Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Neyman smooth-type goodness-of-fit tests in complex surveys

Yan Lu, Lang Zhou, Guoyi Zhang & Ronald Christensen

To cite this article: Yan Lu, Lang Zhou, Guoyi Zhang & Ronald Christensen (21 Jan 2026): Neyman smooth-type goodness-of-fit tests in complex surveys, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2616882

To link to this article: https://doi.org/10.1080/24754269.2026.2616882

Published online: 21 Jan 2026.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2616882

Taylor & Francis
Taylor & Francis Group

Neyman smooth-type goodness-of-fit tests in complex surveys

Yan Lu[a], Lang Zhou[b], Guoyi Zhang[a] and Ronald Christensen[a]

[a]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, USA; [b]Kite Pharma Inc., A Gilead Company, Santa Monica, CA, USA

ABSTRACT

In this study, we extend Neyman smooth-type goodness-of-fit tests to complex survey settings involving categorical data, by incorporating design-consistent estimators under the survey framework. This extension is implemented through data-driven, nonparametric order selection methods. We examine the asymptotic properties of the proposed estimators and demonstrate, through simulations, that our methods improve statistical power while maintaining strong control over Type I error, particularly in detecting subtle yet systematic differences across categories. We also illustrate the practical utility of our approach using data from the National Youth Tobacco Survey (NYTS).

ARTICLE HISTORY
Received 1 February 2025
Revised 27 May 2025
Accepted 11 January 2026

KEYWORDS
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 are ubiquitous in sociological and economic research. A key challenge lies in developing goodness-of-fit (GOF) tests that account for design complexities such as stratification, clustering and unequal sampling. Early contributions include Wald's test (Wald, 1943), which requires covariance matrix estimation, and Fay's jackknifed chi-squared tests (Fay, 1979, 1985) for complex designs. However, these methods often demand detailed survey information that is rarely accessible in practice. Rao and Scott (1981, 1984) addressed this by proposing limited-information tests for multiway tables, while Bedrick (1983) and Rao and Scott (1987) leveraged marginal design effects for approximate GOF assessments. Recent work by Kim et al. (2019) introduced bootstrap approximations for weighted likelihood ratio statistics, and Lu (2014) extended Rao–Scott corrections to dual-frame surveys. Building on Skinner's foundational work on design-based covariance estimation (Skinner, 1989) and categorical data analysis (Skinner, 2019), Jamil et al. (2025) advanced GOF testing for binary factor models under complex sampling by integrating pairwise likelihood estimation with survey weights, demonstrating robust performance of limited-information Pearson and Wald-type tests.

The Neyman smooth test framework (Neyman, 1937) offers an alternative paradigm for GOF evaluation, particularly for its decomposability into orthogonal components. Lancaster (n.d.) and Rayner et al. (1985) established connections between Pearson's chi-squared test and Neyman smooth tests. Building on this foundation, Rayner and Best (1986) extended

CONTACT Yan Lu yanlu@unm.edu Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131-0001, USA

the framework to location-scale families, broadening the framework to accommodate continuous distributions beyond categorical settings. Eubank (1997) introduced a data-driven extension by incorporating order selection techniques to determine the optimal number of components, enhancing the practical utility of smooth tests for detecting model deviations in i.i.d. settings (detailed in Eubank, 1999). A comprehensive overview of Neyman smooth-type goodness of fit (GOF) tests can be found in Rayner and Best (1989, 1990) and Rayner et al. (2009). Although composite likelihood methods (Lindsay, 1988; Varin, 2008) and weighted estimation frameworks (Muthén & Satorra, 1995; Skinner, 1989) have been developed to address the challenges of structured and complex data, the application of Neyman-type smooth GOF tests in complex surveys remains underdeveloped.

In this study, we extend Eubank's order-selection approach (Eubank, 1999) to complex survey designs under the design-based framework of Särndal (2003), which employs triangular-array superpopulation assumptions without parametric constraints. Our contributions include (1) integrating design-consistent estimators into Neyman smooth-type tests, (2) developing data-driven order selection criteria tailored for survey data and (3) establishing asymptotic properties under stratified multistage sampling. Simulations and an empirical application demonstrate improvements over existing methods, particularly in scenarios with subtle yet systematic differences across groups.

This paper is organized as follows. Section 2 introduces the proposed Neyman smooth-type GOF tests for complex surveys. Section 3 derives their asymptotic properties, while Section 4 evaluates performance against established benchmarks. An application to educational survey data is presented in Section 5, followed by conclusions and future directions in Section 6.

2. Neyman smooth-type GOF tests in complex surveys

In this section, we extend the Neyman smooth-type GOF test proposed by Eubank (1997) to the context of complex surveys. We begin by introducing the necessary notation and formulating the research problem. We then propose two Neyman smooth-type GOF tests that incorporate data-driven order selection, tailored for use in complex survey designs.

2.1. Notation and research problem

Let $Y = (Y_1, Y_2, \ldots, Y_K)^\top$ follow a size n multinomial distribution. Let $\boldsymbol{p}_0 = (p_{01}, p_{02}, \ldots, p_{0(K-1)})^\top$ be the underlying probability vector with $p_{0K} = 1 - \sum_{k=1}^{K-1} p_{0k}$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_{K-1})^\top$. A general hypothesis of interest is

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

For each observation j, let $y_j(k) = 1$ if the outcome falls in category k, and 0 otherwise. Let w_j be the sampling weight associated with observation j based on a specified survey design. The design-based estimator of the proportion in category k is given by $\hat{p}_k = \sum_{j=1}^{n} w_j y_j(k)/N$, where $N = \sum_{j=1}^{n} w_j$ is the estimated population size.

The estimators $\hat{p}_k$, for $k = 1, 2, \ldots, K-1$, form a set of purely design-based but jointly consistent and asymptotically normal estimators of the population proportion vector $\boldsymbol{p}$ for category k. Let $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{K-1})^\top$. For survey data, a weighted version of the chi-squared statistic used to measure the distance between observed and expected proportions

under H_0 is given by

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

Equation (2) reduces to the classical Pearson chi-squared test statistic,

$$X^2 = \sum_{i=1}^{K}\frac{(Y_i - e_i)^2}{e_i},$$

when the sampling weights satisfy $w_i = N/n$ under a simple random sample (SRS) design.

Under the null hypothesis (1), the weighted-up X^2 in (2) is distributed asymptotically as a linear sum of $\delta_1 W_1 + \cdots + \delta_{K-1} W_{K-1}$ (Rao & Scott, 1981) instead of a χ^2_{K-1} random variable, where W_i's are i.i.d. χ^2_1 random variables. The weights δ_i's are eigenvalues of the design effect matrix $\boldsymbol{P}^{-1}\boldsymbol{V}$ under H_0, where $\boldsymbol{P} = D(\boldsymbol{p}_0) - \boldsymbol{p}_0\boldsymbol{p}_0^\top$, $D(\boldsymbol{p}_0)$ is a $(K-1)\times(K-1)$ matrix with kth diagonal element p_{0k} and off-diagonal entries 0, and $\boldsymbol{V}/n$ is the covariance matrix of $\hat{\boldsymbol{p}}$.

