University of New Mexico
UNM Digital Repository

Mathematics & Statistics ETDs | Electronic Theses and Dissertations

6-9-2016

Neyman Smooth-Type Goodness of Fit Tests in Complex Surveys

Lang Zhou

Follow this and additional works at: https://digitalrepository.unm.edu/math_etds

Recommended Citation
Zhou, Lang. "Neyman Smooth-Type Goodness of Fit Tests in Complex Surveys." (2016).
https://digitalrepository.unm.edu/math_etds/57

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Mathematics & Statistics ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.

Lang Zhou

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Yan Lu, Chair

Guoyi Zhang, Co-Chair

Gabriel Huerta

Helen Wearing

Ronald Christensen

Neyman Smooth-Type Goodness of Fit Tests in Complex Surveys

by

Lang Zhou

B.S., Xiamen University, 2006
M.S., University of New Mexico, 2013

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2016

Dedication

To my wife, Ling Xu.

To my parents, Zhengzhong Zhou and Zhanying He.

Acknowledgments

I would like to thank my advisor, Professor Yan Lu, for her support and excellent guidance. I also would like to thank my co-advisor, Professor Guoyi Zhang, for his advice and help.

I particularly would like to thank Professor Ronald Christensen for his insightful comments.

I wish to thank Professor Gabriel Huerta and Professor Helen Wearing for their helpful comments.

Neyman Smooth-Type Goodness of Fit Tests in Complex Surveys

by

Lang Zhou

B.S., Xiamen University, 2006
M.S., University of New Mexico, 2013

Ph.D., Statistics, University of New Mexico, 2016

Abstract

In our study, we have extended the Neyman smooth-type goodness of fit tests by Eubank (1997) from simple random sample to complex surveys (Methodologies have been provided for complex surveys, and theorems have been provided only for stratified random samples.) by incorporating consistent estimators under the survey design, which is accomplished by a data-driven nonparametric order selection method. Simulation results show that these proposed methods potentially improve the statistical power while controlling the type I error very well compared to those commonly used existing test procedures, especially for the cases with slow-varying probabilities. We also derived the large sample properties of the test statistics in stratified sampling. Several practical examples are provided to illustrate the usage and advantages of our proposed methods.

KEY WORDS: Goodness of fit, Neyman smooth, Order selection, The first order and second order corrected tests, Stratification, Clustering, Simple Random Sample.

Contents

List of Figures

List of Tables

Glossary

K	Total number of categories in multinomial distributions.
k	Index of the kth category in multinomial distributions, range of k from 1 to K.
n	Observed sample size.
$\tilde{n}$	Kish's effective sample size.
O_k	Observed counts in kth category of multinomial data.
E_k	Expected counts in kth category of multinomial data.
X^2	Test statistic of goodness of fit tests for multinomial data.
H_0	Null hypothesis of goodness of fit tests for multinomial data.
H_0^*	Null hypothesis of Neyman smooth-type goodness of fit tests for multinomial data.
H_1	Alternative hypothesis of goodness of fit tests for multinomial data.
H_1^*	Alternative hypothesis of Neyman smooth-type goodness of fit tests for multinomial data.
$p(k)$	Real proportion of kth category of multinomial data.

Glossary

$\tilde{p}(k)$	Estimated proportion of kth category of multinomial data from Simple Random Sample.
$\hat{p}(k)$	Estimated proportion of kth category of multinomial data from Complex Surveys.
$p_0(k)$	Hypothesized proportion of kth category of multinomial data.
	X^2_{SRS} Test statistic of goodness of fit tests for multinomial data from Simple Random Sample.
χ^2_k	Central chi-squared distribution with k degrees of freedom.
α	Level of significance, which is a real number between 0 and 1.
$\chi^2_k(1-\alpha)$	The $1-\alpha$ quantile of central chi-squared distribution with k degrees of freedom.
$\mathbf{O}$	Vector of observed counts in multinomial data.
$\mathbf{E}$	Vector of expected counts in multinomial data.
$\mathbf{p}_0$	Vector of hypothesized proportions in multinomial data.
$\mathbf{p}$	Vector of real proportions in multinomial data.
$\tilde{\mathbf{p}}$	Vector of estimated proportions in multinomial data from Simple Random Sample.
$\hat{\mathbf{p}}$	Vector of estimated proportions in multinomial data.
$y_j(k)$	An indicator function with value 1 if jth observation is in kth category of multinomial data, 0 otherwise.
$\mathrm{E}(\cdot)$	Function of expected values.

$\mathrm{var}(\cdot)$	Function of variance.
$\mathrm{cov}(\cdot)$	Function of covariance.
$\mathbf{P}$	Product of covariance matrix of $\tilde{\mathbf{p}}$ and n.
$\mathbf{P}_0$	Product of covariance matrix of $\tilde{\mathbf{p}}$ and n under null hypothesis.
w_j	Weight of jth observation.
X_W^2	Test statistic of Wald test.
δ_i	The ith eigenvalue of $\mathbf{P}_0\mathbf{V}$ under null hypothesis.
$\delta_.$	The mean value of all δ_i's.
X_C^2	Test statistic of goodness of fit tests with the first order correction.
X_S^2	Test statistic of goodness of fit tests with the second order correction.
$\hat{a}^2$	A real number used in goodness of fit tests with the second order correction.
$f(k)$	A funtion of $p(k)$ and $p_0(k)$.
$\tilde{f}(k)$	Estimator of $f(k)$ under Simple Random Sample.
$\hat{f}(k)$	Estimator of $f(k)$ under Complex Surveys.
β_j	The jth Fourier coefficients of $f(k)$.
b_j	The jth Fourier coefficients of $\tilde{f}(k)$ or $\hat{f}(k)$, also the estimator of β_j.
$x_j(k)$	A basis function with respect to j and k.

δ_{ji}	An indicator function with value 1 if $j = i$ and 0 if $j \neq i$.
$\mathbf{V}$	Product of covariance matrix of $\hat{\mathbf{p}}$ and n, or product of covariance matrix of the vector of b_j's and n.
q	Order parameter, which is an integer between 0 and $K - 1$.
$M(q)$	A function for maximizing criterion with respect to q.
W	Test statistic in Neyman smooth-type goodness of fit tests.
W_0	Test statistic in Neyman smooth-type goodness of fit tests under null hypothesis.
$\hat{q}$	Estimator of q.
$\hat{q}_\alpha$	Estimator of q for level of significance α.
a_α	A positive number.
Z	Standard normal random variable.
v_{ij}	A function of $x_j(k)$, $x_j(k)$ and $p(k)$ for all $k = 1, \cdots, K$.
$P(\cdot)$	Probability function.
N_{K-1}	A $K - 1$ multivariate normal distribution.
c_r	A function used in our proposed test.
d_r	A function used in our proposed test.
$R(r)$	A function with respect to an integer r.
D	An expression used in the theorem proofs.
$O(\cdot)$	Indicator of decay rate to 0, which is the same as $\cdot$.

Glossary

$o(\cdot)$	Indicator of decay rate to 0, which is faster as $\cdot$.
A	A subset.
B	A subset.
S	A sequence of sets.
C	A finite constant.
W_r	A random variable dependent with r.
ICC	An statistic measurement indicating Intraclass Correlation Coefficient.
cos	Cosine function.
π	The number π.
Φ	The cumulative distribution function of standard normal random variable.
GOF	Goodness of Fit.
SRS	Simple Random Sampling.
psu	Primary Sampling Unit.
ssu	Secondary Sampling Unit.

Chapter 1

Introduction

In finite population sampling, probability sampling is used to select a sample from a directory or map of units called a sampling frame. The selected sample is used to make inferences about the finite population. Most large surveys such as the National Victimization Survey (NCVS) (Dodge & Lentzner, 1980) and the Current Population Survey (CPS) (U.S. Census Bureau, 2002) adopt complex sampling designs. For a simple example, we may consider each state as a stratum, and counties within the states as clusters or primary sampling units (psu). We then take a simple random sample (SRS) of the counties within each state respectively to create a stratified cluster sample. The complex designs take advantage of stratification and clustering, and are usually more efficient than a SRS under a fixed cost.

As Rao and Thomas (1988) noted, "the need to perform statistical analyses of categorical data is frequently encountered in quantitative sociological research". For example, a company wants to find out levels of satisfaction of one of their products. Survey questionnaires are sent out to the customers according to some certain sampling design. Levels of satisfaction may vary from "Very Dissatisfied", "Dissatisfied", "OK", "Satisfied" to "Very Satisfied". Categorical data analyses based on these survey data can be used to draw statistical conclusions, and therefore to provide guidance for business decisions and policy making. Pearson's chi-squared test (Pearson, 1900) and the likelihood ratio test are both well known for assessing model fit and testing hypotheses of interest for categorical data. Cressie and Read (1984) describes various goodness of fit (GOF) tests for multinomial data. However, these methods rely on the assumption of independence, and the data is obtained by SRS from a finite population, where the independence assumption is usually met. Most large scale surveys are complex with stratification and clustering, where the independence assumption is violated. In Lohr (2010, pg. 401, 407), examples 10.1 and 10.4 are given to illustrate the effects of clustering. A husband and wife in the same household are considered as a cluster. If we ignore the cluster effects and consider the husband and wife as two independent individuals who give exactly the same answers to the survey questions, the observed counts in each cell will be doubled and the resulting Pearson's test statistic will be twice the value of the test statistic if we consider a household as one unit. Consequently, we have a low p-value and are more likely to reject the null hypothesis than we should do.

GOF tests in complex surveys have been studied for a while. Wald's test (Wald, 1943) is one of the earliest methods proposed to assess model fit in complex surveys. Brier (1978) proposed a model for clustering, and studied general hypothesis on cell probabilities. Fay (1979, 1985) proposed a jackknifed chi-squared test for complex surveys. Both Wald's test and Fay's jackknifed test require detailed survey information from which the covariance matrix can be estimated. Such detailed information is often not available in practice. Rao and Scott (1981, 1984) proposed corrections to chi-squared tests for assessing GOF and testing independence in two-way and multi-way tables. Bedrick (1983) and Rao and Scott (1987) also studied the use of limited information on cell and marginal design effects to provide approximate tests. Thomas, Singh, and Roberts (1996) "described a Monte Carlo study of competing procedures for testing row-column independence in a two-way table under clustering". Lu and Lohr (2010) and Lu (2014) extended chi-squared tests to dual frame surveys.

One of the problems of the GOF tests in the literature is that they are usually not sensitive to the slow-varying probability problem. For example, we throw an unfair die repeatedly for 150 times, with numbers $1-6$ appearing for 23, 26, 26, 25, 23, 27 times respectively. With this outcome, it is difficult for those existing methods to detect the fact that the die is not fair, unless a larger sample size is attempted. Eubank (1997) introduced an interesting example: some abnormal outcomes were observed for the Pick 3 game from some Texas Lottery machines. The machine was taken off-line and subjected to 150 testing draws. The ten balls (numbered 0 to 9) didn't pop up randomly as expected, but with the higher numbered balls having lower chances of selection as depicted in Figure 1.1. Unfortunately, Pearson's chi-squared tests failed to identify this non-randomness problem at a regular level of significance. This required new methods of GOF tests for detecting the slow varying frequencies, which motivated one of Eubank's research projects.

The Neyman smooth-type test incorporated with order selection is one of the solutions to the slow varying problem. The Neyman smooth-type test was first proposed by Neyman (1937). Lancaster (1969) discussed the decomposition of the Pearson's chi-squared test statistic. Rayner, Best, and Dodds (1985) assessed the similarities and differences between Pearson's chi-squared test and the Neyman smooth test. Rayner and Best (1986) then extended the Neyman smooth-type tests for location-scale families. Further, Rayner and Best (1989, 1990) and Rayner, Thas, and Best (2009) provided a comprehensive overview of Neyman smooth-type GOF tests. Ledwina (1994) proposed a data-driven order selection method for Neyman smooth-type GOF test. Inglot, Jurlewicz, and Ledwina (1990) discuss the situation when the number of categories go to infinity. Fan and Huang (2001) proposed several GOF tests for "examining the adequacy of a family of parametric models against

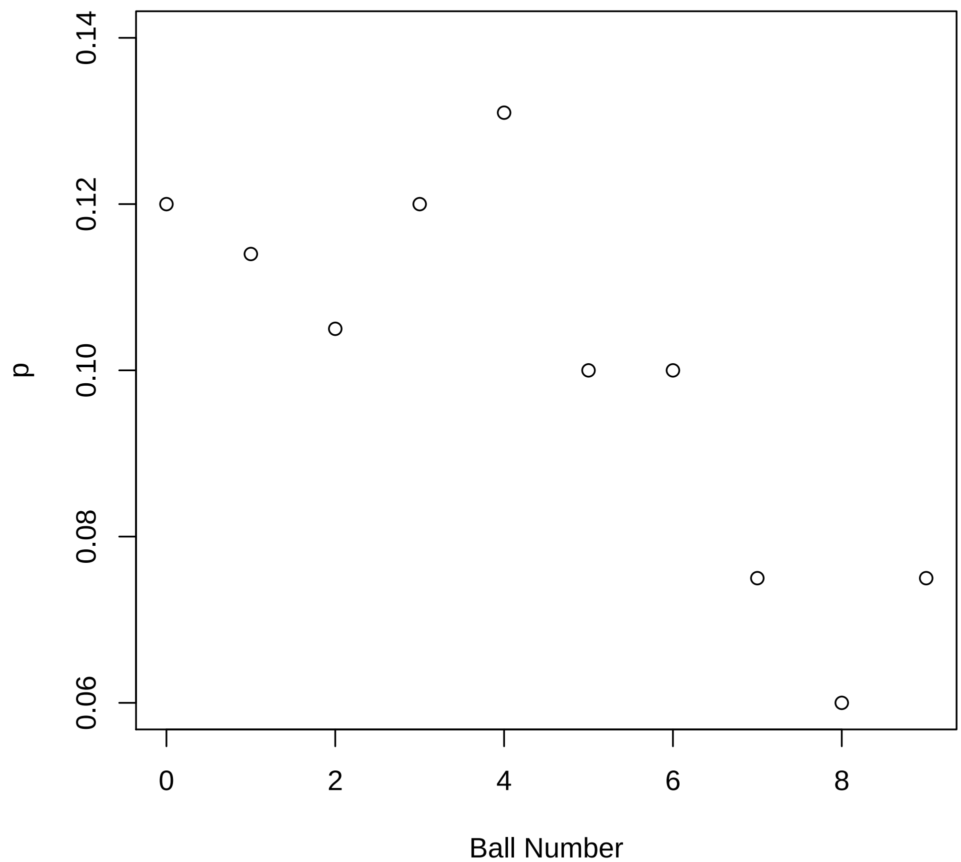

Figure 1.1: Estimated proportions of selected balls of the Pick 3 game.

large nonparametric alternatives".

Order selection can be incorporated with Neyman smooth-type GOF tests. The chi-squared test statistics can be decomposed into ordered components. Since the main information is usually contained in the first few components, not all of the components are necessary to construct the test statistic. Thus, order selection is needed to determine the optimal number of components. In this case, since the last few components only carry negligible information of the data, they can be omitted, which results in more degrees of freedom for the tests. Order selection is well discussed in Eubank and Hart (1992) and a comprehensive discussion can be found in Eubank (1999). Eubank (1997) introduced Neyman smooth-type GOF tests incorporated with order selection, which successfully detected the Pick 3 machine problem at level 0.1 by using the 150 draws. Eubank's methods showed more statistical power than existing chi-squared type tests, especially for the cases with slow varying probabilities.

Our research problem is inspired by Eubank (1997)'s work. In complex surveys, the existing GOF tests are not sensitive to slow varying probabilities either. For example, we are interested in the hypothesis that there is no difference in age groups of nonwhite families who support legalized abortion in Section 5.1 of Chapter 5. Both first order and second order corrected tests (Rao & Scott, 1981, 1984) fail to reject the null hypothesis, although we observed a decreasing trend of the rate of supporting legalized abortion from older age groups, as shown in Figure 5.1. We will extend Eubank's work to the field of complex surveys. One challenge of our research is that we no longer have the independence assumption as in SRS, because most of the survey data are correlated due to clustering. We need to incorporate the correlation and the survey weights into the estimators, order selection procedures and asymptotic properties derivation etc. Our proposed GOF tests have the advantage that they only require a small or moderate sample size for detecting the differences, which is essential in many practical examples when a large sample size is not feasible. For example, a new treatment recently developed is in phase 3 of clinical trials. We observe that the effects of the new treatment are not obvious compared to the existing one, but definitely show improvement. In order to determine statistical significant treatment effects, the budget may not be enough for recruiting more objects for the clinical trial. Another example is discussed in Section 5.2 of Chapter 5. If we are interested in research on some minorities, such as Asians, American Indians, or Alaska Natives, their sample sizes are usually limited. In fact, there are only 973

Asians and 338 American Indians/Alaska Natives in that data, whose total number of observations is $22,007$. The proposed tests work well for complex survey data with slow varying probabilities, and also work well with non-slow varying probability data. Our methods provide greater statistical power than existing GOF tests in complex surveys while controlling the Type I error at the pre-specified level.

This dissertation is organized as follows. In Chapter 2, we give a background review of some commonly used GOF tests for SRS and complex surveys. In Chapter 3, we develop the Neyman smooth-type GOF test incorporated with order selection for use in complex surveys. We also investigate the large sample properties of the proposed estimators for stratified random samples. In Chapter 4, we perform simulation studies to evaluate our methods and to compare our methods with some existing test procedures. In Chapter 5, we use some examples to illustrate our methods. Finally we give conclusions and further research work in Chapter 6.

Chapter 2

Background

Consider a random experiment of tossing a coin with possible outcomes heads or tails. Suppose that we want to investigate whether a coin is "fair" or not, i.e., to test if the probability of getting a head is 50%. The test statistics are usually some measurement of the "distance" between the observed counts and the expected counts under the hypothesis. If the distance is large enough, we reject the null hypothesis. Such tests are usually mentioned as goodness of fit tests (GOF tests).

In this chapter, we will review some classical GOF tests, such as Pearson's chi-squared test and the likelihood ratio test for use in independent data. We then review Neyman smooth-type GOF tests in SRS, with introduction of the Fourier transformation and order selection. Since most of the survey data are correlated, the independence assumption in SRS is usually violated. Several corrected GOF tests in complex surveys will also be reviewed, such as the Wald test, the first order and second order corrected tests.

2.1 Chi-Squared GOF Tests in SRS

In a GOF test of multi-outcome data, the statement to be tested is called a null hypothesis. The test statistics are usually used to measure the "distance" between the expected counts under the null hypothesis and the observed counts. If the "distance" is large enough, we reject the null hypothesis.

A variety of test statistics, such as Pearson's chi-squared test and likelihood ratio test etc., have been developed based on this idea. It has been discovered that these test statistics follow a chi-squared or a function of chi-squared distribution. As a result, these tests are often called chi-squared type GOF tests. One concept in the statistical tests is known as the Type I error. The Type I error is an event that the null hypothesis is rejected while it is true. The pre-specified tolerance of rejecting a true hypothesis is then measured by the probability of the Type I error, which is well known as level of significance α. The most widely used α values are 0.01, 0.05, and 0.1. Notice that this value should be determined before the tests.

Suppose a random experiment has been repeated for n times with K possible different outcomes. The Pearson's test statistic, denoted as X^2, is of the form

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \cdots + \frac{(O_K - E_K)^2}{E_K} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}, \tag{2.1}$$

where K is the total number of categories, O_k is the observed number of outcomes that fall into the kth category, and E_k is the expected number of outcomes that fall into the kth category under the null hypothesis.

2.1.1 Pearson's Chi-Squared Test

Pearson's chi-squared GOF test (Pearson, 1900) is one of the most commonly used chi-squared type tests for multinomial data. In survey data, it can also be applied to SRS where the independence assumption is usually met. Suppose there are K categories and a total of n outcomes for multinomial data. The null hypothesis is as follows

$$H_0 : p(k) = p_0(k), \text{ for } k = 1, \cdots, K, \tag{2.2}$$

where $p(k)$ and $p_0(k)$ are the unknown and hypothesized probability that an outcome may fall into the kth category, respectively. Let O_k and E_k denote the observed and expected counts of outcomes under the null hypothesis (2.2) in the kth category, the estimated proportion of the kth category is $\tilde{p}(k) = \frac{O_k}{n}$, and the expected counts of the kth category is $E_k = np_0(k)$ for $k = 1, \cdots, K$. Based on the "distance" test statistic X^2, the Pearson's chi-squared test statistic is

$$X^2_{SRS} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = \sum_{k=1}^{K} \frac{(n\tilde{p}(k) - np_0(k))^2}{np_0(k)} = n \sum_{k=1}^{K} \frac{(\tilde{p}(k) - p_0(k))^2}{p_0(k)} \tag{2.3}$$

If H_0 is true, this test statistic follows a central chi-squared distribution with $(K-1)$ degrees of freedom, that is

$$X^2_{SRS} \sim \chi^2_{K-1}, \text{ under } H_0 \text{ (Pearson, 1900)}.$$

Therefore, for a pre-specified level of significance α, the value of the test statistic X^2_{SRS} is compared with the $(1-\alpha)$ quantile of the chi-squared distribution with centrality parameter 0 and degrees of freedom $(K-1)$, which is denoted as $\chi^2_{K-1}(1-\alpha)$. One may reject H_0 at level α, if

$$X^2_{SRS} \;>\; \chi^2_{K-1}(1-\alpha).$$

2.1.2 Likelihood Ratio Test

Another widely used GOF test for multinomial data is the likelihood ratio test (LRT). The test statistic is defined as

$$\Lambda(\mathbf{O}) = \frac{\sup\{L(\mathbf{p}|\mathbf{O}) : \mathbf{p} \in \Theta_0\}}{\sup\{L(\mathbf{p}|\mathbf{O}) : \mathbf{p} \in \Theta\}},$$

where $L(\cdot)$ indicates likelihood function, $\mathbf{p}$ and $\mathbf{O}$ are vector of probabilities and vector of observed counts for multinomial data, Θ_0 and Θ are the parameter spaces under the null hypothesis and the general parameter space. For the null hypothesis (2.2) and multinomial data, the numerator and denominator are

$$\sup\{L(\mathbf{p}|\mathbf{x}) : \mathbf{p} \in \Theta_0\} = \frac{n!}{O_1! \cdots O_K!} \prod_{k=1}^{K} p_0(k)^{O_k}$$

and

$$\sup\{L(\mathbf{p}|\mathbf{x}) : \mathbf{p} \in \Theta\} = \frac{n!}{O_1! \cdots O_K!} \prod_{k=1}^{K} \tilde{p}(k)^{O_k}.$$

Consequently, the test statistic becomes

$$\Lambda(\mathbf{O}) = \frac{\sup\{L(\mathbf{p}|\mathbf{O}) : \mathbf{p} \in \Theta_0\}}{\sup\{L(\mathbf{p}|\mathbf{O}) : \mathbf{p} \in \Theta\}} = \prod_{k=1}^{K} \left[\frac{p_0(k)}{\tilde{p}(k)}\right]^{O_k}.$$

By applying the function $-2\ln(\cdot)$, the LR test statistic for GOF test in SRS is given by

$$\begin{aligned} G^2_{SRS} &= -2\ln(\Lambda(\mathbf{O})) = -2\sum_{k=1}^{K} O_k \ln\left[\frac{p_0(k)}{\tilde{p}(k)}\right] \\ &= 2n\sum_{k=1}^{K} \tilde{p}(k) \ln\left[\frac{\tilde{p}(k)}{p_0(k)}\right]. \end{aligned}$$

It is well known that the asymptotic distribution of G^2_{SRS} is χ^2_{K-1} under the null hypothesis (2.2). Therefore, H_0 should be rejected at level α, if

$$G^2_{SRS} \quad > \quad \chi^2_{K-1}(1-\alpha).$$

2.1.3 Matrix Form of Chi-Squared Test Statistics

Suppose there are K categories and a total of n outcomes in a multinomial data set. Define

$$\mathbf{O} \;=\; \begin{pmatrix} O_1 \\ O_2 \\ \vdots \\ O_{K-1} \end{pmatrix}, \; \mathbf{p}_0 = \begin{pmatrix} p_0(1) \\ p_0(2) \\ \vdots \\ p_0(K-1) \end{pmatrix}, \; \mathbf{p} = \begin{pmatrix} p(1) \\ p(2) \\ \vdots \\ p(K-1) \end{pmatrix}, \; \tilde{\mathbf{p}} = \begin{pmatrix} \tilde{p}(1) \\ \tilde{p}(2) \\ \vdots \\ \tilde{p}(K-1) \end{pmatrix} \tag{2.4}$$

where $\mathbf{O}$ denotes a $(K-1)\times 1$ vector of observed counts O_i (the counts of category i), and $\mathbf{p}_0$, $\mathbf{p}$, $\tilde{\mathbf{p}}$ represent the vector of hypothesized, underlying, and estimated proportions of a multinomial data set, respectively. All of these vectors are $(K-1)\times 1$ dimension, because the proportion of the Kth category can be a function of the previous $K-1$ ones with the following relationship,

$$p_0(K) = 1 - p_0(1) - p_0(2) - \cdots - p_0(K-1) = 1 - \sum_{k=1}^{K-1} p_0(k).$$

Similarly, $p(K) = 1 - \sum_{k=1}^{K-1} p(k)$, $\tilde{p}(K) = 1 - \sum_{k=1}^{K-1} \tilde{p}(k)$, and $O_K = n - n\sum_{k=1}^{K-1} \tilde{p}(k)$ since $O_k = n\tilde{p}(k)$, for $k = 1, \cdots, K-1$.

Based on these definitions, the null hypothesis (2.2) can be written as

$$H_0 : \mathbf{p} = \mathbf{p}_0, \tag{2.5}$$

and its corresponding alternative is

$$H_1 : \mathbf{p} \neq \mathbf{p}_0.$$

Now, define a series of independent random variables as follows

$$y_j(k) = \begin{cases} 1 & \text{if outcome } j \text{ is in category } k, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } j = 1, \cdots, n \text{ and } k = 1, \cdots, K.$$

Notice the observed counts in the kth category can be $O_k = \sum_{j=1}^{n} y_j(k)$. Hence the estimated proportion of the kth category can be written as

$$\tilde{p}(k) = \frac{O_k}{n} = \frac{\sum_{j=1}^{n} y_j(k)}{n}, \text{ for } k = 1, \cdots, K.$$

It can be seen that $\tilde{p}(k)$ is an unbiased estimator of the unknown parameter $p(k)$ for all $k = 1, \cdots, K-1$,

$$\mathrm{E}(\tilde{p}(k)) = \mathrm{E}\left(\frac{\sum_{j=1}^{n} y_j(k)}{n}\right) = \frac{\sum_{j=1}^{n} \mathrm{E}[y_j(k)]}{n} = \frac{np(k)}{n} = p(k)$$

In addition, the variance of $\tilde{p}(k)$ is

$$\begin{aligned} \mathrm{var}(\tilde{p}(k)) &= \mathrm{var}\left(\frac{\sum_{j=1}^{n} y_j(k)}{n}\right) = \frac{\sum_{j=1}^{n} \mathrm{var}[y_j(k)]}{n} \\ &= \frac{np(k)(1-p(k))}{n^2} = \frac{p(k)(1-p(k))}{n}, \end{aligned}$$

because all $y_j(k)$'s are independent from each other for $j = 1, \cdots, n$. Now, we work on the covariance between $\tilde{p}(k)$ and $\tilde{p}(l)$, for all $k \neq l$ and $k, l = 1, \cdots, K$.

$$\begin{aligned} \mathrm{cov}(\tilde{p}(k), \tilde{p}(l)) &= \mathrm{cov}\left(\frac{O_k}{n}, \frac{O_l}{n}\right) = \mathrm{cov}\left(\frac{\sum_{j=1}^{n} y_j(k)}{n}, \frac{\sum_{j=1}^{n} y_j(l)}{n}\right) \\ &= \frac{1}{n^2}\mathrm{cov}\left(\sum_{j=1}^{n} y_j(k), \sum_{j=1}^{n} y_j(l)\right) \end{aligned}$$

Notice that the ith and jth observations are independent, for $i \neq j$, so their covariance is always 0. By this result, the covariance equation can be simplified as

$$\mathrm{cov}(\tilde{p}(k), \tilde{p}(l)) = \frac{1}{n^2} \sum_{j=1}^{n} \mathrm{cov}\left[y_j(k), y_j(l)\right]$$

Now, we consider the covariance $\mathrm{cov}[y_j(k), y_j(l)]$ for $j = 1, \cdots, n$,

$$\begin{aligned}
\mathrm{cov}(y_j(k), y_j(l)) &= \mathrm{E}\left[y_j(k) y_j(l)\right] - \mathrm{E}\left[y_j(k)\right] \mathrm{E}\left[y_j(l)\right] \\
&= -\mathrm{E}\left[y_j(k)\right] \mathrm{E}\left[y_j(l)\right] \\
&= -p(k)p(l),
\end{aligned}$$

because the jth observation can not fall into both the kth and lth category at the same time, i.e., $\mathrm{E}\left[y_j(k) y_j(l)\right] = 0$ is always true. So the covariance between $\tilde{p}(k)$ and $\tilde{p}(l)$ is

$$\mathrm{cov}(\tilde{p}(k), \tilde{p}(l)) = \frac{1}{n^2} n[-p(k)p(l)] = -\frac{p(k)p(l)}{n}, \text{ for } k \neq l.$$

We now define

$$\mathbf{P} = \begin{pmatrix} p(1)(1-p(1)) & -p(1)p(2) & \cdots & -p(1)p(K-1) \\ -p(2)p(1) & p(2)(1-p(2)) & \cdots & -p(2)p(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ -p(K-1)p(1) & -p(K-1)p(2) & \cdots & p(K-1)(1-p(K-1)) \end{pmatrix}$$

a $(K-1) \times (K-1)$ matrix. One can easily see that

$$\mathrm{E}(\tilde{\mathbf{p}}) = \mathbf{p} \text{ and } \mathrm{cov}(\tilde{\mathbf{p}}) = \frac{\mathbf{P}}{n}.$$

We also notice that, under the null hypothesis (2.2), $p(k)$ equals the hypothesized proportion $p_0(k)$ for $k = 1, \cdots, K$. Define

$$\mathbf{P}_0 = \begin{pmatrix} p_0(1)(1-p_0(1)) & -p_0(1)p_0(2) & \cdots & -p_0(1)p_0(K-1) \\ -p_0(2)p_0(1) & p_0(2)(1-p_0(2)) & \cdots & -p_0(2)p_0(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ -p_0(K-1)p_0(1) & -p_0(K-1)p_0(2) & \cdots & p_0(K-1)(1-p_0(K-1)) \end{pmatrix}$$

a $(K-1)\times(K-1)$ matrix, then $\mathbf{p}_0$ and $\mathbf{P}_0/n$ are the expectation and the covariance matrix of $\tilde{\mathbf{p}}$ under the null hypothesis (2.5). The Pearson's chi-squared test statistic can be written in matrix form as

$$\begin{aligned} X^2_{SRS} &= n\sum_{k=1}^{K}\frac{(\tilde{p}(k)-p_0(k))^2}{p_0(k)} \\ &= n(\tilde{\mathbf{p}}-\mathbf{p}_0)^{\mathrm{T}}\mathbf{P}_0^{-1}(\tilde{\mathbf{p}}-\tilde{\mathbf{p}}_0). \end{aligned} \tag{2.6}$$

We now verify (2.6). We define $D(\mathbf{p}_0)$ a $(K-1)\times(K-1)$ matrix with kth diagonal element $p_0(k)$ and off-diagonal entries 0. Then, $\mathbf{P}_0$ can be written as

$$\mathbf{P}_0 = D(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^{\mathrm{T}}.$$

According to property B. 56 of Christensen (2011), the inverse of $\mathbf{P}_0$ is

$$\begin{aligned} \mathbf{P}_0^{-1} &= \left[D(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^{\mathrm{T}}\right]^{-1} \\ &= D\left(\frac{1}{\mathbf{p}_0}\right) + D\left(\frac{1}{\mathbf{p}_0}\right)\mathbf{p}_0\left[1-\mathbf{p}_0^{\mathrm{T}}D\left(\frac{1}{\mathbf{p}_0}\right)\mathbf{p}_0\right]^{-1}\mathbf{p}_0^{\mathrm{T}}D\left(\frac{1}{\mathbf{p}_0}\right) \\ &= D\left(\frac{1}{\mathbf{p}_0}\right) + \mathbf{J}\left[1-\mathbf{p}_0^{\mathrm{T}}\mathbf{J}\right]^{-1}\mathbf{J}^{\mathrm{T}}, \end{aligned}$$

where $\mathbf{J}$ is a $(K-1)$ vector with all elements 1. Therefore, (2.6) can be verified as follows.

$$\begin{aligned}
X^2_{SRS} &= n(\tilde{\mathbf{p}}-\mathbf{p}_0)^{\mathrm{T}}\mathbf{P}_0^{-1}(\tilde{\mathbf{p}}-\tilde{\mathbf{p}}_0) \\
&= n(\tilde{\mathbf{p}}-\mathbf{p}_0)^{\mathrm{T}}\left\{D\left(\frac{1}{\mathbf{p}_0}\right)+\mathbf{J}\left[1-\mathbf{p}_0^{\mathrm{T}}\mathbf{J}\right]^{-1}\mathbf{J}^{\mathrm{T}}\right\}(\tilde{\mathbf{p}}-\tilde{\mathbf{p}}_0) \\
&= n(\tilde{\mathbf{p}}-\mathbf{p}_0)^{\mathrm{T}}D\left(\frac{1}{\mathbf{p}_0}\right)(\tilde{\mathbf{p}}-\tilde{\mathbf{p}}_0)+\frac{n(\tilde{\mathbf{p}}-\mathbf{p}_0)^{\mathrm{T}}\mathbf{J}\mathbf{J}^{\mathrm{T}}(\tilde{\mathbf{p}}-\mathbf{p}_0)}{1-\mathbf{p}_0^{\mathrm{T}}\mathbf{J}} \\
&= n\sum_{k=1}^{K-1}\frac{(\tilde{p}(k)-p_0(k))^2}{p_0(k)}+n\frac{\left(\sum_{k=1}^{K-1}\tilde{p}(k)-\sum_{k=1}^{K-1}p_0(k)\right)^2}{1-\sum_{k=1}^{K-1}p_0(k)} \\
&= n\sum_{k=1}^{K-1}\frac{(\tilde{p}(k)-p_0(k))^2}{p_0(k)}+n\frac{(p_0(K)-\tilde{p}(K))^2}{p_0(K)} \\
&= n\sum_{k=1}^{K}\frac{(\tilde{p}(k)-p_0(k))^2}{p_0(k)}
\end{aligned}$$

2.2 Neyman Smooth-Type GOF Tests in SRS

The Neyman smooth-type GOF tests provide directional tests by taking advantage of Fourier transformation so that the Pearson's chi-squared test statistic can be decomposed into several components.

