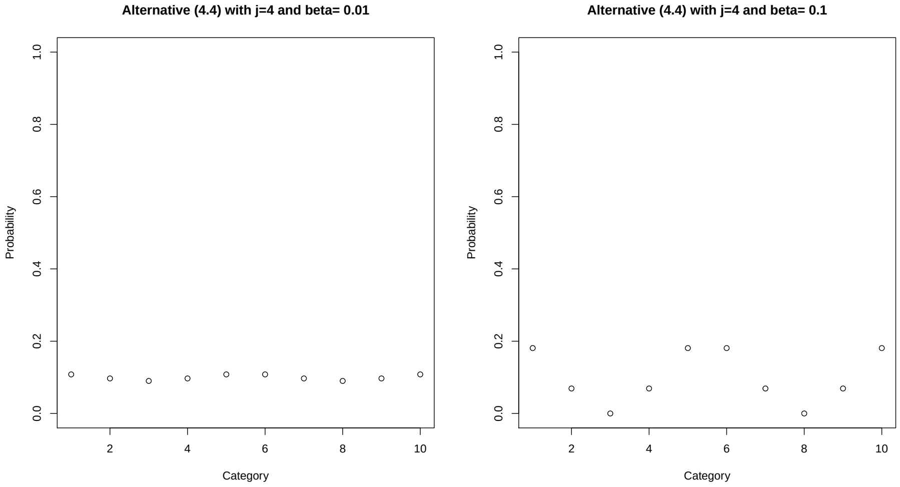
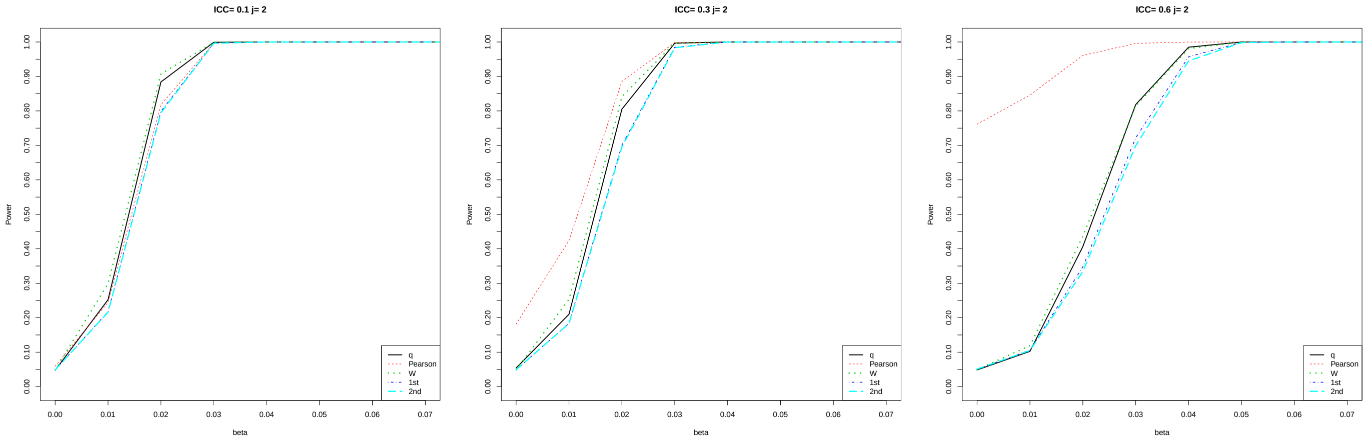


Figure 4: Probabilities in simulation studies generated by alternative (21) with $j = 4$ for $\beta = 0.01$ (left) and $\beta = 0.1$ (right). Probabilities vary slowly when $\beta = 0.01$, but vary greatly when $\beta = 0.1$.