Define the basis vectors $\boldsymbol{x}_i = (x_i(1), x_i(2), \ldots, x_i(K))^\top$, for $i = 1, 2, \ldots, K-1$, and let F be a $(K-1)$-dimensional subspace of $\mathbb{R}^K$ that satisfies the orthogonality constraint:

$$\sum_{k=1}^{K} x_j(k)\sqrt{p_{0k}} = 0, \quad \text{for } j = 1, \ldots, K-1. \tag{3}$$

If, in addition, the basis vectors satisfy the orthonormality condition:

$$\boldsymbol{x}_j^\top \boldsymbol{x}_i = \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{if } j \neq i, \end{cases} \quad \text{for } i, j = 1, \ldots, K-1, \tag{4}$$

then the set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{K-1}\}$ forms an orthonormal basis for the subspace $\mathcal{F}$. Note that the choice of basis functions for a given hypothesis is not unique. Eubank (1999, p. 75) provides a detailed discussion on how to construct such basis functions, while Eubank (1997) discusses how to select them to improve power against specific alternatives. Let

$$\hat{f}(k) = \frac{\hat{p}_k - p_{0k}}{\sqrt{p_{0k}}}$$

be the normalized deviation for category k, with associated Fourier coefficients defined by

$$\hat{b}_j = \sum_{k=1}^{K}\hat{f}(k)x_j(k).$$

Since $E[\hat{p}_k] = p_k$, it follows that $\hat{f}(k)$ is an unbiased estimator of

$$f(k) = \frac{p_k - p_{0k}}{\sqrt{p_{0k}}},$$

with corresponding Fourier coefficients

$$\beta_j = \sum_{k=1}^{K} f(k)x_j(k).$$

Define $\boldsymbol{x}_K = (\sqrt{p_{01}}, \sqrt{p_{02}}, \ldots, \sqrt{p_{0K}})^\top$, and let

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K) \quad \text{and} \quad \boldsymbol{X}_{[K]} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{K-1}).$$

The orthogonality condition in Equation (3) can equivalently be expressed as

$$\boldsymbol{x}_j^\top \boldsymbol{x}_K = 0, \quad \text{for } j = 1, \ldots, K-1.$$

Let

$$\widehat{\boldsymbol{F}} = (\hat{f}(1), \hat{f}(2), \ldots, \hat{f}(K))^\top, \quad \boldsymbol{F} = (f(1), f(2), \ldots, f(K))^\top,$$
$$\hat{\boldsymbol{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{K-1})^\top, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{K-1})^\top,$$

so that the following relationships hold:

$$\hat{\boldsymbol{b}} = \boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}}, \quad \boldsymbol{F}^\top \boldsymbol{x}_K = 0, \quad \boldsymbol{\beta} = \boldsymbol{X}_{[K]}^\top \boldsymbol{F}.$$

Note that under the null hypothesis H_0, the vector $\widehat{\boldsymbol{F}}$ satisfies the constraint $\boldsymbol{x}_K^\top \widehat{\boldsymbol{F}} = 0$, which forces $\widehat{\boldsymbol{F}} \in \mathcal{F}$, the subspace spanned by $\boldsymbol{X}_{[K]}$. The projection matrix acts as an identity operator on $\mathcal{F}$: $\boldsymbol{X}_{[K]} \boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}} = \widehat{\boldsymbol{F}}$. Parseval's relation (Arfken, 1985, p. 425) ensures that its squared norm equals the squared norm of its Fourier coefficients. Thus

$$n\widehat{\boldsymbol{F}}^\top \widehat{\boldsymbol{F}} = n\|\widehat{\boldsymbol{F}}\|^2 = n\|\boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}}\|^2 = n\widehat{\boldsymbol{F}}^\top \boldsymbol{X}_{[K]} \boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}}.$$

Now expand the full matrix $\boldsymbol{X}\boldsymbol{X}^\top$ to $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{X}_{[K]}\boldsymbol{X}_{[K]}^\top + \boldsymbol{x}_K \boldsymbol{x}_K^\top$. Under H_0, $\boldsymbol{x}_K^\top \widehat{\boldsymbol{F}} = 0$, and

$$\widehat{\boldsymbol{F}}^\top \boldsymbol{X}\boldsymbol{X}^\top \widehat{\boldsymbol{F}} = \widehat{\boldsymbol{F}}^\top \boldsymbol{X}_{[K]} \boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}} + \widehat{\boldsymbol{F}}^\top \boldsymbol{x}_K \boldsymbol{x}_K^\top \widehat{\boldsymbol{F}} = \widehat{\boldsymbol{F}}^\top \boldsymbol{X}_{[K]} \boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}} = \widehat{\boldsymbol{F}}^\top \widehat{\boldsymbol{F}}.$$

Therefore, the weighted-up X^2 defined in (2) can be written as

$$\begin{aligned} X^2 &= n \sum_{k=1}^{K} \frac{(\hat{p}_k - p_{0k})^2}{p_{0k}} = n\widehat{\boldsymbol{F}}^\top \widehat{\boldsymbol{F}} = n\widehat{\boldsymbol{F}}^\top \boldsymbol{X}\boldsymbol{X}^\top \widehat{\boldsymbol{F}} \\ &= n\hat{\boldsymbol{b}}^\top \hat{\boldsymbol{b}} = n \sum_{j=1}^{K} \hat{b}_j^2 = n \sum_{j=1}^{K-1} \hat{b}_j^2, \end{aligned}$$

where the last equality follows from the fact that $\hat{b}_K = \widehat{\boldsymbol{F}}^\top \boldsymbol{x}_K = 0$.

Note that $f(k) = 0$ under the null hypothesis $\boldsymbol{p} = \boldsymbol{p}_0$. Let $\hat{\boldsymbol{b}}_q = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_q)^\top$, $\boldsymbol{\beta}_q = (\beta_1, \beta_2, \ldots, \beta_q)^\top$, and define $\boldsymbol{X}_q = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q)$.

A Neyman smooth-type GOF test statistic is given by $\sum_{j=1}^{q} n\hat{b}_j^2$, for some $q = 1, 2, \ldots, K-1$. If the true order is

$$q_0 = \max\{k \leq K-1 : \beta_k \neq 0\},$$

then an estimator of q_0 can be obtained by minimizing the following criterion:

$$\sum_{k=1}^{K}\left(\sum_{j=1}^{q}\hat{b}_j x_j(k) - f(k)\right)^2 = (\boldsymbol{X}_q\boldsymbol{X}_q^\top\widehat{\boldsymbol{F}} - \boldsymbol{F})^\top(\boldsymbol{X}_q\boldsymbol{X}_q^\top\widehat{\boldsymbol{F}} - \boldsymbol{F})$$
$$= \hat{\boldsymbol{b}}_q^\top\hat{\boldsymbol{b}}_q - 2\hat{\boldsymbol{b}}_q^\top\boldsymbol{\beta}_q + \boldsymbol{\beta}^\top\boldsymbol{\beta}. \tag{5}$$

In this study, we propose statistical tests for the hypothesis H_0 versus the alternative H_α based on the criterion in Equation (5), and investigate the properties of the resulting estimators.

2.2. Proposed test W

In this section, we propose a general test statistic W for testing H_0 under complex survey designs. Since $\boldsymbol{\beta}^\top\boldsymbol{\beta}$ does not depend on q, minimizing the criterion in Equation (5) is equivalent to maximizing the following:

$$M(q) = -\hat{\boldsymbol{b}}_q^\top\hat{\boldsymbol{b}}_q + 2\hat{\boldsymbol{b}}_q^\top\boldsymbol{\beta}_q.$$

Let $\delta_\cdot = \sum_{i=1}^{K-1}\delta_i/(K-1)$ be the average design effect, and let $\hat{\delta}_\cdot$ be its estimator. It can be shown that $(K-1)\delta_\cdot = n\sum_{i=1}^{K} V_{ii}/p_{0i}$, so $\delta_\cdot$ can be estimated using the diagonal elements of the estimated covariance matrix $\boldsymbol{V}$.