2.2.1 Fourier Transformation

According to Fourier transformation, some functions can be written as a combination of *cosine* and *sine* functions, i.e., a sum of the trigonometric functions. This relationship is built according to Euler's formula (Euler, 1743),

$$e^{2\pi i\theta} = \cos(2\pi\theta) + i\sin(2\pi\theta)$$

where i is the complex number with $i^2 = -1$.

Let $f(x)$ be a function that is not constantly zero on its domain. If x is continuous, its continuous Fourier coefficients can be written as

$$c_t = \int_{-\infty}^{\infty} f(x)e^{-itx}dx.$$

If x is discrete, its discrete Fourier coefficient is of the form

$$c_t = \sum_{x=-\infty}^{\infty} f(x)e^{-itx}.$$

On the other hand, the inverse Fourier transformation of c_t is

$$f(x) = \sum_{t=-\infty}^{\infty} c_t e^{itx},$$

under appropriate conditions, according to Fourier inversion theorem in Fourier (1822, pg. 525) and Fourier and Freeman (1878, pg. 408). In addition, $f(x)$ and c_t are also considered to be a Fourier integral pair or Fourier transform pair in Rahman (2011, pg. 10).

2.2.2 Decomposing Pearson's Chi-Squared Test Statistic

We first introduce one of the general decompositions, which has been well studied in Lancaster (1969), Nair (1987, 1988), Rayner et al. (1985), Rayner and Best (1986), Rayner and Best (1989) and so on. Suppose there are K categories and n observations in multinomial data from SRS.

Define a series of basis functions $h_1(\cdot), \cdots, h_{K-1}(\cdot)$, with orthogonality condition

$$\sum_{k=1}^{K-1} h_j(O_k)h_i(O_k)\tilde{p}(k) \quad = \quad \delta_{ji}$$

where δ_{ji} is an indicator function with value 1 if $j = i$ and value 0 if $j \neq i$. Let

$$L_j = \sum_{k=1}^{K} h_j(O_k),$$

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

$$X^2_{SRS} = \sum_{j=1}^{K-1} L_j^2.$$

Another decomposition is introduced in Eubank (1997). For $j = 1, \cdots, K-1$, let the basis function $x_j(k)$ satisfy the following orthogonality conditions

$$\sum_{k=1}^{K} x_j(k)x_i(k) = \delta_{ji} = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i, \end{cases}$$

and

$$\sum_{k=1}^{K} x_j(k)\sqrt{p_0(k)} = 0,$$

Let

$$\tilde{f}(k) = \frac{\tilde{p}(k) - p_0(k)}{\sqrt{p_0(k)}}, \text{ for } k = 1, \cdots, K, \tag{2.7}$$

with associated (discrete) generalized Fourier coefficients

$$b_j = \sum_{k=1}^{K} \tilde{f}(k)x_j(k), \text{ for } j = 1, \cdots, K-1. \tag{2.8}$$

It is trivial to show that $\tilde{f}(k)$ is an unbiased estimator of

$$f(k) = \frac{p(k) - p_0(k)}{\sqrt{p_0(k)}}, \text{ for } k = 1, \cdots, K, \tag{2.9}$$

with associated Fourier coefficients

$$\beta_j = \sum_{k=1}^{K} f(k) x_j(k), \text{ for } j = 1, \cdots, K-1, \tag{2.10}$$

because $\tilde{p}(k)$ is an unbiased estimator of $p(k)$. By Parseval's relation (Arfken, 1985, pg. 425), the test statistic of Pearson's chi-squared test statistic can be re-organized as

$$X^2_{SRS} = n \sum_{k=1}^{K} \frac{(\tilde{p}(k) - p_0(k))^2}{p_0(k)} = n \sum_{k=1}^{K} (\tilde{f}(k))^2 = \sum_{j=1}^{K-1} n b_j^2, \tag{2.11}$$

which is a sum of ordered terms. For example, nb_1^2 is the most important term that includes the most information from the data, while nb_{K-1}^2 is the least important term that may not carry as much information as previous terms.

2.2.3 Order Selection

Order selection is an important technique in nonparametric regression. We now use the polynomial regression as an example to introduce the idea of order selection. Suppose there are n pairs of data points, $(x_1, y_1), \cdots, (x_n, y_n)$ with x_i and y_i the realizations of covariates and responses. Assume one wants to fit a polynomial regression model for this data set, that is,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i, \; i = 1, \cdots n,$$

where ϵ_i's are independent random variables with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2 < \infty$, and β_i's are regression coefficients. The next problem is to select an appropriate value for order p by some appropriate criteria. One of the most commonly used methods is the mean squared error (MSE). Let's use $p = 2$ as an example to illustrate the order selection process by MSE criterion. In this case, the regression model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, and MSE is given by $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2\right)^2$. By using cross-validation method, the predicted value of y_i, i.e., $\hat{y}_i$ is $\hat{\beta}_{0(j)} - \hat{\beta}_{1(j)} x_i - \hat{\beta}_{2(j)} x_i^2$, where $\hat{\beta}_{i(j)}$'s are calculated using all observations except the jth one. For $j = 1, \cdots, n$, there are n MSE's calculated and the average MSE can be computed for $p = 2$. Similarly, average MSE's for $p = 3$, $p = 4$, $\cdots$ can also be computed. The selected order p is the one associated with the smallest average MSE.

MSE and cross validation method can be applied to many order selection problems, including GOF tests in regression. However, it may be computationally expensive if the data set is large. Eubank and Hart (1992) developed a novel order selection criterion for GOF test in regression, which considers the regression model

$$y_j = f(x_j) + \epsilon_j,\ j = 1, \cdots, n,$$

for observed data $(x_1, y_1), \cdots, (x_n, y_n)$, where $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$ and f is an arbitrary function. The null hypothesis of interest is

$$H_0 : f(x) = \sum_{j=1}^{p} \beta_j t_j(x), \tag{2.12}$$

where β_j's are unknown regression coefficients and t_j's are some unknown functions. Its corresponding alternative hypothesis is

$$H_1 : f(x) = \sum_{j=1}^{p} \beta_j t_j(x) + g(x),$$

where g is a function that is not a combination of the t_j's. Next, Eubank and Hart (1992) defines the estimates of g as b_{jn} with Fourier coefficients b_j for $j = 1, \cdots, n-p$, so that the null hypothesis (2.12) becomes

$$H_0^* : b_1 = \cdots = b_{n-p} = 0.$$

As a result, the test statistic $\hat{q}$ is the maximizer of the criterion

$$J(q) = \sum_{j=1}^{q} b_{jn}^2 - \frac{q a_\alpha \hat{\sigma}^2}{n}, \text{ for } q = 1, \cdots, n-p, \tag{2.13}$$

where $H(0) = 0$, $\hat{\sigma}^2$ is the estimator of $\sigma^2 = \text{var}(\epsilon_i)$, a_α is a constant that is determined by the level of significance α. In practice, a_α is the solution of (3.14) or (3.15). The null hypothesis H_0^* is rejected at level α if $\hat{q} > 0$ is found.

Eubank and Hart (1992) also mention that if $a_\alpha = 2$ is selected for (2.13), $\hat{q}$ is equivalent to the minimizer of an unbiased estimator of the risk function

$$R(q) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_q(x_i) - g(x_i))^2\right].$$

This approach is different from previous work, such as Eubank and Spiegelman (1990), because it is data-driven. This method is later chosen to be the order selection method in Eubank (1997) and in this dissertation.

2.2.4 Neyman Smooth-Type GOF tests with Order Selection

In Eubank (1997), some abnormal outcomes were observed for the Pick 3 game from the Texas Lottery machines. The game was described by Eubank as follows. For each game, 3 machines are used. Each has 10 ping-pong-type balls numbered from 0 to 9. In every machine, there is a mixing chamber and a vertical tube. When the game starts, balls enter in the vertical tube one by one, with ball numbered 0 at the bottom and ball numbered 9 at the top. Then, all the balls go to the mixing chamber and are mixed by air. Finally, only one ball is selected to go through the exit hole. A player wins the game if the pre-chosen numbers match the numbers selected randomly by the machines in daily draws. If the machines are working properly, the probability of any ball that is selected is supposed to be around 0.1, that is, the null hypothesis is $H_0 : p(1) = \cdots = p(10) = 0.1$. However, the owner of the machines found that one particular machine was not performing as expected. Thus, this machine was taken offline for further investigation. According to 150 draws, the results are shown in Figure 1.1, where the X-axis is the number of the balls (10 discrete numbers from 0 to 9), and the Y-axis is the observed probability of each ball obtained through the 150 draws. If the machine worked properly, the anticipated observed probabilities of the 10 balls should be randomly distributed around 0.1. However, it is observed that balls with larger numbers (7, 8, and 9) tend to have lower probabilities of selection, while small number balls (0, 1, 3, and 4) are more likely to be chosen. Besides those values of observed probabilities, one can also observe that there is an overall decreasing trend for the estimated probabilities, instead of randomly distributed around 0.1. In addition, this figure also illustrates that the selection probabilities are close to each other (range from 0.06 to 0.13), which can be described as slow varying probabilities despite the selection probability trend.

As shown in Eubank (1997), it was found that the Pearson's chi-squared test statistic is $X^2_{SRS} = 7.73$, which failed to identify this non-randomness problem at level of 0.1 or 0.05, because the critical values of χ^2_9 are 14.68 and 16.92 for $\alpha = 0.1$ and $\alpha = 0.05$, respectively. In fact, the p-value is approximately 0.48, which can't reject the null hypothesis even at level 0.48. This is a contradiction against what was observed. Thus, Eubank (1997) proposed a data-driven Neyman smooth-type goodness of fit tests incorporated with order selection, which successfully rejected the null hypothesis between level 0.05 and 0.1, i.e., detected the non-randomness problem. We now review these methods.

Suppose there are K categories and n observations for a multinomial data from SRS. The null hypothesis is the same as (2.2). The Pearson's chi-squared test statistic is then decomposed as in (2.11). According to Eubank (1997), one of the choices of the basis function for $X^2_{SRS} = \sum_{j=1}^{K-1} nb_j^2$ is

$$x_j(k) = \sqrt{\frac{2}{K}}\cos\left(\frac{j\pi(k-0.5)}{K}\right), \text{ for } j = 1, \cdots, K-1, \tag{2.14}$$

The basis function (2.14) is chosen for the particular null hypothesis $H_0 : p(1) = \cdots = p(10) = 0.1$. Eubank (1999, pg. 75) gives a detailed description of how to construct this basis function.

Notice that, $f(k)$ in (2.9) is 0 under H_0, for $k = 1, \cdots, K$. Therefore, the null hypothesis (2.2) is now equivalent to $H_0^* : \beta_1 = \cdots = \beta_{K-1} = 0$, with the corresponding alternative $H_1^* : \beta_q \neq 0$ and $\beta_{q+1} = \cdots = \beta_{K-1} = 0$ for $q = 1, \cdots, K-1$, according to Lehmann (1986, sec. 8.8 and ex. 37 in pg. 495). It was found that

$$\sqrt{n}\begin{pmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \\ \vdots \\ b_{K-1} - \beta_{K-1} \end{pmatrix} \to N_{K-1}(\mathbf{0}, \mathbf{V}), \text{ as } n \to \infty,$$

where $\mathbf{V} = \{v_{ij} - \beta_i\beta_j\}$ is a $(K-1) \times (K-1)$ covariance matrix with $(v_{ij} - \beta_i\beta_j)$ as the ijth entry and $v_{ij} = \sum_{k=1}^{K} x_i(k)x_j(k)p(k)/p_0(k)$ for $i, j = 1, \cdots, K-1$. The test statistic of order q is $X_q^2 = \sum_{j=1}^{q} nb_j^2$, for $q = 1, \cdots, K-1$. The test can be conducted by comparing X_q^2 with χ_q^2.

Suppose the underlying order is q_0. The next task is to find out a good estimator of q_0 according to data. Eubank (1997) argued that an appropriate value of q_0 should be the minimizer of the criterion $\sum_{k=1}^{K}(f_q(k)-f(k))^2$, or equivalently, the maximizer of $M(q)=-\sum_{j=1}^{q}b_j^2+2\sum_{j=1}^{q}\beta_j b_j$ with respect to q, where $f_q=\sum_{j=1}^{q}b_j x_j$.

According to Hart (1985), an unbiased estimator of $M(q)$ is

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

where $\tilde{v}_{jj}=\sum_{k=1}^{K}x_j(k)^2\tilde{p}(k)/p_0(k)$ for $j=1,\cdots,K-1$. If $\tilde{q}$ is the maximizer of $\tilde{M}(q)$ with $\tilde{M}(0)=0$, the first test statistic (Eubank, 1997) is given by $W=(X_{\tilde{q}}^2-\tilde{q})/\sqrt{2\tilde{q}}$ for $\tilde{q}\neq 0$ with $W=0$ if $\tilde{q}=0$. The distribution of W under H_0^*, say W_0, is obtained through simulation. Then, the level α test is performed by comparing the value of W and the $1-\alpha$ quantile of W_0.

Another estimator of q_0 at level α, denoted as $\tilde{q}_\alpha$, is the maximizer of

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

with $\tilde{M}_\alpha(0)=0$, where $\tilde{v}_{jj}=\sum_{k=1}^{K}x_j^2(k)\tilde{p}(k)/p_0(k)$, for $j=1,\cdots,K-1$, and the number a_α is chosen so that it is the solution of $1-\alpha=\exp\left\{-\sum_{k=1}^{\infty}\frac{P(\chi_k^2>ka_\alpha)}{k}\right\}$ according to Eubank and Hart (1992) or $P\left(\max_{1\le k\le K-1}\left[\frac{1}{k}\sum_{j=1}^{k}Z_j^2\right]\geq a_\alpha\right)=\alpha$ according to Eubank (1997). For example, if $\alpha=0.05$ is chosen, the corresponding solution is $a_{0.05}=4.18$. Note that χ_k^2 is a central chi-squared random variable with k degrees of freedom, and the Z_j's are independent standard normal random variables. Because of the transformed null and alternative hypothesis H_0^* and H_1^*, the null hypothesis should be rejected if $\tilde{q}_\alpha>0$, for example, if $\tilde{q}_\alpha=1$ is found, at least we know that $\beta_1\neq 0$ and thus the null hypothesis does not hold.

Now, recall the previous Pick 3 game example. It was found that $\tilde{q}=1$, $X_{\tilde{q}}^2=X_1^2=5.186$ and $W=2.96$. The critical values of W_0 are 2.99 and 2.3 for level 0.05 and 0.1, respectively. It was also found that $\tilde{q}_{0.05}=0$ and $\tilde{q}_{0.1}=0$. Both methods suggest that the null hypothesis should be rejected at a level between 0.05 and 0.1, that is, the 10 balls didn't pop up randomly. This is consistent with what we observed from the 150 draws.

In Eubank (1997), both Neyman smooth-type GOF tests and order selection are utilized to get rid of redundant components of the test statistics, so that we have more degrees of freedom for the test. As a result, these methods are more sensitive in detecting slow varying probabilities in multinomial data.

2.3 Chi-Squared GOF Tests in Complex Surveys

A complex survey usually consists of stratification and clustering in order to increase the precision of the estimator and to decrease the survey cost. Data from a complex survey is usually correlated, where the assumption of independence is no longer met. The existing GOF tests need to be modified to accommodate the correlated structure of the survey data. In this section, we will review a few sampling designs as well as some GOF tests for use in complex surveys.

2.3.1 Sampling Designs

Sampling designs play an important role in data collection and related statistical analyses. The frequently used and popular designs include simple random sampling, stratified random sampling and cluster sampling.

For an SRS, suppose that there are N students in the University of New Mexico (UNM). We randomly select n students from the population (N students). Therefore, each student represents N/n students, called sampling weight of the selected student.

For some other type of surveys, for example, we want to survey the attitude towards immigration among the students in UNM. If we use an SRS, we may miss the minorities such as those international students. We may divide the students into several strata such as local students (New Mexico students), with population a total of N_1, students from other areas of the country (with population a total of N_2), and international students (with population a total of N_3). We then take an SRS from each stratum respectively with sample sizes of n_1, n_2, and n_3. This is called stratified random sampling. Sampling weights for the three strata are N_1/n_1, N_2/n_2, and N_3/n_3, respectively.

For cluster sampling, we illustrate one example from Lohr (2010, pg. 171). A student is interested in estimating the average GPA of all students in his dormitory. There are 100 suites in the dorm and every suite contains 4 students. A sample of 20 students is determined to conduct the survey. If SRS is used, all 400 students should be listed and 20 should be randomly picked up from the list. In this case, the student investigator may need to go to many suites to complete the survey. Therefore, cluster sampling is used. Each suite is considered as a primary sampling unit (psu) and students in each suite are secondary sampling units (ssu). The student investigator randomly selects 5 suites and conducts surveys of all 4 students in the chosen suite. The sampling strategy in this example is called cluster sampling, which usually decreases precision, but is less expensive.

Lohr (2010, pg. 407) gives an example about the effects of clustering on GOF tests. In that example, the husband and wife give exactly the same answer to the survey questionnaire, thus they are perfectly correlated. If we ignore the clustering effects and treat husband and wife as independent individuals, the resulting test statistic X^2_{SRS} in (2.6) will be doubled. Therefore, we are more likely to reject the null hypothesis than we should do. This requires new methods for GOF tests in complex surveys.

2.3.2 Wald Test

The Wald test (Wald, 1943) is named after Abraham Wald. It can be considered as a correction to Pearson's chi-squared GOF test that can be used in both SRS and complex surveys. Recall vectors in (2.4)

$$\mathbf{p}_0 = \begin{pmatrix} p_0(1) \\ p_0(2) \\ \vdots \\ p_0(K-1) \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p(1) \\ p(2) \\ \vdots \\ p(K-1) \end{pmatrix}, \quad \hat{\mathbf{p}} = \begin{pmatrix} \hat{p}(1) \\ \hat{p}(2) \\ \vdots \\ \hat{p}(K-1) \end{pmatrix}.$$

with the restriction that $\sum_{k=1}^{K} p_0(k) = \sum_{k=1}^{K} p(k) = \sum_{k=1}^{K} \hat{p}(k) = 1$. $\hat{\mathbf{p}}$ is calculated based on the sampling design. Denote $\mathbf{V}$ the $(K-1) \times (K-1)$ covariance matrix of $\hat{\mathbf{p}}$ and $\hat{\mathbf{V}}$ the estimator of $\mathbf{V}$. Rao and Scott (1981) discussed several approaches to calculated $\hat{\mathbf{V}}$. The Wald test statistic is of the matrix form

$$X_W^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)^{\mathrm{T}} \hat{\mathbf{V}}^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0),$$

whose asymptotic distribution under the null hypothesis (2.2) is a central chi-squared distribution with $(K-1)$ degrees of freedom. The Wald test statistic is equivalent to Pearson's chi-squared test statistic if $\mathbf{V} = \mathbf{P}_0$. Therefore, Pearson's chi-squared test statistic may be considered as a special case of the Wald test statistic. Wald test can be used in both complex surveys and simple random sample. The major drawback of the Wald test statistic is the instability of the estimator of $\mathbf{V}^{-1}$, which may result in poor control of the Type I error in some situations (Thomas and Rao (1987)).

2.3.3 Corrections to Chi-Squared Test Statistic

In this section, we review the first order and second order corrected chi-squared type GOF tests proposed by Rao and Scott (1981, 1984) and Bedrick (1983) for use in complex surveys. Rao and Thomas (1988) compared many chi-squared type tests in complex surveys and concluded that both first order and second order corrected tests (Rao & Scott, 1981, 1984 and Bedrick, 1983) work well for complex surveys. Let

$$X^2 = n \sum_{k=1}^{K} \frac{(\hat{p}(k) - p_0(k))^2}{p_0(k)}$$

be the test statistic for multinomial data from complex surveys. Notice that the unweighted estimated proportion $\tilde{p}(k)$ is replaced by the weighted estimated proportion $\hat{p}(k)$ to incorporate the design information. The corresponding matrix form of this test statistic is

$$X^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)^{\mathrm{T}} \mathbf{P}_0^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0).$$

Rao and Scott (1981) found that the asymptotic distribution of X^2 under the null hypothesis (2.2) was a linear combination of central chi-squared distribution with 1 degrees of freedom, that is,

$$X^2 \sim \sum_{i=1}^{K-1} \delta_i \chi_1^2,$$

where δ_i's, for $i = 1, \cdots, K-1$ are eigenvalues of the design effects matrix $\mathbf{P}_0^{-1}\mathbf{V}$. If the sampling design is SRS, all the eigenvalues equal 1 under the null hypothesis (2.2) because $\mathbf{V} = \mathbf{P}_0$. As a result, $X^2 = X_{SRS}^2$ follows central chi-squared distribution with $(K-1)$ degrees of freedom under the null hypothesis (2.2). At this point, the critical values need to be simulated, which brings difficulty in the calculation. Thus, Rao and Scott (1981), Bedrick (1983), and Rao and Scott (1984) proposed the first order and second order corrected tests to accelerate the test procedures.

For the first order corrected test, let

$$\delta_. \;=\; \frac{\sum_{i=1}^{K-1} \delta_i}{K-1}$$

denote the average of the $(K-1)$ eigenvalues of the design effects matrix $\mathbf{P}_0^{-1}\mathbf{V}$. Since the expected value of $\sum_{i=1}^{K-1} \delta_i \chi_1^2$ is $\sum_{i=1}^{K-1} \delta_i = (K-1)\delta_.$, the expected value of $X^2/\delta_.$ is $(K-1)$ under the null hypothesis (2.2). In practice, $\delta_.$ can be estimated by

$$\hat{\delta}_. \;=\; \frac{n}{K-1} \sum_{k=1}^{K} \frac{\hat{V}_{kk}}{p_0(k)},$$

where $\hat{V}_{kk} = \hat{p}(k)(1-\hat{p}(k))$ is the kth diagonal element of the matrix $\hat{\mathbf{V}}$, for $k = 1, \cdots, K-1$. Hence, under the null hypothesis (2.2), the test statistic of the first order corrected test is

$$X_C^2 = \frac{X^2}{\hat{\delta}_.}, \text{ with } \mathrm{E}(X_C^2) = K-1. \tag{2.15}$$

Therefore, by matching the mean of the test statistic, under the null hypothesis (2.2), $X^2/\delta_.$ can be compared with a central chi-squared distribution with $(K-1)$ degrees of freedom.

In addition, the second order test statistic is given by

$$X_S^2 = \frac{X^2}{\hat{\delta}_.(1+\hat{a}^2)} = \frac{X_C^2}{(1+\hat{a}^2)}, \tag{2.16}$$

where

$$\hat{a}^2 = \sum_{i=1}^{K-1} \hat{\delta}_i^2/[(K-1)\hat{\delta}_.^2] - 1 \tag{2.17}$$

and $\sum_{i=1}^{K-1} \hat{\delta}_i^2$ can be calculated by

$$\sum_{i=1}^{K-1} \hat{\delta}_i^2 = n^2 \sum_{i=1}^{K} \sum_{i=1}^{K} \hat{V}_{ij}/p_0(i)p_0(j),$$

if the full estimated covariance matrix is known. The test statistic of the second order corrected test is then compared with central chi-squared distribution with $(K-1)/(1+\hat{a}^2)$ degrees of freedom under the null hypothesis (2.2).

Chapter 3

Neyman Smooth-Type GOF Tests in Complex Surveys

In complex survey, observations are usually correlated due to clustering, where the independence assumption is no longer met. Therefore, Pearson's chi-squared test, the likelihood ratio test, and Neyman smooth-type GOF tests, etc. don't have $\chi^2(K-1)$ distributions. A number of methods were proposed to account for the survey design information when testing goodness of fit, such as Wald test, Fay's Jackknived chi-squared test, the Rao-Scott approaches, and so on. But none of the research address the slow varying probability problem corresponding to the null hypothesis (3.1). In this chapter, we extend Neyman smooth-type GOF test (Eubank, 1997) to complex surveys. We first propose two GOF tests in complex surveys in Section 3.1. Next, we discuss the asymptotic properties of the estimators in Section 3.2 for stratified sampling.

3.1 Proposed Tests in Complex Surveys

Consider a complex survey, n observations are classified into K categories according to certain factor levels. In this section, we first give the framework of the tests, then we propose two Neyman smooth-type GOF tests with order selection for use in complex surveys.

3.1.1 Framework

Suppose the null hypothesis of interest is

$$H_0 : p(k) = p_0(k) = \frac{1}{K}, \; k = 1, \cdots, K, \tag{3.1}$$

that is, the probability of each category $p(k)$ is a fixed number $p_0(k)$ with restriction $\sum_{k=1}^{K} p_0(k) = 1$. Note that observations from a survey data are weighted, i.e., each observation y_i represents w_i observations.

We first define the basis function $x_j(k)$ that satisfies the following orthogonality conditions:

$$\sum_{k=1}^{K} x_j(k)x_i(k) = \delta_{ji} = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i, \end{cases} \quad i, j = 1, \cdots, K-1 \tag{3.2}$$

and

$$\sum_{k=1}^{K} x_j(k)\sqrt{p_0(k)} = 0, \; j = 1, \cdots, K-1. \tag{3.3}$$

There are many choices of the basis function. Following Eubank (1997), the basis function (3.4) for the null hypothesis (3.1) is used in our proposed tests,

$$x_j(k) = \sqrt{\frac{2}{K}} \cos\left(\frac{j\pi(k-0.5)}{K}\right), \; k = 1, \cdots, K \text{ and } j = 1, \cdots, K-1. \tag{3.4}$$

In general, it is possible to chose different basis functions for an arbitrary null hypothesis $H_0 : p(k) = p_0(k)$ for $k = 1, \cdots, K$. The basis functions should satisfy the orthogonality conditions (3.2) and (3.3). To do that, 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 $\mathbf{p}_0 = (0.5, 0.3, 0.2)^{\mathrm{T}}$, 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 (3.2) and (3.3). In our research, we focus on the simple uniform hypothesis (3.1) and use the basis function (3.4) for $j = 1, \cdots, K - 1$ and $k = 1, \cdots, K$. It is easy to check that the basis function (3.4) satisfy the orthogonality conditions (3.2) and (3.3).

Now, let

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

and w_j be the sampling weight of $y_j(k)$, for $j = 1, \cdots, n$ and $k = 1, \cdots, K$. The general form of the consistent estimated proportion of the kth category is

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

Let

$$\hat{f}(k) = \frac{\hat{p}(k) - p_0(k)}{\sqrt{p_0(k)}}, \tag{3.5}$$

with associated (discrete) generalized Fourier coefficients

$$b_j = \sum_{k=1}^{K} \hat{f}(k) x_j(k), \ j = 1, \cdots, K - 1. \tag{3.6}$$

By the fact that $\mathrm{E}[\hat{p}(k)] = p(k)$, we can prove that $\hat{f}(k)$ is an unbiased estimator of

$$f(k) = \frac{p(k) - p_0(k)}{\sqrt{p_0(k)}}, \tag{3.7}$$

with associated Fourier coefficients

$$\beta_j = \sum_{k=1}^{K} f(k) x_j(k). \tag{3.8}$$

As a result of Parseval's relation (Arfken, 1985, pg. 425), the GOF test statistic in complex surveys can be re-written as

$$X^2 = n \sum_{k=1}^{K} \frac{(\hat{p}(k) - p_0(k))^2}{p_0(k)} = n \sum_{k=1}^{K} (\hat{f}(k))^2 = \sum_{j=1}^{K-1} n b_j^2,$$

which consists of $(K-1)$ orthogonal components. Notice that, $f(k)$ is 0 under the null hypothesis (3.1). By sec. 8.8 and ex. 37 of Lehmann (1986, pg. 495), the null hypothesis (3.1) is equivalent to

$$H_0^* : \beta_1 = \cdots = \beta_{K-1} = 0, \tag{3.9}$$

and its corresponding alternative becomes

$$H_1^* : \beta_q \neq 0 \text{ and } \beta_{q+1} = \cdots = \beta_{K-1} = 0, \text{ for } q = 1, \cdots, K-1. \tag{3.10}$$

The proposed Neyman smooth-type GOF test statistic is

$$X_q^2 = \sum_{j=1}^{q} n b_j^2, \; q = 1, \cdots, K-1,$$

where order q is selected by data.

If the underlying order is q_0, an optimal estimator of q_0 should be the minimizer of the criterion

$$\sum_{k=1}^{K} \left(f_q(k) - f(k)\right)^2, \text{ where } f_q = \sum_{j=1}^{q} b_j x_j.$$

This is equivalent to maximize

$$M(q) = -\sum_{j=1}^{q} b_j^2 + 2\sum_{j=1}^{q} \beta_j b_j$$

with respect to q (Eubank, 1997). In the next two sections, we will discuss the two proposed GOF tests for complex surveys.

3.1.2 Proposed Test W

We first propose an estimator for q_0, say $\hat{q}$, which is the maximizer of the following equation

$$\hat{M}(q) = \frac{n+1}{n-1}\sum_{j=1}^{q} b_j^2 - \frac{2\hat{\delta}_{\cdot}}{n-1}\sum_{j=1}^{q} \hat{v}_{jj}, \; q = 1, \cdots, K-1, \tag{3.11}$$

where $\hat{M}(0) = 0$, $\hat{v}_{jj} = \sum_{k=1}^{K} x_j(k)^2 \hat{p}(k)/p_0(k)$, for $j = 1, \cdots, K-1$, and $\hat{\delta}_{\cdot}$ is the estimator of average of the eigenvalues of the design effects matrix $\mathbf{P}_0^{-1}\mathbf{V}$ introduced in Section 2.3.3.

Because of the transformed hypotheses H_0^* (3.9) and H_1^* (3.10), $\hat{q}$ may be a natural test statistic and the null hypothesis can be rejected if $\hat{q} > 0$ is obtained through data. For example, if $\hat{q} = 1$, at least we know that $\beta_1 \neq 0$, thus reject of H_0^*. However, as shown in Spitzer (1956) and Zhang (1992), and also by simulation, the limiting probability of the Type I error of such a test is

$$\lim_{K\to\infty} \lim_{n\to\infty} P(\hat{q} > 0 | q_0 = 0) = 0.29,$$

which implies that the level of significance is approximately 0.29.

In order to perform the test at a pre-specified level α, our first proposed test statistic based on Eubank (1997) is

$$W = \begin{cases} \frac{X_{\hat{q}}^2 - \hat{q}}{\sqrt{2\hat{q}}}, & \hat{q} > 0, \\ 0, & \hat{q} = 0, \end{cases} \tag{3.12}$$

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

Since the chance of overselection for q_0 is almost 30% if $q_0 = 0$, there is a high chance that $X_{\hat{q}}^2$ is a large number. For example, if $q_0 = 0$ is true and the selected $\hat{q}$ is $K-1$, then X_{K-1}^2 may be relatively much larger than the true value $X_0^2 = 0$ and thus can be considered as an outliers. For this reason, the test statistic W in (3.12) is a normalized $X_{\hat{q}}^2$ so that potential outliers may be avoid.

The distribution of W under null model (3.1) or (3.9) is obtained through simulations, which is denoted as W_0. For an arbitrary pre-specified level of significance α, the test can be performed by comparing the value of W and the $1-\alpha$ quantile of W_0.

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

We now change the Equation (3.11) slightly as the following

$$\hat{M}_\alpha(q) = \frac{n+1}{n-1} \sum_{j=1}^{q} b_j^2 - \frac{a_\alpha \hat{\delta}_.}{n-1} \sum_{j=1}^{q} \hat{v}_{jj}, \; q = 1, \cdots, K-1 \tag{3.13}$$

where $\hat{M}_\alpha(0) = 0$ and $\hat{v}_{jj} = \sum_{k=1}^{K} x_j(k)^2 \hat{p}(k)/p_0(k)$, for $j = 1, \cdots, K-1$. The second proposed estimator of q_0, say $\hat{q}_\alpha$, is the maximizer of Equation (3.13).

Comparing with the maximizing criterion (3.11), one may notice that the constant $2\hat{\delta}_{.}$ is replaced by the constant $a_\alpha \hat{\delta}_{.}$, which is a function of the level of significance α and the estimated average of the eigenvalues $\hat{\delta}_{.}$. That is because we want to conduct the test using the maximizer $\hat{q}_\alpha$ directly for a pre-specified level of significance α. Recall that the limiting probability of the Type I error for the proposed estimator $\hat{q}$ is 0.29. Thus, a_α is used to control the Type I error in this test. According to Eubank and Hart (1992) and Eubank (1997), the value of a_α, for a given α, is the solutions of the equations

$$1 - \alpha = \exp\left\{ -\sum_{k=1}^{\infty} \frac{P(\chi_k^2 > k a_\alpha)}{k} \right\} \tag{3.14}$$

or

$$P\left(\max_{1 \le k \le K-1} \left[\frac{1}{k} \sum_{j=1}^{k} Z_j^2 \right] \ge a_\alpha \right) = \alpha \tag{3.15}$$

where χ_k^2 indicates 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 (3.14). However, we noticed by simulations that a_α converges to the desired value quickly for $K > 10$. Finally, the level α test is conducted by rejecting the null hypothesis (3.1) or (3.9) if $\hat{q}_\alpha > 0$ is obtained.

3.2 Asymptotic Properties of Estimators in Stratified Sampling

In Section 3.1, we proposed two Neyman smooth-type GOF tests incorporated with order selection for use in complex surveys. In this section, we will examine the asymptotic properties of these two estimators in stratified sampling. We will first discuss the effective sample size, then we discuss the limiting distribution of the Fourier coefficients and the asymptotic properties of the estimators. The asymptotic properties of the estimators in cluster sampling will be examined in the future.