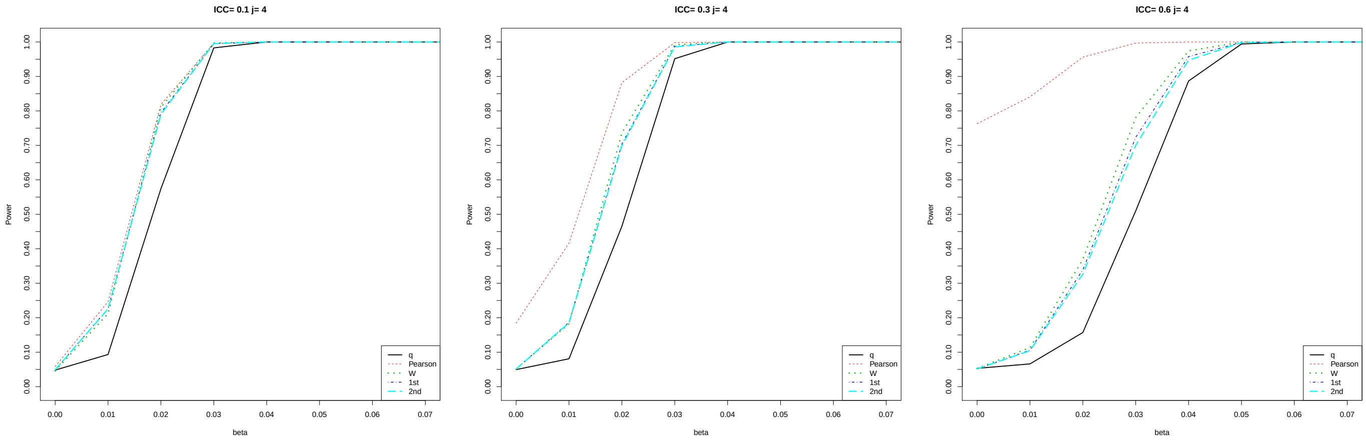
(a) (b) (c)

Figure 5: The power curves of selected methods for simulated complex survey data with respect to the alternative (21) $p(k) = \frac{1}{10} + \beta cos\left(\frac{j\pi(k-0.5)}{10}\right)$, for $k = 1, \cdots, 10$ with $j = 2$.

0.1 are selected with step 0.01. Figure 3 and 4 plot the probability vectors for $j = 2$ with $\beta = 0.01$ and $\beta = 0.1$ and $j = 4$ with $\beta = 0.01$ and $\beta = 0.1$.

Figures 5 and 6 plot the statistical powers of the five tests versus β for ICC $=$ $0.1(Figures5a$ and $6a), 0.3(Figures5b$ and $6b), 0.6(Figures5c$ and $6c)$ for the alternative (21) with $j = 2$ and $j = 4$ respectively.

Notice that null hypothesis (19) is recovered when $\beta = 0$. It can be seen that all the tests except Pearson's chi-squared test are able to control the level of significance at the nominal level ($\alpha = 0.05$) very well. For $j = 2$, both proposed tests have higher statistical power than the first order and second order corrected tests, when the underlying probabilities are varying slowly ($\beta \leq 0.03$). All of the four tests demon-



(a) (b) (c)

Figure 6: The power curves of selected methods for simulated complex survey data with respect to the alternative (21) $p(k) = \frac{1}{10} + \beta cos\left(\frac{j\pi(k-0.5)}{10}\right)$, for $k = 1, \cdots, 10$ with $j = 4$.

strate very similar empirical statistical powers when the underlying probabilities vary greatly ($\beta > 0.03$). In this setting, the two proposed tests are both competitive, and the proposed test W performs slightly better than $\hat{q}_{0.05}$.

The $j = 4$ case generates highly varying underlying probabilities as can be seen from Figure 4. One can see that the proposed test W is competitive, but the statistical power of test $\hat{q}_{0.05}$ has decreased. As ICC goes up, test W becomes the best in regarding statistical power. These results show that the proposed test W is stable in both slow varying and non-slow varying cases. When the underlying probabilities are varying slowly, the proposed test W performs better than the first order and second order corrected tests. When the underlying probabilities are varying greatly,