In complex surveys, we approximate the design-based covariance matrix as

$$\widehat{V}(\hat{\boldsymbol{p}}) = \hat{\delta}_\cdot \cdot \widehat{V}_{\mathrm{SRS}}(\hat{\boldsymbol{p}}),$$

where $\widehat{V}_{\mathrm{SRS}}(\hat{\boldsymbol{p}})$ is the covariance matrix under simple random sampling. Following Kish (1965), we define the effective sample size as $\tilde{n} = n/\hat{\delta}_\cdot$.

Adapting the criterion proposed by Eubank (1997), we define the following maximization criterion for complex surveys:

$$\widehat{M}(q) = \frac{\tilde{n}+1}{\tilde{n}-1}\sum_{j=1}^{q}\hat{b}_j^2 - \frac{2}{\tilde{n}-1}\sum_{j=1}^{q}\hat{v}_{jj}, \quad q = 1, \ldots, K-1, \tag{6}$$

with the convention that $\widehat{M}(0) = 0$.

In a simple random sample (SRS), Equation (11) shows that

$$\mathrm{Var}(\hat{b}_j) = \left(\sum_{k=1}^{K} x_j^2(k)\frac{p_k}{p_{0k}} - \beta_j^2\right)\Big/\tilde{n},$$

where $\tilde{n} = n/(1-n/N)$ is the finite population correction-adjusted sample size. Note that under an SRS, the design-based covariance matrix satisfies

$$\frac{\boldsymbol{V}}{n} \approx \frac{(1-n/N)\boldsymbol{P}}{n}.$$

Under the null hypothesis H_0, we have $\boldsymbol{P}_0^{-1}\boldsymbol{V} = \boldsymbol{P}_0^{-1}(1 - n/N)\boldsymbol{P}_0 = (1 - n/N)\boldsymbol{I}_K$. Therefore, the average eigenvalue $\delta.$ of the matrix $\boldsymbol{P}_0^{-1}\boldsymbol{V}$ under an SRS is approximately $1 - n/N$, implying that $\tilde{n} = n/\hat{\delta}.$.

Let $\hat{q}$ be the estimate of q_0 by maximizing Criterion (6). $\hat{q}$ is a natural test statistic, which rejects H_0 if $\hat{q} > 0$. However, as shown in 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. Following Eubank (1997), we propose a standardized test statistic W as follows:

$$W = \begin{cases} \dfrac{X_{\hat{q}}^2 - \hat{q}}{\sqrt{2\hat{q}}}, & \text{if } \hat{q} > 0, \\ 0, & \text{if } \hat{q} = 0, \end{cases} \tag{7}$$

where $X_{\hat{q}}^2 = \sum_{j=1}^{\hat{q}} \tilde{n}\hat{b}_j^2$, for some $q = 1, \ldots, K-1$ and $X_{\hat{q}}^2 = 0$ for $\hat{q} = 0$.

The distribution of W under H_0 denoted by W_0 can be 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.

2.3. 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 Eubank (1997) is modified by replacing sample size n with the effective size $\tilde{n}$.

$$\widehat{M}_\alpha(q) = \frac{\tilde{n}+1}{\tilde{n}-1}\sum_{j=1}^{q}\hat{b}_j^2 - \frac{a_\alpha}{\tilde{n}-1}\sum_{j=1}^{q}\hat{v}_{jj}, \quad q = 1, \ldots, K-1, \tag{8}$$

with the convention of $\widehat{M}_\alpha(0) = 0$ and $\hat{v}_{jj} = \sum_{k=1}^{K} x_j^2(k)\hat{p}_k/p_{0k}$. The proposed test statistic $\hat{q}_\alpha$ is the maximizer of Criterion (8).

Recall that the limiting probability of the Type I error for the proposed estimator $\hat{q}$ is 0.29. In the proposed test $\hat{q}_\alpha$, a_α in Equation (8) is used to control the Type I error at a specified level. According to Eubank and Hart (1992) and Eubank (1997), a_α is the solution of the equation

$$1 - \alpha = \exp\left\{-\sum_{j=1}^{\infty}\frac{P(\chi_j^2 > ja_\alpha)}{j}\right\} \tag{9}$$

or the solution of the following equation:

$$P\left(\max_{1\le k\le K-1}\left[\frac{1}{k}\sum_{j=1}^{k}Z_j^2\right] \ge a_\alpha\right) = \alpha, \tag{10}$$

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. Note that a large K approximation is needed for Equation (9). We observed that a_α converges to the desired value quickly when $K > 10$. An approximate level α test is conducted by rejecting H_0 if $\hat{q}_\alpha > 0$.

3. Properties of the proposed estimators

In Section 2, we proposed two Neyman smooth-type goodness-of-fit (GOF) tests that incorporate order selection for application in complex surveys. In this section, we first derive the distribution of the Fourier coefficients $\hat{b}_j$. We then investigate the asymptotic properties of the proposed estimators.

3.1. Limiting distribution of the Fourier coefficients $\hat{b}_j$'s

Theorem 3.1 (Asymptotic Normality of Fourier Coefficients): *Assume the following regularity conditions hold.*

(C1) ***Superpopulation Framework**: Let $\{\mathcal{U}_t\}_{t=1}^{\infty}$ be a sequence of nested finite populations where $\mathcal{U}_t \subset \mathcal{U}_{t+1}$, following Isaki and Fuller (1982). The population size $N_t \to \infty$ and sample size $n_t \to \infty$ such that $n_t/N_t \to \psi \in (0,1)$ as $t \to \infty$. This ensures that the superpopulation grows in a stable asymptotic framework.*

(C2) ***Moment Conditions**: For the binary indicator variable $y_j(k) \in \{0,1\}$ (1 if unit j is in category k, 0 otherwise), there exists $\delta > 0$ such that*

$$\sup_{t \geq 1} \frac{1}{N_t} \sum_{j \in \mathcal{U}_t} |y_j(k) - p_{0k}|^{2+\delta} < \infty \quad \text{for all } k = 1, \ldots, K,$$

where $p_{0k} = \Pr(y_j(k) = 1)$ under the null hypothesis. For binary $y_j(k)$, this reduces to bounding the Bernoulli moments:

$$p_{0k}(1 - p_{0k})^{2+\delta} + (1 - p_{0k}) p_{0k}^{2+\delta} < \infty,$$

which holds if p_{0k} is bounded away from 0 and 1.

(C3) ***Design Consistency**: The covariance estimator $\widehat{\boldsymbol{V}}_t$ satisfies*

$$\widehat{\boldsymbol{V}}_t - \boldsymbol{V} = o_P(1) \quad \text{as } t \to \infty,$$

where $\boldsymbol{V} = \lim_{t \to \infty} \mathrm{Var}(\sqrt{n_t}(\hat{\boldsymbol{p}}_t - \boldsymbol{p}))$.

Note that $\boldsymbol{V}$ can only be specified for certain designs such as SRS without replacement and stratified random sampling, etc. (Rao & Scott, 1981).

Then, under any sampling design satisfying (C1)–(C3), the Fourier coefficient estimator $\hat{\boldsymbol{b}}_t$ satisfies

$$\sqrt{\tilde{n}_t}(\hat{\boldsymbol{b}}_t - \beta) \xrightarrow{d} N_{K-1}(0, \boldsymbol{V}) \quad \text{as } t \to \infty,$$

where $\tilde{n}_t = n_t/\delta$ is the effective sample size adjusted for design effects.