3.2.1 Effective Sample Size

In this section, we will investigate the relationship between the effective sample size and the design effects matrix $\mathbf{P}_0^{-1}\mathbf{V}$, where $\mathbf{P}_0$ and $\mathbf{V}$ are defined in (2.6) and (B.5). Recall that the Kish's effective sample size introduced in (B.3) is as follows,

$$\tilde{n} \quad = \quad \frac{\left(\sum_{j=1}^{n} w_j\right)^2}{\sum_{j=1}^{n}, w_j^2}.$$

We want to build a relationship between the Kish's effective sample size $\tilde{n}$ and the observed sample size n. We start with the test statistic of the first order corrected test $X_C^2 = \frac{X^2}{\delta_.}$. The following equations can be derived,

$$\begin{aligned}
\mathrm{E}\left(\frac{X^2}{\delta_.}\right) &= \mathrm{E}\left(\frac{n}{\delta_.}\sum_{k=1}^{K}\frac{\left(\hat{p}(k)-p_0(k)\right)^2}{p_0(k)}\right) \\
&= \frac{n}{\delta_.}\sum_{k=1}^{K}\frac{\mathrm{E}\left(\hat{p}(k)-p_0(k)\right)^2}{p_0(k)} \\
&= \frac{n}{\delta_.}\sum_{k=1}^{K}\frac{\mathrm{E}\left[\left(\hat{p}(k)-p(k)\right)+\left(p(k)-p_0(k)\right)\right]^2}{p_0(k)} \\
&= \frac{n}{\delta_.}\sum_{k=1}^{K}\left\{\frac{\mathrm{E}\left[\hat{p}(k)-p(k)\right]^2}{p_0(k)}+\frac{\mathrm{E}\left[p(k)-p_0(k)\right]^2}{p_0(k)}\right. \\
&\quad \left. +2\frac{\mathrm{E}\left[\left(\hat{p}(k)-p(k)\right)\left(p(k)-p_0(k)\right)\right]}{p_0(k)}\right\}.
\end{aligned}$$

Since $\hat{p}(k)$ is an unbiased estimator of $p(k)$, we have

$$\begin{aligned}
\mathrm{E}\left[(\hat{p}(k)-p(k))(p(k)-p_0(k))\right] &= (p(k)-p_0(k))\mathrm{E}(\hat{p}(k)-p(k)) \\
&= (p(k)-p_0(k))[\mathrm{E}(\hat{p}(k))-p(k)] \\
&= (p(k)-p_0(k))[p(k)-p(k)] \\
&= 0.
\end{aligned}$$

The following equations also hold

$$\begin{aligned}
\mathrm{E}\left[\hat{p}(k)-p(k)\right]^2 &= \mathrm{var}(\hat{p}(k)) \\
&= \frac{p(k)(1-p(k))}{\tilde{n}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{E}\left(\frac{X^2}{\delta_.}\right) &= \frac{n}{\delta_.}\sum_{k=1}^{K}\frac{\mathrm{var}(\hat{p}(k))+(p(k)-p_0(k))^2}{p_0(k)} \\
&= \frac{n}{\delta_.}\sum_{k=1}^{K}\frac{\frac{p(k)(1-p(k))}{\tilde{n}}+(p(k)-p_0(k))^2}{p_0(k)}.
\end{aligned}$$

Under the null hypothesis (3.1) $H_0 : p(k) = p_0(k)$ for $k = 1, \cdots, K$, we have

$$(p(k)-p_0(k))^2 = (p_0(k)-p_0(k))^2 = 0,$$

and

$$\frac{p(k)(1-p(k))}{\tilde{n}} = \frac{p_0(k)(1-p_0(k))}{\tilde{n}}.$$

As a result, under the null hypothesis (3.1), we have

$$\begin{aligned}
\mathrm{E}\left(\frac{X^2}{\delta_.}\right) &= \frac{n}{\delta_.}\sum_{k=1}^{K}\frac{p_0(k)(1-p_0(k))}{\tilde{n}}\frac{1}{p_0(k)} \\
&= \frac{n}{\delta_.}\sum_{k=1}^{K}\frac{(1-p_0(k))}{\tilde{n}} \\
&= \frac{n}{\delta_.}\frac{1}{\tilde{n}}\sum_{k=1}^{K}(1-p_0(k)) \\
&= \frac{n}{\delta_.}\frac{1}{\tilde{n}}(K-\sum_{k=1}^{K}p_0(k)) \\
&= \frac{n}{\delta_.}\frac{1}{\tilde{n}}(K-1).
\end{aligned}$$

Under the null hypothesis (3.1), the asymptotic distribution of X^2 is a linear combination of χ_1^2,

$$X^2 \sim \sum_{j=1}^{K-1}\delta_j\chi_1^2,$$

where χ_1^2 indicates independently central chi-squared random variable with 1 degree of freedom. We now have

$$\begin{aligned}
\mathrm{E}(X^2) &= \mathrm{E}\left[\sum_{j=1}^{K-1}\delta_j\chi_1^2\right] \\
&= \delta_1\mathrm{E}(\chi_1^2)+\cdots+\delta_{K-1}\mathrm{E}(\chi_1^2) \\
&= \delta_1+\cdots+\delta_{K-1} \\
&= \sum_{j=1}^{K-1}\delta_j = (K-1)\delta_..
\end{aligned}$$

Hence,

$$\mathrm{E}\left(\frac{X^2}{\delta_.}\right) = \frac{\mathrm{E}\,(X^2)}{\delta_.} = K - 1.$$

Therefore, under the null hypothesis (3.1), it can be proved that

$$\begin{aligned} \mathrm{E}\left(\frac{X^2}{\delta_.}\right) &= \frac{n}{\delta_.}\frac{1}{\tilde{n}}(K-1) \\ &= K - 1. \end{aligned}$$

By solving the equation above, it can be derived that

$$\tilde{n} = \frac{n}{\delta_.}. \tag{3.16}$$

This relationship plays a crucial role when proving the theorems proposed in this chapter. In practice, our tests can be conducted using the observed sample size n with the estimator of $\delta_.$. In proofs of the asymptotic properties of the test statistics, Kish's effective sample size is used. The connection between $\tilde{n}$ and n is a bridge for the theorems and the proposed test procedures.

3.2.2 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*

$$\pi_{it} = p(\textit{psu } i \textit{ is in the sample, using population } \mathcal{U}_t)$$

and

$$\pi_{ijt} = p(\textit{psu } i \textit{ and psu } j \textit{ are both in the sample, using population } \mathcal{U}_t)$$

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*

$$0 < c_2 < \pi_{it} < c_1 < 1 \tag{3.17}$$

for all i *and any superpopulation in the sequence. Also assume there exists an* α_t *with* $\alpha_t = o(1)$ *such that*

$$\pi_{it}\pi_{jt} - \pi_{ijt} \leq \alpha_t \pi_{it}\pi_{jt}. \tag{3.18}$$

Let w_j *denote the sampling weight of the* j*th observation, for* $j = 1, \cdots, n$ *and denote the Kish's effective sample size as* $\tilde{n} = \frac{(\sum_{j=1}^n w_j)^2}{\sum_{j=1}^n w_j^2}$. *We have*

$$\sqrt{\tilde{n}} \begin{pmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \\ \vdots \\ b_{K-1} - \beta_{K-1} \end{pmatrix} \to N_{K-1}(\mathbf{0}, \mathbf{V}), \text{ as } \tilde{n} \to \infty,$$

where $\mathbf{V} = \{v_{ij} - \beta_i\beta_j\}$ *is a* $(K-1) \times (K-1)$ *covariance matrix with* ij*th entry* $(v_{ij} - \beta_i\beta_j)$ *and*

$$v_{ij} = \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p(k)}{p_0(k)}, \; i, j = 1, \cdots, K-1.$$

Proof. In this proof, we will first derive the variance-covariance matrix of the b_j's. Next, we use the finite population Central Limit Theorem (CLT) to show the asymptotic distribution.

In stratified sampling, the estimated proportion of the kth category is

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

where w_j is the sampling weight of the jth observation and $y_j(k)$'s are uncorrelated random variables defined as follows,

$$y_j(k) = \begin{cases} 1 & \text{If } j\text{th observation falls in the } k\text{th category,} \\ 0 & \text{Otherwise,} \end{cases} \qquad j = 1, \cdots, n.$$

In stratified sampling, it has been shown in Appendix B that

$$\mathrm{E}[\hat{p}(k)] = p(k), \ \mathrm{var}[\hat{p}(k)] = \frac{p(k)(1 - p(k))}{\tilde{n}},$$

and

$$\mathrm{cov}(\hat{p}(k), \hat{p}(l)) = -\frac{p(k)p(l)}{\tilde{n}}, \ k \neq l.$$

Now, in order to decompose the chi-squared test statistic to perpendicular components, we need a basis function that satisfies the orthogonality conditions (3.2) and (3.3). Following Eubank (1997), we choose equation (3.4) for the null hypothesis (3.1), which satisfies both of the orthogonality conditions and changes sign j times for each $k = 1, \cdots, K - 1$.

It can be shown that $\hat{f}(k)$ defined in equation (3.5) is an unbiased estimator of $f(k)$ in equation (3.7) for $k = 1, \cdots, K$, and the associated (discrete) generalized Fourier coefficients b_j in equation (3.6) of $\hat{f}(k)$ is also unbiased for β_j in equation (3.8), the associated (discrete) generalized Fourier coefficients of $f(k)$ for $j = 1, \cdots, K-1$.

$$\begin{aligned} \mathrm{E}[\hat{f}(k)] &= \mathrm{E}\left[\frac{\hat{p}(k) - p_0(k)}{\sqrt{p_0(k)}}\right] = \frac{\mathrm{E}[\hat{p}(k)] - p_0(k)}{\sqrt{p_0(k)}} \\ &= \frac{p(k) - p_0(k)}{\sqrt{p_0(k)}} = f(k), \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}(b_j) &= \mathrm{E}\left[\sum_{k=1}^{K} \hat{f}(k)x_j(k)\right] = \sum_{k=1}^{K} \mathrm{E}[\hat{f}(k)]x_j(k) \\ &= \sum_{k=1}^{K} f(k)x_j(k) = \beta_j. \end{aligned}$$

Now, we want to derive the covariance matrix $\mathbf{V}$ of b_j's. Suppose there are a total of K categories in multinomial data under stratified sampling, so that there are $(K-1)$ b_j's, that are all consistent with their corresponding parameters β_j's, that is, $\mathrm{E}[b_j] = \beta_j$ for $j = 1, \cdots, K-1$. Since b_j can be expanded as

$$\begin{aligned} b_j &= \sum_{k=1}^{K} \hat{f}(k)x_j(k) \\ &= \hat{f}(1)x_j(1) + \cdots + \hat{f}(K)x_j(K) \\ &= \frac{\hat{p}(1) - p_0(1)}{\sqrt{p_0(1)}}x_j(1) + \cdots + \frac{\hat{p}(K) - p_0(K)}{\sqrt{p_0(K)}}x_j(K), \end{aligned}$$

the variance of b_j can be written as

$$\begin{aligned} \mathrm{var}(b_j) &= \mathrm{var}\left[\frac{\hat{p}(1) - p_0(1)}{\sqrt{p_0(1)}}x_j(1) + \cdots + \frac{\hat{p}(K) - p_0(K)}{\sqrt{p_0(K)}}x_j(K)\right] \\ &= \frac{[x_j(1)]^2}{p_0(1)}\mathrm{var}\left[\hat{p}(1) - p_0(1)\right] + \cdots + \frac{[x_j(K)]^2}{p_0(K)}\mathrm{var}\left[\hat{p}(K) - p_0(K)\right] \\ &\quad +2\sum_{k \neq l} \mathrm{cov}\left[\frac{\hat{p}(k) - p_0(k)}{\sqrt{p_0(k)}}x_j(k), \frac{\hat{p}(l) - p_0(l)}{\sqrt{p_0(l)}}x_j(l)\right] \\ &= \sum_{m=1}^{K} \frac{[x_j(m)]^2}{p_0(m)}\mathrm{var}\left[\hat{p}(m) - p_0(m)\right] \\ &\quad +2\sum_{k \neq l} \mathrm{cov}\left[\frac{\hat{p}(k) - p_0(k)}{\sqrt{p_0(k)}}x_j(k), \frac{\hat{p}(l) - p_0(l)}{\sqrt{p_0(l)}}x_j(l)\right] \end{aligned}$$

Note that

$$\text{var}\,[\hat{p}(m) - p_0(m)] = \text{var}\,[\hat{p}(m)] = \frac{p(m)(1-p(m))}{\tilde{n}},\ m = 1, \cdots K, \tag{3.19}$$

and

$$\begin{aligned}
& \text{cov}\left[\frac{\hat{p}(k) - p_0(k)}{\sqrt{p_0(k)}} x_j(k), \frac{\hat{p}(l) - p_0(l)}{\sqrt{p_0(l)}} x_j(l)\right] \\
= & \frac{x_j(k)x_j(l)}{\sqrt{p_0(k)p_0(l)}} \text{cov}[\hat{p}(k), \hat{p}(l)] \\
= & -\frac{x_j(k)x_j(l)}{\sqrt{p_0(k)p_0(l)}} \frac{p(k)p(l)}{\tilde{n}},\ k \neq l. \qquad (3.20)
\end{aligned}$$

(3.19) and (3.20) results in

$$\text{var}(b_j) = \sum_{m=1}^{K} \frac{[x_j(m)]^2}{p_0(m)} \frac{p(m)(1-p(m))}{\tilde{n}} - 2\sum_{k \neq l} \frac{x_j(k)x_j(l)}{\sqrt{p_0(k)p_0(l)}} \frac{p(k)p(l)}{\tilde{n}}.$$

By expanding $p(m)(1 - p(m))$, $\text{var}(b_j)$ turns out to be

$$\begin{aligned}
\text{var}(b_j) = & \frac{1}{\tilde{n}}\left\{\sum_{m=1}^{K} \frac{[x_j(m)]^2}{p_0(m)} p(m)\right\} - \frac{1}{\tilde{n}}\left\{\sum_{m=1}^{K} \frac{[x_j(m)]^2}{p_0(m)} (p(m))^2\right\} \\
& -\frac{1}{\tilde{n}}\left\{2\sum_{k \neq l} \frac{x_j(k)x_j(l)}{\sqrt{p_0(k)p_0(l)}}\right\},
\end{aligned}$$

and again, the last two terms form a complete square. Thus,

$$\text{var}(b_j) = \frac{1}{\tilde{n}}\left\{\sum_{k=1}^{K} \frac{[x_j(k)]^2}{p_0(k)} p(k) - \left[\sum_{k=1}^{K} \frac{x_j(k)}{\sqrt{p_0(k)}} p(k)\right]^2\right\}.$$

By the second orthogonality condition,

$$0 = \sum_{k=1}^{K} x_j(k)\sqrt{p_0(k)} = \sum_{k=1}^{K} \frac{x_j(k)}{\sqrt{p_0(k)}} p_0(k),\ \text{for } p_0(k) \neq 0,$$

the variance maintains the same if 0 is subtracted

$$
\begin{aligned}
\operatorname{var}(b_j) &= \frac{1}{\tilde{n}}\left\{\sum_{k=1}^{K}\frac{[x_j(k)]^2}{p_0(k)}p(k) - \left[\sum_{k=1}^{K}\frac{x_j(k)}{\sqrt{p_0(k)}}p(k) - 0\right]^2\right\} \\
&= \frac{1}{\tilde{n}}\left\{\sum_{k=1}^{K}\frac{[x_j(k)]^2}{p_0(k)}p(k) - \left[\sum_{k=1}^{K}\frac{x_j(k)}{\sqrt{p_0(k)}}p(k) - \sum_{k=1}^{K}\frac{x_j(k)}{\sqrt{p_0(k)}}p_0(k)\right]^2\right\} \\
&= \frac{1}{\tilde{n}}\left\{\sum_{k=1}^{K}\frac{[x_j(k)]^2}{p_0(k)}p(k) - \left[\sum_{k=1}^{K}\frac{x_j(k)}{\sqrt{p_0(k)}}(p(k) - p_0(k))\right]^2\right\}.
\end{aligned}
$$

By the definitions of

$$
\sum_{k=1}^{K}\frac{[x_j(k)]^2}{p_0(k)}p(k) = v_{jj}
$$

and

$$
\begin{aligned}
\sum_{k=1}^{K}\frac{x_j(k)}{\sqrt{p_0(k)}}(p(k) - p_0(k)) &= \sum_{k=1}^{K}\frac{p(k) - p_0(k)}{\sqrt{p_0(k)}}x_j(k) \\
&= \sum_{k=1}^{K} f(k)x_j(k) \\
&= \beta_j,
\end{aligned}
$$

the variance is then summarized as

$$
\begin{aligned}
\operatorname{var}(b_j) &= \frac{1}{\tilde{n}}\left\{\sum_{k=1}^{K}\frac{[x_j(k)]^2}{p_0(k)}p(k) - \left[\sum_{k=1}^{K}\frac{x_j(k)}{\sqrt{p_0(k)}}(p(k) - p_0(k))\right]^2\right\} \\
&= \frac{1}{\tilde{n}}(v_{jj} - \beta_j^2),
\end{aligned}
$$

and the variance of b_j is derived for all $j = 1, \cdots, K-1$.

The next step is to derive the covariance between b_i and b_j for $i \neq j$ and $i, j = 1, \cdots, K-1$.

$$\begin{aligned}
\operatorname{cov}(b_i, b_j) &= \operatorname{cov}\left(\sum_{k=1}^{K} \hat{f}(k) x_i(k), \sum_{k=1}^{K} \hat{f}(k) x_j(k)\right) \\
&= \operatorname{cov}\left(\sum_{k=1}^{K} \frac{\hat{p}(k)-p_0(k)}{\sqrt{p_0(k)}} x_i(k), \sum_{k=1}^{K} \frac{\hat{p}(k)-p_0(k)}{\sqrt{p_0(k)}} x_j(k)\right) \\
&= \sum_{m=1}^{K} \operatorname{cov}\left(\frac{\hat{p}(m)-p_0(m)}{\sqrt{p_0(m)}} x_i(k), \frac{\hat{p}(m)-p_0(m)}{\sqrt{p_0(m)}} x_j(k)\right) \\
&\quad + \sum_{k \neq l} \operatorname{cov}\left(\frac{\hat{p}(k)-p_0(k)}{\sqrt{p_0(k)}} x_i(k), \frac{\hat{p}(l)-p_0(l)}{\sqrt{p_0(l)}} x_j(l)\right),
\end{aligned}$$

with the first term

$$\begin{aligned}
& \sum_{m=1}^{K} \operatorname{cov}\left(\frac{\hat{p}(m)-p_0(m)}{\sqrt{p_0(m)}} x_i(m), \frac{\hat{p}(m)-p_0(m)}{\sqrt{p_0(m)}} x_j(m)\right) \\
= & \frac{x_i(m) x_j(m)}{p_0(m)} \sum_{m=1}^{K} \operatorname{cov}\left(\hat{p}(m)-p_0(m), \hat{p}(m)-p_0(m)\right) \\
= & \frac{x_i(k) x_j(m)}{p_0(m)} \sum_{m=1}^{K} \operatorname{var}[\hat{p}(m)] \\
= & \sum_{m=1}^{K} \frac{x_i(m) x_j(m)}{p_0(m)} \frac{p(m)(1-p(m))}{\tilde{n}},
\end{aligned}$$

and the second term

$$\begin{aligned}
& \sum_{k \neq l} \operatorname{cov}\left(\frac{\hat{p}(k)-p_0(k)}{\sqrt{p_0(k)}} x_i(k), \frac{\hat{p}(l)-p_0(l)}{\sqrt{p_0(l)}} x_j(l)\right) \\
= & \sum_{k \neq l} \frac{x_i(k) x_j(l)}{\sqrt{p_0(k) p_0(l)}} \operatorname{cov}(\hat{p}(k), \hat{p}(l)) \\
= & -\sum_{k \neq l} \frac{x_i(k) x_j(l)}{\sqrt{p_0(k) p_0(l)}} \frac{p(k) p(l)}{\tilde{n}}.
\end{aligned}$$

Consequently, the covariance is as follows,

$$\begin{aligned}&\operatorname{cov}(b_i, b_j)\\ = \quad & \sum_{k=1}^{K} \frac{x_i(m)x_j(m)}{p_0(m)} \frac{p(m)(1-p(m))}{\tilde{n}} - \sum_{k \neq l} \frac{x_i(k)x_j(l)}{\sqrt{p_0(k)p_0(l)}} \frac{p(k)p(l)}{\tilde{n}}.\end{aligned}$$

By opening the parentheses of $p(m)(1-p(m))$, we observe that

$$\begin{aligned}\operatorname{cov}(b_i, b_j) \quad = \quad & \frac{1}{\tilde{n}} \sum_{m=1}^{K} \frac{x_i(m)x_j(m)p(m)}{p_0(m)} - \left\{ \frac{1}{\tilde{n}} \sum_{m=1}^{K} \frac{x_i(m)x_j(m)(p(m))^2}{p_0(m)} \right.\\ & \left. + \frac{1}{\tilde{n}} \sum_{k \neq l} \frac{x_i(k)x_j(l)}{\sqrt{p_0(k)p_0(l)}} p(k)p(l) \right\},\end{aligned}$$

whose terms in the curly brackets can be a product of two factors. As a result, the covariance is

$$\begin{aligned}\operatorname{cov}(b_i, b_j) \quad = \quad & \frac{1}{\tilde{n}} \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p(k)}{p_0(k)}\\ & - \frac{1}{\tilde{n}} \left[\sum_{k=1}^{K} \frac{p(k)}{\sqrt{p_0(k)}} x_i(k)) \right] \left[\sum_{k=1}^{K} \frac{p(k)}{\sqrt{p_0(k)}} x_j(k) \right].\end{aligned}$$

By the second orthogonality condition again,

$$0 = \sum_{k=1}^{K} x_j(k) \sqrt{p_0(k)} = \sum_{k=1}^{K} \frac{p_0(k)}{\sqrt{p_0(k)}} x_j(k)$$

for $p_0(k) \neq 0$, it turns out that

$$\begin{aligned}
\operatorname{cov}(b_i, b_j) &= \frac{1}{\tilde{n}} \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p(k)}{p_0(k)} \\
&\quad - \frac{1}{\tilde{n}} \left[\sum_{k=1}^{K} \frac{p(k)}{\sqrt{p_0(k)}} x_i(k) - \sum_{k=1}^{K} \frac{p_0(k)}{\sqrt{p_0(k)}} x_i(k) \right] \\
&\quad \times \left[\sum_{k=1}^{K} \frac{p(k)}{\sqrt{p_0(k)}} x_j(k) - \sum_{k=1}^{K} \frac{p_0(k)}{\sqrt{p_0(k)}} x_j(k) \right] \\
&= \frac{1}{\tilde{n}} \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p(k)}{p_0(k)} \\
&\quad - \frac{1}{\tilde{n}} \left[\sum_{k=1}^{K} \frac{p(k) - p_0(k)}{\sqrt{p_0(k)}} x_i(k) \right] \left[\sum_{k=1}^{K} \frac{p(k) - p_0(k)}{\sqrt{p_0(k)}} x_j(k) \right]
\end{aligned}$$

Since $v_{ij} = \sum_{k=1}^{K} x_i(k)x_j(k)p(k)/p_0(k)$ and (3.8), the covariance is finalized as

$$\operatorname{cov}(b_i, b_j) = \frac{1}{\tilde{n}} \left\{ v_{ij} - \beta_i \beta_j \right\}, \text{ for all } i \neq j, i, j = 1, \cdots, K-1.$$

Now we want to show the asymptotic distributions of the b_j's. We will first show the conditions in the consistency and asymptotic normality theorem of Isaki and Fuller (1982) are met for the samples. Conditions (3.17) and (3.18) imply that the inclusion probabilities π_{it} and joint probabilities π_{ijt} are uniformly bounded away from 0, and the inverses of the inclusion probabilities $(\pi_{it})^{-1}$ are uniformly bounded away from 0. This ensures that every psu in a sampling frame has a positive probability of being included in the sample.

In our situation, observed values y_i's are 0 or 1. Therefore, under conditions (3.17) and (3.18), $|(\pi_i)^{-1} y_i|$ is bounded. Thus, the condition in Lemma 1 and 2 of Isaki and Fuller (1982) are met. As a result, b_j's are consistent and asymptotically normal with

$$\sqrt{\tilde{n}}\begin{pmatrix} b_1-\beta_1 \\ b_2-\beta_2 \\ \vdots \\ b_{K-1}-\beta_{K-1} \end{pmatrix} \rightarrow N_{K-1}(\mathbf{0},\mathbf{V}),\ \text{as } \tilde{n}\rightarrow\infty,$$

where the ijth entry of the $(K-1)\times(K-1)$ covariance matrix $\mathbf{V}$ is $(v_{ij}-\beta_i\beta_j)$ with

$$v_{ij}=\sum_{k=1}^{K}\frac{x_i(k)x_j(k)p(k)}{p_0(k)},\ \text{for } i,j=1,\cdots,K-1.$$

In addition, we replace $\tilde{n}$ with $n/\delta_.$ by (3.16), and the Theorem 1 can be written as

$$\sqrt{\frac{n}{\delta_.}}\begin{pmatrix} b_1-\beta_1 \\ b_2-\beta_2 \\ \vdots \\ b_{K-1}-\beta_{K-1} \end{pmatrix} \rightarrow N_{K-1}(\mathbf{0},\mathbf{V}),\ \text{as } n\rightarrow\infty,$$

which is equivalent to

$$\sqrt{n}\begin{pmatrix} (b_1-\beta_1)/\sqrt{\delta_.} \\ (b_2-\beta_2)/\sqrt{\delta_.} \\ \vdots \\ (b_{K-1}-\beta_{K-1})/\sqrt{\delta_.} \end{pmatrix} \rightarrow N_{K-1}(\mathbf{0},\mathbf{V}),\ \text{as } n\rightarrow\infty. \tag{3.21}$$

If the sampling design is SRS, $\delta_.$ is 1, and this theorem can go back to the statement in Eubank (1997). The proof of the this theorem is thus completed. □

3.2.3 Asymptotic Distribution of X_q^2 Under H_0^*

Theorem 2. *Under stratified sampling, the conditions of Theorem 1, and the null hypothesis H_0 (3.1) or H_0^* (3.9), the Neyman smooth-type GOF test statistic $X_q^2 = \sum_{j=1}^{q} \tilde{n} b_j^2$ is asymptotically distributed as the central chi-squared distribution with q degrees of freedom.*

Proof. According to Theorem 1, the asymptotic distribution of the vector $\tilde{n}(b_1 - \beta_1, \cdots, b_{K-1} - \beta_{K-1})^{\mathrm{T}}$ is a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V} = \{v_{ij} - \beta_i\beta_j\}$ where $v_{ij} = \sum_{k=1}^{K} \frac{x_j(k)x_j(k)p(k)}{p_0(k)}$ for $i, j = 1, \cdots, K-1$ under stratified sampling. Since all β_j's are 0 and $p(k) = p_0(k)$ under the null hypothesis, by the orthogonality condition (3.2) , it can be calculated that

$$
\begin{aligned}
v_{jj} &= \sum_{k=1}^{K} \frac{x_j(k)^2 p(k)}{p_0(k)} = \sum_{k=1}^{K} \frac{x_j(k)^2 p_0(k)}{p_0(k)} = \sum_{k=1}^{K} x_j(k)^2 \\
&= 1, \text{ for } i = j,
\end{aligned}
$$

and

$$
\begin{aligned}
v_{ij} &= \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p(k)}{p_0(k)} = \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p_0(k)}{p_0(k)} = \sum_{k=1}^{K} x_i(k)x_j(k) \\
&= 0, \text{ for } i \neq j.
\end{aligned}
$$

As a result, the entries on the diagonal of the covariance matrix $\mathbf{V}$ are

$$v_{jj} - \beta_j^2 = 1,$$

and those off the diagonal are

$$v_{ij} - \beta_i\beta_j = 0.$$

Thus, under the null hypothesis $H_0^* : \beta_1 = \cdots \beta_{K-1} = 0$,

$$\sqrt{\tilde{n}} \begin{pmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \\ \vdots \\ b_{K-1} - \beta_{K-1} \end{pmatrix} = \sqrt{\tilde{n}} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{K-1} \end{pmatrix} \rightarrow N_{K-1}(\mathbf{0}, \mathbf{V}), \text{ as } \tilde{n} \rightarrow \infty,$$

where the $(K-1) \times (K-1)$ covariance matrix is

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Therefore, $\sqrt{\tilde{n}} b_j$ is asymptotically distributed as a standard normal distribution, that is

$$\sqrt{\tilde{n}} b_j \rightarrow N(0,1), \text{ as } \tilde{n} \rightarrow \infty, \text{ for } j = 1, \cdots, K-1.$$

Also, since $\text{cov}(\sqrt{\tilde{n}} b_i, \sqrt{\tilde{n}} b_j) = 0$, $\sqrt{\tilde{n}} b_i$ and $\sqrt{\tilde{n}} b_j$ are asymptotically independent normal random variables, which implies that $\tilde{n} b_i^2$ and $\tilde{n} b_j^2$ are asymptotically independent central chi-squared random variables with 1 degree of freedom, for $i \neq j$ and $i, j = 1, \cdots, K-1$.

Consequently, the Neyman smooth-type test statistic $X_q^2 = \sum_{j=1}^{q} \tilde{n} b_j^2$ is asymptotically distributed as central chi-squared distribution with q degrees of freedom, that is,

$$X_q^2 = \sum_{j=1}^{q} \tilde{n} b_j^2 \rightarrow \chi_q^2, \text{ as } \tilde{n} \rightarrow \infty$$

under the null hypothesis (3.9) and stratified sampling.

Furthermore, consider the asymptotic distribution (3.21). Under the null hypothesis (3.9), $\mathbf{V}$ is an $(K-1) \times (K-1)$ identity covariance matrix. Therefore, $\sqrt{n}(b_j/\sqrt{\delta_.})$'s are asymptotically independent standard normal random variables, for $j = 1, \cdots, K-1$. As a result, $n(b_j^2/\delta_.)$'s are asymptotically independent central chi-squared random variables with 1 degree of freedom, for $j = 1, \cdots, K-1$. So, under the null hypothesis (3.9),

$$\sum_{j=1}^{q} n \frac{b_j^2}{\delta_.} \to \chi_q^2, \text{ as } n \to \infty,$$

which is the same result as proposed in Theorem 2 because $n/\delta_. = \tilde{n}$. Under the null hypothesis (3.9) and stratified sampling, this result can also be written as

$$\frac{\sum_{j=1}^{q} n b_j^2}{\delta_.} \to \chi_q^2, \text{ as } n \to \infty,$$

if the sampling design is SRS, i.e. $\delta_. = 1$, the asymptotic distribution goes back to the one in Eubank (1997). □

3.2.4 Asymptotic Properties of $\hat{q}$

As discussed in Section 3.1.1, suppose the unknown order of the test statistic $X_q^2 = \sum_{j=1}^{q} \tilde{n} b_j^2$ is denoted by q_0. A good estimation of q is the minimizer of $\sum_{k=1}^{K} (f_q(k) - f(k))^2$, where $f_q = \sum_{j=1}^{q} b_j x_j$, or equivalently, according to Eubank (1997), the maximizer of the criterion $M(q) = -\sum_{j=1}^{q} b_j^2 + 2\sum_{j=1}^{q} \beta_j b_j$. Hart (1985) shows that an unbiased estimator of $M(q)$ is

$$\hat{M}(q) = \frac{\tilde{n}+1}{\tilde{n}-1} \sum_{j=1}^{q} b_j^2 - \frac{2}{\tilde{n}-1} \sum_{j=1}^{q} \hat{v}_{jj}, \text{ for } q = 1, \cdots, K-1, \tag{3.22}$$

with $\hat{M}(0) = 0$ and $\hat{v}_{jj} = \sum_{k=1}^{K} \frac{x_j(k)^2 \hat{p}(k)}{p_0(k)}$. The estimator of q_0, denoted as $\hat{q}$, is obtained by the maximizer of the criterion $\hat{M}(q)$.

In this section, we derive the asymptotic properties of $\hat{q}$ used in our proposed test W in Section 3.1.2 under stratified sampling.

Theorem 3. *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 (3.9),

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

In addition, under alternative hypothesis (3.10),

$$\lim_{\tilde{n}\to\infty} P(\hat{q} < q_0) = 0, \text{ and } \lim_{\tilde{n}\to\infty} P(\hat{q} = q_0 + r) = P(r^* = r),\ r = 0, \cdots, K - q_0 - 1, \tag{3.23}$$

where r^* is the maximizer of the criterion,

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

with $R(0) = 0$, and $(Z_1, \cdots, Z_{K-q_0-1})^{\mathrm{T}}$ a vector of normal random variables with mean $\mathbf{0}$ and covariance

$$\operatorname{cov}(Z_i, Z_j) = \frac{v_{(i+q_0)(j+q_0)}}{\sqrt{v_{(i+q_0)(i+q_0)} v_{(j+q_0)(j+q_0)}}}.$$

Proof. Let q^* be the maximizer of the criterion

$$M^*(q) = \sum_{j=1}^{q} (\tilde{n} b_j^2 - a v_{jj}), \tag{3.24}$$

where a is a pre-specified positive number that may increase with $\tilde{n}$. As argued in the appendix of Eubank (1997), the maximizer $\hat{q}$ of $\hat{M}(q)$ in (3.11) has the same limiting distribution as q^* when $a = 2$. We will prove the theorem by three cases: case 1, $q < q_0$ for $q_0 > 0$; case 2, $q \geq q_0$ for $q_0 = 0$; and case 3, $q \geq q_0$ for $q_0 > 0$.