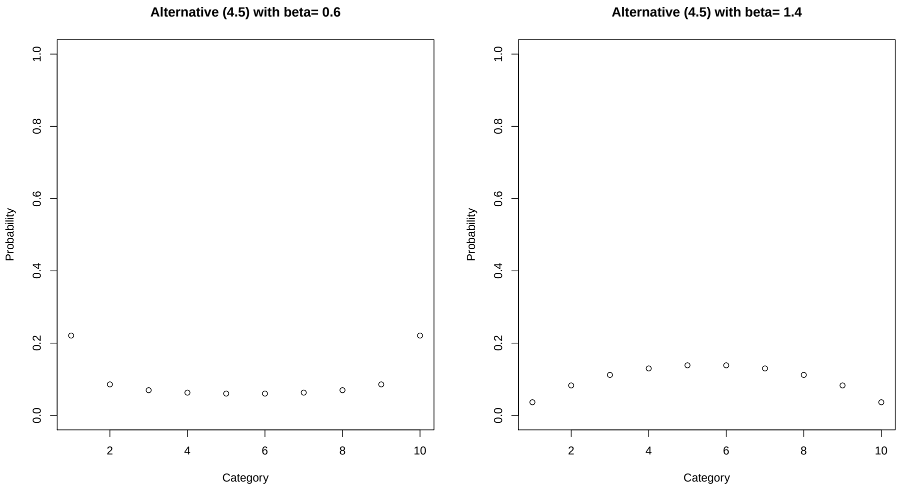
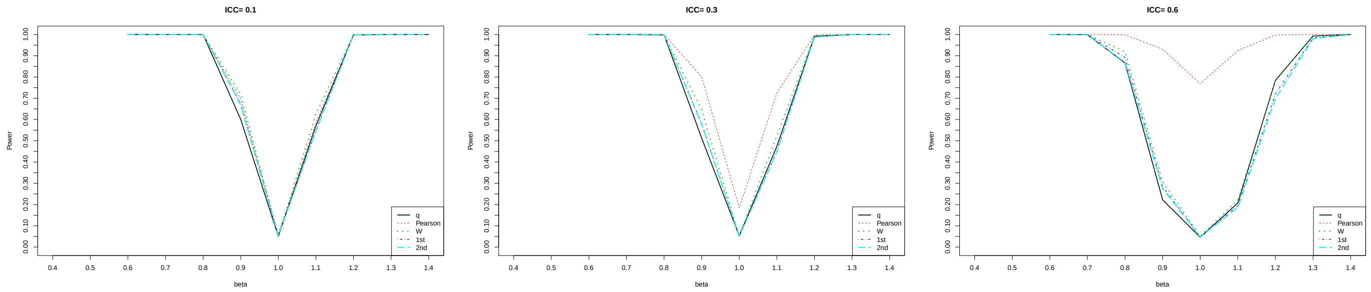


Figure 7: Probabilities in simulation studies generated by alternative (4.5) for $\beta = 0.6$ (left) and $\beta = 1.4$ (right). Maximum probabilities are $p(1)$ and $p(10)$ for $\beta = 0.6$, and maximum probabilities are $p(5)$ and $p(6)$ for $\beta = 1.4$.

the proposed test W is as good as the existing approaches.

5.3.3 Simulation Results by Alternative (22)

For alternative (22), $p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right]$, for $k = 1, \cdots, 10$. β is selected from 0.6 to 1.4 with step 0.1, where null hypothesis (19) is recovered when $\beta = 1$. Alternative (22) simulates a set of moderately slow varying probabilities with bell curves. Notice that, the maximum probability usually occurs at the first and the last categories when β is between 0.6 and 1.0. For example, when $\beta = 0.6$, the maximum probabilities are $p(1)$ and $p(10)$, which are both around 0.221 as shown in the left graph of Figure 7. On the other hand, the largest probabilities occur in the middle categories when β is between 1.0 and 1.4. For example, when $\beta = 1.4$, $p(5) = p(6) = 0.139$ are the maximum probabilities, which is shown in the right graph of Figure 7.



(a) (b) (c)

Figure 8: The power curves of selected methods for simulated complex survey data with respect to the alternative (22) $p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right]$, for $k = 1, \cdots, 10$.

Figure 8 plots the empirical powers of the five tests versus β with ICC$= 0.1(Figure 8a)$, $0.3(Figure 8b)$, and $0.6(Figure 8c)$ respectively. Note that null hypothesis (19) is simulated when $\beta = 1$, which is located in the middle of the graphs. Again, Pearson's chi-squared test is not able to control the nominal level of significance $\alpha = 0.05$. All other four tests can control the desired Type I error.

The proposed test W outperforms the first order and second order corrected tests, and it is superior than the proposed test $\hat{q}_{0.05}$ when $\beta < 1$. When $\beta > 1$, the proposed test $\hat{q}_{0.05}$ becomes almost as good as the proposed test W, and both of them have higher statistical power than first order and second order corrected tests. Though all methods (not including Pearson's chi-squared test) perform similarly with each other, the proposed test $\hat{q}_{0.05}$ works best under alternative (22).