Proof: We first prove the theorem under simple random sampling (SRS) without replacement. Then we extend the proof to the complex survey design case. For clarity, we omit the index t throughout the proof.

Let $\boldsymbol{p}^* = (\boldsymbol{p}^\top, p_K)^\top, \boldsymbol{p}_0^* = (\boldsymbol{p}_0^\top, p_{0K})^\top, \hat{\boldsymbol{p}}^* = (\hat{\boldsymbol{p}}^\top, \hat{p}_K)^\top$ and $\boldsymbol{P}^* = D(\boldsymbol{p}^*) - \boldsymbol{p}^*(\boldsymbol{p}^*)^\top$. In an SRS without replacement,

$$\mathrm{Var}(\hat{\boldsymbol{p}}^*) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{N}{N-1}\boldsymbol{P}^* \approx \frac{1}{n}\left(1 - \frac{n}{N}\right)\boldsymbol{P}^* = \frac{1}{\tilde{n}}\boldsymbol{P}^*,$$
$$\sqrt{\tilde{n}}(\hat{\boldsymbol{p}}^* - \boldsymbol{p}^*) \xrightarrow{d} N\left(0, D(\boldsymbol{p}^*) - \boldsymbol{p}^*(\boldsymbol{p}^*)^\top\right), \quad \tilde{n} \to \infty,$$

and

$$\sqrt{\tilde{n}}(\widehat{\boldsymbol{F}} - \boldsymbol{F}) \xrightarrow{d} N\left(0, D\left(\frac{\boldsymbol{p}^*}{\boldsymbol{p}_0^*}\right) - \left(\frac{\boldsymbol{p}^*}{\sqrt{\boldsymbol{p}_0^*}}\right)\left(\frac{\boldsymbol{p}^*}{\sqrt{\boldsymbol{p}_0^*}}\right)^\top\right).$$

Therefore,

$$\begin{aligned}\sqrt{\tilde{n}}(\hat{\boldsymbol{b}} - \boldsymbol{\beta}) &= \sqrt{\tilde{n}}\left(\boldsymbol{X}_{[K]}^\top \widehat{\boldsymbol{F}} - \boldsymbol{X}_{[K]}^\top \boldsymbol{F}\right) \\ &\xrightarrow{d} N\left(0, \boldsymbol{X}_{[K]}^\top D\left(\frac{\boldsymbol{p}^*}{\boldsymbol{p}_0^*}\right)\boldsymbol{X}_{[K]} - \boldsymbol{X}_{[K]}^\top\left(\frac{\boldsymbol{p}^*}{\sqrt{\boldsymbol{p}_0^*}}\right)\left(\frac{\boldsymbol{p}^*}{\sqrt{\boldsymbol{p}_0^*}}\right)^\top \boldsymbol{X}_{[K]}\right) \\ &= N\left(0, \boldsymbol{X}_{[K]}^\top D\left(\frac{\boldsymbol{p}^*}{\boldsymbol{p}_0^*}\right)\boldsymbol{X}_{[K]} - \boldsymbol{\beta}\boldsymbol{\beta}^\top\right). \end{aligned} \tag{11}$$

If the design is complex, since the regularity conditions of CLT for means are assumed, let $\mathrm{Var}(\hat{\boldsymbol{p}}^*) = \boldsymbol{V}^*/\tilde{n}$, and

$$\sqrt{\tilde{n}}(\hat{\boldsymbol{p}}^* - \boldsymbol{p}^*) \xrightarrow{d} N\left(0, \boldsymbol{V}^*\right), \quad \tilde{n} \to \infty.$$

Therefore,

$$\sqrt{\tilde{n}}(\hat{\boldsymbol{b}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(0, \boldsymbol{X}_{[K]}^\top D\left(\frac{1}{\sqrt{\boldsymbol{p}_0^*}}\right)\boldsymbol{V}^* D\left(\frac{1}{\sqrt{\boldsymbol{p}_0^*}}\right)\boldsymbol{X}_{[K]}\right) \tag{12}$$

as $\tilde{n} \to \infty$. 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) give detailed formula for $\boldsymbol{V}^*$. ■

3.2. Asymptotic properties of $\hat{q}$

In this section, we state the asymptotic properties of $\hat{q}$ which is the maximizer of Criterion (6) and is used in the proposed test W in Section 2.2.

Theorem 3.2: *Following Eubank (1999, p. 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 $c_0 = d_0 = 1$, *and* $\sum_r^*$ *denotes the sum extending over all r-tuples of integers* $(N_1, \ldots, N_r)$, *such that* $N_1 + 2N_2 + \cdots + rN_r = r$. *Under the null hypothesis* (1),

$$\lim_{\tilde{n}\to\infty} \Pr(\hat{q} = q) = c_q d_{K-1-q}, \quad \text{for } q = 0, \ldots, K-1.$$

In addition, under the alternative hypothesis H_α,

$$\lim_{\tilde{n}\to\infty} \Pr\left(\hat{q} < q_0\right) = 0$$

and

$$\lim_{\tilde{n}\to\infty} \Pr\left(\hat{q} = q_0 + r\right) = \Pr\left(r^* = r\right), \quad r = 0, \ldots, K - q_0 - 1,$$

where r^* *is the maximizer of the criterion,*

$$R(r) = \sum_{j=1}^{r} \nu_{(j+q_0)(j+q_0)}(Z_j^2 - 2), \quad \text{for } r = 1, \ldots, K - q_0 - 1, \tag{13}$$

with $R(0) = 0$, *and* $(Z_1, \ldots, Z_{K-q_0-1})^\top$ *is a vector of normal random variables with mean* $\mathbf{0}$ *and correlation* $\mathrm{Corr}(Z_i, Z_j) = \nu_{(i+q_0)(j+q_0)} / \sqrt{\nu_{(i+q_0)(i+q_0)}\nu_{(j+q_0)(j+q_0)}}$.

***Proof*:** The proof of Theorem 3.2 follows from the proof of Theorem 1 in Eubank (1997). When constructing the test statistics, we assume $V(\hat{\boldsymbol{p}}) = \delta \cdot V_{\mathrm{SRS}}(\hat{\boldsymbol{p}})$. The maximizing criterion (6) is obtained by replacing the sample size n with the effective sample size $\tilde{n} = n/\delta$.. These steps allow for a straightforward extension of the results to complex survey designs. For readers interested in the technical details, we refer them to Section 3.2.4 of Zhou's dissertation (Zhou, 2016). ■

There are several useful conclusions from Theorem 3.2. First, the limiting probability that $\hat{q}$ is under-selected (choose an order that is less than q_0) goes to 0, under both null and alternative hypotheses. Second, the limiting probability that $\hat{q}$ is over-selected (choose an order that is greater than q_0) is not negligible under both null and alternative hypotheses. If the maximizing criterion with $a = 2$ is taken in Theorem 3.2, it is known that the limiting probability of the Type I error is 0.29 as $K \to \infty$ and $\tilde{n} \to \infty$. As a result of Theorem 3.2, the following corollary can be derived.

Corollary 3.3: *Under both null* H_0 *and alternative* H_α,

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

and r^* *is the maximizer of Criterion* (13), *in which the vector of normal random variables* $(Z_1, \ldots, Z_{K-q_0-1})^\top$ *is the same as that defined in Theorem 3.2. In addition, for any fixed finite constant C,* $\lim_{\tilde{n}\to\infty} \Pr(X_{\hat{q}}^2 \geq C \mid q_0 \neq 0) = 1$.