Case 1: we first work on the underselection case $q < q_0$ for $q_0 > 0$. Since q^* is the maximizer of $M^*(q)$, we have

$$\begin{aligned}
P(q^* = q) &= P(M^*(q) \geq M^*(l), \quad l = 0, \cdots, K-1) \\
&= P\left(\sum_{j=1}^{q} \tilde{n} b_j^2 - \sum_{j=1}^{l} \tilde{n} b_j^2 \geq \sum_{j=1}^{q} a v_{jj} - \sum_{j=1}^{l} a v_{jj}, \quad l = 0, \cdots, K-1 \right) \\
&= P\left(\sum_{j=1}^{q} (\tilde{n} b_j^2 - a v_{jj}) - \sum_{j=1}^{l} (\tilde{n} b_j^2 - a v_{jj}) \geq 0, \quad l = 0, \cdots, K-1 \right) \\
&= P\Bigg(\sum_{j=l+1}^{q} (\tilde{n} b_j^2 - a v_{jj}) \geq 0, \quad l = 0, \cdots, q-1, i.e., l < q, \\
&\qquad \sum_{j=q+1}^{l} (\tilde{n} b_j^2 - a v_{jj}) \leq 0, \quad l = q+1, \cdots, K-1, i.e., l > q \Bigg).
\end{aligned} \tag{3.25}$$

Notice fact that $P(A \cap B) \leq P(A)$ for any subsets A and B, which results in

$$\begin{aligned}
P(q^* = q) &\leq P\left(\sum_{j=q+1}^{l} (\tilde{n} b_j^2 - a v_{jj}) \leq 0, \quad l = q+1, \cdots, K-1 (\text{i.e.}, l > q) \right) \\
&= P\left(\sum_{j=q+1}^{q_0} (\tilde{n} b_j^2 - a v_{jj}) \leq 0, \quad l = q_0 > q \right) \\
&= P\left(\sum_{j=q+1}^{q_0} \tilde{n} b_j^2 \leq \sum_{j=q+1}^{q_0} a v_{jj}, \quad q < q_0 \right)
\end{aligned}$$

Since all $\tilde{n}b_j^2$'s are non-negative,

$$P\left(\sum_{j=q+1}^{q_0} \tilde{n}b_j^2 \leq \sum_{j=q+1}^{q_0} av_{jj}, \quad q < q_0\right) \;\leq\; P\left(\tilde{n}b_{q_0}^2 \leq \sum_{j=q+1}^{q_0} av_{jj}, \quad q < q_0\right).$$

Now, let $D = \sum_{j=q+1}^{q_0} av_{jj}$ for $q < q_0$. According to Theorem 1,

$$\sqrt{\tilde{n}}(b_{q_0} - \beta_{q_0}) \;\rightarrow\; N(0, v_{q_0q_0}), \text{ as } \tilde{n} \rightarrow \infty.$$

Consequently, for $q < q_0$,

$$\begin{aligned}
& P\left(\tilde{n}b_{q_0}^2 \leq \sum_{j=q+1}^{q_0} av_{jj}\right) = P\left(\tilde{n}b_{q_0}^2 \leq D\right) \\
= \;& P\left(-\sqrt{D} \leq \sqrt{\tilde{n}}b_{q_0} \leq \sqrt{D}\right) \\
= \;& P\left(-\sqrt{D} - \sqrt{\tilde{n}}\beta_{q_0} \leq \sqrt{\tilde{n}}b_{q_0} - \sqrt{\tilde{n}}\beta_{q_0} \leq \sqrt{D} - \sqrt{\tilde{n}}\beta_{q_0}\right) \\
= \;& P\left(-\sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}} \leq \frac{\sqrt{\tilde{n}}b_{q_0} - \sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}} \leq \sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}}\right) \\
= \;& P\left(-\sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}} \leq Z \leq \sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}}\right) \\
\leq \;& P\left(Z \leq \sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}}\right).
\end{aligned}$$

At this point, for $q < q_0$, we have

$$P(q^* = q) \leq \;\; P\left(nb_{q_0}^2 \leq D\right) \;\; \leq P\left(Z \leq \sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}}\right).$$

Now, we need to use a result given in Feller (1968, pg. 175), which states that, for $x > 0$,

$$\frac{(x^{-1} - x^{-3})\exp(-\frac{x^2}{2})}{\sqrt{2\pi}} \leq P(Z > x) \leq \frac{x^{-1}\exp(-\frac{x^2}{2})}{\sqrt{2\pi}}.$$

Thus, by the symmetry of the normal distribution, we have

$$\begin{aligned}
& P\left(Z \leq \sqrt{\frac{D}{v_{q_0q_0}}} - \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}}\right) \\
= \; & P\left(Z \geq \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}} - \sqrt{\frac{D}{v_{q_0q_0}}}\right) \\
\leq \; & \frac{x^{-1}\exp(-\frac{x^2}{2})}{\sqrt{2\pi}},
\end{aligned}$$

where

$$x = \frac{\sqrt{\tilde{n}}\beta_{q_0}}{\sqrt{v_{q_0q_0}}} - \sqrt{\frac{D}{v_{q_0q_0}}}.$$

for our case. Therefore, for $q < q_0$,

$$P(q^* = q) \quad = \quad O\left(\tilde{n}^{-\frac{1}{2}}\exp\left(-\frac{\tilde{n}\beta_{q_0}^2}{2v_{q_0q_0}}\right)\right),$$

which decays to 0 exponentially as $\tilde{n} \to \infty$. As a result, for $q < q_0$ and $q_0 > 0$, we have

$$\lim_{\tilde{n}\to\infty} P(q* = q) = 0.$$

Case 1 is proved, which implies that the limiting probability of underselecting q_0 is 0 under alternative hypothesis.

Case 2: We then move on to the case with $q \geq q_0$ under null hypothesis ($q_0 = 0$), In this proof, Lemma 2.1 in Eubank (1999, pg. 54) is needed, which states that
If $\{A_n\}$ and $\{B_n\}$ are sequences of sets with $P(B_n) \to 1$ as $n \to \infty$, then $P(A_n) - P(A_n \cap B_n) \to 0$, as $n \to \infty$.
Now, define

$$A_{\tilde{n}} \quad = \quad \{M^*(q) \geq M^*(l), \quad l = q_0, \cdots, K-1\}$$

and

$$B_{\tilde{n}} = \{M^*(q) \geq M^*(l), \quad l = 0, \cdots, q_0 - 1\}.$$

It is trivial that $P(B_{\tilde{n}}) \to 1$ as $\tilde{n} \to \infty$, since $M^*(q)$ is definitely larger than $M^*(l)$ for $l = 0, \cdots, q_0 - 1$ with $q \geq q_0$. As shown in the proof of the case 1, and according to the lemma above, it turns out that

$$\begin{aligned}
\lim_{\tilde{n}\to\infty} P(q^* = q) &= \lim_{\tilde{n}\to\infty} P(A_{\tilde{n}} \cap B_{\tilde{n}}) \\
&= \lim_{\tilde{n}\to\infty} P(A_{\tilde{n}}).
\end{aligned}$$

Thus, the probability of set $A_{\tilde{n}}$ can be decomposed as follow,

$$\begin{aligned}
P(A_{\tilde{n}}) &= P\left(M^*(q) \geq M^*(l), \quad l = q_0, \cdots, K-1\right) \\
&= P\left(\sum_{j=1}^{q}(\tilde{n}b_j^2 - av_{jj}) \geq \sum_{j=1}^{l}(\tilde{n}b_j^2 - av_{jj}), \quad l = q_0, \cdots, K-1\right) \\
&= P\left(\sum_{j=l+1}^{q}(\tilde{n}b_j^2 - av_{jj}) \geq 0, \quad l = q_0, \cdots, q-1,\right. \\
&\qquad \left.\sum_{j=q+1}^{l}(\tilde{n}b_j^2 - av_{jj}) \leq 0, \quad l = q+1, \cdots, K-1\right) \\
&= P\left(\sum_{j=l+1}^{q}\tilde{n}b_j^2 \geq \sum_{j=l+1}^{q} av_{jj}, \quad l = q_0, \cdots, q-1\right) \\
&\quad \times P\left(\sum_{j=q+1}^{l}\tilde{n}b_j^2 \leq \sum_{j=q+1}^{l} av_{jj}, \quad l = q+1, \cdots, K-1\right).
\end{aligned}$$

The last step of the equation works, because all $\sqrt{\tilde{n}}b_j$'s are independent standard normal random variables under null hypothesis by Theorem 2.

To accomplish the proof, we now define a sequence of sets S_j's as follows

$$S_j = \sum_{i=1}^{j}(\chi^2_{1i} - a)$$

where χ^2_{1i}'s are independent central chi-squared distributions with 1 degree of freedom. Then, after re-indexing the expression of $P(A_{\tilde{n}})$, it can be seen that

$$\begin{aligned} P(A_{\tilde{n}}) &= P\left(S_j \geq 0, \quad j = 1, \cdots, q - q_0\right) \\ &\quad \times P\left(S_j \leq 0, \quad j = 0, \cdots, K - 1 - q\right). \end{aligned}$$

As a result, by Spitzer (1956, pg. 329-330) and Shibata (1976), it turns out that

$$\begin{aligned} \lim_{\tilde{n}\to\infty} P(q^* = q) &= \lim_{\tilde{n}\to\infty} P(A_n) \\ &= \lim_{\tilde{n}\to\infty} P\left(S_j \geq 0, \quad j = 1, \cdots, q - q_0\right) \\ &\quad \times \lim_{\tilde{n}\to\infty} P\left(S_j \leq 0, \quad j = 0, \cdots, K - 1 - q\right) \\ &= c_{q-q_0} d_{K-1-q} \\ &= c_q d_{K-1-q}, \text{ for } q = 0, \cdots, K - 1, \end{aligned}$$

where c_r and d_r are defined as

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

and

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

with $c_0 = d_0 = 1$. The last step of the equation holds, since $q_0 = 0$ and $q \geq q_0$ are conditions in this case. Case 2 is then proved as a special case with $a = 2$, which provides the limiting probabilities of overselection and correct selection under the null model.

Case 3: our last case of the proof of this theorem involves the conditions $q \geq q_0$ and $q_0 > 0$. Let

$$q = q_0 + r, \text{ for } r = 0, 1, \cdots, K - q_0 - 1.$$

Then, $M^*(q)$ can be decomposed into two terms,

$$\begin{aligned} M^*(q) &= \sum_{j=1}^{q} \left(\tilde{n}b_j^2 - av_{jj}\right) \\ &= \sum_{j=1}^{q_0+r} \left(\tilde{n}b_j^2 - av_{jj}\right) \\ &= \sum_{j=1}^{q_0} \left(\tilde{n}b_j^2 - av_{jj}\right) + \sum_{j=q_0+1}^{q_0+r} \left(\tilde{n}b_j^2 - av_{jj}\right), \end{aligned}$$

where the first term is for $r = 0$ and the second term is for $r = 1, \cdots, K - q_0 - 1$. Note that the first term $\sum_{j=1}^{q_0} \left(\tilde{n}b_j^2 - av_{jj}\right)$ is actually fixed with respect to q_0, since q_0 is a constant.

Consequently, maximizing $M^*(q)$ is now equivalent to maximize the second term $\sum_{j=q_0+1}^{q_0+r} \left(\tilde{n}b_j^2 - av_{jj}\right)$ with respect to r, which implies that the maximizer r^* needs to be found so that the second term reaches its maximum. Therefore, for $r = 1, \cdots, K - q_0 - 1$, we organize the second term into the following form,

$$\begin{aligned} \text{second term} &= \sum_{j=q_0+1}^{q_0+r} \left(\tilde{n}b_j^2 - av_{jj}\right) \\ &= \sum_{j=q_0+1}^{q_0+r} v_{jj} \left(\frac{\tilde{n}b_j^2}{v_{jj}} - a\right). \end{aligned}$$

Now, recall that the conditions of this case are $q \geq q_0$ and $q_0 > 0$ which imply that $\beta_{q_0} = 0$ and $\beta_{q_0+1} = \beta_{q_0+2} = \cdots = \beta_{K-1} = 0$. According to Theorem 1, it can be obtained that

$$\sqrt{\tilde{n}} \begin{pmatrix} b_{q_0+1} \\ b_{q_0+2} \\ \vdots \\ b_{K-1} \end{pmatrix} \to N_{K-q_0-1}(\mathbf{0}, \mathbf{V}),$$

as $\tilde{n} \to \infty$, where $\mathbf{V} = \{v_{ij}\}$ is a $(K - q_0 - 1) \times (K - q_0 - 1)$ covariance matrix with

$$v_{ij} = \sum_{k=1}^{K} \frac{x_i(k)x_j(k)p(k)}{p_0(k)}, \text{ for } i, j = q_0 + 1, \cdots, K - 1.$$

Let

$$Z_j = \frac{\sqrt{\tilde{n}} b_j}{\sqrt{v_{jj}}}, \text{ for } j = q_0 + 1, \cdots, K - 1,$$

and define

$$R(r) = \sum_{j=q_0+1}^{q_0+r} v_{jj} \left(\frac{\tilde{n} b_j^2}{v_{jj}} - a \right), \text{ for } r = 1, \cdots, K - q_0 - 1.$$

Then, $R(r)$ is of the form

$$\begin{aligned} R(r) &= \sum_{j=q_0+1}^{q_0+r} v_{jj} \left(\frac{\tilde{n} b_j^2}{v_{jj}} - a \right) \\ &= \sum_{j=1}^{r} v_{(j+q_0)(j+q_0)} \left(\left[\frac{\sqrt{\tilde{n}} b_{(j+q_0)}}{\sqrt{v_{(j+q_0)(j+q_0)}}} \right]^2 - a \right) \\ &= \sum_{j=1}^{r} v_{(j+q_0)(j+q_0)} \left(Z_j^2 - a \right), \text{ for } r = 1, \cdots, K - q_0 - 1, \end{aligned}$$

with

$$\mathrm{E}(Z_j) = \mathrm{E}\left(\frac{\sqrt{\tilde{n}}b_j}{\sqrt{v_{jj}}}\right) = 0,$$

$$\mathrm{var}(Z_j) = \mathrm{var}\left(\frac{\sqrt{\tilde{n}}b_j}{\sqrt{v_{jj}}}\right) = 1,$$

and

$$\begin{aligned}
\mathrm{cov}(Z_i, Z_j) &= \mathrm{cov}\left(\frac{\sqrt{\tilde{n}}b_i}{\sqrt{v_{ii}}}, \frac{\sqrt{\tilde{n}}b_j}{\sqrt{v_{jj}}}\right) \\
&= \frac{v_{(i+q_0)(j+q_0)}}{\sqrt{v_{(i+q_0)(i+q_0)}v_{(j+q_0)(j+q_0)}}}.
\end{aligned}$$

Notice that if $r = 0$, $M^*(q) = M^*(q_0)$ is fixed for $q = q_0$, so it is reasonable to set $R(0) = 0$. Therefore, it is verified that

$$\begin{aligned}
\lim_{\tilde{n}\to\infty} P(q^* = q_0 + r) &= \lim_{\tilde{n}\to\infty} P(q^* - q_0 = r) \\
&= P(r^* = r), \text{ for } r = 0, 1, \cdots, K - q_0 - 1,
\end{aligned}$$

where r^* is the maximizer of function R

$$R(r) = \sum_{j=1}^{r} v_{(j+q_0)(j+q_0)} \left(Z_j^2 - a\right)$$

with $R(0) = 0$. When $a = 2$ is chosen, case 3 is proved. Similar to case 2, this part of the theorem also provides the limiting probability of overselection and correct selection, but under alternative hypothesis.

This proof provides theoretical support to our proposed test W in Section 3.1.2 under stratified sampling. But we use the observed sample size n in the test, instead of the Kish's effective sample size $\tilde{n}$. Based on (3.16) $\tilde{n} = \frac{n}{\delta_{.}}$, we now build the connection between this theorem and the proposed test W.

Note that, $\hat{q}$ has the same asymptotic properties as q^*, the maximizer of (3.24). If q^* is the maximizer of (3.24), it is also the maximizer of

$$\begin{aligned}
\delta_{.} M^*(q) &= \delta_{.} \sum_{j=1}^{q} (\tilde{n} b_j^2 - a v_{jj}) = \sum_{j=1}^{q} (\delta_{.} \tilde{n} b_j^2 - \delta_{.} a v_{jj}) \\
&= \sum_{j=1}^{q} (n b_j^2 - \delta_{.} a v_{jj})
\end{aligned}$$

because $\delta_{.}$ is always a positive real number. We consider the general maximizing criterion (3.25), which can be written as

$$\begin{aligned}
P(q^* = q) &= P(M^*(q) \geq M^*(l), \quad l = 0, \cdots, K-1) \\
&= P(\delta_{.} M^*(q) \geq \delta_{.} M^*(l), \quad l = 0, \cdots, K-1) \\
&= P\left(\sum_{j=1}^{q} (n b_j^2 - \delta_{.} a v_{jj}) - \sum_{j=1}^{l} (n b_j^2 - \delta_{.} a v_{jj}) \geq 0, \quad l = 0, \cdots, K-1 \right) \\
&= P\left(\sum_{j=l+1}^{q} (n b_j^2 - \delta_{.} a v_{jj}) \geq 0, \quad l = 0, \cdots, q-1, i.e., l < q, \right. \\
&\qquad \left. \sum_{j=q+1}^{l} (n b_j^2 - \delta_{.} a v_{jj}) \leq 0, \quad l = q+1, \cdots, K-1, i.e., l > q \right).
\end{aligned}$$

Therefore, the following steps of the proof still hold if we replace $\tilde{n}$ and a by n and $\delta_{.} a$ respectively. As a result, when $a = 2$, the maximizer of (3.11) and (3.22) are essentially the same, and we use (3.11) for the proposed test W because the observed sample size n is more accessible. $\square$

As a result of Theorem 3, the following corollary can be revealed.

Corollary 1. *Under both null (3.9) and alternative (3.10) hypotheses,*

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

and r^* is the maximizer of the criterion (3.23), in which the vector of normal random variables $(Z_1, \cdots, Z_{K-q_0-1})^{\mathrm{T}}$ is also defined in Theorem 3.

In addition, for any fixed, finite constant C,

$$\lim_{\tilde{n}\to\infty} P\left(X_{\hat{q}}^2 \geq C | q_0 \neq 0\right) = 1.$$

Proof. According to Theorem 3, the limiting probability of underselection is 0 under both null and alternative hypotheses, in other words, asymptotically it is almost impossible that the selected $\hat{q}$ is less than q_0 in any situation.

Since $\hat{q} \geq q_0$, let

$$\hat{q} \;=\; q_0 + r, \text{ for } r = 0, 1, \cdots, K - q_0 - 1.$$

If $r = 0$, then $\hat{q} = q_0$, hence

$$X_{\hat{q}}^2 - X_{q_0}^2 = 0.$$

If $r > 0$, we have

$$\begin{aligned}
X_{\hat{q}}^2 - X_{q_0}^2 &= \sum_{j=1}^{q_0+r} \tilde{n} b_j^2 - \sum_{j=1}^{q_0} \tilde{n} b_j^2 = \sum_{j=q_0+1}^{q_0+r} \tilde{n} b_j^2 \\
&= \sum_{j=q_0+1}^{q_0+r} v_{jj} \left(\frac{\sqrt{\tilde{n}} b_j}{\sqrt{v_{jj}}}\right)^2 \\
&= \sum_{j=q_0+1}^{q_0+r} v_{jj} Z_j^2
\end{aligned}$$

if we define $Z_j = \frac{\sqrt{\tilde{n}} b_j}{\sqrt{v_{jj}}}$ for $j = q_0 + 1, \cdots, K-1$. Again, for a fixed constant $q_0 > 0$, it is provided by the alternative hypothesis that $\beta_{q_0} \neq 0$ and $\beta_{q_0+1} = \cdots = \beta_{K-1} = 0$. Then, by Theorem 1,

$$\sqrt{\tilde{n}} \begin{pmatrix} b_{q_0+1} - \beta_{q_0+1} \\ b_{q_0+2} - \beta_{q_0+2} \\ \vdots \\ b_{K-1} - \beta_{K-1} \end{pmatrix} = \sqrt{\tilde{n}} \begin{pmatrix} b_{q_0+1} \\ b_{q_0+2} \\ \vdots \\ b_{K-1} \end{pmatrix} \to N_{K-q_0-1}(\mathbf{0}, \mathbf{V}), \text{ as } \tilde{n} \to \infty,$$

where $\mathbf{V} = \{v_{ij}\}$ is the $(K - q_0 - 1) \times (K - q_0 - 1)$ covariance matrix with

$$v_{ij} = \sum_{k=1}^{K} \frac{x_i(k) x_j(k) p(k)}{p_0(k)}, \text{ for } i, j = q_0 + 1, \cdots, K-1.$$

Therefore, after re-indexing, $(Z_1, \cdots, Z_{K-q_0-1})^{\mathrm{T}}$ is a vector of random variables with mean $\mathbf{0}$ and covariance

$$\begin{aligned} \operatorname{cov}(Z_i, Z_j) &= \operatorname{cov}\left(\frac{\sqrt{\tilde{n}} b_i}{\sqrt{v_{ii}}}, \frac{\sqrt{\tilde{n}} b_j}{\sqrt{v_{jj}}} \right) \\ &= \frac{v_{(i+q_0)(j+q_0)}}{\sqrt{v_{(i+q_0)(i+q_0)} v_{(j+q_0)(j+q_0)}}}. \end{aligned}$$

So, If we define $W_{r^*} = \sum_{j=1}^{r^*} v_{(j+q_0)(j+q_0)} Z_j^2$ for $r^* > 0$ and $W_{r^*} = 0$ for $r^* = 0$, it turns out that

$$\begin{aligned} X_{\hat{q}}^2 - X_{q_0}^2 &= \sum_{j=q_0+1}^{q_0+r} v_{jj} \left(\frac{\sqrt{\tilde{n}} b_j}{\sqrt{v_{jj}}} \right)^2 \\ &\xrightarrow{d} \sum_{j=1}^{r} v_{(j+q_0)(j+q_0)} Z_j^2 \\ &= W_{r^*}, \text{ as } \tilde{n} \to \infty, \text{ for } q_0 \geq 0. \end{aligned}$$

The first part of the corollary is then proved, which indicates that the overselection chance is not negligible for $a = 2$ in both null and alternative hypotheses.

Now, according to the first part of the corollary, we have

$$X^2_{\hat{q}} \xrightarrow{d} X^2_{q_0} + W_{r^*}.$$

where W_r is a random variable that does not depend on sample size. For $q_0 \neq 0$, $X^2_{q_0} = \tilde{n} \sum_{j=1}^{q_0} b_j^2 \to \infty$ as $\tilde{n} \to \infty$, which results in $X^2_{\hat{q}} \to \infty$ as $\tilde{n} \to \infty$. Therefore, for any fixed, finite constant C,

$$P(X^2_{\hat{q}} \geq C | q_0 \neq 0) \to 1, \text{ as } \tilde{n} \to \infty.$$

The proof of this corollary is then accomplished. The second part of the corollary implies that the power of the test goes to 1 when the sample size is large enough. $\square$

3.2.5 Asymptotic Properties of $\tilde{q}$

In Section 3.1.2, we mention that the limiting probability of the Type I error is about 0.29 for $a = 2$ through simulation, which is already verified theoretically by Theorem 3 and Corollary 1. One reason may be that $a = 2$ is not large enough. In this sense, a may be considered as a penalty term, which controls the Type I error of the tests. Therefore, we now consider another maximizing criterion

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

for $q = 1, \cdots, K-1$ and $\tilde{M}(q) = 0$ for $q = 0$ with

$$\hat{v}_{jj} = \sum_{k=1}^{K} \frac{x_j^2(k)\hat{p}(k)}{p_0(k)},$$

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 $\tilde{q}$ (maximizer of $\tilde{M}(q)$) is consistent with q_0 in any situation if $a_{\tilde{n}}$ is large enough.

Theorem 4. *If $a_{\tilde{n}} = o(\sqrt{\tilde{n}})$ and $a_{\tilde{n}} > 2\ln(\ln(\tilde{n}))$, we have*

$$\tilde{q} \xrightarrow{P} q_0, \ \textit{for } q_0 \geq 0.$$

Proof. The proof of this theorem utilizes the general maximizing criterion (3.24)

$$M^*(q) \quad = \quad \sum_{j=1}^{q}(\tilde{n}b_j^2 - av_{jj})$$

with q^* as the maximizer. As shown in (3.25) (in the proof of Theorem 3)

$$\begin{aligned} P(q^* = q) \quad = \quad P\Bigg(& \sum_{j=l+1}^{q}(\tilde{n}b_j^2 - av_{jj}) \geq 0, \quad l = q_0, \cdots, q-1, \\ & \sum_{j=q+1}^{l}(\tilde{n}b_j^2 - av_{jj}) \leq 0, \quad l = q+1, \cdots, K-1 \Bigg). \end{aligned}$$

For $q_0 = 0$ and $q \geq 1$, since $v_{jj} = 1$ for all j's, by the law of the iterated logarithm,

$$P(q^* = q) \leq P(\tilde{n}b_q^2 \geq a) \to 0, \text{ as } \tilde{n} \to \infty \text{ and } a > 2\ln(\ln(\tilde{n})).$$

This result implies that, under the null hypothesis (3.9), $P(q^* = q) \to 0$ for $q \geq 1$, as $\tilde{n} \to \infty$ and $a > 2\ln(\ln(\tilde{n}))$. This limiting probability is equivalent to $P(q^* = 0) \to 1$ under the null hypothesis, as $\tilde{n} \to \infty$ and $a > 2\ln(\ln(\tilde{n}))$.

Now, for $q_0 > 0$, i.e., under an alternative hypothesis (3.10), by the law of iterated logarithm, we also have

$$P(q^* = q) \leq P\left(\tilde{n}b_q^2 \geq av_{jj}\right) = P(Z_q^2 \geq a) \to 0, \text{ for } \tilde{n} \to \infty \text{ and } a > 2\ln(\ln(\tilde{n})).$$

Thus, for $q > q_0 > 0$,

$$P(q^* = q_0) = 1 - P(q* = q) \to 1, \text{ for } \tilde{n} \to \infty \text{ and } a > 2\ln(\ln(\tilde{n})).$$

This theorem implies that the maximizer $\tilde{q}$ is always a consistent estimator of q_0 when the effective sample size is large enough.

Now, consider the relationship between $\tilde{n}$ and n in (3.16). As shown in the proof of Theorem 3 in Section 3.2.4, the maximizer of $M^*(q)$ in (3.25) is the same as the maximizer of $\delta_. M^*(q)$, which results in the same maximizer of (3.11) and (3.22). Therefore, the maximizer $\tilde{q}$ of

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

is the same as the maximizer of

$$\tilde{M}'(q) = \frac{n+1}{n-1} \sum_{j=1}^{q} b_j^2 - \frac{\delta_. a_n}{n-1} \sum_{j=1}^{q} \hat{v}_{jj}.$$

In both of the above maximizing criteria, $\tilde{q}$ is consistent with the underlying order q_0 for $n \to \infty$ and $a > 2\ln(\ln(n))$. □

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

Before the Theorem 5 is illustrated, we first examine Theorem 3 in Section 3.2.4 again, from which there are several useful conclusions. First, the limiting probability that $\hat{q}$ is underselected goes to 0, under both null (3.9) and alternative (3.10) hypotheses. Second, the limiting probability that $\hat{q}$ is overselected is not negligible in both null and alternative hypotheses. If the maximizing criterion with $a = 2$ is taken in Theorem 3, it is known that the limiting probability of the Type I error is

$$P(\hat{q} \neq 0 | q_0 = 0) \to 0.29, \text{ as } K \to \infty \text{ and } \tilde{n} \to \infty.$$

Also, Theorem 4 in Section 3.2.5 indicates that under both null and alternative hypotheses, the estimator $\tilde{q}$ is consistent with q_0 , as $K \to \infty$, $\tilde{n} \to \infty$ and $a > 2\ln(\ln(\tilde{n}))$. This implies that

$$P(\tilde{q} = 0 | q_0 = 0) \to 1, \text{ as } K \to \infty, \ \tilde{n} \to \infty \text{ and } a > 2\ln(\ln(\tilde{n})). \tag{3.26}$$

Thus, the limiting probability of the Type I error is, for $a > 2\ln(\ln(\tilde{n}))$,

$$P(\hat{q} \neq 0 | q_0 = 0) = 1 - P(\tilde{q} = 0 | q_0 = 0) \to 0, \text{ as } K \to \infty, \text{ and } \tilde{n} \to \infty. \tag{3.27}$$

Based on (3.26) and (3.27), it is reasonable that there exists an value a_α such that the limiting probability of the Type I error is between 0 and 0.29, that is, for a pre-specified level of significance $0 \leq \alpha \leq 0.29$,

$$P(\hat{q}_\alpha \neq 0 | q_0 = 0) \to \alpha, \text{ as } K \to \infty \text{ and } \tilde{n} \to \infty,$$

where the estimator $\hat{q}_\alpha$ is determined by a_α. Therefore, Theorem 5 gives the asymptotic behaviors of $\hat{q}_\alpha$ in Section 3.1.3.

Theorem 5. *Let $\hat{q}_\alpha$ be maximizer of the criterion*

$$\hat{M}_\alpha(q) = \frac{\tilde{n}+1}{\tilde{n}-1} \sum_{j=1}^{q} b_j^2 - \frac{a_\alpha}{\tilde{n}-1} \sum_{j=1}^{q} \hat{v}_{jj}, \ \textit{for } q = 1, \cdots, K-1, \tag{3.28}$$

where $\hat{M}_\alpha(0) = 0$,

$$\hat{v}_{jj} = \sum_{k=1}^{K} \frac{x_j^2(k)\hat{p}(k)}{p_0(k)}, \ \textit{for } j = 1, \cdots, K-1,$$

and a_α is chosen so that it is the solution of

$$1 - \alpha = \exp\left\{ - \sum_{k=1}^{\infty} \frac{P(\chi_k^2 > k a_\alpha)}{k} \right\} \tag{3.29}$$

or

$$P\left(\max_{1 \le k \le K-1} \left[\frac{1}{k} \sum_{j=1}^{k} Z_j^2 \right] \ge a_\alpha \right) = \alpha. \tag{3.30}$$

The limiting probabilities of the maximizer $\hat{q}_\alpha$ are shown as follows, for $\tilde{n} \to \infty$,

$$P(\hat{q}_\alpha > 0 | q_0 = 0) \to \alpha$$

and

$$P(\hat{q}_\alpha > 0 | q_0 \ne 0) \to 1$$

Proof. As discussed above, when $a_\alpha = 2$, the limiting probability of the Type I error is 0.29, and when a_α grows with the effective sample size with an appropriate rate, the limiting probability of the Type I error reduces to 0. Thus, there exists an appropriate value for the penalty term a_α so that the limiting probability of the Type I error of the test is a value between 0 and 0.29.

Based on Theorem 3.1 of Eubank and Hart (1992) and random walk theory in Spitzer (1956), if a_α is the solution of

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

then

$$P(\hat{q}_\alpha = 0 | q_0 = 0) = 1 - \alpha, \text{ as } \tilde{n} \to \infty,$$

which results in

$$P(\hat{q}_\alpha > 0|q_0 = 0) \to \alpha, \text{ as } \tilde{n} \to \infty.$$

Furthermore, consider the general maximizing criterion (3.24),

$$\begin{aligned} M^*(q) &= \sum_{j=1}^{q}(\tilde{n}b_j^2 - av_{jj}) \\ &= \sum_{j=1}^{q} v_{jj}\left[\left(\frac{\sqrt{\tilde{n}}b_j}{\sqrt{v_{jj}}}\right)^2 - a\right] \end{aligned}$$

Under the null hypothesis (4.1), $v_{jj} = 1$ by the orthogonality condition (3.2), and $\sqrt{n}b_j$'s are asymptotically independent standard normal random variables for all $j = 1, \cdots, q$ by Theorem 1 in Section 3.2.2 Thus,

$$\begin{aligned} M^*(q) &= \sum_{j=1}^{q}\left[\left(\sqrt{\tilde{n}}b_j\right)^2 - a\right] \\ &\to \sum_{j=1}^{q}\left[Z_j^2 - a\right], \text{ as } \tilde{n} \to \infty \\ &= \sum_{j=1}^{q} Z_j^2 - qa = \frac{1}{q}\sum_{j=1}^{q} Z_j^2 - a \end{aligned}$$

where Z_j denotes $\sqrt{n}b_j$. Since the limiting probability of underselection is 0 according to Theorem 3 in Section 3.2.4 and Theorem 4 in Section 3.2.5, overselection is the only source for the Type I error. Under null hypothesis (3.1), since the maximizer must be a value greater than 0 for overselection, $M^*(q)$ must be equal or greater than 0 (if $M^*(q) < 0$, the maximizer is $q = 0$ with $M(0) = 0$). Therefore,

$$\max_{1 \le q \le K-1} \left(\frac{1}{q} \sum_{q=1}^{k} Z_j^2 \right) - a \ge 0, \text{ or equivalently, } \max_{1 \le q \le K-1} \left(\frac{1}{q} \sum_{q=1}^{k} Z_j^2 \right) \ge a$$

Since it is desired that the overselection is controlled in a pre-specified level, we restrict the probability as

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

where we replace q by k and a by a_α. Consequently, if a_α is the solution of

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

for a pre-specified level α, the limiting probability of the Type I error is

$$P(\hat{q}_\alpha > 0 | q_0 = 0) = \alpha, \text{ as } \tilde{n} \to \infty.$$

This result can also be found in Eubank (1997). The solutions of a_α are the uniform for (3.29) and (3.30), which is verified by solving them numerically.

In addition, under the alternative hypothesis (3.10), i.e., $q_0 > 0$, since the limiting probability of underselection is still 0 according to Theorem 3 in Section 3.2.4, we have $P(\hat{q}_\alpha < q_0 | q_0 \neq 0) \to 0$ or $P(\hat{q}_\alpha = 0 | q_0 > 0) \to 0$, as $\tilde{n} \to \infty$, As a result,,

$$\begin{aligned} P(\hat{q}_\alpha \ge q_0 | q_0 > 0) &= P(\hat{q}_\alpha > 0 | q_0 > 0) = 1 - P(\hat{q}_\alpha = 0 | q_0 > 0) \\ &= 1 - P(\hat{q}_\alpha = 0 | q_0 \neq 0) \\ &\to 1, \text{ as } \tilde{n} \to \infty, \end{aligned}$$

or equivalently

$$P(\hat{q}_\alpha > 0 | q_0 \neq 0) \to 1, \text{ as } \tilde{n} \to \infty.$$

and the theorem is proved, which implies that the probability of overselection can be controlled within level α when $q_0 = 0$, and the estimator $\hat{q}_\alpha$ is consistent to the real order q_0 when $q_0 > 0$, at the same time.

Similarly, because of the relationship between $\tilde{n}$ and n that $\tilde{n} = \frac{n}{\delta_.}$, if $\hat{q}_\alpha$ is the maximizer of (3.24), it is the maximizer of $\delta_. M^*(q) = \delta_. \sum_{j=1}^{q}(\tilde{n}b_j^2 - av_{jj}) = \sum_{j=1}^{q}(nb_j^2 - \delta_. av_{jj})$ for $\delta_. > 0$. So, if the observed sample size n is used (more accessible), $\hat{\delta}_. a_\alpha$ in (3.13) is used in the proposed test $\hat{q}_\alpha$. □

Chapter 4

Simulation Studies

To evaluate our proposed methods, in this chapter, we proceed with limited simulation studies. We consider the general null hypothesis as in (4.1) and three alternatives as in (4.3), (4.4) and (4.5). First, we show that our proposed tests control the Type I error at the pre-specified level well. In other words, the rate of incorrect conclusion against null hypothesis does not exceed the pre-determined tolerance. Second, we compare the empirical statistical powers of our proposed tests with some existing methods. Finally, we examine the influence of the data patterns on the proposed tests.

4.1 Simulation Set Up

In our simulation studies, the level of significance $\alpha = 0.05$ is chosen for all of the simulation settings. To compare with the simulation studies in Eubank (1997), $K = 10$ are selected as the total number of categories in multinomial data under complex surveys. 50 clusters (psus) are considered, and 15 individuals (ssus) are sampled within each cluster. Thus, there are a total of 750 units that are observed.