5.3.4 Summary of Simulation Results

In this section, we summarize the findings of simulation studies. First, both proposed tests, first and second corrected tests control Type I error at the pre-specified level of significance under all settings. Second, comparing with first and second order corrected tests, the proposed tests substantially improve the empirical statistical powers when the underlying probabilities vary slowly. Third, the proposed test W shows a great stability in statistical powers, and performs competitively with other methods in cases of non-slow varying probability. Fourth, the proposed test $\hat{q}_\alpha$ work best under alternative (22). Finally, as a result of the comparison between alternative (20) and (21), we conclude that the proposed test $\hat{q}_\alpha$ is the most powerful one for slow varying probabilities, but it is not as stable as the proposed test W for non-slow varying probability cases. In another perspective of view, the proposed test $\hat{q}_\alpha$ is more sensitive in detecting small differences among the underlying probabilities, but the proposed test W is more stable for various cases. In practice, the selected approach should be determined by the characteristics of the multinomial data.

6 Application

In this section, we apply the proposed Neyman smooth-type GOF tests in complex surveys to real life problems. For comparison purpose, the results of GOF tests, such as Pearson's chi-squared test, the first order and second order corrected tests (Rao

& Scott, 1981, 1984), are also reported. Data is from the National Youth Tobacco Survey (NYTS). We are interested in testing the difference among the severity groups on tobacco usage for Asian and American Indian/Alaska Native students.

6.1 Data Description

The NYTS is to provide data support for research related to the use of tobacco among middle and high school students. A variety of tobaccos are included, such as cigarettes, cigars, hookahs, electronic cigarettes, and so on. NYTS started in 1999, and continued in 2000, 2002, 2004, 2006, 2011, 2012, 2013, 2014 until now. The Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA) have participated in the management of NYTS since 2011.

The 2014 NYTS is a stratified three-stage clustered sample (Office on Smoking and Health (2014)). 16 strata were created in U.S. based on predominant minority (non-Hispanic Black and Hispanic) and the factor of urban/nonurban. A psu is defined as a county, a combination of several small counties, or part of a large county. More detailed information on the psu can be obtained from the Office on Smoking and Health (2014, pg. 7). Middle schools and high schools were considered as ssus in each psu. In each selected school, 1 or 2 classes were selected for every grade. All students in the selected classes were eligible for the interview. Sampling was done without replacement. Figure 9 shows a concise survey design flow chart of NYTS. More details can be found in Chapter 2 of Office on Smoking and Health (2014).

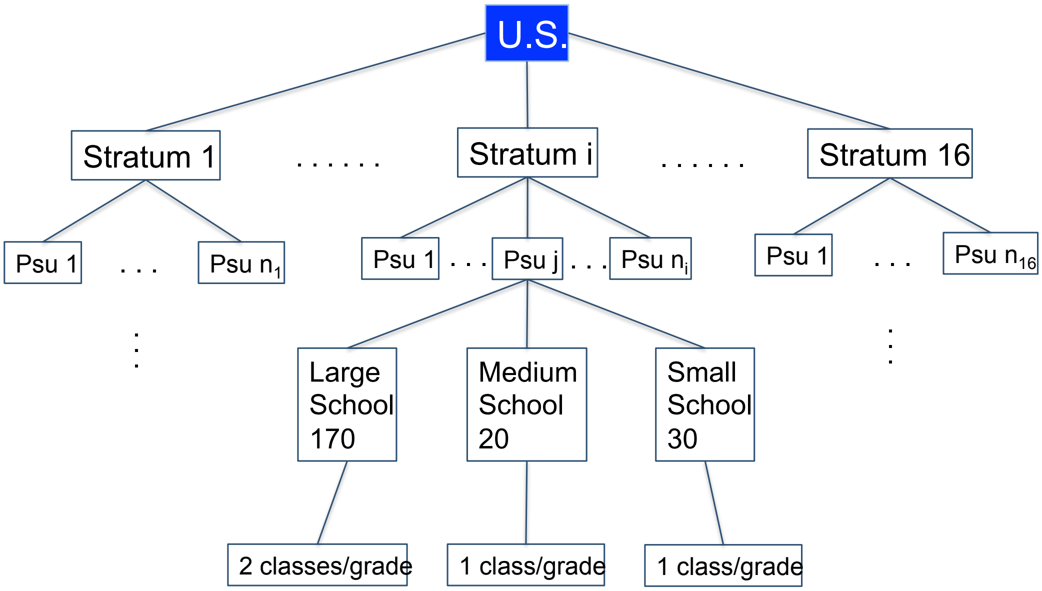


Figure 9: NYTS 2014 sampling design chart.