Recall Criterion (6), $\widehat{M}(q) = \frac{\tilde{n}+1}{\tilde{n}-1}\sum_{j=1}^{q}\hat{b}_j^2 - \frac{2}{\tilde{n}-1}\sum_{j=1}^{q}\hat{v}_{jj}$. The limiting probability of the Type I error for this case is about 0.29 for this case. Now let us consider another maximizing criterion:

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

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

Theorem 3.4: *If $a_{\tilde{n}} = o(\sqrt{\tilde{n}})$ and $a_{\tilde{n}} > 2\ln(\ln(\tilde{n}))$, we have $\hat{q}_{a_{\tilde{n}}} \xrightarrow{P} q_0$, for $q_0 \geq 0$.*

Proof: The proof of Theorem 3.4 follows from proof of Theorem 2 in Eubank (1997). When constructing the test statistics, we assume $V(\hat{\boldsymbol{p}}) = \delta \cdot V_{\text{SRS}}(\hat{\boldsymbol{p}})$. The maximizing criterion (8) is obtained by replacing the sample size n with the effective sample size $\tilde{n} = n/\delta$.. These steps allow for a straightforward extension of the results to complex survey designs. For readers interested in the technical details, we refer them to Section 3.2.5 of Zhou's dissertation (Zhou, 2016). ■

Theorem 3.4 indicates that the limiting probability of Type I error goes to 0 when the effective sample size $\tilde{n}$ is large enough and the penalty term $a_{\tilde{n}}$ grows with sample size $\tilde{n}$ at an appropriate rate.

3.3. Asymptotic properties of $\hat{q}_\alpha$

In this section, we examine properties of the direct test statistic $\hat{q}_\alpha$ (the maximizer of Criterion (8) in Section 2.3). For a pre-specified level of significance α, it is reasonable that there exists a value a_α such that $\Pr(\hat{q}_\alpha \neq 0 \mid q_0 = 0) \to \alpha$ as $K \to \infty$ and $\tilde{n} \to \infty$, where the estimator $\hat{q}_\alpha$ is determined by a_α. Theorem 3.5 gives the asymptotic behavior of $\hat{q}_\alpha$.

Theorem 3.5: *Let $\hat{q}_\alpha$ be the maximizer of Criterion* (8),

$$\widehat{M}_\alpha(q) = \frac{\tilde{n}+1}{\tilde{n}-1}\sum_{j=1}^{q}\hat{b}_j^2 - \frac{a_\alpha}{\tilde{n}-1}\sum_{j=1}^{q}\hat{v}_{jj}, \quad \text{for } q = 1,\ldots,K-1,$$

where $\widehat{M}_\alpha(0) = 0$, $\hat{v}_{jj} = \sum_{k=1}^{K} x_j^2(k)\hat{p}_k/p_{0k}$, for $j = 1,\ldots,K-1$, and a_α is the solution of Equations (9) *or* (10). *As $\tilde{n} \to \infty$, we have*

$$\Pr\left(\hat{q}_\alpha > 0 \mid q_0 = 0\right) \to \alpha \quad \text{and} \quad \Pr\left(\hat{q}_\alpha > 0 | q_0 \neq 0\right) \to 1.$$

Proof: The proof of Theorem 3.5 follows from proof of Theorem 3 in Eubank (1997). When constructing the test statistics, we assume $V(\hat{\boldsymbol{p}}) = \delta \cdot V_{\text{SRS}}(\hat{\boldsymbol{p}})$. The maximizing criterion (8) is obtained by replacing the sample size n with the effective sample size $\tilde{n} = n/\delta$.. These steps allow for a straightforward extension of the results to complex survey designs. For readers interested in the technical details, we refer them to Section 3.2.6 of Zhou's dissertation (Zhou, 2016). ■

4. Simulation studies

In this section, we present simulation studies to evaluate the performance of the proposed methods. Across all settings considered, the proposed tests effectively control the Type I error at the nominal level, while also demonstrating higher empirical power compared to the first- and second-order corrected tests.

4.1. Simulation set up

Consider data with $K = 10$ categories arising from a complex survey design. The hypothesis of interest in the simulation studies is

$$H_0 : p_1 = \cdots = p_{10} = 0.1. \tag{14}$$

The basis function is

$$x_j(k) = \sqrt{\frac{2}{K}} \cos\left(\frac{j\pi\,(k-0.5)}{K}\right), \quad k = 1, \ldots, K \text{and } j = 1, \ldots, K-1. \tag{15}$$

The simulation studies are conducted under the following settings: (a) Level of significance is set at $\alpha = 0.05$; (b) The basis function defined in Equation (15) is used; (c) Each simulated sample consists of 50 clusters (also referred to as primary sampling units, or PSUs), with 15 individuals (secondary sampling units, or SSUs) sampled from each cluster, yielding a total of 750 observations per sample; (d) Intraclass Correlation Coefficients (ICCs) are set to 0.1, 0.3, and 0.6 to represent low, medium, and high levels of within-cluster dependence, respectively, following the discussion in Lohr (2021, pp. 174–176). When ICC $= 0$, all observations are uncorrelated, corresponding to a simple random sample (SRS). When ICC $= 1$, SSUs within the same cluster are perfectly correlated, meaning individuals within a cluster providing identical responses to the variable of interest; (e) Following Eubank (1997), three alternative hypotheses to Equation (14) are considered. They are

$$p(k) = \frac{1}{10} + \beta(k - 5.5)/10, \quad \text{for } k = 1, \ldots, 10, \tag{16}$$

$$p(k) = \frac{1}{10} + \beta \cos\left(\frac{j\pi\,(k-0.5)}{10}\right), \quad \text{for } k = 1, \ldots, 10, \tag{17}$$

and

$$p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right], \quad \text{for } k = 1, \ldots, 10, \tag{18}$$

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

We present simulation results for Alternative (16) along with a summary of all simulation scenarios. For Alternatives (16) and (17), the parameter β controls the degree of departure from the null model. When $\beta = 0$, the null hypothesis is recovered. Since the effect of β is symmetric about zero, we report only the results for $\beta > 0$. For Alternative (18), the null model corresponds to $\beta = 1$. Values of $\beta > 1$ that are larger (smaller) than 1 produce more (less) probability masses for the centrally numbered categories.

4.2. Simulation steps

To generate a sample of clustered multinomial responses, we follow steps from Skinner and Rao (1996). For a given vector of probabilities, within each PSU, we randomly permute the categories, so that the probabilities of categories are different each time. The cumulative probabilities of the first nine permuted categories are calculated, and the quantiles of these cumulative probabilities are found under standard normal distribution. A clustered standard normal random variable is created by $\mu + N(0, 1 - \text{ICC})$, where μ is generated from $N(0, \text{ICC})$. 15 generated values are then compared with the previously calculated 9 quantiles. For example, if the value is less than the first quantile, this SSU is categorized to the first permuted category. If a value is less than the fifth quantile but greater than the fourth quantile, then the corresponding SSU is grouped to the fifth permuted category.

For the proposed test W, $\hat{q}$ is obtained via maximizing Equation (6). To estimate $\delta_\cdot$, 100,000 multinomial data from a complex design under the null hypothesis (14) are generated. The covariance matrix $\boldsymbol{V}$ is estimated by these 100,000 samples. $\hat{\delta}_\cdot$ is obtained by averaging the eigenvalues of the matrix $\boldsymbol{P}_0^{-1}\widehat{\boldsymbol{V}}$. Test statistic is calculated by (7). The empirical distribution of W_0 (W under the null hypothesis (14)) is obtained through the 100,000 iterations and is used to find the critical value at certain level of significance.