As discussed before, observations in complex surveys are not independent, and the dependency is reflected by ICC (Intraclass Correlation Coefficient, Lohr, 2010, pg. 174-176) within each cluster. In order to model the correlation among ssus in a cluster, ICC is set to be 0.1, 0.3, and 0.6 to show the performance of the proposed test procedures under low, medium, and high levels of correlation, respectively. Notice that if ICC is 0, all the observations are uncorrelated as in SRS. If ICC is 1, the ssus in 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.

The generation of multinomial data under complex surveys involves four inputs, number of psu, number of ssu, ICC, and a set of given probabilities of categories in our simulation studies. The total population is calculated by the product of the number of psu and the number of ssu. In each of the psu, we first randomly permute the categories with their probabilities such that the order of the categories does not determine the clustering. With the permuted categories, the cumulative probabilities of the first 9 categories are calculated. After that, the quantiles of these cumulative probabilities are found under standard normal distribution. Then, a clustered standard normal random variable is created by $N(0, \text{ICC}) + N(0, 1 - \text{ICC})$, which represent standard normal random variables with mean 0 and standard deviation ICC and $1 - \text{ICC}$ respectively. After that, 15 quantiles of this clustered standard normal random variable are generated randomly, because we have 15 ssus in each psu. These 15 values are then compared with the previously calculated 9 cumulative probabilities. For example, if one of the values is less the first cumulative probability, this ssu is categorized to the first permuted category. If another value is less than the fifth cumulative probability, then the corresponding ssu is grouped to the fifth permuted category. If the value is greater than all 9 cumulative probability, the corresponding ssu goes to the last permuted category. Such a process is repeated for all psus, hence all the observations will be eventually categorized.

The null hypothesis in the simulation studies for $K = 10$ is

$$H_0 : p(1) = \cdots = p(10) = 0.1, \tag{4.1}$$

which is equivalent to

$$H_0^* : \beta_1 = \cdots = \beta_9 = 0. \tag{4.2}$$

For the proposed test W, $\hat{q}$ is obtained such that it maximizes equation (3.11), then the test statistic is calculated by (3.12). The critical value is obtained by simulation of the same process under the null hypothesis (4.1). For the proposed test $\hat{q}_\alpha$, the test statistic $\hat{q}_\alpha$ is the maximizer of the equation (3.13).

In order to find the estimator of $\delta_.$, say $\hat{\delta}_.$, $100,000$ complex multinomial data under the null hypothesis (4.1) are generated. For each of the generated data, the sample covariance matrix is calculated. Then, covariance matrix (B.5) under complex surveys and null hypothesis (4.1) is estimated by the average of the $100,000$ covariance matrix ($\hat{\mathbf{V}}$). Note that, the covariance matrix under SRS and the null hypothesis (4.1) is known. Therefore, $\hat{\delta}_.$ is obtained by averaging the eigenvalues of the matrix $\mathbf{P}_0^{-1}\mathbf{V}$.

For the proposed test W in Section 3.1.2, the empirical distribution of W_0 (W under the null hypothesis 4.2) is obtained through $100,000$ iterations, and then the critical value is calculated at the pre-specified level of significance α. For proposed test $\hat{q}_\alpha$, instead of finding the solution of a_α ($a_{0.05}$ in our cases) for each estimated $\hat{\delta}_.$, the values of a_α with several commonly used levels of significance in SRS are obtained through $10,000$ iterations for (3.14) and (3.15), and then the values of a_α in complex surveys can be obtained by the product of $\hat{\delta}_.$ and the values of a_α in SRS. The advantage of doing this is that only the estimation of $\delta_.$ needs to be calculated for each value of ICC's, which is much faster than calculating corresponding a_α for every simulated data. According to Eubank and Hart (1992) and our verifications, the values of a_α in SRS can be summarized in Table 4.1. Notice that the solution of Equation 3.14 requires large K approximations. But in fact, solution converges to the same value as long as $K > 10$. Also, we verified these values numerically and it is found that both equations (3.14) and (3.15) give the same solution for a_α.

α	0.01	0.05	0.10	0.20	0.29
a_α	6.74	4.18	3.22	2.38	2

Table 4.1: Values of a_α of corresponding level of significance under SRS.

In our simulation studies, Pearson's chi-squared test, the first order and second order corrected tests are included to compare with both of our proposed tests. Pearson's chi-squared GOF test is one of the GOF tests for SRS. The first order and second order corrected tests are two of the commonly used GOF tests in complex surveys, which are efficient (Rao & Thomas, 1988) compared with other approaches and only require relatively small amount of computation. Thus, they are good competitors of our proposed tests.

Following Eubank (1997), three alternatives of (4.1) are used to examine our proposed tests. They are

$$p(k) = \frac{1}{10} + \beta(k - 5.5)/10, \text{ for } k = 1, \cdots, 10, \tag{4.3}$$

$$p(k) = \frac{1}{10} + \beta\cos(\frac{j\pi(k-0.5)}{10}), \text{ for } k = 1, \cdots, 10, \tag{4.4}$$

and

$$p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right], \text{ for } k = 1, \cdots, 10. \tag{4.5}$$

Following are the steps of our simulation for an arbitrary alternative.

1. We generate $100,000$ multinomial data under complex surveys and the null hypothesis (4.1). For each generated data, the $(K-1)$ vector of estimated proportions $\hat{\mathbf{p}} = (\hat{p}(1), \cdots, \hat{p}(K-1))$ is recorded. The estimated mean of $\hat{\mathbf{p}}$ is calculated by averaging the $100,000$ generated $\hat{\mathbf{p}}$'s and denoted as $\bar{\mathbf{p}}$. Then, the estimated covariance matrix $\hat{\mathbf{V}}$ is obtained by $(\hat{\mathbf{p}} - \bar{\mathbf{p}})(\hat{\mathbf{p}} - \bar{\mathbf{p}})^{\mathrm{T}}/(100,000-1)$. The eigenvalues of the the matrix $\mathbf{P}_0^{-1}\mathbf{V}$ are calculated using the `eigen()` function in `R`. In the meanwhile, $\hat{a}^2$ can be calculated by (2.17).

2. Under the null hypothesis (4.1), we search for the maximizer of the equation (3.11) to get $\hat{q}$ and then a value of W_0 is obtained using (3.12). This process is repeated $100,000$ time and the empirical distribution of W_0 is discovered. Therefore, the 95% quantile of W_0 is found out.

3. Under the given alternative, Pearson's chi-squared test statistic, the first order and the second order corrected test statistics are calculated by (2.3), (2.15) and (2.16), respectively. Next, by searching all $q = 1, \cdots, K-1$, $\hat{q}$ is found out if it maximizes equation (3.11). W is calculated by equation (3.12). We then search $\hat{q}_{0.05}$ among $q = 1, \cdots, K-1$ to maximize equation (3.13), where $a_{0.05} = 4.18$ is chosen from Table 4.1 and $\hat{\delta}_{.}$ is from Step 1.

4. We compare the test statistics in Step 3 with their corresponding rejection criteria. The test statistics of Pearson's chi-square test statistic and 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, 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.

6. Step 1-5 can be repeated if other alternatives are given.

4.2 Simulation Results

In this section, we report the simulation results. Empirical power comparisons are given by three alternatives (4.3), (4.4), and (4.5). Alternative (4.3) generates slow varying probabilities, alternative (4.4) is generates bot slow varying and non-slow varying probabilities, and alternative (4.5) focuses on certain data patterns.

4.2.1 Simulation Results by Alternative (4.3)

We examine the simulation results starting with alternative (4.3). The null hypothesis is true when $\beta = 0$. For β from 0 to 0.14 with step 0.01, 15 sets of probabilities (including the null hypothesis) are generated as in Table 4.2 and Figure A.1. Two sample probabilities are plotted in Figure 4.1 for $\beta = 0.01$ and $\beta = 0.14$. These probabilities are treated as the underlying parameters to generate the multinomial data. One can verify that the sum of $p(1)$ through $p(10)$ is 1. The generated probabilities are very similar when $\beta = 0.01$, and become moderately slow varying as β increases.

Figures 4.2, 4.3, and 4.4 plot the empirical powers of the five tests, our 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 (4.3) under ICC $= 0.1, 0.3, 0.6$, respectively. As discussed previously, ICC $= 0$ means that all the observations are independent, and ICC $= 1$ stands for the perfect correlation among observations within the same cluster. Thus, for ICC varies from 0.1 to 0.3, and to 0.6, observations within the same cluster are more correlated.

β \ $p(k)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.0955	0.0965	0.0975	0.0985	0.0995
0.02	0.091	0.093	0.095	0.097	0.099
0.03	0.0865	0.0895	0.0925	0.0955	0.0985
0.04	0.082	0.086	0.09	0.094	0.098
0.05	0.0775	0.0825	0.0875	0.0925	0.0975
0.06	0.073	0.079	0.085	0.091	0.097
0.07	0.0685	0.0755	0.0825	0.0895	0.0965
0.08	0.064	0.072	0.08	0.088	0.096
0.09	0.0595	0.0685	0.0775	0.0865	0.0955
0.10	0.055	0.065	0.075	0.085	0.095
0.11	0.0505	0.0615	0.0725	0.0835	0.0945
0.12	0.046	0.058	0.07	0.082	0.094
0.13	0.0415	0.0545	0.0675	0.0805	0.0935
0.14	0.037	0.051	0.065	0.079	0.093
β \ $p(k)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.1005	0.1015	0.1025	0.1035	0.1045
0.02	0.101	0.103	0.105	0.107	0.109
0.03	0.1015	0.1045	0.1075	0.1105	0.1135
0.04	0.102	0.106	0.11	0.114	0.118
0.05	0.1025	0.1075	0.1125	0.1175	0.1225
0.06	0.103	0.109	0.115	0.121	0.127
0.07	0.1035	0.1105	0.1175	0.1245	0.1315
0.08	0.104	0.112	0.12	0.128	0.136
0.09	0.1045	0.1135	0.1225	0.1315	0.1405
0.10	0.105	0.115	0.125	0.135	0.145
0.11	0.1055	0.1165	0.1275	0.1385	0.1495
0.12	0.106	0.118	0.13	0.142	0.154
0.13	0.1065	0.1195	0.1325	0.1455	0.1585
0.14	0.107	0.121	0.135	0.149	0.163

Table 4.2: Simulated probabilities of 10 categories generated by alternative (4.3) for β from 0 to 0.14 with step 0.01.

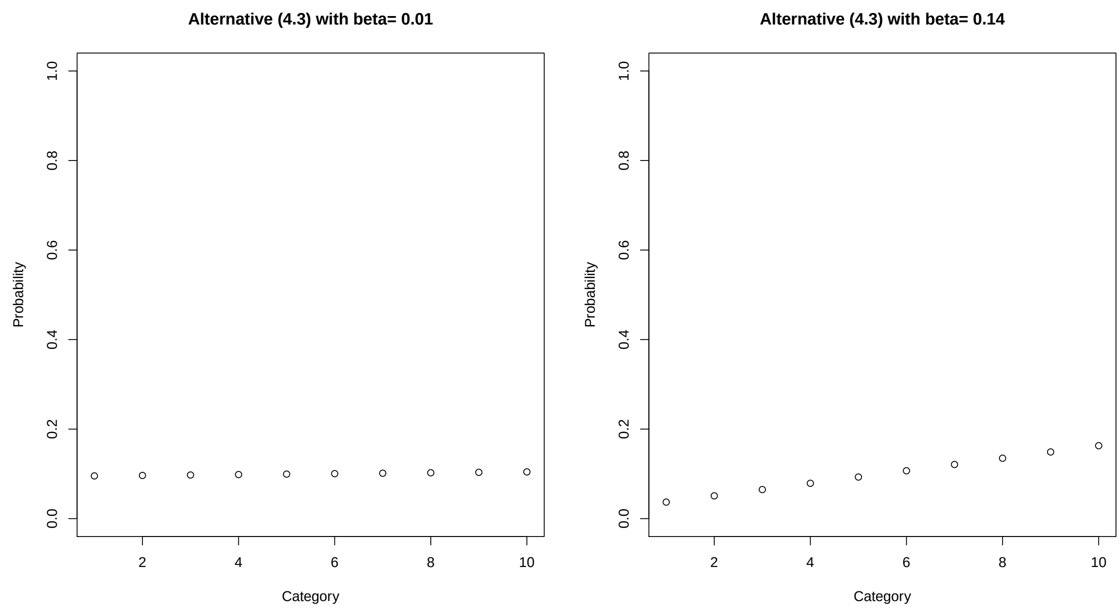

Figure 4.1: Probabilities in simulation studies generated by alternative (4.3) for $\beta = 0.01$ (left) and $\beta = 0.14$ (right). Probabilities vary slowly when $\beta = 0.01$, but vary great when $\beta = 0.14$.

We first look at the control of the Type I error, which is equivalent to the powers of the tests under the null hypothesis (4.2) given by $\beta = 0$. Pearson's chi-squared test shows poor control of the Type I error at the pre-determined level $\alpha = 0.05$, especially with higher value of ICC. The larger the ICC is, the more off the Type I error is for Pearson's chi-squared test. Notice that 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 $= 0.6$. This is evident that GOF tests for multinomial data in SRS should not be directly applied to multinomial data in complex surveys, and thus their empirical statistical powers in the following points should not be considered to be compared with other approaches.

We then examine the empirical powers of our proposed tests and the first order and second order corrected tests. For the three figures 4.2, 4.3, and 4.4, both of our proposed tests (using W and $\hat{q}_{0.05}$) are superior to 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 diverge enough ($\beta > 0.07$), our proposed tests are as good as the first order and second order corrected tests with regard to empirical statistical powers. Meanwhile, one can see that our proposed tests and the first order and second order corrected tests control the Type I error at the desired level (0.05 in this simulation). In addition, the test $\hat{q}_{0.05}$ has the best empirical statistical power , and the test W has the second best empirical statistical power for alternative 4.3. In sum, our proposed tests have great improved statistical powers compared with the first order and second order corrected tests, especially when the underlying probabilities vary slowly for multinomial data in complex surveys.

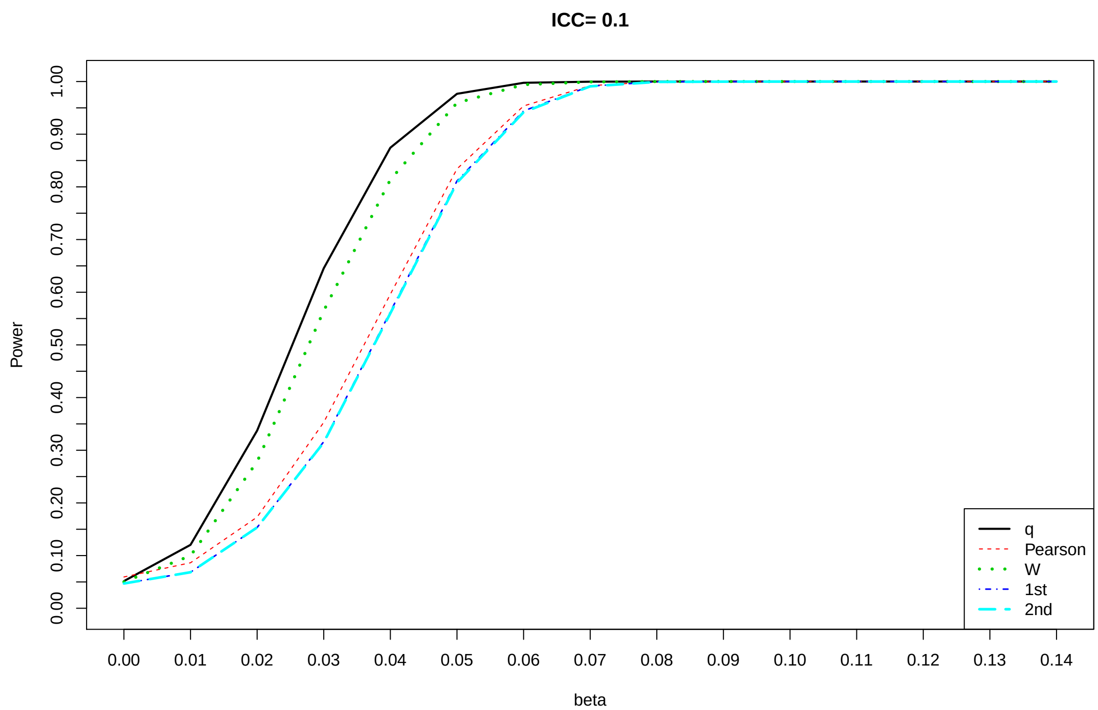

Figure 4.2: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.3) $p(k) = \frac{1}{10} + \beta(k - 5.5)/10$, for $k = 1, \cdots, 10$. The simulations consist of 50 psus, 15 ssus and 10 categories, with ICC 0.1. The Type I error of Pearson's test is a little off and all other four tests control the Type I error well. The proposed test $\hat{q}_{0.05}$ has the best empirical power, followed by the proposed test W. Both of our proposed tests are better than the first order and second order corrected tests when the probabilities vary slowly. Powers of all tests converge to 1 when the probabilities vary greatly.

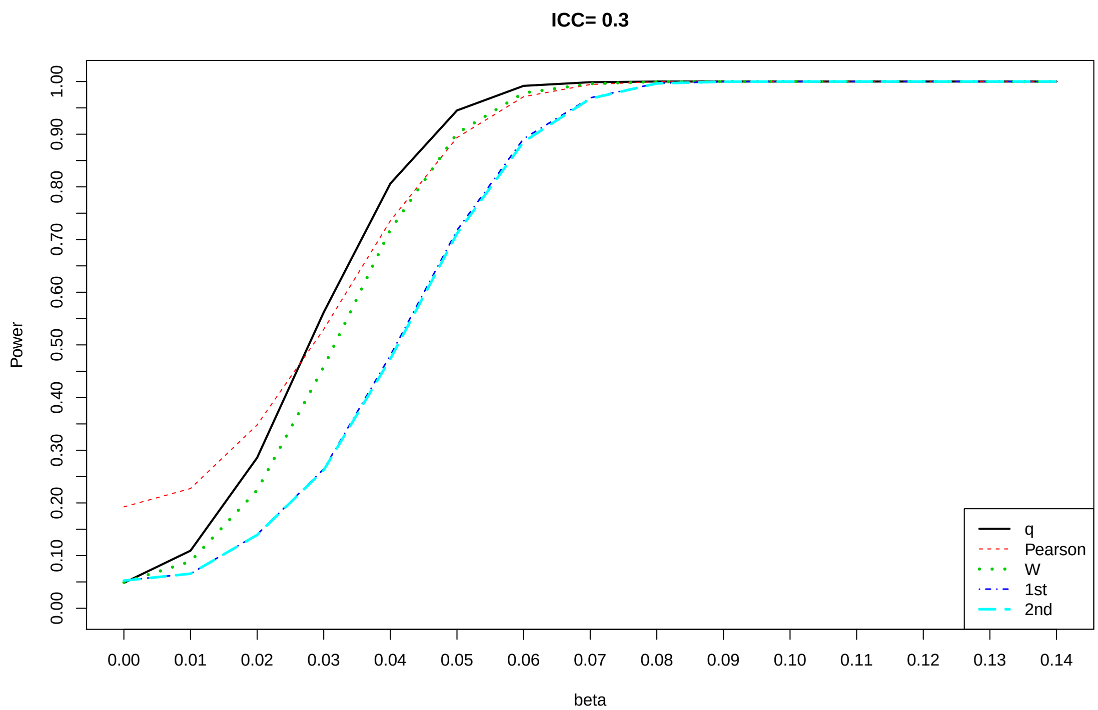

Figure 4.3: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.3) $p(k) = \frac{1}{10} + \beta(k - 5.5)/10$, for $k = 1, \cdots, 10$. The simulations consist of 50 psus, 15 ssus and 10 categories, with ICC 0.3. Pearson's test doesn't control the Type I error well and all other four tests control the Type I error well. The proposed test $\hat{q}_{0.05}$ has the best empirical power, followed by the proposed test W. Both of our proposed tests are better than the first order and second order corrected tests when the probabilities vary slowly. Powers of all tests converge to 1 when the probabilities vary greatly.

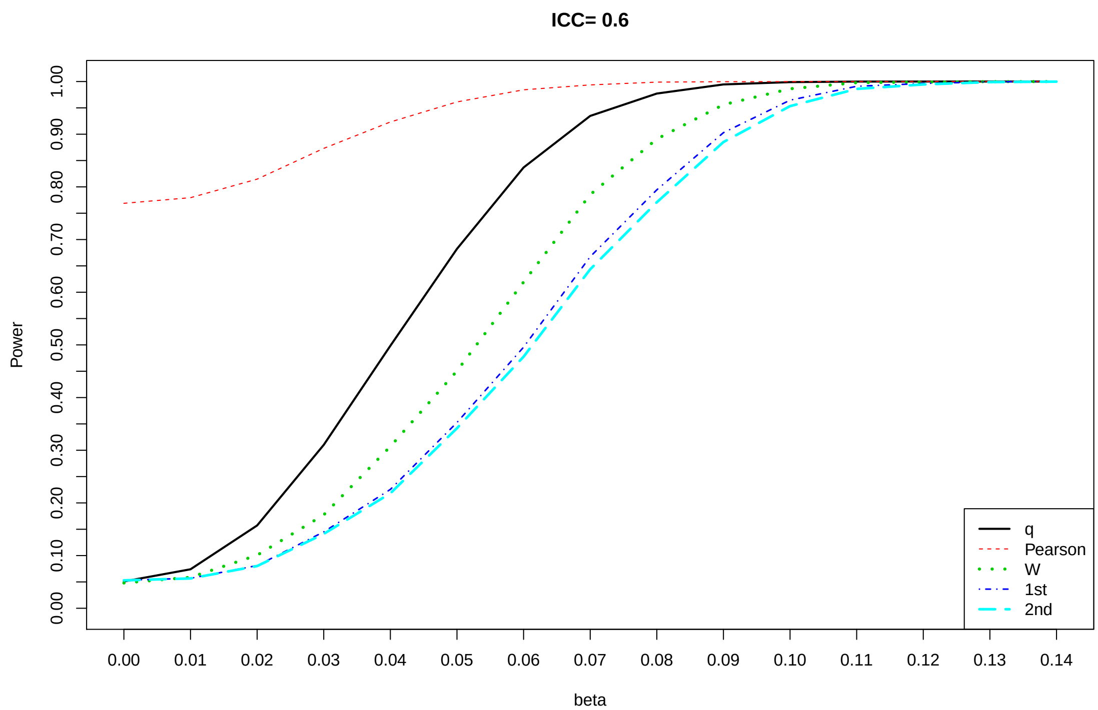

Figure 4.4: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.3) $p(k) = \frac{1}{10} + \beta(k - 5.5)/10$, for $k = 1, \cdots, 10$. The simulations consist of 50 psus, 15 ssus and 10 categories, with ICC 0.6. Pearson's test is not able to control the Type I error and all other four tests control the Type I error well. The proposed test $\hat{q}_{0.05}$ has the best empirical power, followed by the proposed test W. Both of our proposed tests are better than the first order and second order corrected tests when the probabilities vary slowly. Powers of all tests converge to 1 when the probabilities vary greatly.

4.2.2 Simulation Results by Alternative (4.4)

For alternative (4.4), j is the parameter to govern how different the underlying probabilities are. With $j = 1$, the underlying probabilities (Table 4.3 and Figure A.2) are very close to each other. In contrast, the underlying probabilities vary dramatically with $j = 9$ (Table 4.6 and A.5). For $j = 1$, the empirical statistical power curves are similar to those in figures plotted with alternative (4.3). As a result, we choose $j = 2$ and $j = 4$ to examine the performance of our proposed tests with medium and high varying underlying probabilities. For each selected j, 11 value of β from 0 to 0.1 are used with step 0.01. The exact values of probabilities are shown in Table 4.4 and Table 4.5 for $j = 2$ and $j = 4$, respectively. Two sample probabilities are plotted in Figure 4.5 and 4.6 for $j = 2$ and $j = 4$. In addition, Figure A.3 and A.4 show all probabilities for $j = 2$ and $j = 4$ respectively. It can also be verified that the sum of generated probabilities is 1 for each β and each j.

For the alternative (4.4) with $j = 2$, figures 4.7, 4.8, and 4.9 plot the powers of the five tests versus β ranged from 0.00 to 0.07 for ICC $= 0.1, 0.3, 0.6$, respectively. Notice that the null hypothesis (4.1) is true when $\beta = 0$. It can be seen that our proposed tests are able to control the level of significance at the nominal level ($\alpha = 0.05$), so do the first order and second order corrected tests. However, not surprisingly, the Pearson's chi-squared test fails to control the Type I error at the desired level when ICC $= 0.3, 0.6$, i.e., observations within the cluster are correlated, and the independence assumption is no longer met.

For $j = 2$, both of our proposed tests outperform the first order and second order corrected tests, when the underlying probabilities are varying slowly ($\beta \leq 0.03$). All of the four tests demonstrate very similar empirical statistical powers when the underlying probabilities vary greatly ($\beta > 0.03$). In this setting, the two proposed tests are both competitive, with the proposed test W slightly better than

β \ $p(k)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.109876883	0.108910065	0.107071068	0.104539905	0.101564345
0.02	0.119753767	0.11782013	0.114142136	0.10907981	0.103128689
0.03	0.12963065	0.126730196	0.121213203	0.113619715	0.104693034
0.04	0.139507534	0.135640261	0.128284271	0.11815962	0.106257379
0.05	0.149384417	0.144550326	0.135355339	0.122699525	0.107821723
0.06	0.1592613	0.153460391	0.142426407	0.12723943	0.109386068
0.07	0.169138184	0.162370457	0.149497475	0.131779335	0.110950413
0.08	0.179015067	0.171280522	0.156568542	0.13631924	0.112514757
0.09	0.188891951	0.180190587	0.16363961	0.140859145	0.114079102
0.10	0.198768834	0.189100652	0.170710678	0.14539905	0.115643447
β \ $p(k)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.098435655	0.095460095	0.092928932	0.091089935	0.090123117
0.02	0.096871311	0.09092019	0.085857864	0.08217987	0.080246233
0.03	0.095306966	0.086380285	0.078786797	0.073269804	0.07036935
0.04	0.093742621	0.08184038	0.071715729	0.064359739	0.060492466
0.05	0.092178277	0.077300475	0.064644661	0.055449674	0.050615583
0.06	0.090613932	0.07276057	0.057573593	0.046539609	0.0407387
0.07	0.089049587	0.068220665	0.050502525	0.037629543	0.030861816
0.08	0.087485243	0.06368076	0.043431458	0.028719478	0.020984933
0.09	0.085920898	0.059140855	0.03636039	0.019809413	0.011108049
0.10	0.084356553	0.05460095	0.029289322	0.010899348	0.001231166

Table 4.3: Simulated probabilities of 10 categories generated by alternative (4.4) with $j = 1$ for β from 0 to 0.1 with step 0.01.

the proposed test $\hat{q}_{0.05}$, but both of them show higher empirical statistical powers than the first order and second order corrected tests.

Next, we investigate the simulation results with $j = 4$ for alternative (4.4). Figures 4.10, 4.11, and 4.12 plot the empirical powers of the five tests versus β in alternative (4.4) with ICC 0.1, 0.3, and 0.6, respectively. The case of the null hypothesis (4.1) is simulated when $\beta = 0$. Both of our proposed tests and first order and second

β \ $p(k)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.109510565	0.105877853	0.1	0.094122147	0.090489435
0.02	0.11902113	0.111755705	0.1	0.088244295	0.08097887
0.03	0.128531695	0.117633558	0.1	0.082366442	0.071468305
0.04	0.138042261	0.12351141	0.1	0.07648859	0.061957739
0.05	0.147552826	0.129389263	0.1	0.070610737	0.052447174
0.06	0.157063391	0.135267115	0.1	0.064732885	0.042936609
0.07	0.166573956	0.141144968	0.1	0.058855032	0.033426044
0.08	0.176084521	0.14702282	0.1	0.05297718	0.023915479
0.09	0.185595086	0.152900673	0.1	0.047099327	0.014404914
0.10	0.195105652	0.158778525	0.1	0.041221475	0.004894348
β \ $p(k)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.090489435	0.094122147	0.1	0.105877853	0.109510565
0.02	0.08097887	0.088244295	0.1	0.111755705	0.11902113
0.03	0.071468305	0.082366442	0.1	0.117633558	0.128531695
0.04	0.061957739	0.07648859	0.1	0.12351141	0.138042261
0.05	0.052447174	0.070610737	0.1	0.129389263	0.147552826
0.06	0.042936609	0.064732885	0.1	0.135267115	0.157063391
0.07	0.033426044	0.058855032	0.1	0.141144968	0.166573956
0.08	0.023915479	0.05297718	0.1	0.14702282	0.176084521
0.09	0.014404914	0.047099327	0.1	0.152900673	0.185595086
0.10	0.004894348	0.041221475	0.1	0.158778525	0.195105652

Table 4.4: Simulated probabilities of 10 categories generated by alternative (4.4) with $j = 2$ for β from 0 to 0.1 with step 0.01.

order corrected tests control the Type I error well, while the Pearson's chi-squared tests can not even get the correct pre-specified level of significance $\alpha = 0.05$.

Since $j = 4$ generates highly varying underlying probabilities as seen in Table 4.5, the results of power comparison are different from those in $j = 2$. One can see that our proposed test W is now competitive with the first order and second order corrected tests, but the power of the test $\hat{q}_{0.05}$ has decreased. In addition, as the

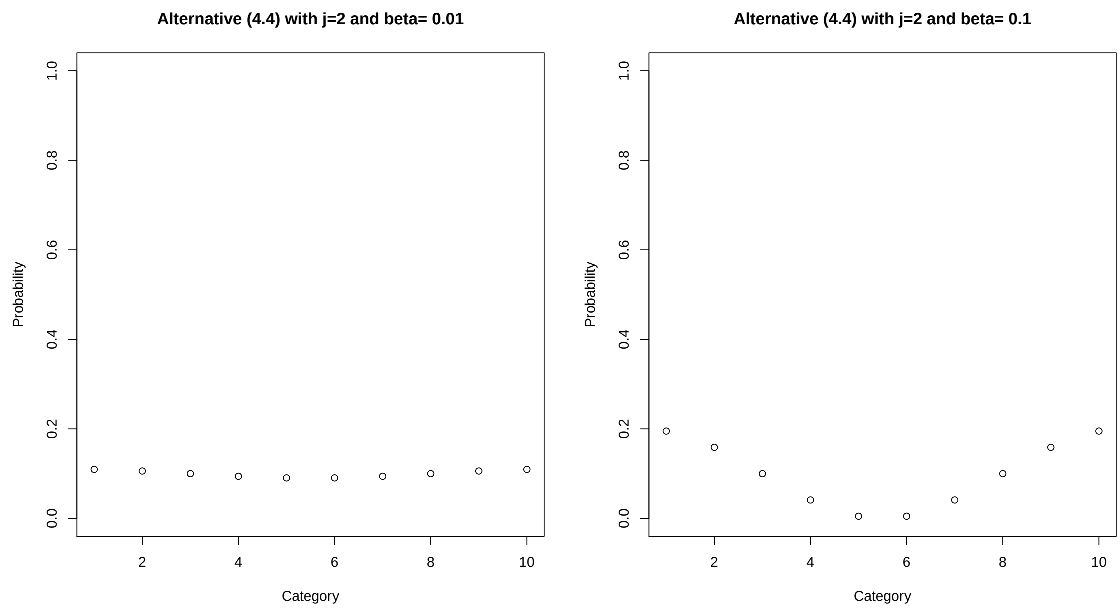

Figure 4.5: Probabilities in simulation studies generated by alternative (4.4) with $j = 2$ for $\beta = 0.01$ (left) and $\beta = 0.1$ (right). Probabilities vary slowly when $\beta = 0.01$, but vary greatly when $\beta = 0.1$.

ICC goes up, our test W becomes the best. These results show that our proposed test W is very stable in both slow varying and non-slow varying cases. When the underlying probabilities are varying slowly, our proposed test W is superior to the first order and second order corrected tests. When the underlying probabilities are varying greatly, method based on W is at least as good as the existing approaches.

β \ $p(k)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.10809017	0.09690983	0.09	0.09690983	0.10809017
0.02	0.11618034	0.09381966	0.08	0.09381966	0.11618034
0.03	0.12427051	0.09072949	0.07	0.09072949	0.12427051
0.04	0.13236068	0.08763932	0.06	0.08763932	0.13236068
0.05	0.14045085	0.08454915	0.05	0.08454915	0.14045085
0.06	0.14854102	0.08145898	0.04	0.08145898	0.14854102
0.07	0.15663119	0.07836881	0.03	0.07836881	0.15663119
0.08	0.16472136	0.07527864	0.02	0.07527864	0.16472136
0.09	0.172811529	0.072188471	0.01	0.072188471	0.172811529
0.10	0.180901699	0.069098301	0	0.069098301	0.180901699
β \ $p(k)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.10809017	0.09690983	0.09	0.09690983	0.10809017
0.02	0.11618034	0.09381966	0.08	0.09381966	0.11618034
0.03	0.12427051	0.09072949	0.07	0.09072949	0.12427051
0.04	0.13236068	0.08763932	0.06	0.08763932	0.13236068
0.05	0.14045085	0.08454915	0.05	0.08454915	0.14045085
0.06	0.14854102	0.08145898	0.04	0.08145898	0.14854102
0.07	0.15663119	0.07836881	0.03	0.07836881	0.15663119
0.08	0.16472136	0.07527864	0.02	0.07527864	0.16472136
0.09	0.172811529	0.072188471	0.01	0.072188471	0.172811529
0.10	0.180901699	0.069098301	0	0.069098301	0.180901699

Table 4.5: Simulated probabilities of 10 categories generated by alternative (4.4) with $j = 4$ for β from 0 to 0.1 with step 0.01.