There are about 81 questions in the questionnaire. Students were required to answer these questions using pencils. The collected data was trimmed and the sampling weights of the individuals were calculated based on the sampling design and nonresponse adjustments. Office on Smoking and Health (2014)[chap 4] gives a detailed description on how to calculate the sampling weight. The 2014 NYTS data consists of 157 variables (including weight variable) and a total of 22, 007 observations.

6.2 Severity Differences Among Asian Students Smokers

In this example of application, we focus on Asian students who smoked during the past 30 days before they were surveyed. 25 out of 973 Asian students reported that they smoked in the past 30 days. In addition, they also reported the number of cigarettes smoked per day, which was categorized into 5 levels, < 1/day (light smokers), 1/day (moderately light smokers), $2-5$/day (medium smokers), $6-10$/day (moderately heavy smokers), and ≥ 11/day (heavy smokers).

We are interested in examining the differences of proportions regarding smoking severity among these students. The null hypothesis is

$$H_0 : p(1) = \cdots = p(5) = \frac{1}{5}. \tag{23}$$

The observed, weighted counts and the estimated proportions of the 5 groups are shown in Table 1. Figure 10 plot the estimated proportions $\hat{p}(k)$'s.

Number Group	< 1	1	2-5	6-10	≥ 11	Total
Counts	7	6	8	2	2	25
Weighted Counts	6840.261	5818.418	6595.912	1391.909	1907.703	22554.2
$\hat{p}(k)$	0.303	0.258	0.292	0.062	0.085	1

Table 1: Observed and weighted data for Asian students who reported smoking during the past 30 days of the survey from NYTS 2014. Proportions $\hat{p}(k)$ is calculated using weighted data. "Number group" are the number of cigarettes smoked per day.

The observed sample size in Table 1 is $n = 25$. We find $\hat{\delta}_. = 1.21677$ and $\hat{a} \approx 0$. The test statistics of the first order and second order corrected tests are

$$X_C^2 = \frac{X^2}{\hat{\delta}_.} = \sum_{k=1}^{K} n \frac{(\hat{p}(k) - p_0(k))^2}{p_0(k)} / \hat{\delta}_. = 5.62$$

and, because $\hat{a} \approx 0$,

$$X_S^2 = \frac{X^2}{\hat{\delta}_.(1 + \hat{a}^2)} = \frac{X_C^2}{(1 + \hat{a}^2)} \approx 5.62.$$

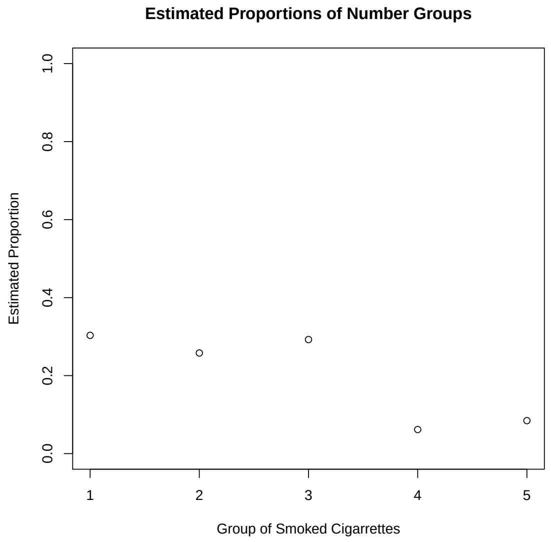


Figure 10: Estimated proportions of the 5 smoking groups using weighted counts in Table 1.

Compared to $\chi^2(5)$ at significance level of 0.05, P-value =0.23 for both tests. We fail to reject the null hypothesis (23).

Next, we use the proposed methods to test the hypothesis. Follow simulation step2, we can simulate the empirical distribution of W_0 and find 95% quantile of W_0. By searching all $q = 1, \cdots, 5-1$, $\hat{q}$ is the one that maximizes equation (13). We found that

$$\hat{q} = 1,\ W = 2.99,\ \text{and p} - \text{value} = 0.039.$$

For the proposed test $\hat{q}_\alpha$, we have

$$\hat{q}_{0.05} = 1 \text{ and p} - \text{value} = 0.033.$$

Both W and $\hat{q}_\alpha$ tests reject hypothesis (23) at level of 0.05.