The proposed test $\hat{q}_\alpha$ is the maximizer of Equation (8). Instead of calculating $a_{0.05}$ for each estimated $\hat{\delta}_\cdot$, we approximate $a_{0.05}$ by a product of $\hat{\delta}_\cdot$ and the values of $a_{0.05}$ estimated under an SRS, which is 4.18 (Eubank & Hart, 1992). By this method, we only need to estimate $\delta_\cdot$ for each setting, which is faster than estimating the corresponding $a_{0.05}$ for each simulated sample. Note that the solution of Equation (9) requires $K > 10$ approximations.

We compare Type I error and statistical power of the proposed tests with those from Pearson's chi-squared test, the first-order corrected test (FC) and second-order corrected test (SC) (Rao & Scott, 1981, 1984). Three alternatives (16), (17) and (18) are used to illustrate possible influence on Type I error and statistical power from different data patterns.

The FC statistic compares the observed value of $X^2/\hat{\delta}_\cdot$ to $\chi^2_{K-1}(\alpha)$. The SC statistic is $X^2_S = X^2/[\hat{\delta}_\cdot(1 + \hat{a}^2)]$, where

$$\hat{a}^2 = \sum_{i=1}^{K-1} \hat{\delta}_i^2/[(K-1)\hat{\delta}_\cdot^{\,2}] - 1. \tag{19}$$

This statistic is approximately a chi-squared random variable on $\nu = (K-1)(1+\hat{a}^2)$ degrees of freedom. If the design effects of the categories are all similar, the FC and SC behave similarly. Otherwise, the SC statistics approximate the distribution of X^2 better. Below are the simulation steps for an arbitrary alternative.

(1) Generate 100,000 samples of clustered multinomial under the null hypothesis (14). For each generated sample, the estimated proportion $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{K-1})$ is calculated. The estimated mean of $\hat{\boldsymbol{p}}$, denoted as $\bar{\boldsymbol{p}}$, is calculated by averaging the 100,000 $\hat{\boldsymbol{p}}$'s. The estimated covariance matrix $\widehat{\boldsymbol{V}}/\sqrt{n}$ is obtained by $(\hat{\boldsymbol{p}} - \bar{\boldsymbol{p}})(\hat{\boldsymbol{p}} - \bar{\boldsymbol{p}})^\top/(100{,}000 - 1)$. The eigenvalues of the matrix $\boldsymbol{P}_0^{-1}\widehat{\boldsymbol{V}}$ are calculated using the eigen() function in R. $\hat{a}^2$ is calculated by (19).
(2) Under the null hypothesis (14), we search for $\hat{q}$, the maximizer of Criterion (6), and obtain a value of W_0 by (7). 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, and the FC and SC test statistics are calculated. Next, by searching all $q = 1, \ldots, K-1$, $\hat{q}$ is obtained by maximizing Criterion (6). W is calculated by Equation (7). We then search $\hat{q}_{0.05}$ among $q = 1, \ldots, K-1$ to maximize Equation (8), where $a_{0.05} = 4.18 * \hat{\delta}$.
(4) We compare the test statistics in Step 3 with their corresponding rejection criteria. The Pearson's chi-square test statistic, and the FC and SC 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 it is 0.
(5) Steps 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.

4.3. Simulation results

Type I error rates and empirical power comparisons for the different tests are reported under Alternative (16), which generates slowly varying probabilities across categories. For values of β ranging from 0 to 0.14 in increments of 0.01, a total of 15 sets of probability vectors (including the null model at $\beta = 0$) are generated according to Equation (16). When $\beta = 0.01$, the category probabilities are nearly uniform, and as β increases, the differences across categories vary slowly to moderately.

Figure 1 displays the empirical power curves of five tests under Alternative (16), with varying levels of intraclass correlation (ICC $= 0.1, 0.3$ and 0.6). The tests compared include the proposed $\hat{q}_{0.05}$ and W, Pearson's chi-squared goodness-of-fit test, and the first-order (FC) and second-order (SC) corrected tests. Power is plotted as a function of the departure parameter β.

We begin by examining how the tests control the Type I error rate, which corresponds to evaluating their power under Alternative (16) when $\beta = 0$. When ICC $= 0.1$, Pearson's chi-squared test maintains the Type I error rate close to the nominal level of 0.05. However, as the ICC increases, the Type I error rate becomes substantially inflated: it rises to approximately 0.18 when ICC $= 0.3$, and further to about 0.76 when ICC $= 0.6$. This demonstrates that the performance of Pearson's test deteriorates severely with increasing intra-cluster correlation.

In contrast, the other four tests – the proposed $\hat{q}_{0.05}$ and W statistics, as well as the FC and SC tests – maintain Type I error control close to the nominal level across all ICC settings. These results clearly indicate that Pearson's chi-squared test is not appropriate for multinomial data arising from complex survey designs and should not be used without correction.

We now turn to the empirical power comparisons of the five tests. For ICC $= 0.1$, 0.3, and 0.6, both the W and $\hat{q}_{0.05}$ tests exhibit higher power than the FC and SC tests when the underlying probabilities vary slowly (i.e., $\beta \leq 0.07$). When the probabilities vary more substantially ($\beta > 0.07$), the proposed tests perform comparably to the FC and SC tests in terms of empirical power. Among all methods, the $\hat{q}_{0.05}$ test consistently demonstrates the highest empirical power, followed by the W test.

In summary, under all three alternatives, the proposed tests as well as the first- and second-order corrected tests successfully control the Type I error at the pre-specified significance level across all simulation settings. Additionally, the proposed tests offer substantial gains in empirical power compared to the FC and SC tests when the underlying probabilities differ