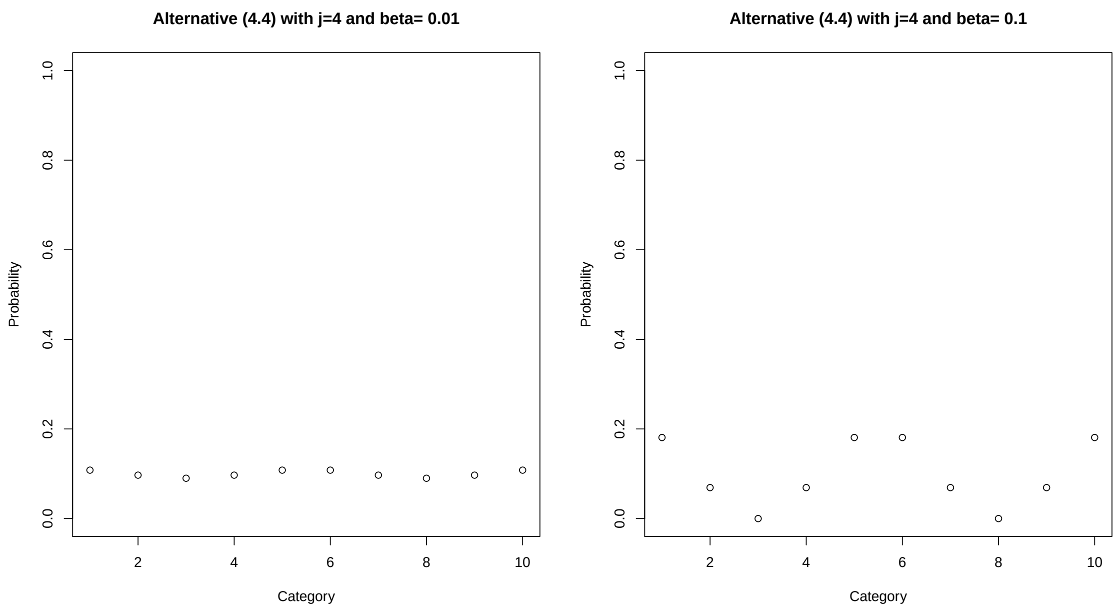

Figure 4.6: Probabilities in simulation studies generated by alternative (4.4) with $j = 4$ for $\beta = 0.01$ (left) and $\beta = 0.1$ (right). Probabilities vary slowly when $\beta = 0.01$, but vary greatly when $\beta = 0.1$.

β \ $p(k)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.101564345	0.095460095	0.107071068	0.091089935	0.109876883
0.02	0.103128689	0.09092019	0.114142136	0.08217987	0.119753767
0.03	0.104693034	0.086380285	0.121213203	0.073269804	0.12963065
0.04	0.106257379	0.08184038	0.128284271	0.064359739	0.139507534
0.05	0.107821723	0.077300475	0.135355339	0.055449674	0.149384417
0.06	0.109386068	0.07276057	0.142426407	0.046539609	0.1592613
0.07	0.110950413	0.068220665	0.149497475	0.037629543	0.169138184
0.08	0.112514757	0.06368076	0.156568542	0.028719478	0.179015067
0.09	0.114079102	0.059140855	0.16363961	0.019809413	0.188891951
0.10	0.115643447	0.05460095	0.170710678	0.010899348	0.198768834
β \ $p(k)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$
0.00	0.1	0.1	0.1	0.1	0.1
0.01	0.090123117	0.108910065	0.092928932	0.104539905	0.098435655
0.02	0.080246233	0.11782013	0.085857864	0.10907981	0.096871311
0.03	0.07036935	0.126730196	0.078786797	0.113619715	0.095306966
0.04	0.060492466	0.135640261	0.071715729	0.11815962	0.093742621
0.05	0.050615583	0.144550326	0.064644661	0.122699525	0.092178277
0.06	0.0407387	0.153460391	0.057573593	0.12723943	0.090613932
0.07	0.030861816	0.162370457	0.050502525	0.131779335	0.089049587
0.08	0.020984933	0.171280522	0.043431458	0.13631924	0.087485243
0.09	0.011108049	0.180190587	0.03636039	0.140859145	0.085920898
0.10	0.001231166	0.189100652	0.029289322	0.14539905	0.084356553

Table 4.6: Simulated probabilities of 10 categories generated by alternative (4.4) with $j = 9$ for β from 0 to 0.1 with step 0.01.

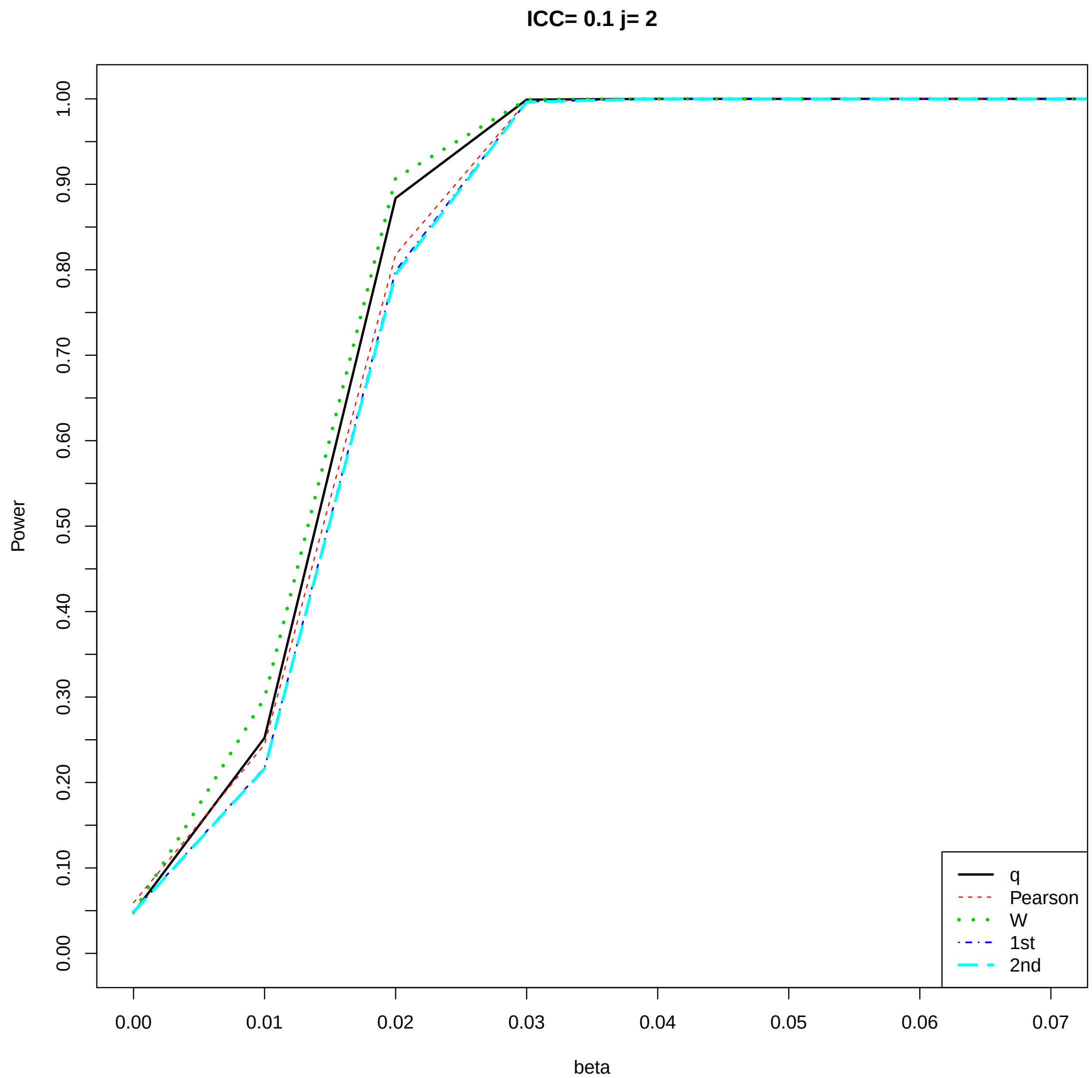

Figure 4.7: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.4) $p(k) = \frac{1}{10} + \beta cos(\frac{j\pi(k-0.5)}{10})$, for $k = 1, \cdots, 10$ with $j = 2$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.1. The Type I error of Pearson's test is a little off and all other four tests control the Type I error well. The proposed test W has the best empirical power, followed by the proposed test $\hat{q}_{0.05}$. Both of our proposed tests are better than the first order and second order corrected tests when the probabilities vary slowly. Powers of all tests converge to 1 when the probabilities vary greatly.

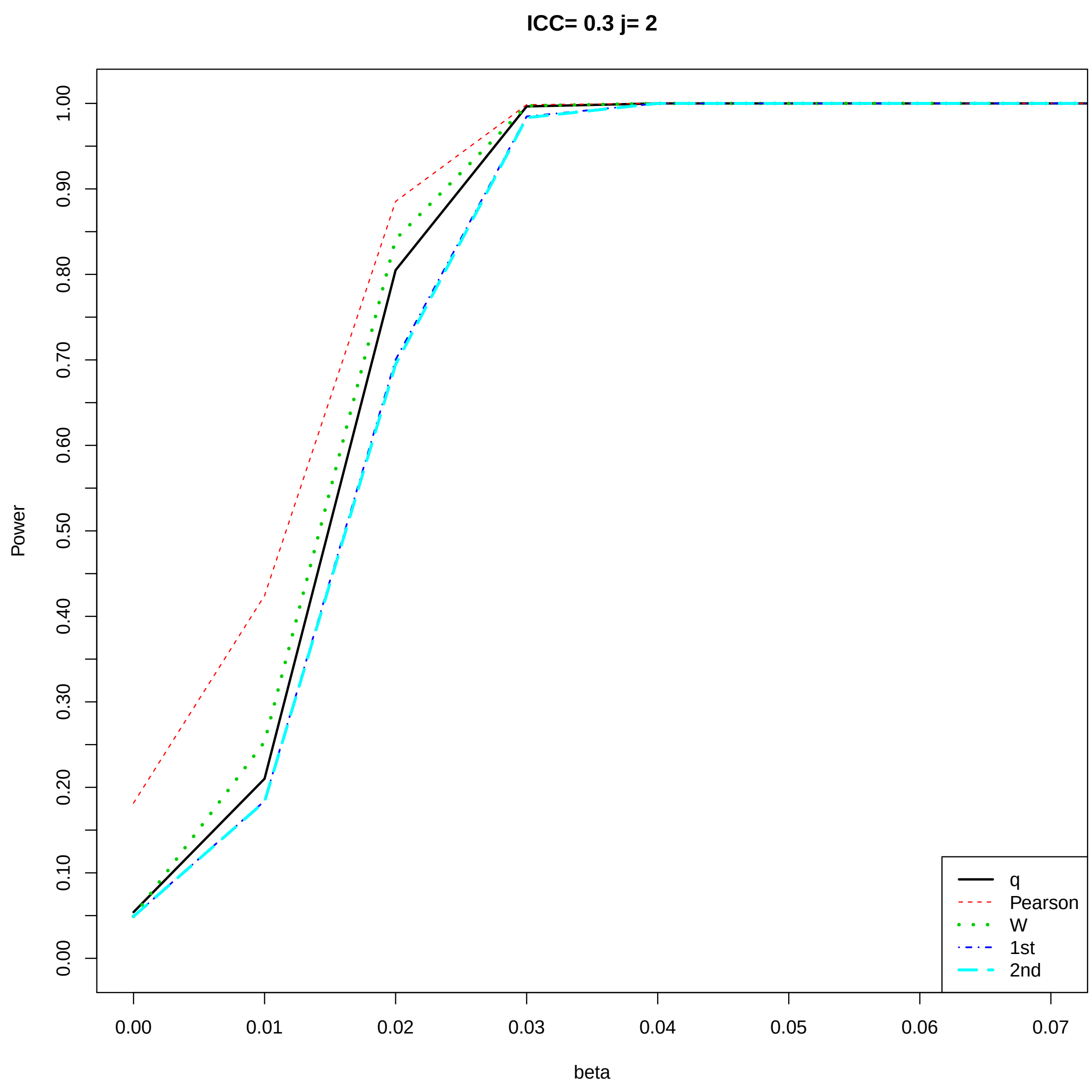

Figure 4.8: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.4) $p(k) = \frac{1}{10} + \beta\cos(\frac{j\pi(k-0.5)}{10})$, for $k = 1, \cdots, 10$ with $j = 2$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.3. Pearson's test doesn't control the Type I error well and all other four tests control the Type I error well. The proposed test W has the best empirical power, followed by the proposed test $\hat{q}_{0.05}$. Both of our proposed tests are better than the first order and second order corrected tests when the probabilities vary slowly. Powers of all tests converge to 1 when the probabilities vary greatly.

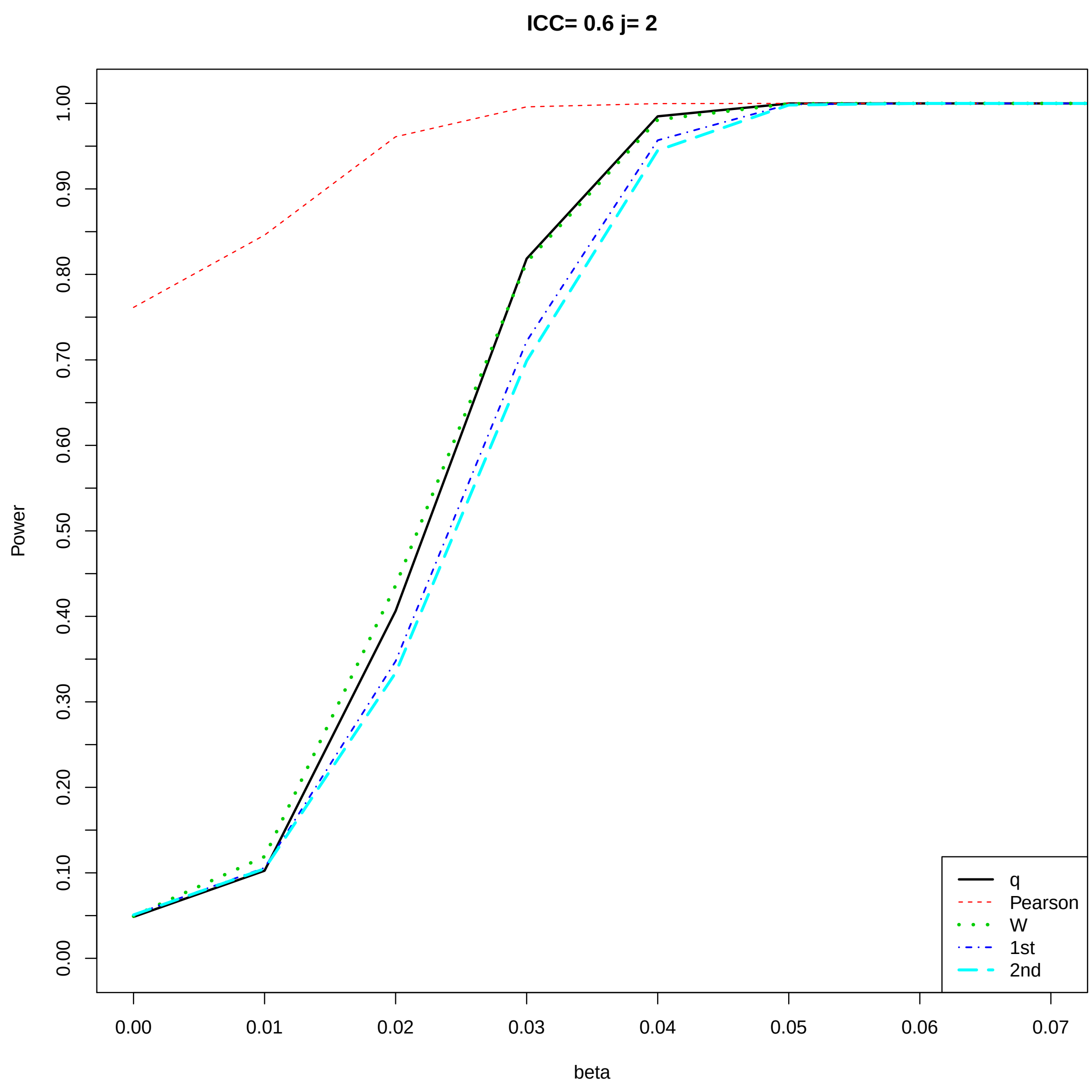

Figure 4.9: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.4) $p(k) = \frac{1}{10} + \beta\cos(\frac{j\pi(k-0.5)}{10})$, for $k = 1, \cdots, 10$ with $j = 2$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.6. Pearson's test is not able control the Type I error and all other four tests control the Type I error well. The proposed test W has the best empirical power, followed by the proposed test $\hat{q}_{0.05}$. Both of our proposed tests are better than the first order and second order corrected tests when the probabilities vary slowly. Powers of all tests converge to 1 when the probabilities vary greatly.

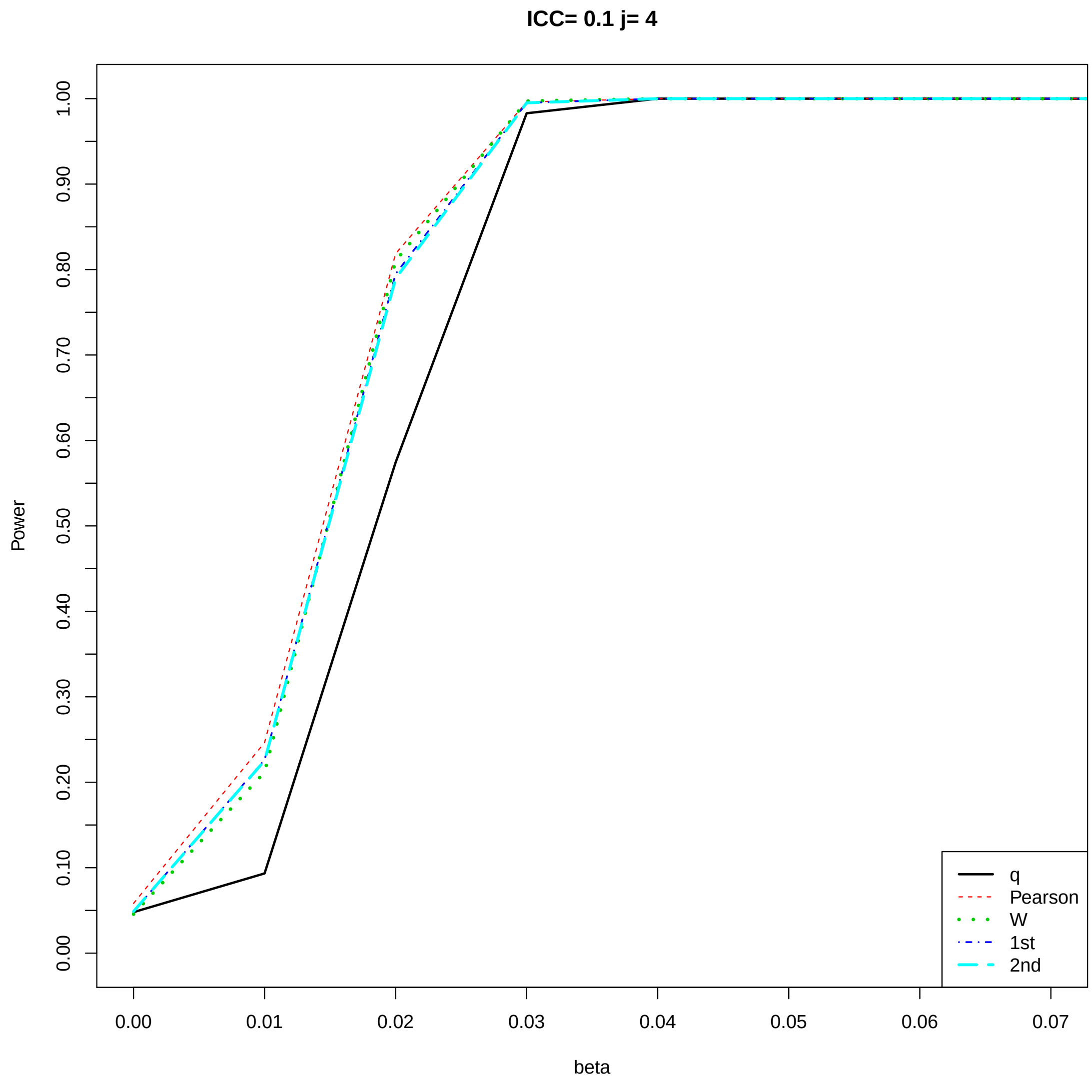

Figure 4.10: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.4) $p(k) = \frac{1}{10} + \beta\cos(\frac{j\pi(k-0.5)}{10})$, for $k = 1, \cdots, 10$ with $j = 4$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.1. The Type I error of Pearson's test is a little off and all other four tests control the Type I error well. The proposed test W is competitive with the first order and second order corrected tests. But the proposed test $\hat{q}_{0.05}$ has the lowest empirical power. Powers of all tests converge to 1 when the probabilities vary greatly.

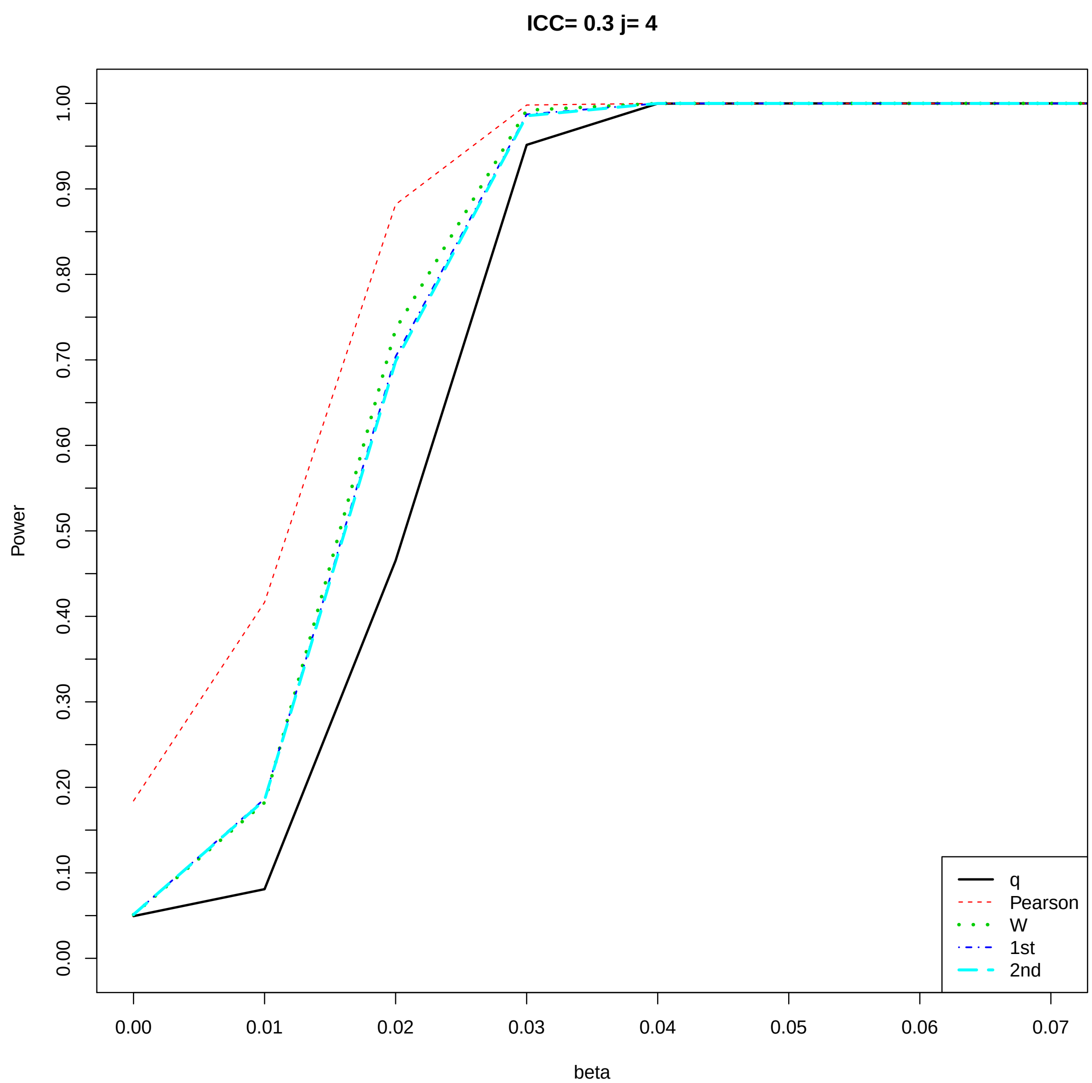

Figure 4.11: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.4) $p(k) = \frac{1}{10} + \beta\cos(\frac{j\pi(k-0.5)}{10})$, for $k = 1, \cdots, 10$ with $j = 4$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.3. Pearson's test doesn't control the Type I error well and all other four tests control the Type I error well. The proposed test W has the best empirical power, but the first order and second order corrected tests are competitive with the proposed test W. But the proposed test $\hat{q}_{0.05}$ has the lowest empirical power. Powers of all tests converge to 1 when the probabilities vary greatly.

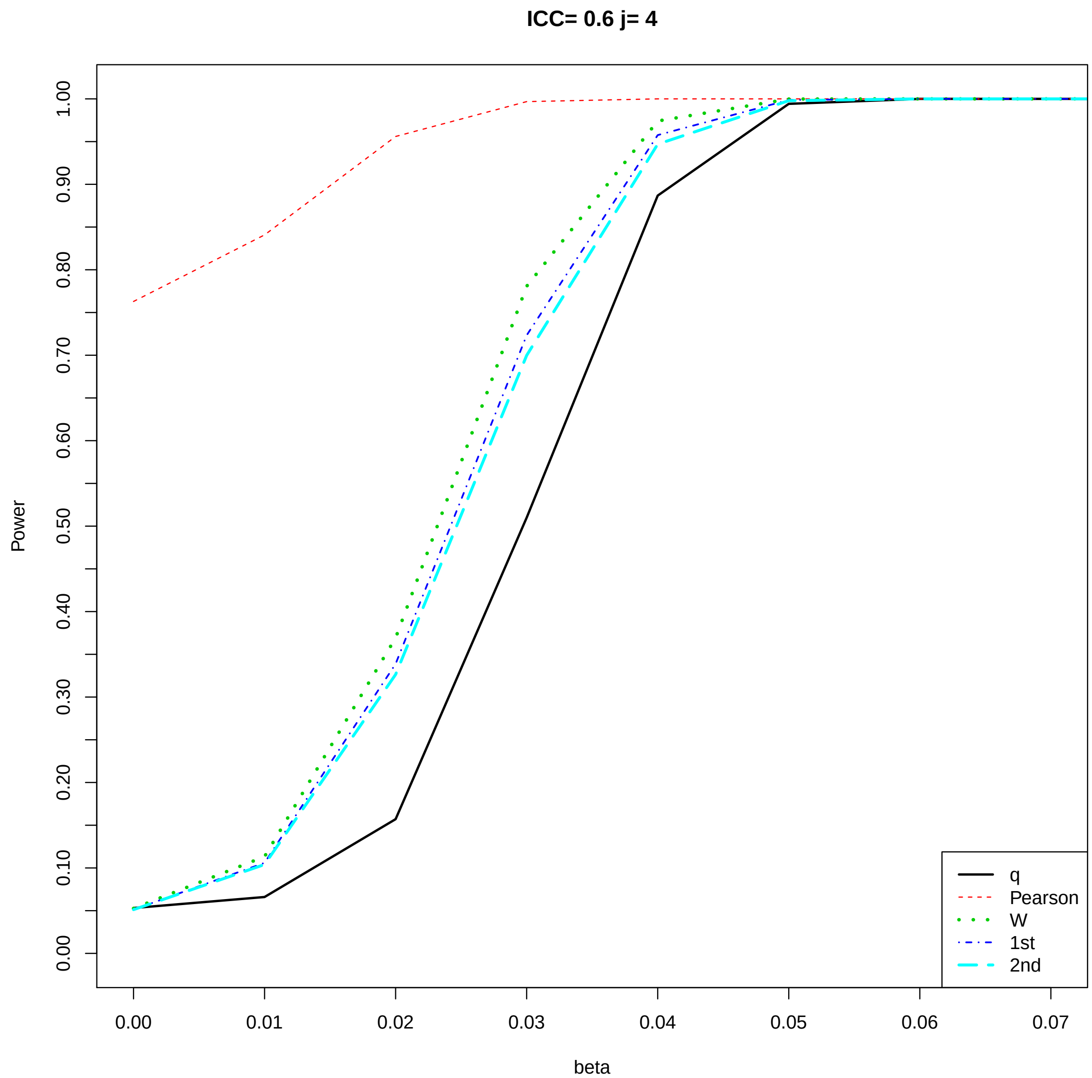

Figure 4.12: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.4) $p(k) = \frac{1}{10} + \beta\cos(\frac{j\pi(k-0.5)}{10})$, for $k = 1, \cdots, 10$ with $j = 4$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.6. Pearson's test is not able to control the Type I error and all other four tests control the Type I error well. The proposed test W has the best empirical power, followed by the first order and second order corrected tests. But the proposed test $\hat{q}_{0.05}$ has the lowest empirical power. Powers of all tests converge to 1 when the probabilities vary greatly.

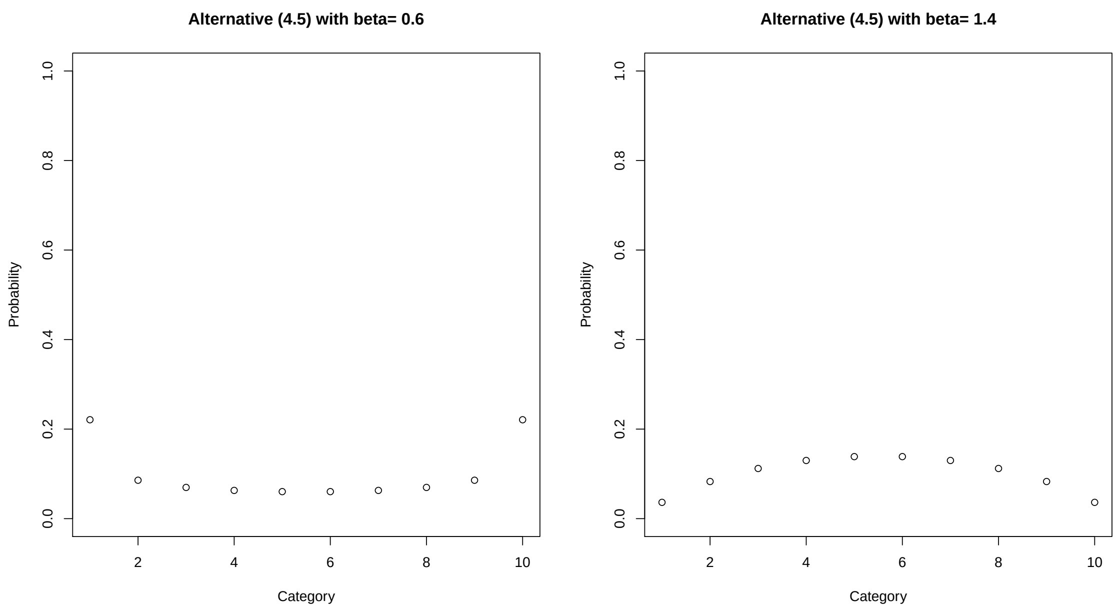

Figure 4.13: Probabilities in simulation studies generated by alternative (4.5) for $\beta = 0.6$ (left) and $\beta = 1.4$ (right). Maximum probabilities are $p(1)$ and $p(10)$ for $\beta = 0.6$, and maximum probabilities are $p(5)$ and $p(6)$ for $\beta = 1.4$.

4.2.3 Simulation Results by Alternative (4.5)

For alternative (4.5), the null hypothesis (4.1) is simulated when $\beta = 1$ for alternative (4.5). β is chosen from 0.6 to 1.4 with step 0.1. $\Phi(\cdot)$ and $\Phi(\cdot)^{-1}$ are Cumulative Distribution Function and inverse Cumulative Distribution Function of standard normal random variable, respectively. There are 9 sets of probabilities generated for the simulated multinomial data listed in Table 4.7, which are fully plotted in Figure A.6. Notice that, the maximum probability usually locates at the first and the last categories for β between 0.6 and 1.0, for example, when $\beta = 0.6$, the maximum probabilities are $p(1)$ and $p(10)$, which are both about 0.221 shown in the left graph of Figure 4.13. On the other side, the largest probabilities show up in a middle categories for β 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 4.13.

β \ $p(k)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$
0.6	0.220967155	0.085821837	0.069728394	0.063072824	0.06040979
0.7	0.184836489	0.093048915	0.078894135	0.072839961	0.070380501
0.8	0.152624684	0.097754501	0.087038736	0.082275563	0.080306516
0.9	0.124373857	0.10001357	0.094090078	0.091340798	0.090181697
1.0	0.1	0.1	0.1	0.1	0.1
1.1	0.07931315	0.09796614	0.104744293	0.108220928	0.10975549
1.2	0.062041195	0.094219511	0.108321974	0.115974971	0.119442349
1.3	0.047855011	0.089098883	0.110753904	0.123237313	0.129054889
1.4	0.036392845	0.082951702	0.11208085	0.129987045	0.138587559
β \ $p(k)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$
0.6	0.06040979	0.063072824	0.069728394	0.085821837	0.220967155
0.7	0.070380501	0.072839961	0.078894135	0.093048915	0.184836489
0.8	0.080306516	0.082275563	0.087038736	0.097754501	0.152624684
0.9	0.090181697	0.091340798	0.094090078	0.10001357	0.124373857
1.0	0.1	0.1	0.1	0.1	0.1
1.1	0.10975549	0.108220928	0.104744293	0.09796614	0.07931315
1.2	0.119442349	0.115974971	0.108321974	0.094219511	0.062041195
1.3	0.129054889	0.123237313	0.110753904	0.089098883	0.047855011
1.4	0.138587559	0.129987045	0.11208085	0.082951702	0.036392845

Table 4.7: Simulated probabilities of 10 categories generated by alternative (4.5) for β from 0.6 to 1.4 with step 0.1.

Figures 4.14, 4.15, and 4.16 plot the empirical powers of the five tests versus β with ICC 0.1, 0.3, and 0.6 separately. Note that the null hypothesis (4.1) is now simulated when $\beta = 1$, which is located in the middle of the graphs. Again, Pearson's chi-squared test is not able to reach the nominal level of significance $\alpha = 0.05$, and when ICC increases, the control of the Type I error is even worse. On the other hand, our proposed tests W and $\hat{q}_{0.05}$ can easily control the desired Type I error.

Alternative (4.5) simulates a set of moderately slow varying probabilities, and two major results can be observed. First, our proposed test W outperforms the first order and second order corrected tests, and it is superior than our proposed test $\hat{q}_{0.05}$

when $\beta < 1$. Second, when $\beta > 1$, our proposed test $\hat{q}_{0.05}$ becomes almost as good as the proposed test W, and both of them are better than first order and second order corrected tests. Overall, all the four tests, W, $\hat{q}_{0.05}$, first order and second order corrected tests are competitive in complex surveys, because the simulated underlying probabilities are neither too slow varying, nor varying dramatically.