In this example, the first order and second order corrected tests fail to reject the null hypothesis (23) at level of significance 0.05, though it was observed that the proportions of Asia students who smoked $6 - 10$ and ≥ 11 cigarettes per day were low. On the other hand, both proposed test W and $\hat{q}_\alpha$ are able to reject the null hypothesis at level 0.05, indicating that the numbers of light, moderately light, medium, moderately heavy, and heavy smokers are different among the Asian students (grades 6-12) in the U.S., which is consistent with what was observed.

7 Conclusion

Categorical data analysis is widely used in complex surveys arising from sociological, behavioral, economical, and medical research studies, where the observations are usually correlated. In this research, we proposed two Neyman smooth-type GOF tests (W and q_α) for use in complex survey multinomial data. These tests control the type I error at specified significance level very well. They also show improved statistical powers compared with some classical methods, especially when the sample size is small and the differences among the estimated proportions of categories are not large (slow varying probabilities). Simulation results show that test $\hat{q}_\alpha$ is the most powerful test for slow varying probability data compared to test W and the first and second order corrected tests. The proposed test W is a stable test, which outperforms the first order and second order tests when the probabilities are slow varying, and is as good as the first order and second order corrected tests when the probabilities are varying greatly.

The application of the idea of Fourier transformation and dimensional reduction (order selection) in survey data provides a broad view of many possible topics. For example, for a nonparametric regression model $y_i = \mu(t_i) + \epsilon_i$ in survey data, we may be interested in testing if $\mu(t_i)$ is a constant or not. We can introduce the basis function with Fourier transformation, extend the classical nonparametric regression estimator to weighted estimators by incorporating survey weights, and a tuning parameter q for dimensional reduction. Many other tests could be derived by a similar procedure as we have done in this research. We expect the proposed tests to be more sensitive and to provide more statistical power in detecting the differences by taking advantage of dimensional reduction when constructing test statistics.

Appendix

Below is the proof of Theorem 1.

Proof. First, we prove the theorem in an SRS without replacement case. Next, we discuss complex design case. In this proof, we define $\mathbf{p}^* = (\mathbf{p}', p_k)'$, $\mathbf{p}_0^* = (\mathbf{p}_0', p_{0k})'$, $\hat{\mathbf{p}}^* = (\hat{\mathbf{p}}', \hat{p}_k)'$, and $\mathbf{P}^* = D(\mathbf{p}^*) - \mathbf{p}^*\mathbf{p}^{*'}$.

In an SRS without replacement,

$$\mathrm{var}(\hat{\mathbf{p}}^*) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{N}{N-1}\mathbf{P}^* \approx \frac{1}{n}\left(1 - \frac{n}{N}\right)\mathbf{P}^* = \frac{1}{\tilde{n}}\mathbf{P}^*.$$

When $\tilde{n} \to \infty$,

$$\sqrt{\tilde{n}}(\hat{\mathbf{p}}^* - \mathbf{p}^*) \to N(\mathbf{0}, D(\mathbf{p}^*) - \mathbf{p}^*\mathbf{p}^{*'})$$

and

$$\sqrt{\tilde{n}}\left(\hat{\mathbf{F}} - \mathbf{F}\right) \to N\left(\mathbf{0}, D\left(\frac{\mathbf{p}^*}{\mathbf{p}_0^*}\right) - \left(\frac{\mathbf{p}^*}{\sqrt{\mathbf{p}_0^*}}\right)\left(\frac{\mathbf{p}^*}{\sqrt{\mathbf{p}_0^*}}\right)'\right).$$

Therefore,

$$\begin{aligned} \sqrt{\tilde{n}}(\mathbf{b} - \boldsymbol{\beta}) &= \sqrt{\tilde{n}}\left(\mathbf{X}'_{[k]}\hat{\mathbf{F}} - \mathbf{X}'_{[k]}\mathbf{F}\right) \\ &\to N(\mathbf{0}, \mathbf{X}'_{[k]}D\left(\frac{\mathbf{p}^*}{\mathbf{p}_0^*}\right)\mathbf{X}_{[k]} - \mathbf{X}'_{[k]}\left(\frac{\mathbf{p}^*}{\sqrt{\mathbf{p}_0^*}}\right)\left(\frac{\mathbf{p}^*}{\sqrt{\mathbf{p}_0^*}}\right)'\mathbf{X}'_{[k]} \\ &= N\left(\mathbf{0}, \mathbf{X}'_{[k]}D\left(\frac{\mathbf{p}^*}{\mathbf{p}_0^*}\right)\mathbf{X}_{[k]} - \boldsymbol{\beta}\boldsymbol{\beta}'\right) \end{aligned} \tag{24}$$