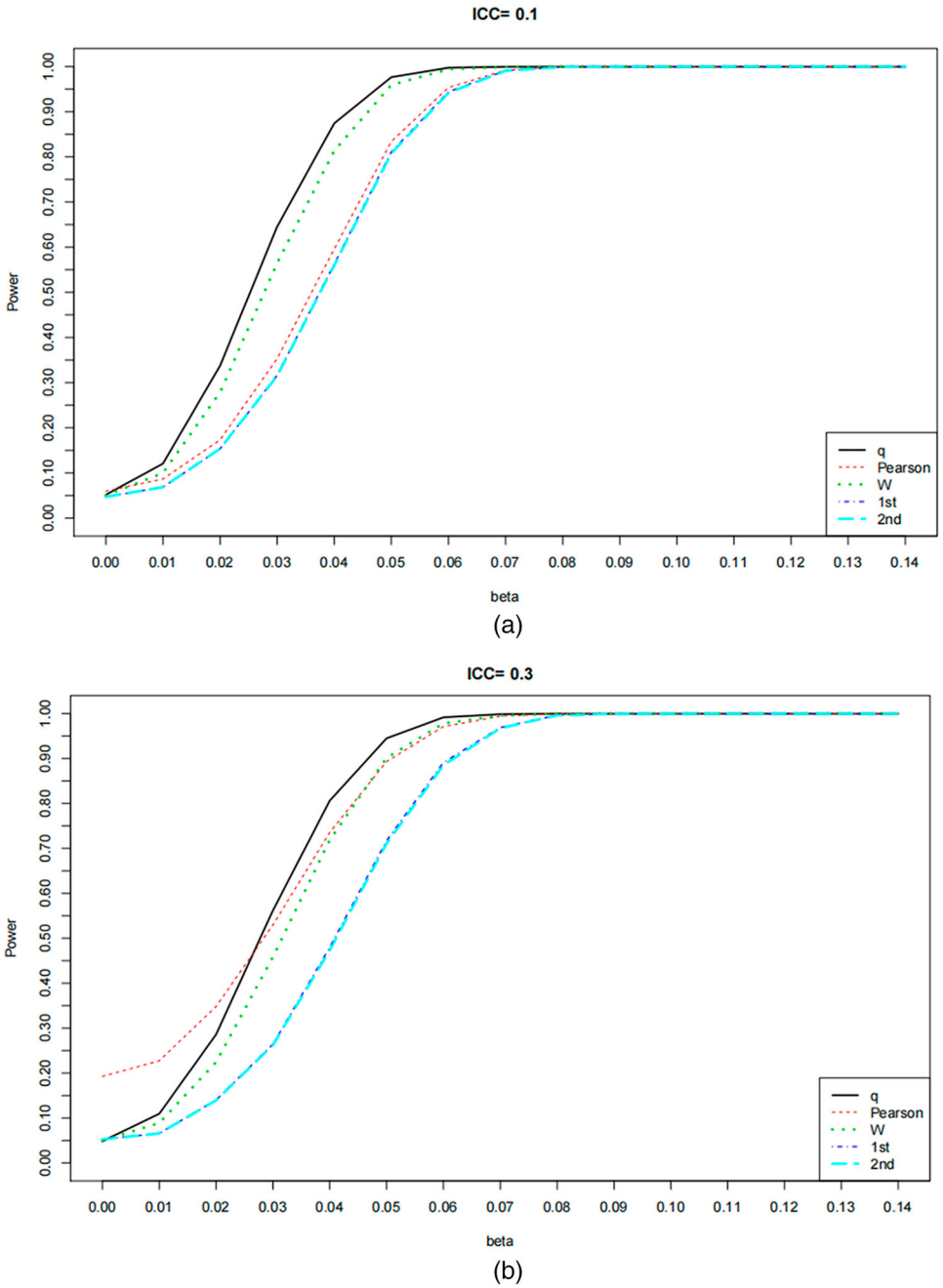

Figure 1. The power curves of selected methods for simulated complex survey data under Alternative (16).

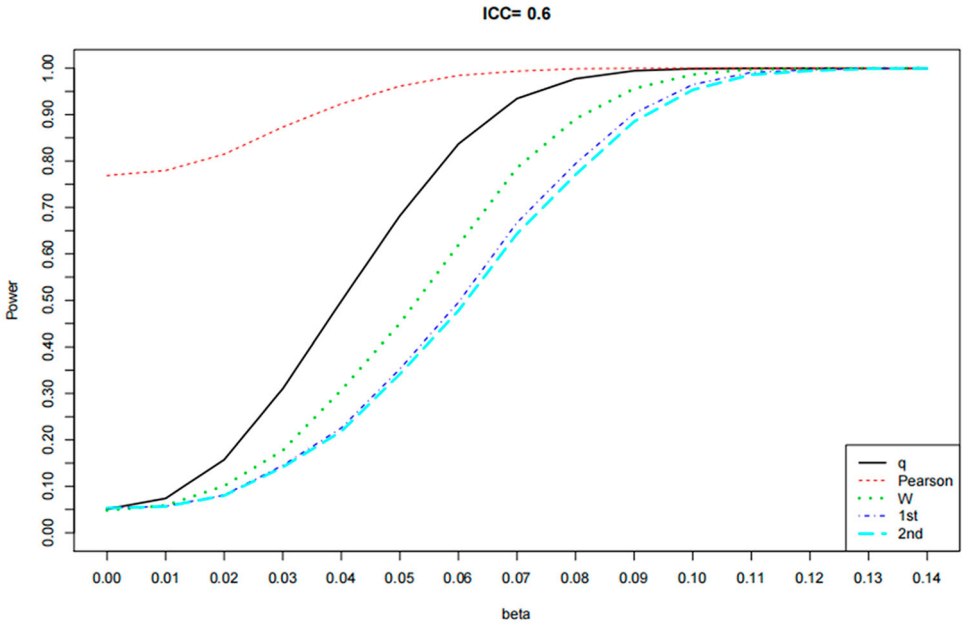

Figure 1. Continued.

only slightly from each other. Specifically, the $\hat{q}_\alpha$ test is particularly sensitive to small deviations in category probabilities, while the W test exhibits greater stability across a range of scenarios.

5. Application

In this section, we apply the proposed Neyman smooth-type GOF tests for complex surveys to a real-world dataset. For comparison purposes, we also report the results of classical GOF tests, including Pearson's chi-squared test and the first- and second-order corrected tests (Rao & Scott, 1981, 1984). The data are drawn from the National Youth Tobacco Survey (NYTS). Our objective is to test for differences in tobacco use severity across groups, specifically among Asian and American Indian/Alaska Native students.

5.1. Data description

NYTS provides data to support research on tobacco use among middle and high school students. The survey covers a wide range of tobacco products, including cigarettes, cigars, hookahs, electronic cigarettes, and others. NYTS began in 1999 and was subsequently conducted in 2000, 2002, 2004, 2006, 2011, 2012, 2013, 2014, and continues to be administered in more recent years. Since 2011, the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA) have jointly managed NYTS.

The 2014 NYTS employed a stratified three-stage clustered sampling design (Office on Smoking and Health, 2014). Sixteen strata were defined based on urbanicity and the proportion of minority populations (non-Hispanic Black and Hispanic) in different regions of the

Table 1. Smoking severity distribution among Asian students (NYTS 2014).

Number Group	< 1	1	2–5	6–10	≥ 11	Total
Counts	7	6	8	2	2	25
Weighted Counts	6840.3	5818.4	6595.9	1391.9	1907.7	22,554.2
$\hat{p}(k)$	0.303	0.258	0.292	0.062	0.085	1

U.S. A PSU was defined as a county, a group of smaller counties, or part of a large county. More details about PSU construction can be found in Office on Smoking and Health (2014, p. 7). Within each PSU, middle and high schools served as SSUs. In each selected school, one or two classes were sampled per grade, and all students in the selected classes were eligible to participate. Sampling was conducted without replacement.

The survey included approximately 81 questions, which students completed using paper-and-pencil instruments. After data collection, the responses were cleaned, and individual sampling weights were computed to account for the complex sampling design and nonresponse adjustments. A detailed description of the weighting procedure is provided in Chapter 4 of Office on Smoking and Health (2014). The 2014 NYTS dataset contains 157 variables (including the weight variable) and a total of 22,007 observations.

5.2. Severity differences among Asian students smokers

We focus on Asian students who reported smoking during the past 30 days. From 973 surveyed Asian students, 25 reported tobacco use. Smoking severity was categorized into five levels based on daily cigarette consumption:

- **Light smokers**: $<$ 1 cigarette/day;
- **Moderately light**: 1 cigarette/day;
- **Medium**: 2 to 5 cigarettes/day;
- **Moderately heavy**: 6 to 10 cigarettes/day;
- **Heavy**: $\geq$ 11 cigarettes/day.

The proportions are calculated using survey weights, with a summary of the data provided in Table 1.

The null hypothesis tests homogeneity across categories:

$$H_0 : p_1 = \cdots = p_5 = \frac{1}{5}. \tag{20}$$

We estimate the average design effect as $\hat{\delta}_\cdot = 1.21677$ with $\hat{a} \approx 0$. The first-order (FC) and second-order (SC) corrected test statistics are calculated as

$$X_C^2 = \frac{X^2}{\hat{\delta}_\cdot} = \sum_{k=1}^{5} n \frac{(\hat{p}_k - p_{0k})^2}{p_{0k}} / \hat{\delta}_\cdot = 5.62,$$

$$X_S^2 = \frac{X_C^2}{1 + \hat{a}^2} \approx 5.62 \quad (\text{since } \hat{a} \approx 0).$$

Using a χ^2 reference distribution with 4 degrees of freedom (5 categories – 1), both tests yield $p = 0.23$, failing to reject the null hypothesis at $\alpha = 0.05$ significance level.