As we mentioned, the maximum of these generated probabilities by alternative (4.5) is usually in category 1 or 10 for $0.6 \leq \beta < 1$, or in a middle category for $1 < \beta \leq 1.4$. Though all methods (not including Pearson's chi-squared test) perform similarly with each other, it can be seen that our proposed test $\hat{q}_{0.05}$ works best when the largest probability is shown in a middle category, though the differences between these two cases are not very obvious.

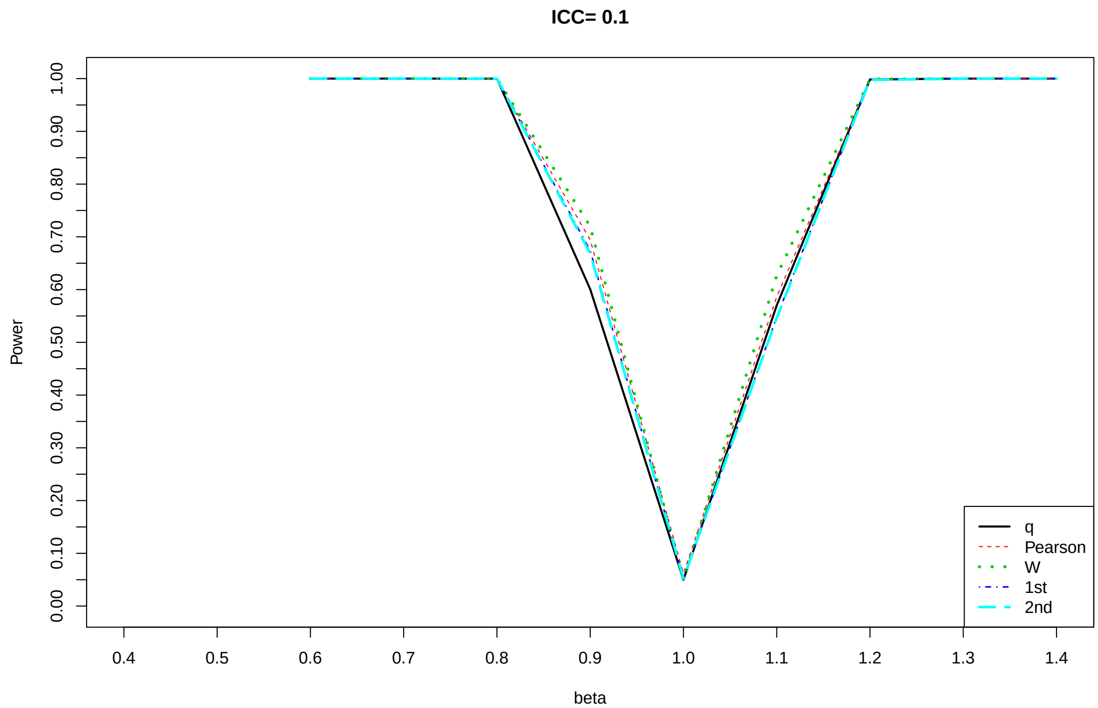

Figure 4.14: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.5) $p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right]$, for $k = 1, \cdots, 10$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.1. Notice that the null hypothesis is obtained when $\beta = 1.0$ in this case. The Type I error of Pearson's test is a little off and all other four tests control the Type I error well. The proposed test W has the best empirical power. The first order and second order corrected tests are better than the proposed test $\hat{q}_{0.05}$ when $\beta < 1$, but the proposed test $\hat{q}_{0.05}$ becomes better than the first order and second order corrected tests when $\beta \geq 1$. Powers of all tests converge to 1 when the probabilities vary greatly.

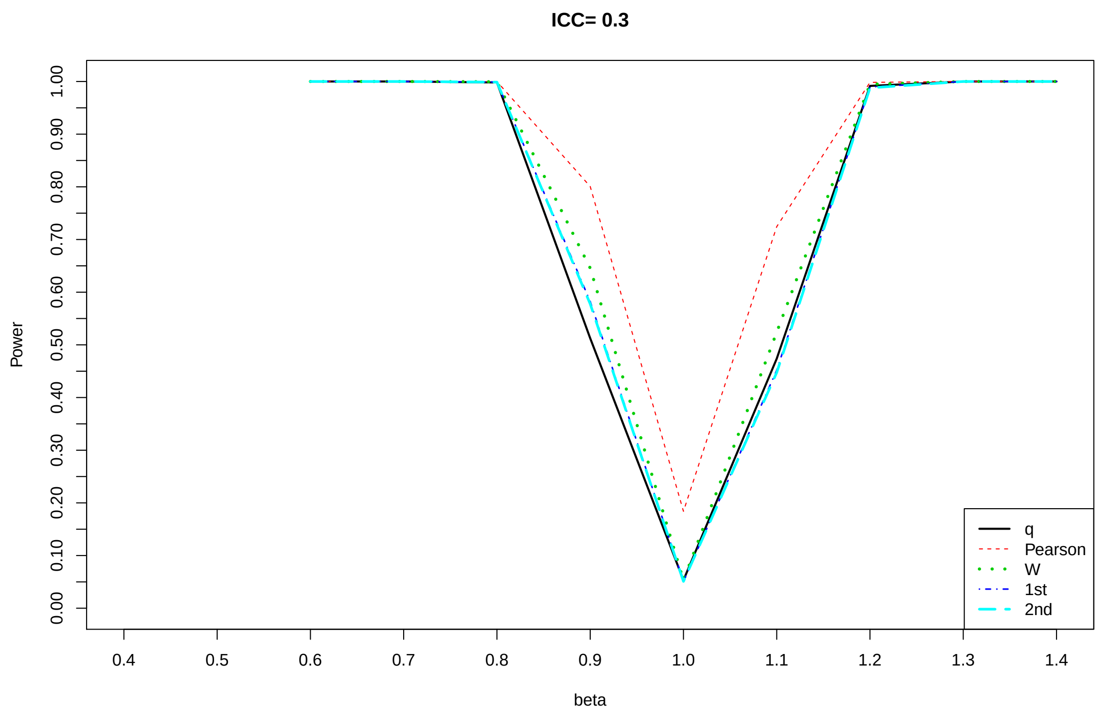

Figure 4.15: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.5) $p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right]$, for $k = 1, \cdots, 10$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.3. Notice that the null hypothesis is obtained when $\beta = 1.0$ in this case. Pearson's test doesn't control the Type I error well and all other four tests control the Type I error well. The proposed test W has the best empirical power. The first order and second order corrected tests are better than the proposed test $\hat{q}_{0.05}$ when $\beta < 1$, but the proposed test $\hat{q}_{0.05}$ becomes better than the first order and second order corrected tests when $\beta \geq 1$. Powers of all tests converge to 1 when the probabilities vary greatly.

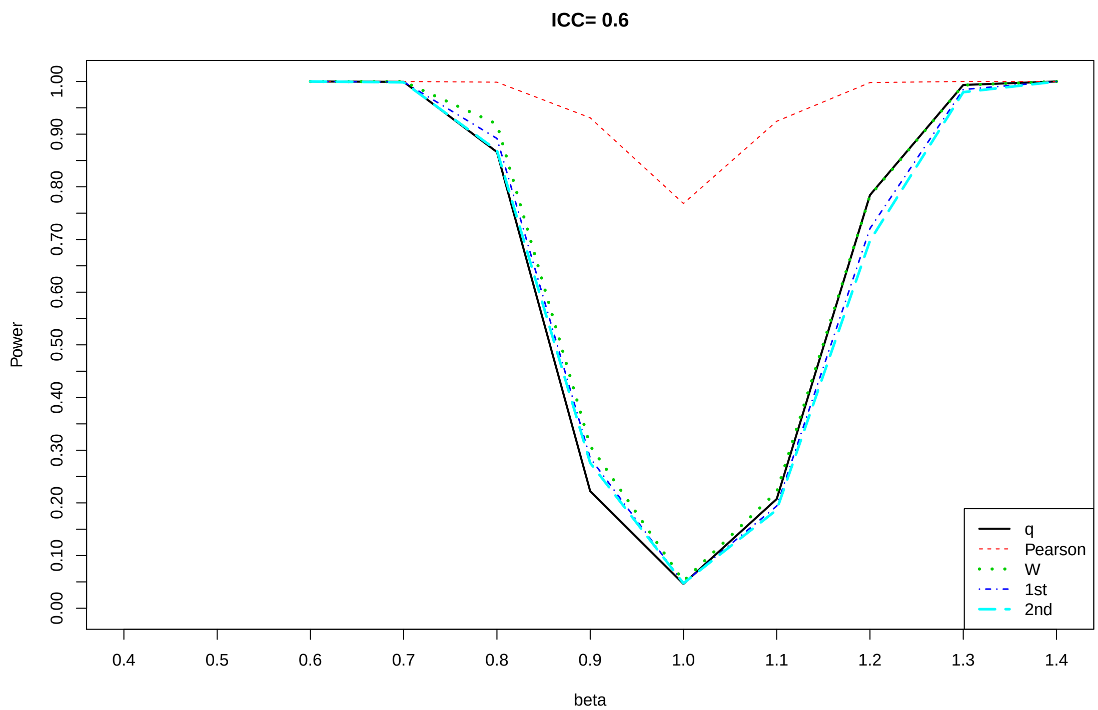

Figure 4.16: The power curves of selected methods for simulated complex survey data with respect to the alternative (4.5) $p(k) = \Phi\left[\beta\Phi^{-1}\left(\frac{k}{10}\right)\right] - \Phi\left[\beta\Phi^{-1}\left(\frac{k-1}{10}\right)\right]$, for $k = 1, \cdots, 10$. The simulations consist of 50 psus, 15 ssus, and 10 categories, with ICC 0.6. Notice that the null hypothesis is obtained when $\beta = 1.0$ in this case. Pearson's test is not able to control the Type I error and all other four tests control the Type I error well. The proposed test W has the best empirical power. The first order and second order corrected tests are better than the proposed test $\hat{q}_{0.05}$ when $\beta < 1$, but the proposed test $\hat{q}_{0.05}$ becomes as good as the proposed test W, both of which are better than the first order and second order corrected tests when $\beta \geq 1$. Powers of all tests converge to 1 when the probabilities vary greatly.

4.3 Summary

To sum up, our proposed tests have been examined by the simulation studies, and have been proved to work well in different settings. First of all, our proposed tests demonstrate strong ability to control the Type I error at the pre-specified level of significance under all settings. Second, our proposed tests substantially improve the empirical statistical powers for multinomial data in complex surveys, comparing with the first order and second order corrected tests (Rao & Scott, 1981, 1984), especially when the underlying probabilities vary slowly. Third, our proposed test W shows a great stability in both slow varying and non-slow varying cases, which usually outperforms the first order and second order corrected tests in cases with slow varying provabilities, and performs competitively in cases with medium or high varying probabilities. Fourth, our proposed test $\hat{q}_\alpha$ work best when the maximum of the underlying probabilities appears in a middle category of the multinomial data. Finally, as a result of the comparison between alternative (4.3) and (4.4), we conclude that our proposed test $\hat{q}_\alpha$ is the most powerful one for slow varying probabilities, but it may not be as stable as the proposed test W when the probabilities do not vary slowly. In another perspective of view, the proposed test $\hat{q}_\alpha$ is more sensitive to detect 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.

Chapter 5

Application

In this chapter, we apply our 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. The first example considers an artificial data set, where the responses are perfectly correlated. The second example is from the National Youth Tobacco Survey (NYTS). We are interested in testing the difference among the age/grade and severity groups on tobacco usage for Asian and American Indian/Alaska Native students.

5.1 Example 1

Data is obtained from Christensen (1997, pg. 111, Table 3.1). All individuals are randomly selected from some certain population. Information about race, sex, age and opinions on legalized abortion is recorded. Detailed description of the factors can be seen in Table 5.1. The original data is listed in Table 5.2. We define 18-25 years as the first category, 26-35 years as the second category, 36-45 years as the third category, 46-55 years as the fourth category, 56-65 years as the fifth category, and 65+ years as the last category. The corresponding counts of each age group for nonwhite females who support legalized abortion is summarized in Table 5.3.

Factor	Levels
Race	White, Nonwhite
Sex	Male, Female
Opinion	Yes=Supports Legalized Abortion
	No=Opposed to Legalized Abortion
	Und=Undecided
Age	18-25, 26-35, 36-45, 46-55, 56-65, 66+ years

Table 5.1: Description of factors from Christensen (1997, pg. 111, Table 3.1).

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Und	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Und	1	4	9	9	10	16
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Und	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Und	1	2	1	1	1	1

Table 5.2: Abortion opinion data from Christensen (1997, pg. 111, Table 3.1).

In this example, we first perform the Pearson's chi-squared test on the data in Table 5.3, to see if there are age group differences for nonwhite females who support legalized abortion. We then modify the data by treating a female's response as

Age	18-25	26-35	36-45	46-55	56-65	66+	Total
Counts	21	25	20	17	14	13	110

Table 5.3: Observed age data for nonwhite females who support legalized abortion (original data is from Christensen (1997, pg. 111, Table 3.1)).

a family's response (response from the female and her husband). The counts in each group are therefore doubled and the husband and wife's responses are perfectly correlated (assume husband and wife are in the same age group). We will use the modified data to illustrate the clustering effects, our proposed GOF tests, and the first order and second order corrected tests.

5.1.1 Pearson's Chi-squared Test in SRS

The null hypothesis of interest is that there is no age group difference for nonwhite females who support legalized abortion. The corresponding counts of people who support legalized abortion within each age group are shown in Table 5.3. The original survey is an SRS, therefore, the independence assumption is met. The null hypothesis can be written as

$$H_0 : p_0(1) = \cdots = p_0(6) = \frac{1}{6}, \tag{5.1}$$

and the corresponding alternative hypothesis is

$$H_1 : \text{at least one } p_0(k) \neq \frac{1}{6},\ k = 1, \cdots, 6. \tag{5.2}$$

Figure 5.1 shows the estimated proportions of all age groups. It can be seen that there is a decreasing trend of supporting legalized abortion, as the nonwhite females get older. Furthermore, all estimated proportions are also very close to each other (they vary between 0.118 and 0.227).

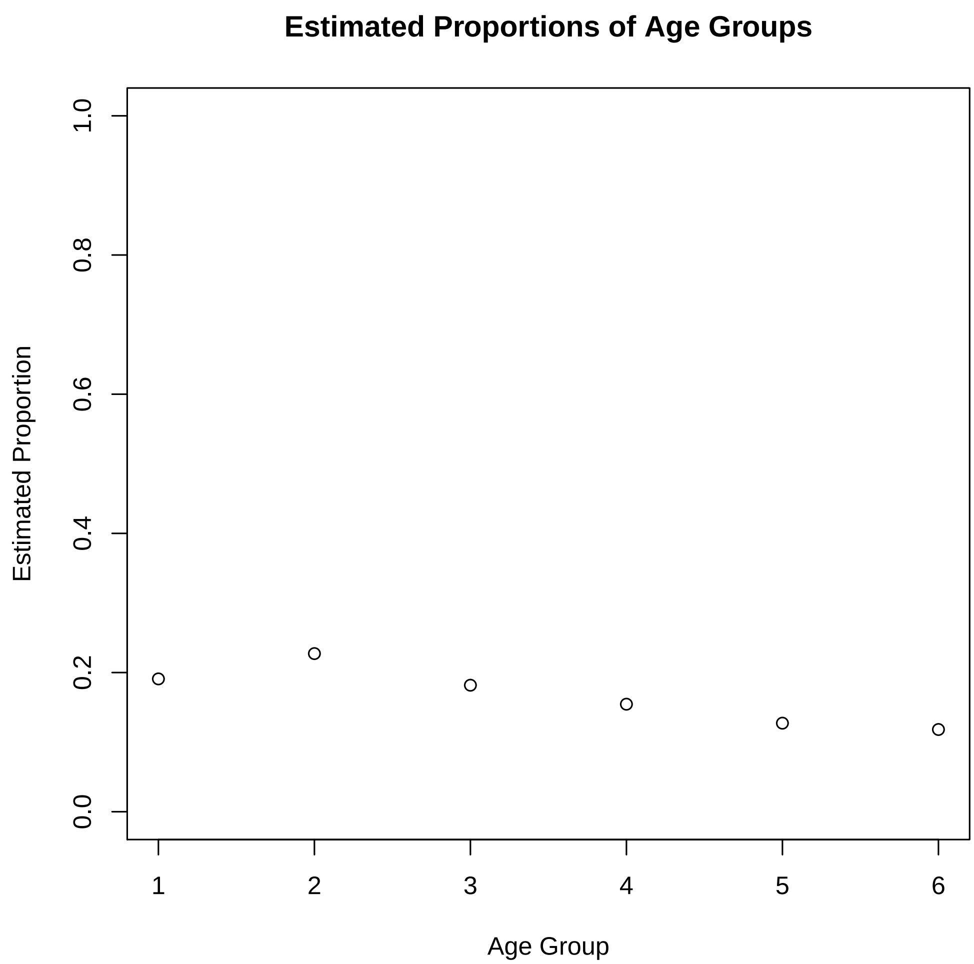

Figure 5.1: Estimated proportions of 6 age groups in Table 5.3.

We first apply Pearson's chi-squared test to test the hypothesis (5.1). The expected counts for each age group under the null hypothesis are shown in Table 5.4. The test statistic is calculated by

Age	18-25	26-35	36-45	46-55	56-65	66+
Expected Counts	18.33	18.33	18.33	18.33	18.33	18.33

Table 5.4: Age data for nonwhite females who support legalized abortion, under the null hypothesis (5.1).

$$X^2_{SRS} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = 5.6364,$$

which is compared with central chi-squared distribution with 5 degrees of freedom. We get the p-value as 0.3432, which fails to reject the null hypothesis at level of significance 0.05 and 0.1. This indicates that there is no difference between the age groups for nonwhite females who support legalized abortion.

5.1.2 Apply Proposed GOF Tests on Correlated Data

Now, suppose the surveys are given to both these females and their spouses, and assume their spouses give the same answers as these females and each couple is in the same age group. Mathematically, this implies that these females and their spouses are perfectly correlated (ICC = 1). The modified data is shown in Table 5.5. Notice that, based on this table, the estimated proportions of all age groups remain

Age	18-25	26-35	36-45	46-55	56-65	66+	Total
Counts	42	50	40	34	28	26	220

Table 5.5: Observed data for nonwhite females and their spouses who support legalized abortion, assuming that both these females and their spouses give identical answers (ICC = 1).

the same, but the sample size and counts of all categories are doubled.

Suppose that we are interested in testing if there are differences in the rate of supporting legalized abortion among the age groups of the nonwhite families. If Pearson's chi-squared test is utilized again, because the sample size is twice as before, the value of the test statistic is also doubled,

$$X^2 = n \sum_{k=1}^{K} \frac{\left(\hat{p}(k) - p_0(k)\right)^2}{p_0(k)} = 11.2728,$$

We find that the p-value is 0.04625, and thus we reject the null hypothesis at level 0.05. In a further investigation, this result is not correct, because each couple forms a natural cluster with ICC 1 and thus the sampling design is no longer SRS. As a result, we are more likely to reject the null hypothesis than we should do. The first order and second order corrected tests may be a better fit for such a situation. We then re-examine the hypothesis test using the first order and second order corrected tests.

In order to get $\hat{\delta}_.$ in (2.15), $100,000$ complex multinomial data under the null hypothesis (5.1) are generated. For each of the generated data, the sample covariance matrix is found out. Then, covariance matrix under the null hypothesis (5.1) is estimated by the average of the $100,000$ covariance matrix ($\hat{\mathbf{V}}$). Note that, the covariance matrix under SRS and the null hypothesis (5.1) is known. Therefore, $\hat{\delta}_.$ can be obtained by averaging the eigenvalues of the matrix $\mathbf{P}_0^{-1}\mathbf{V}$. In addition, $\hat{a}^2$ is calculated by (2.17).

The test statistics of the first order and second order corrected tests are

$$X_C^2 = \frac{X_{SRS}^2}{\hat{\delta}_.} = 5.68865$$

and

$$X_S^2 = \frac{X^2}{\hat{\delta}_.(1+\hat{a}^2)} = \frac{X_C^2}{(1+\hat{a}^2)} = 5.684551.$$

The test statistic of the first order corrected test (X_C^2) is compared with central chi-squared distribution with 5 degrees of freedom and the test statistic of the second order corrected test (X_S^2) is compared with central chi-squared distribution with $5/(1+\hat{a}^2)$ degrees of freedom. The p-values for these two test statistics are both about 0.34, which implies that the null hypothesis should not be rejected by both tests at either level 0.05 or level 0.1. However, we do observe an decreasing trend from age groups 2 to 6. Next, we will use our proposed Neyman smooth-type GOF tests to investigate the problem again.

For our first proposed test W, under the null hypothesis (5.1), we search for the maximizer of the equation (3.11) to get $\hat{q}$ and use (3.12) to find a value of W_0. This process is repeated $100,000$ times and the empirical distribution of W_0 is discovered. $\hat{q}$ is the one that maximizes equation (3.11) for $q = 1, \cdots, K-1$. For the data in Table 5.5, it is found that

$$\hat{q} = 1,$$

which results in, according to (3.12),

$$W = \frac{X_1^2 - 1}{\sqrt{2}} = 6.109525.$$

The simulated critical values of W_0 are 5.226383 and 6.561806 for $\alpha = 0.1$ and $\alpha = 0.05$, respectively. With a p-value of 0.065, the null hypothesis (5.1) should be rejected at a level between 0.1 and 0.05.

For our second proposed test $\hat{q}_{0.05}$, $a_{0.05}$ in (3.15) is found to be

$$\hat{a}_{0.05} = 4.18,$$

and the the test statistic is

$$\hat{q}_{0.05} = 1.$$

As a result, the null hypothesis (5.1) should be rejected at level 0.05 using this method.

In this example, the estimated probabilities of the age groups are slow varying with a sample size of 110. Though the differences among the age groups are observed, existing methods, such as the first order and the second order corrected tests, are not able to reach a consistent conclusion with the observed phenomenon. Our proposed tests successfully detect the heterogeneity among age groups, which confirms the observed phenomenon by using a limited sample size. We conclude that there is significant difference among the six age groups in nonwhite families who support legalized abortion.

5.2 Example 2

The National Youth Tobacco Survey (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, and 2014. The Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA) have involved in the management of NYTS since 2011.

The latest published data is for the year 2014 (Office on Smoking and Health, 2014). According to the Office on Smoking and Health (2014), the 2014 NYTS sampling design involved stratification and three-stage clustering so that students in middle schools and high schools in 50 U. S. states and the District of Columbia could be mostly represented. 16 strata were created in U.S. based on predominant minority (non-Hispanic Black and Hispanic) and the factor urban/nonurban, as shown in Table 5.6 (Office on Smoking and Health, 2014, pg. 9). A psu was 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

Predominant Minority	Urban / Nonurban	Density Group Number	Stratum Code	Student Population	Number of Sample PSUs
Non-Hispanic Black	**Urban**	1	BU1	2,720,181	9
		2	BU2	975,490	3
		3	BU3	908,299	3
		4	BU4	516,712	2
	Nonurban	1	BR1	3,937,157	12
		2	BR2	1,503,403	5
		3	BR3	1,026,612	4
		4	BR4	313,063	2
Hispanic	**Urban**	1	HU1	3,530,556	11
		2	HU2	2,429,442	7
		3	HU3	1,865,988	5
		4	HU4	2,106,242	7
	Nonurban	1	HR1	4,427,215	14
		2	HR2	1,284,402	4
		3	HR3	988,655	3
		4	HR4	523,491	2

Table 5.6: Definitions of 16 strata from Exhibit 2-2 in Office on Smoking and Health (2014, pg. 9).

the selected classes were eligible for the survey. Sampling steps are shown as follows and in Figure 5.2, according to the Office on Smoking and Health (2014, pg. 5).

1. The U.S. is divided into 16 strata according to predominant minority (non-Hispanic Black and Hispanic) and the factor urban/nonurban.

2. A total of 93 psus/counties were randomly chosen from the 16 strata with pre-determined probabilities, which are proportional to the number of students in psus.

3. Schools in each selected psu/county were stratified as large, medium, and small schools. Within each psu, 170 large schools, 20 medium schools, and 30 small

schools were randomly selected.

4. Within each selected large school, 2 classes of each grade (grades 6-12) were randomly selected; within each selected medium or small school, 1 class of each grade (grades 6-12) was randomly selected.

5. All students in selected classes were eligible for the interviews.

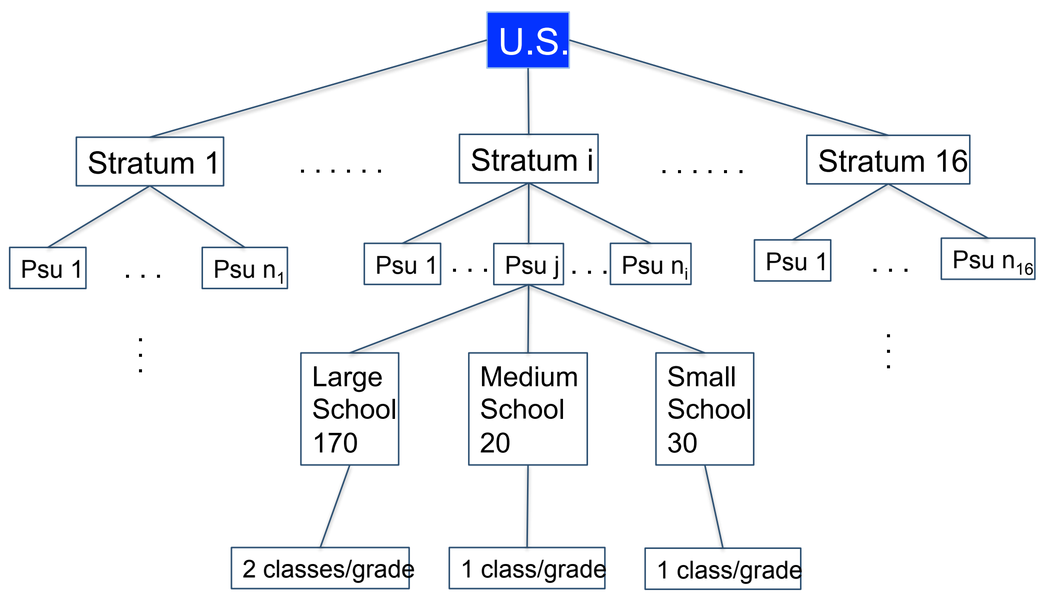

Figure 5.2: NYTS 2014 sampling design chart.

All sampling was without replacement. More details can be found in Chapter 2 of Office on Smoking and Health (2014).

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 the nonresponse adjustments. Office on Smoking and Health (2014) introduces the sampling weights calculation in detail in Chapter 4. The final 2014 NYTS data consists of 157 variables (including weight variable) and a total of 22,007 observations.

5.2.1 Age Differences Among Asian Students Who Smoked

In this section, we focus on Asian students who have tried cigarette smoking before, even a puff or two. There are a total of 973 Asian students who participated in NYTS 2014 and 114 of them admitted that they had cigarette smoking experience before. The age of the first try ranges from ≤ 8 years old to 18 years old. In this example, we define 8 groups for the age of the first try, ≤ 8 years old, $9-10$ years old, $11-12$ years old, 13 years old, 14 years old, 15 years old, 16 years old and $17-18$ years old.

The null hypothesis of interest is that there are no differences in age groups of the first cigarette try, which is seated in (5.3). The observed data from NYTS 2014 is listed in Table 5.7 and the weighted data is shown in Table 5.8.

$$H_0 : p(1) = \cdots = p(8) = \frac{1}{8} \tag{5.3}$$

Age	≤ 8	9-10	11-12	13	14	15	16	17-18	Total
Counts	19	12	17	18	14	13	17	4	114

Table 5.7: Observed data for Asian students who have tried cigarette smoking from NYTS 2014. Age indicates the age of a student's first try of cigarette smoking, even a puff or two.

Age	8	9-10	11-12	13	Total
Counts	18121.18	8111.43	15323.68	20617.98	
Age	14	15	16	17-18	
Counts	13362.30	11764.84	11408.68	2306.368	101016.5

Table 5.8: Weighted data for Asian students who have tried cigarette smoking from NYTS 2014. Age indicates the age of a student's first try of cigarette smoking, even a puff or two.

The estimated proportions of the weighted counts in Table 5.8 are plotted in

Figure 5.3. The estimated proportions $\hat{p}(k)$'s in (B.1) using the weighted counts are given in Table 5.9.

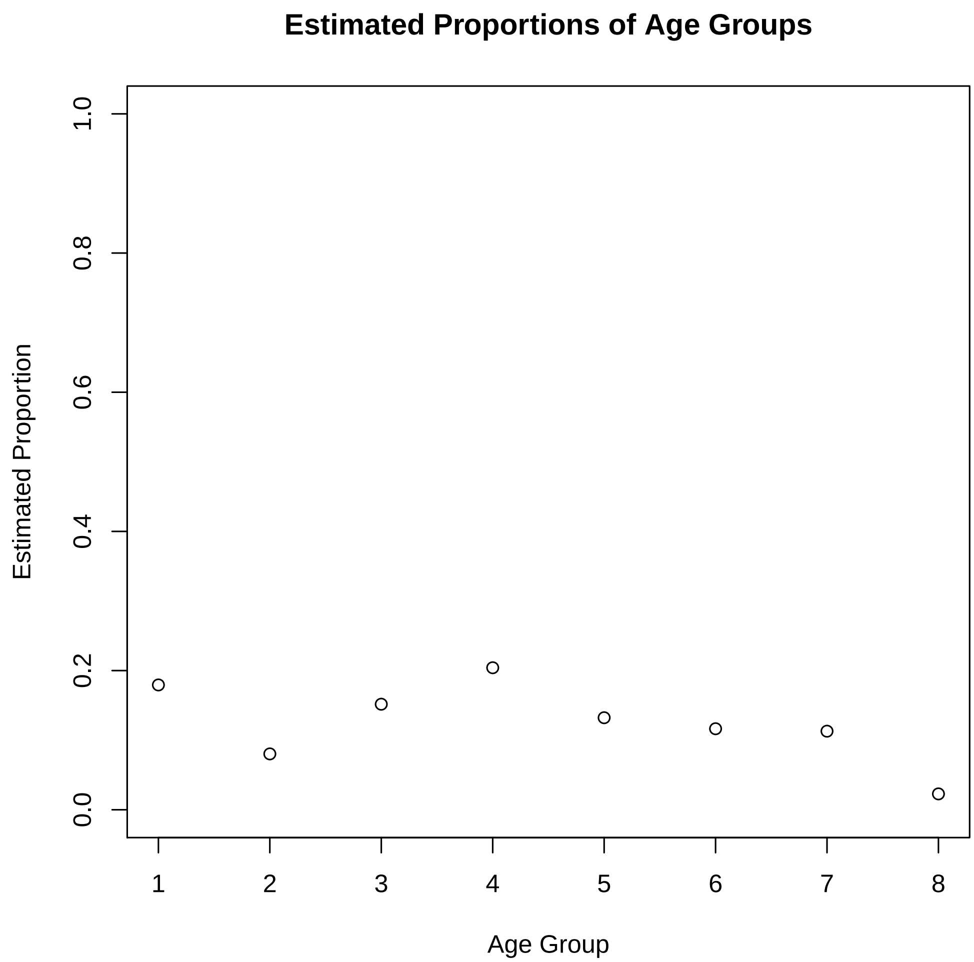

Figure 5.3: Estimated proportions of 8 age groups using weighted data in Table 5.8.

The observed sample size in Table 5.7 and Kish's effective sample size (B.3) are $n = 114$ and $\tilde{n} = 88.96$, respectively. $\hat{\delta}_{\cdot}$ in (2.15) is obtained using (3.16). In addition, $\hat{a}^2$ is calculated by (2.17). With $p_0(k) = \frac{1}{8}$ for $k = 1, \cdots, 8$, the test statistics of the

Age	≤ 8	9-10	11-12	13	14	15	16	17-18	Total
$\hat{p}(k)$	0.179	0.08	0.152	0.204	0.132	0.116	0.113	0.023	1

Table 5.9: Estimated proportions using weighted data in Table 5.8 for Asian students who have tried cigarette smoking from NYTS 2014. Age indicates the age of a student's first try of cigarette smoking, even a puff or two.

first order and second order corrected tests are

$$X_C^2 = \frac{X^2}{\hat{\delta}_\cdot} = \sum_{k=1}^{K} n \frac{(\hat{p}(k) - p_0(k))^2}{p_0(k)} / \hat{\delta}_\cdot = 16.1093$$

and

$$X_S^2 = \frac{X^2}{\hat{\delta}_\cdot(1 + \hat{a}^2)} = \frac{X_C^2}{(1 + \hat{a}^2)} = 16.11.$$

Thus, the null hypothesis (5.3) should be rejected at level $\alpha = 0.05$, because the p-values of both the first order and the second order corrected tests are $0.024 < 0.05$.

Next, we will use our proposed methods to re-examine the data again. For the proposed test W, under the null hypothesis (5.3), we search for the maximizer of the equation (3.11) to get $\hat{q}$ and then a value of W_0 is obtained using (3.12). This process is repeated $100,000$ times and the empirical distribution of W_0 is simulated. Then, by searching all $q = 1, \cdots, K-1$, $\hat{q}$ is found out to be the one that maximizes equation (3.11). For the data in Table 5.7, it is found that

$$\hat{q} = 5,$$

which results in, according to (3.12),

$$W = \frac{X_1^2 - 1}{\sqrt{2}} = 4.92.$$

Compared with the simulated distribution W_0 under the null hypothesis (5.3), the p-value of the proposed test W is 0.0078, which rejects the null hypothesis (5.3) at level $\alpha = 0.05$.

For the proposed test $\hat{q}_\alpha$, $a_{0.022}$ and $a_{0.023}$ in (3.15) are found to be

$$\hat{a}_{0.022} = 5.44 \text{ and } \hat{a}_{0.023} = 5.37.$$

The corresponding test statistics are

$$\hat{q}_{0.022} = 0 \text{ and } \hat{q}_{0.023} = 1,$$

which implies that p-value of the proposed test $\hat{q}_\alpha$ is about 0.023, so the null hypothesis (5.3) should be rejected at level $\alpha = 0.05$.

In this example, all four tests reject the null hypothesis (5.3) at level $\alpha = 0.05$. The p-value of the proposed test $\hat{q}_\alpha$ is similar to those of the first order and the second order corrected tests. However, the p-value of the proposed test W is smaller than those of all other three tests. If the level of significance is chosen to be $\alpha = 0.01$, only our proposed test W is able to reject the null hypothesis.