If the design is complex, let $\mathrm{var}(\hat{\mathbf{p}}^*) = \frac{1}{\hat{n}}\mathbf{V}^*$. We can derive

$$\sqrt{\hat{n}}(\mathbf{b} - \boldsymbol{\beta}) \to N\left(\mathbf{X}'_{[k]}D\left(\frac{1}{\sqrt{\mathbf{p}_0^*}}\right)\mathbf{V}^*D\left(\frac{1}{\sqrt{\mathbf{p}_0^*}}\right)\mathbf{X}_{[k]}\right) \tag{25}$$

as $\hat{n} \to \infty$. For some special case, such as a stratified random sampling with replacement and with proportional allocation, and two stage cluster sampling, with the first stage proportional to psu size and second stage as SRS with replacement, Rao and Scott (1981) gives detailed formula of $\mathbf{V}^*$. □

The proofs of Theorems 2-4 follow from proofs of Theorems 1-3 in Eubank (1997). When constructing the test statistics, we assume that $V(\hat{\mathbf{p}}) = \delta.V_{\mathrm{SRS}}(\hat{\mathbf{p}})$. The maximizing criteria (13) and (15) are mainly replacing sample size n by the effective size $\hat{n} = n/\delta.$. Therefore, proofs of Theorems 2-4 follow from Eubank (1997).

References

Arfken, G. (1985). *Mathematical methods for physicists (3rd ed.).* Orlando, FL: Academic Press.

Bedrick, E. J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika*, *70*, 591-595.

Eubank, R. L. (1997). Testing goodness of fit with multinomial data. *Journal of the American Statistical Association*, *92*, 1084-1093.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing (2nd edition).* CRC Press.

Eubank, R. L., & Hart, J. D. (1992). Testing goodness of fit in regression via order selection criteria. *The Annals of Statistics*, *20*, 1412-1425.

Fay, R. E. (1979). On adjusting the Pearson chi-square statistic for clustered sampling. In *ASA Proceedings of the Social Statistics Section* (pp. 402–406). American Statistical Association.

Fay, R. E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, *80*, 148-157.

Hart, J. D. (1985). On the choice of a truncation point in fourier series density estimation. *Journal of Statistical Computing and Simulation*, *21*, 95-116.

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*, 89-96.

kwang Kim, J., Rao, J. N. K., & Wang, Z. (2019). *Hypotheses testing from com-*

plex survey data using bootstrap weights: A unified approach (Technical Paper). Ithaca, NY: Cornell University.

Lancaster, H. O. (1969). *The chi-squared distribution.* New York: Wiley.

Lehmann, E. L. (1986). *Testing statistical hypotheses (2nd ed.).* New York: Wiley.

Lohr, S. L. (2010). *Sampling: Design and analysis (2nd ed.).* Belmont, CA: Duxbury Press.

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, *20*, 150-199.

Office on Smoking and Health. (2014). *2014 national youth tobacco survey: Methodology report* (Technical Paper). Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. Retrieved from `www.cdc.gov/tobacco/data_statistics/surveys/nyts`

Rao, J. N. K., & 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., & 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. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, *15*, 385–397.

Rayner, J. C. W., & Best, D. J. (1986). Neyman-type smooth tests for location-scale families. *Biometrika*, *73*, 437-446.

Rayner, J. C. W., & Best, D. J. (1989). *Smooth tests of goodness of fit.* New York: Oxford University Press.

Rayner, J. C. W., & Best, D. J. (1990). Smooth tests of goodness-of-fit: An overview. *International Statistical Review*, *58*, 9-17.

Rayner, J. C. W., Best, D. J., & Dodds, K. G. (1985). The construction of the simple x^2 and neyman smooth goodness of fit tests. *Statistica Neerlandica*, 35-50.

Rayner, J. C. W., Thas, O., & Best, D. J. (2009). *Smooth tests of goodness of fit: Using r (2nd ed.).* New York: Wiely.

Spitzer, F. (1956). A combinatorial lemma with its applications to probability theory. *Transactions of the American Mathematical Society*, *82*, 323-339.

Thomas, D. R., Singh, A. C., & Roberts, G. R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, *64*, 295–311.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426-482.

Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, *87*, 732-737.