Next, we use the proposed methods to test the hypothesis. Following Simulation Step 2, we can simulate the empirical distribution of W_0 and find the 95% quantile of W_0. By searching all $q = 1, \ldots, 5-1$, $\hat{q}$ is the one that maximizes Equation (6). We found that $\hat{q} = 1$, $W = 2.99$, and p-value $= 0.039$. For the proposed test $\hat{q}_\alpha$, we found $\hat{q}_{0.05} = 1$ and p-value $= 0.033$. Both W and $\hat{q}_\alpha$ tests reject hypothesis (20) at level of 0.05.

5.2.1. Practical implications

While traditional methods did not detect significant disparities under their modelling assumptions ($p = 0.23$), our proposed approach identified significant distributional differences among smoking levels ($p = 0.039$) across Asian students (Grades 6–12) in the U.S. This suggests that our method has improved sensitivity to subtle yet systematic differences across groups. The practical implications are as follows.

- Prevention Opportunities: The observed 30.3% prevalence of light smoking indicates a window for early intervention. Schools may consider implementing peer counselling or awareness programs specifically targeting light and moderate smokers, accompanied by rigorous impact evaluation.
- Targeted Support Needs: Although the proportion of heavy smokers is relatively low (8.5%), this group may require access to tailored cessation services. These interventions should be culturally adapted and responsive to community-specific barriers commonly reported within Asian populations.
- Methodological Implications: Simulation results show that our method offers higher statistical power while maintaining appropriate Type I error rates, especially in detecting subtle distributional differences missed by traditional FC/SC tests.

6. Conclusion and future research

In this study, we proposed two Neyman smooth-type goodness-of-fit (GOF) tests, $\hat{q}_\alpha$ and W, for analysing multinomial data under complex survey designs. These tests maintain nominal significance levels ($\alpha = 0.05$) across diverse designs, including stratified and clustered sampling with high intra-class correlations (ICC). They also show improved statistical powers, when comparing with some existing methods, particularly in scenarios with small sample sizes or subtle distributional deviations. The proposed test W offers robust performance across both subtle and pronounced deviations. Applied to the National Youth Tobacco Survey (NYTS), the proposed tests detected significant disparities in smoking severity among Asian students ($p < 0.05$) that traditional methods missed ($p = 0.23$).

The success of this framework suggests promising extensions. First, adapting the methodology to stratified designs could involve stratum-specific order selection or weighted criteria (e.g., $\widehat{M}(q) = \sum_h w_h \widehat{M}_h(q)$) to unify order estimation across strata. Second, the principles of Fourier transformation and dimension reduction could generalize to other inferential tasks, such as testing loglinear models or row-column independence in contingency tables. Finally, developing open-source software (e.g., R/Python packages) would facilitate adoption by practitioners.

By bridging finite population asymptotics with Neyman's smooth-test framework, this work advances categorical data analysis in complex surveys. Future research should explore integrations with machine learning for automated dimension reduction and applications to high-dimensional sparse tables–areas where traditional χ^2-based methods often falter.

Acknowledgments

The authors sincerely thank the referees for their careful reading of the manuscript, as well as for their valuable comments and constructive suggestions, which have significantly improved the quality of the manuscript. Special thanks go to Shuya Cai for her meticulous proofreading, which has greatly enhanced the formal clarity and overall presentation of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Arfken, G. (1985). *Mathematical Methods for Physicists* (3rd ed.). Academic Press.

Bedrick, E. J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika*, *70*(3), 591–595. https://doi.org/10.1093/biomet/70.3.591

Eubank, R. L. (1997). Testing goodness of fit with multinomial data. *Journal of the American Statistical Association*, *92*(439), 1084–1093. https://doi.org/10.1080/01621459.1997.10474064

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). CRC Press.

Eubank, R. L., & Hart, J. D. (1992). Testing goodness of fit in regression via order selection criteria. *The Annals of Statistics*, *20*(3), 1412–1425. https://doi.org/10.1214/aos/1176348775

Fay, R. E. (1979). On adjusting the Pearson chi-square statistic for clustered sampling. In *Proceedings of the American Statistical Association, 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*(389), 370–375.

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*(377), 89–96. https://doi.org/10.1080/01621459.1982.10477770

Jamil, H., Moustaki, I., & Skinner, C. J. (2025). Pairwise likelihood estimation and limited information goodness-of fit test statistics for binary factor analysis models under complex survey sampling. *British Journal of Mathematical and Statistical Psychology*, *78*(1), 258–285. https://doi.org/10.1111/bmsp.v78.1

Kim, J., Rao, J. N. K., & Wang, Z. (2019). *Hypotheses Testing from Complex Survey Data Using Bootstrap Weights: A Unified Approach* (Technical Paper No. 265). Iowa State University.

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

Lancaster, H. O. (n.d.). *The Chi-squared Distribution*. Wiley.

Lindsay, B. G. (1988). Composite likelihood methods. *Statistical Inference from Stochastic Processes*, *80*, 221–239. https://doi.org/10.1090/conm/080

Lohr, S. L. (2021). *Sampling: Design and Analysis* (3rd ed.). Chapman and Hall/CRC.

Lu, Y. (2014). Chi-squared tests in dual frame surveys. *Survey Methodology*, *40*(2), 323–334.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. https://doi.org/10.2307/271070

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, *20*(3–4), 149–199.

Office on Smoking and Health (2014). *2014 National Youth Tobacco Survey: Methodology Report* (Technical Paper). 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.

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*(374), 221–230. https://doi.org/10.1080/01621459.1981.10477633

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*(1), 46–60. https://doi.org/10.1214/aos/1176346391

Rao, J. N. K., & Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, *15*(1), 385–397. https://doi.org/10.1214/aos/1176350273

Rayner, J. C. W., & Best, D. J. (1986). Neyman-type smooth tests for location-scale families. *Biometrika*, *73*(2), 437–446. https://doi.org/10.1093/biomet/73.2.437

Rayner, J. C. W., & Best, D. J. (1989). *Smooth Tests of Goodness of Fit*. Oxford University Press.

Rayner, J. C. W., & Best, D. J. (1990). Smooth tests of goodness of fit: An overview. *International Statistical Review*, *58*(1), 9–17. https://doi.org/10.2307/1403470

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*, *39*(1), 35–50. https://doi.org/10.1111/stan.1985.39.issue-1

Rayner, J. C. W., Thas, O., & Best, D. J. (2009). *Smooth Tests of Goodness of Fit* (2nd ed.). Wiley.

Sárndal, C.-E. (2003). *Model Assisted Survey Sampling*. Springer.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–75). Wiley.

Skinner, C. J. (2019). Analysis of categorical data for complex surveys. *International Statistical Review*, *87*(S1), 64–78. https://doi.org/10.1111/insr.v87.S1

Skinner, C. J., & Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, *91*(433), 349–356. https://doi.org/10.1080/01621459.1996.10476695

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, *92*(1), 1–28. https://doi.org/10.1007/s10182-008-0060-7

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. https://doi.org/10.1090/tran/1943-054-03

Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, *87*(418), 732–737. https://doi.org/10.1080/01621459.1992.10475275

Zhou, L. (2016). *Neyman Smooth-type Goodness of Fit Tests in Complex Surveys* [Unpublished doctoral dissertation]. University of New Mexico.