5.2.2 Grade Differences Among American Indian and Alaska Native Students Who Smoked

In this section, the target population is American Indian and Alaska Native (AIAN) students who have tried cigarette smoking before, even a puff or two. A total of 338 students participated in NYTS 2014 and 77 of them have had cigarette smoking experience before. These students were in grades 6-12 and each grade forms a category in our data analysis. For example, grade 6 is the first category, and grade 12 is the 7th category.

The null hypothesis of interest is that there are no differences in grades for these students, which can be written mathematically as in (5.4). The observed counts are shown in Table 5.10 and the weighted counts are listed in Table 5.11.

$$H_0 : p(1) = \cdots = p(7) = \frac{1}{7} \tag{5.4}$$

Grade	6	7	8	9	10	11	12	Total
Counts	11	8	17	10	19	6	6	77

Table 5.10: Observed data for AIAN students who have tried cigarette smoking from NYTS 2014. Grade indicates the current grade when a student was in the survey.

Grade	6	7	8	9	Total
Counts	3198.14	2094.63	5560.12	6843.18	
Grade	10	11	12		
Counts	14394.36	6383.73	6157.99		44632.14

Table 5.11: Weighted data for AIAN students who have tried cigarette smoking from NYTS 2014. Grade indicates the current grade when a student was in the survey.

Figure 5.4 plots the estimated proportions of the weighted data in Table 5.11. The estimated proportions $\hat{p}(k)$'s in (B.1) using the weighted counts are given in Table 5.12.

Grade	6	7	8	9	10	11	12	Total
$\hat{p}$	0.072	0.047	0.125	0.153	0.323	0.143	0.138	1

Table 5.12: Estimated proportions using weighted data in Table 5.11 for AIAN students who have tried cigarette smoking from NYTS 2014. Grade indicates the current grade when a student was in the survey

Following the steps shown in Section 5.2.1, the test statistics of the first order and second order corrected tests are

$$X_C^2 = X_S^2 = 16.47,$$

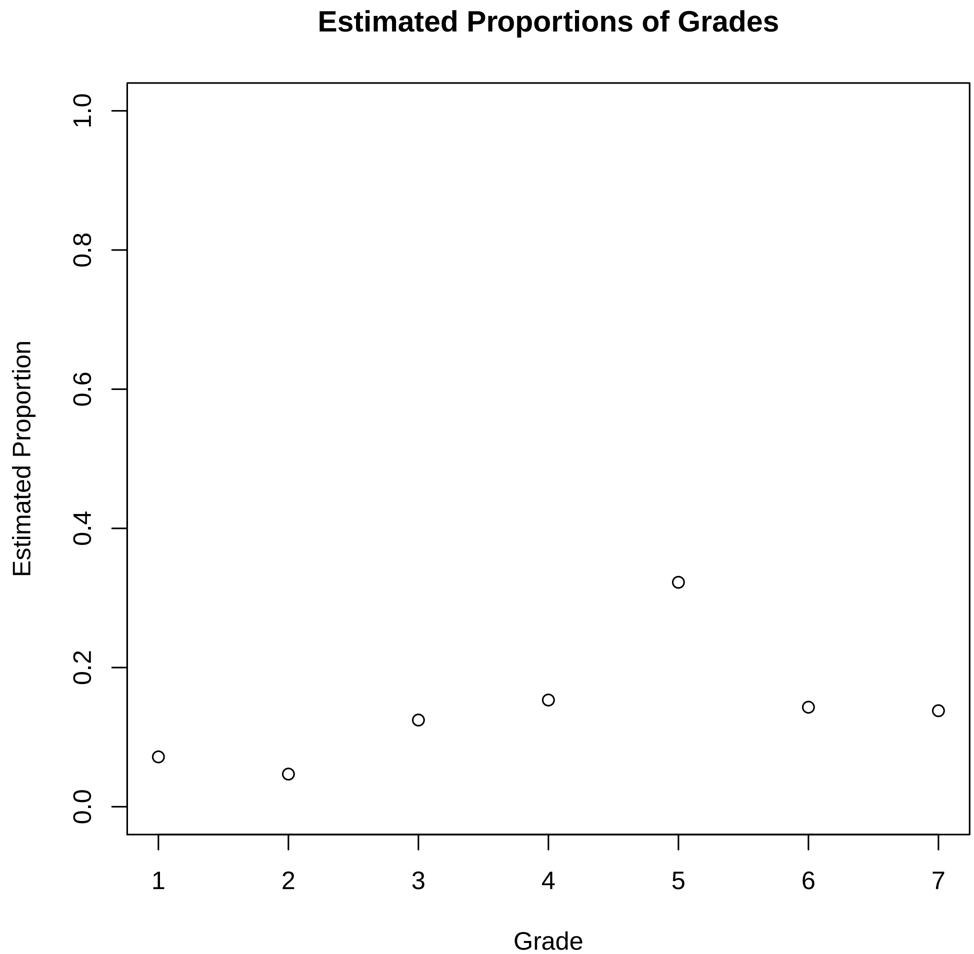

Figure 5.4: Estimated proportions of 7 grades in Table 5.10.

which results in the p-values of 0.011 for both of the tests.

In addition, for the proposed test W, it is found that

$$\hat{q} = 3 \text{ and } W = 6.62$$

which results in a p-value of 0.0017.

For the proposed test $\hat{q}_\alpha$, with $a_{0.012} = 6.44$ and $a_{0.011} = 6.58$, the test statistics are

$$\hat{q}_{0.012} = 1 \text{ and } \hat{q}_{0.011} = 0,$$

equivalently, the p-value is about 0.012.

In this example, the null hypothesis (5.4) is rejected by all four tests at level $\alpha = 0.05$. The p-value of the proposed test $\hat{q}_\alpha$ is close to those of the first order and second order corrected tests. But the proposed test W has a smaller p-value than the other three tests. Only our proposed test W can reject the null hypothesis at level $\alpha = 0.01$.

5.2.3 Severity Differences Among Asian Students Smokers

In this example, we focus on Asian students who smoked during the past 30 days in the survey. 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 smoking severity among these students. The null hypothesis is

$$H_0 : p(1) = \cdots = p(5) = \frac{1}{5}. \tag{5.5}$$

The observed and the weighted counts of the 5 levels are shown in Table 5.13 and 5.14.

The estimated proportions of the weighted counts in Table 5.14 are listed in Table 5.15 and plotted in Figure 5.5.

Number Group	< 1	1	2-5	6-10	≥ 11	Total
Counts	7	6	8	2	2	25

Table 5.13: Observed data for Asian students who reported smoking during the past 30 days of the survey from NYTS 2014. Number group indicates the number of cigarettes smoked per day.

Number Group	< 1	1	2-5	Total
Counts	6840.261	5818.418	6595.912	
Number Group	6-10	≥ 11		
Counts	1391.909	1907.703		22554.2

Table 5.14: Weighted data for Asian students who reported smoking during the past 30 days of the survey from NYTS 2014. Number group indicates the number of cigarettes smoked per day.

Number Group	< 1	1	2-5	6-10	≥ 11	Total
$\hat{p}(k)$	0.303	0.258	0.292	0.062	0.085	1

Table 5.15: Estimated proportions using weighted data in Table 5.14 for Asian students who reported smoking during the past 30 days of the survey from NYTS 2014. Number group indicates the number of cigarettes smoked per day.

Following the steps shown in Section 5.2.1, the test statistics of the first order and second order corrected tests are

$$X_C^2 = X_S^2 = 5.62,$$

with p-values 0.23 for both tests.

For the proposed test W, it is found that

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

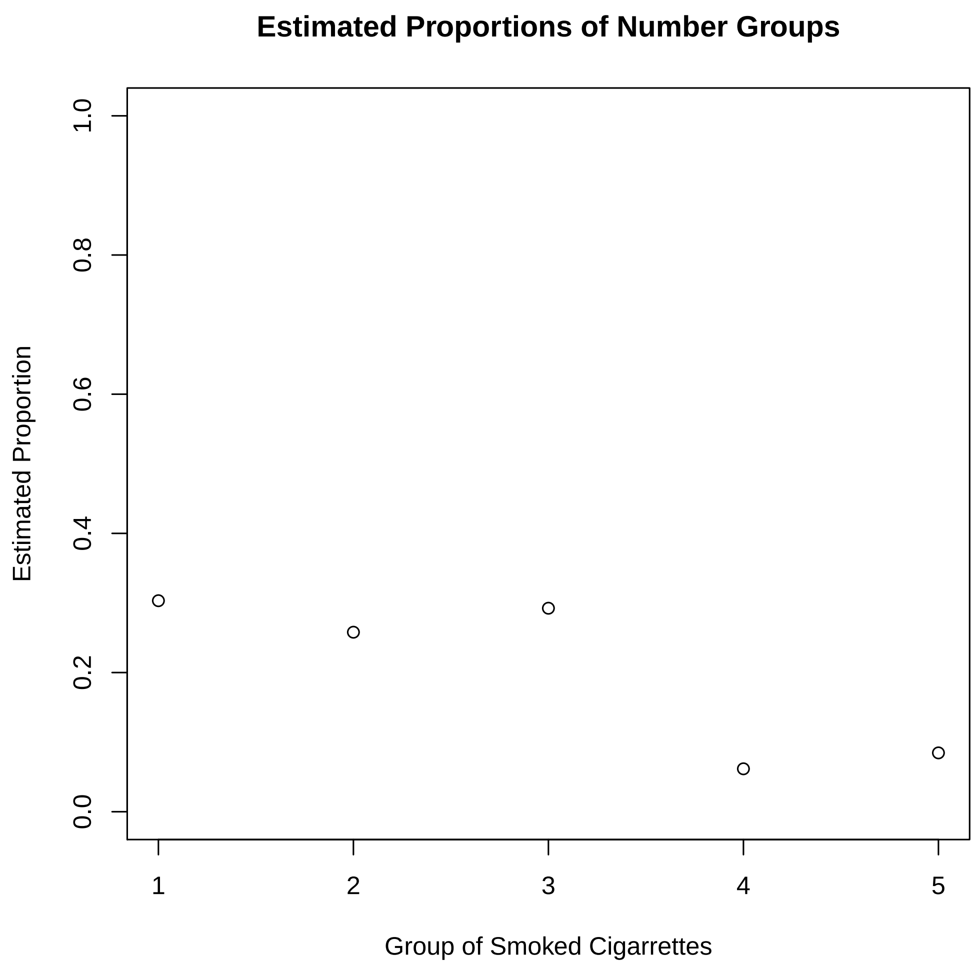

Figure 5.5: Estimated proportions of the 5 smoking severity levels using weighted data in Table 5.14.

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

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

In this example, the first order and second order corrected tests fail to reject the null hypothesis (5.5) 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 of our 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 to what was observed.

Chapter 6

Conclusion

Categorical data analysis is widely used in sociological, behavioral, economical, and medical research studies. For categorical data, one of the interests is to measure the "goodness of fit" between the observed frequencies and the frequencies under the null hypothesis. Pearson's chi-squared test is one of the most commonly used tools for such purpose. It is noticed that simple random sampling may not be suitable for complicated practical situations. Therefore, sampling designs, such as stratification and clustering, are often used to accommodate the needs of data collection that fits the real situations. A complex survey is considered as a combination of several sampling designs, where observations are usually correlated. Existing GOF tests, which require the assumption of independence, no longer work for the complex survey data. Rao and Scott (1981, 1984) proposed the first order and second order corrected tests for use in complex categorical data. Rao and Thomas (1988) reviewed 25 GOF tests for use in complex surveys. They found that the Rao-Scott approaches are efficient.

Another problem of the classical GOF tests is that they may not provide enough statistical power in order to detect differences among categories, especially when the probabilities of categories vary slowly with respect to the null hypothesis (3.1) and the sample size is relatively small. Eubank (1997) proposed a Neyman smooth-type GOF tests by incorporating the order selection. Eubank's proposed tests show improvements over classical GOF tests in detecting some types of slow-varying alternatives. However, Eubank's tests are also only applied to data under the independence assumption.

As we have discussed before, in complex surveys, data are often correlated. For example, we are interviewing husband and wife for their opinion of legalized abortion. The couple usually will have the same opinion and therefore, they are perfectly correlated. Further, as we can see from Figure 5.1, the rate at which legalized abortion is supported vary slowly among the different age groups in nonwhite families. Such data will require new methods for detecting slow varying differences among the groups with correlated responses.

In our research, we have proposed two Neyman smooth-type GOF tests for use in complex surveys. These methods show improved statistical powers, compared with classical methods, especially when the sample size is relatively small and the differences of the estimated proportions of categories are not great. In the procedures of our approaches, chi-squared type test statistic is decomposed into ordered orthogonal components with the first few terms carrying the principal information of the data. Then, order selection is utilized to choose the dominant components such that more degrees of freedom are released. Hence, our proposed tests are more sensitive to the cases of slow varying probabilities. Simulation results reveal that the test using $\hat{q}_\alpha$ is the most powerful test for the data with slow varying probabilities compared to some existing methods. The proposed test W is a stable test, which outperforms the first order and second order tests when the probabilities are slow varying, or is as good as the first order and second order corrected tests when the probabilities are varying greatly. We also investigated the asymptotic properties of the proposed estimators in stratified sampling or unequal probability sampling with one observation from each psu. First, the asymptotic distribution of the test statistics (b_j's and X_q^2) are found. Second, asymptotic properties of $\hat{q}$ are derived. Finally, asymptotic properties of $\hat{q}_\alpha$ are examined so that the test using order directly at level α is developed. Several real-life data examples are used to illustrate our proposed methods.

Our research can be applied to broader areas, where the GOF tests of $H_0 : p(1) = \cdots = p(k) = \frac{1}{K}$ for correlated observations are of interest. Our methods are potentially more powerful than the existing methods, which means that our methods need smaller sample sizes to reach similar power than the existing methods. This is important in many real applications. For example, in clinical trials, the experimental unit and the experiment itself may be too expensive to increase the sample size.

Our proposed methods in this dissertation can be considered as the first order correction to the tests proposed by Eubank (1997). One of the future research directions is to integrate the second order corrected test with the Neyman smooth-type GOF test. We are also interested in working on the general cases with the null hypothesis $H_0 : p(k) = p_0(k)$, where $p_0(k)$ is an arbitrary probability. Another future research is to derive the asymptotic properties of the estimators in cluster sampling. Since our proposed methods can be implemented in any programming language, we plan to develop an R package first, and then SAS macro and Python package for use by the readers. We also plan to extend our research to assess GOF of multi-dimensional multinomial data.

The idea of the Fourier transformation and order selection can also be extended to test for no effect in nonparametric regression in survey data. By introducing the basis function, the estimator of the regression parameters is the weighted least square estimator with the tuning parameter λ, and weight matrix $\mathbf{W}$, where $\lambda = 0$ means no effect. A similar procedure in this dissertation can be applied to choose the tuning parameter λ to complete the test.

Appendix A

Probabilities Generated by Alternatives (4.3), (4.4) and (4.5)

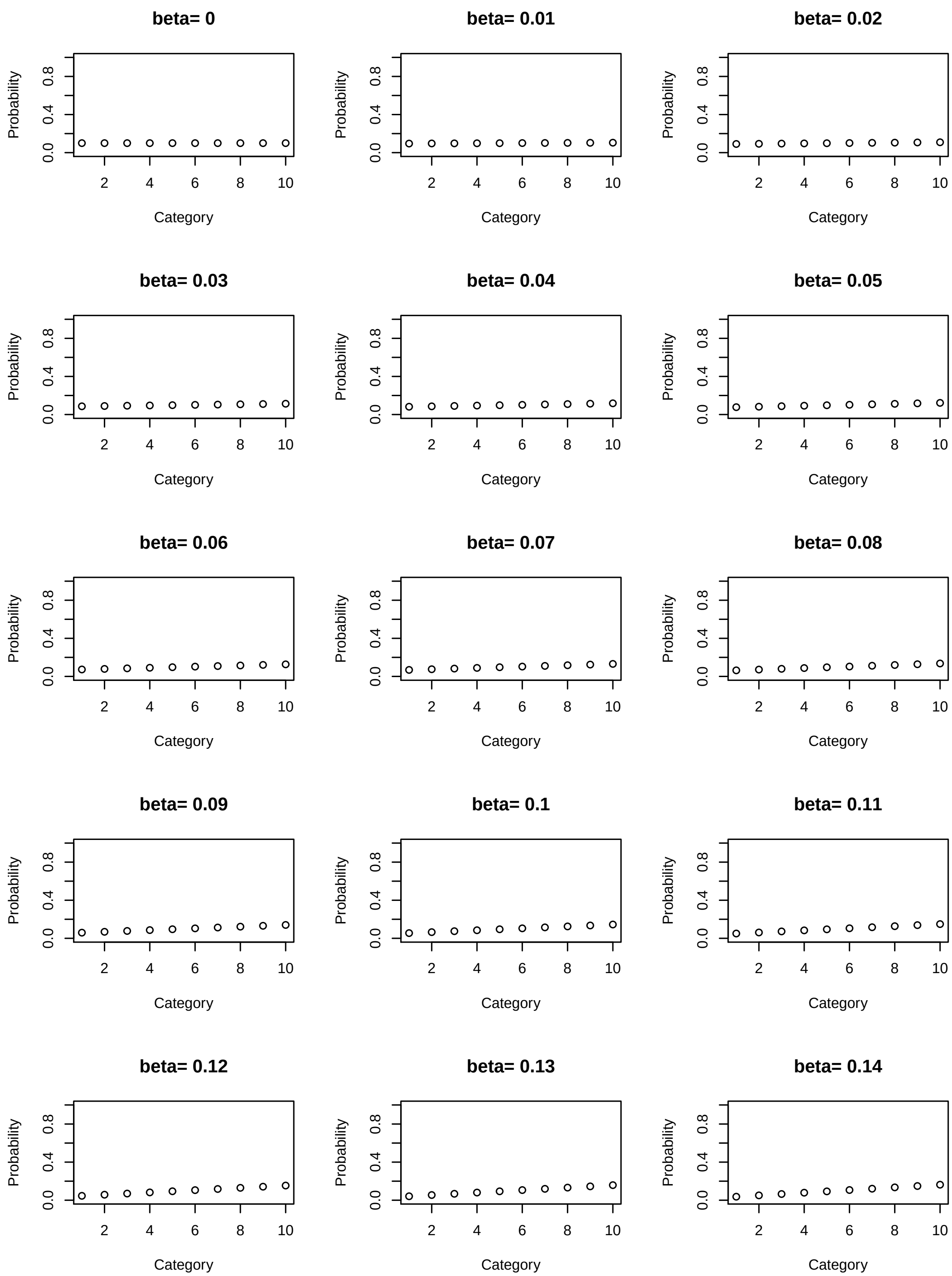

Figure A.1: Probabilities in simulation studies generated by alternative (4.3) for β range from 0 to 0.14 with step 0.01. Probabilities vary greater when *beta* increases.

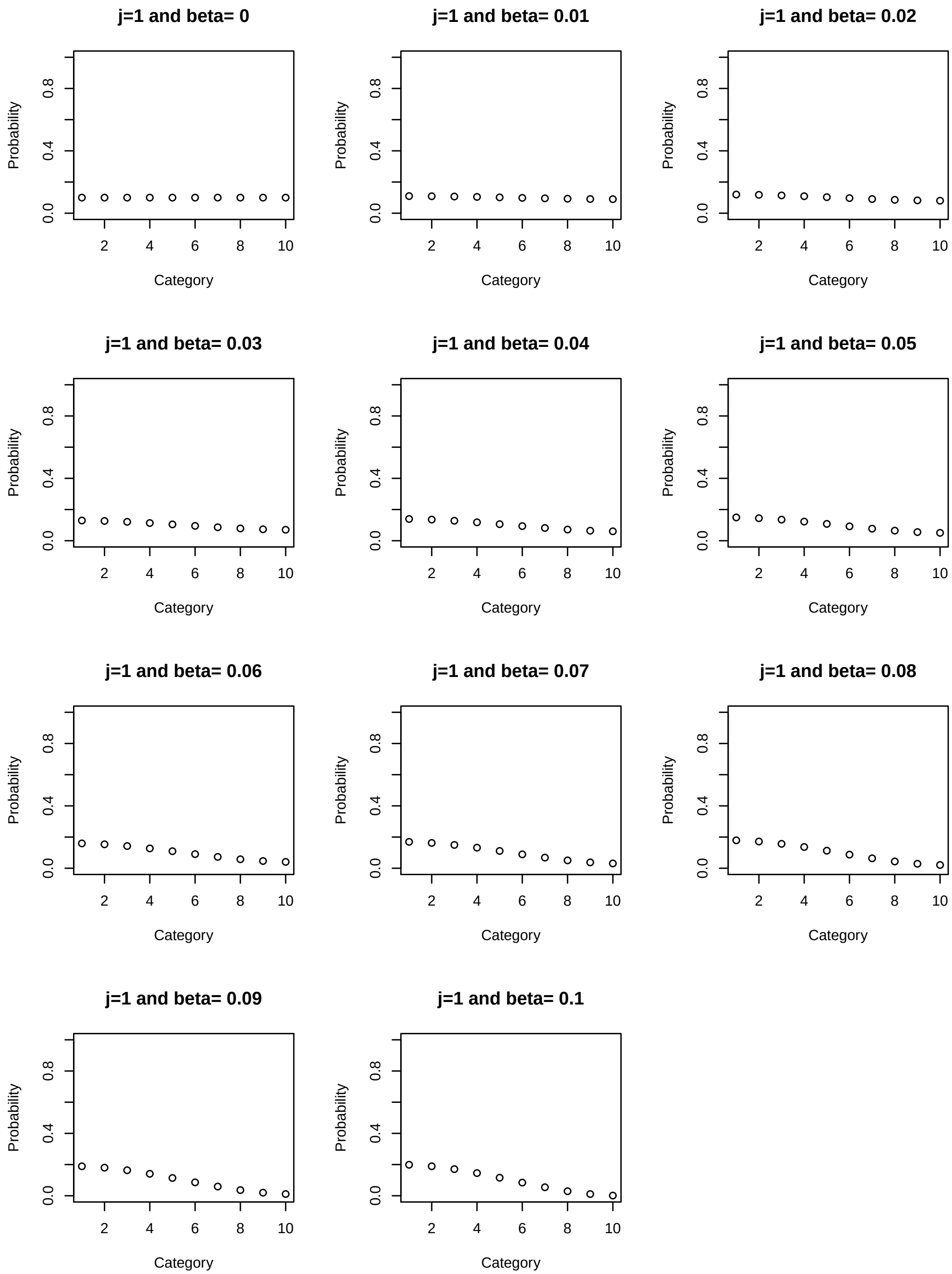

Figure A.2: Probabilities in simulation studies generated by alternative (4.4) for β range from 0 to 0.1 with step 0.01 when $j = 1$. Probabilities vary greater when *beta* increases.

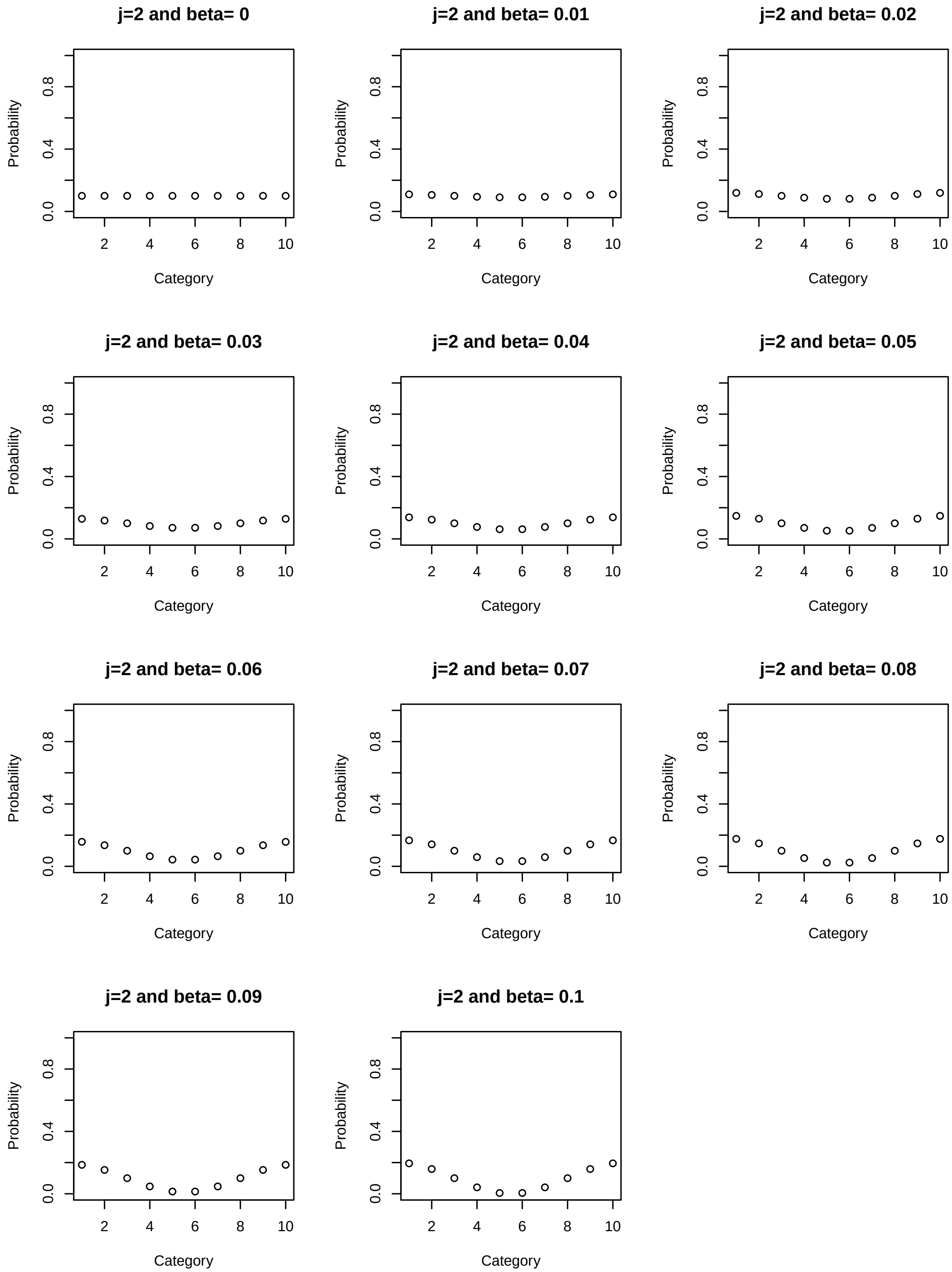

Figure A.3: Probabilities in simulation studies generated by alternative (4.4) for β range from 0 to 0.1 with step 0.01 when $j = 2$. Probabilities vary greater when *beta* increases.

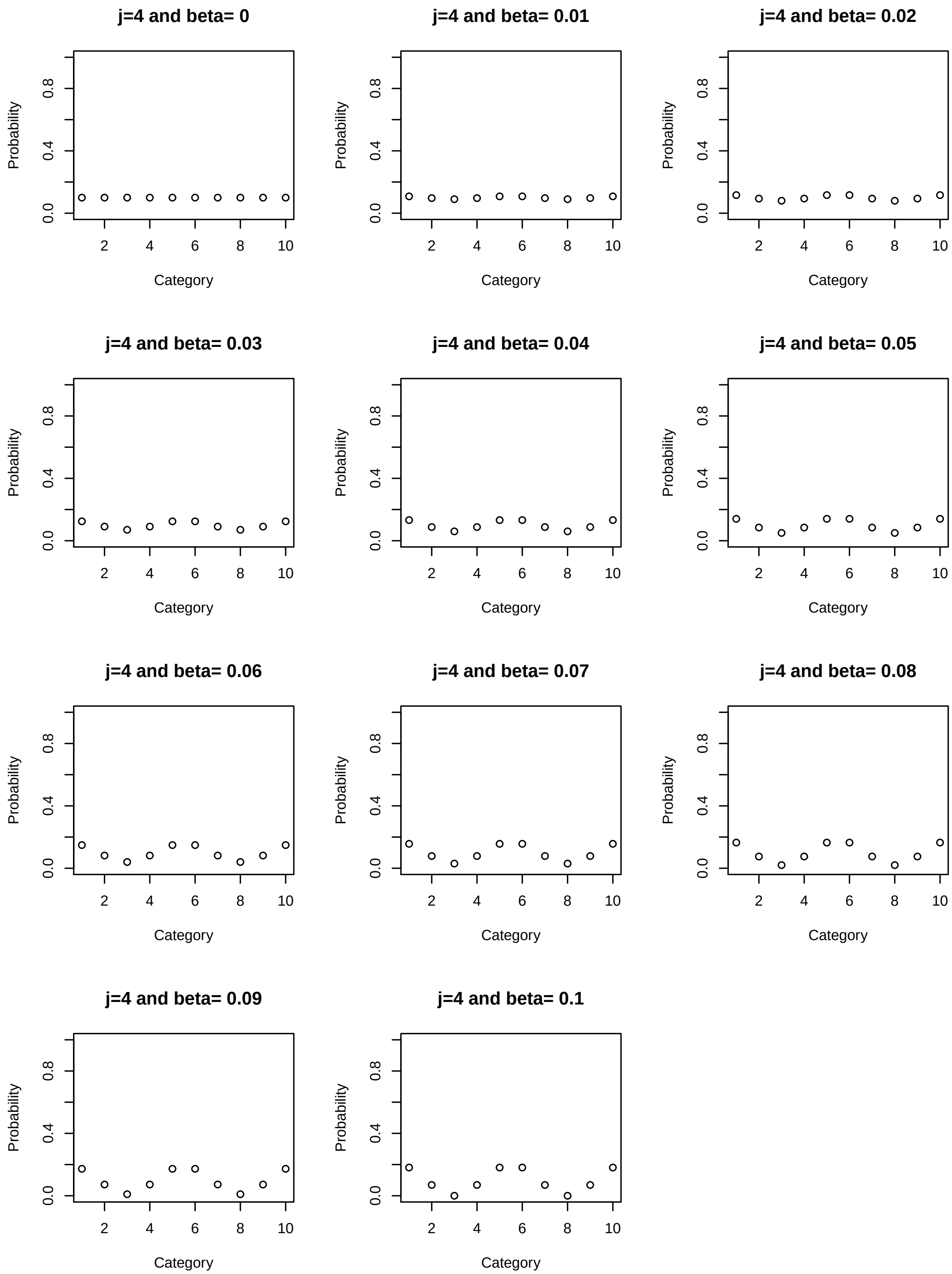

Figure A.4: Probabilities in simulation studies generated by alternative (4.4) for β range from 0 to 0.1 with step 0.01 when $j = 4$. Probabilities vary greater when *beta* increases.

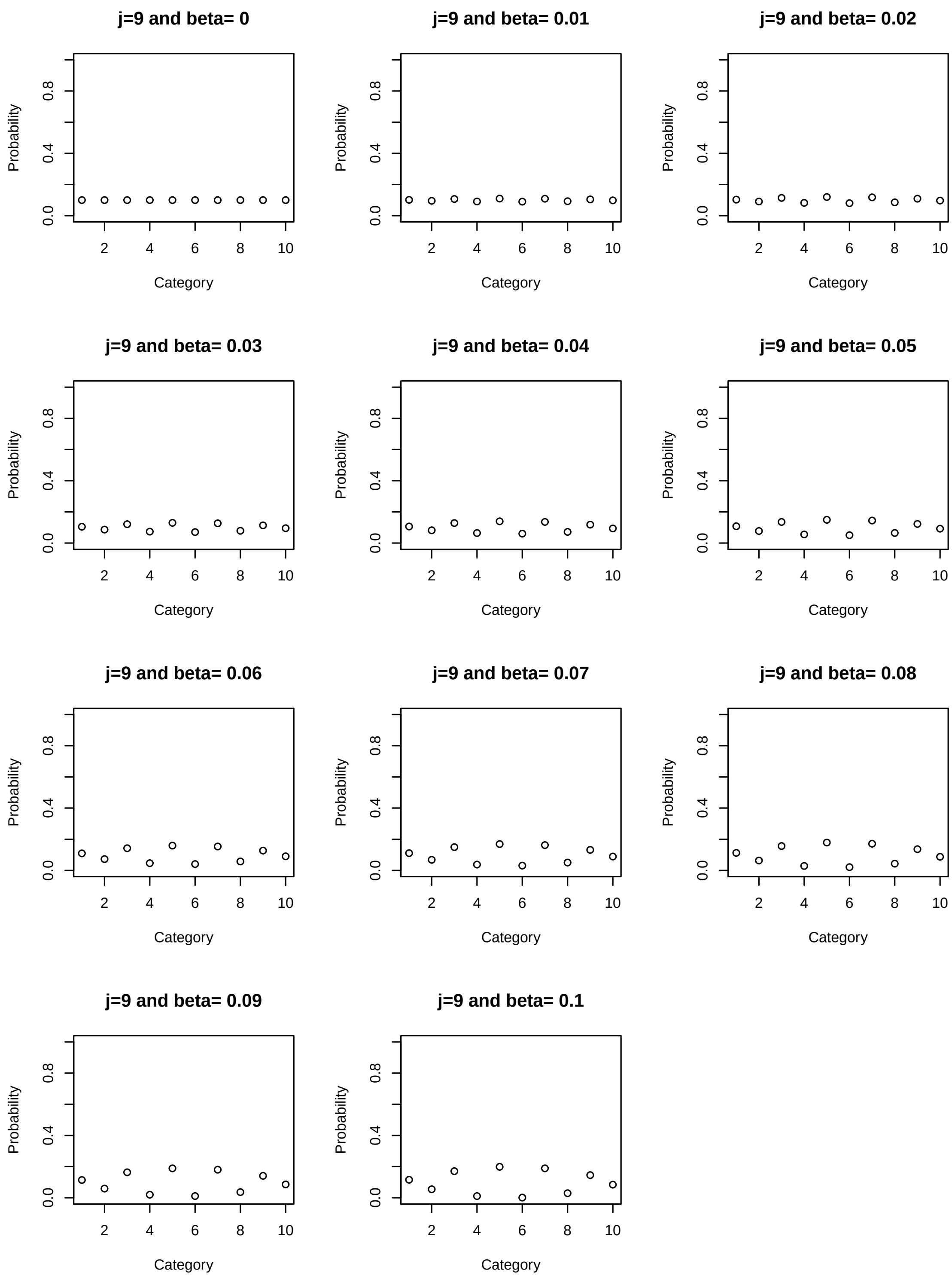

Figure A.5: Probabilities in simulation studies generated by alternative (4.4) for β range from 0 to 0.1 with step 0.01 when $j = 9$. Probabilities vary greater when *beta* increases.

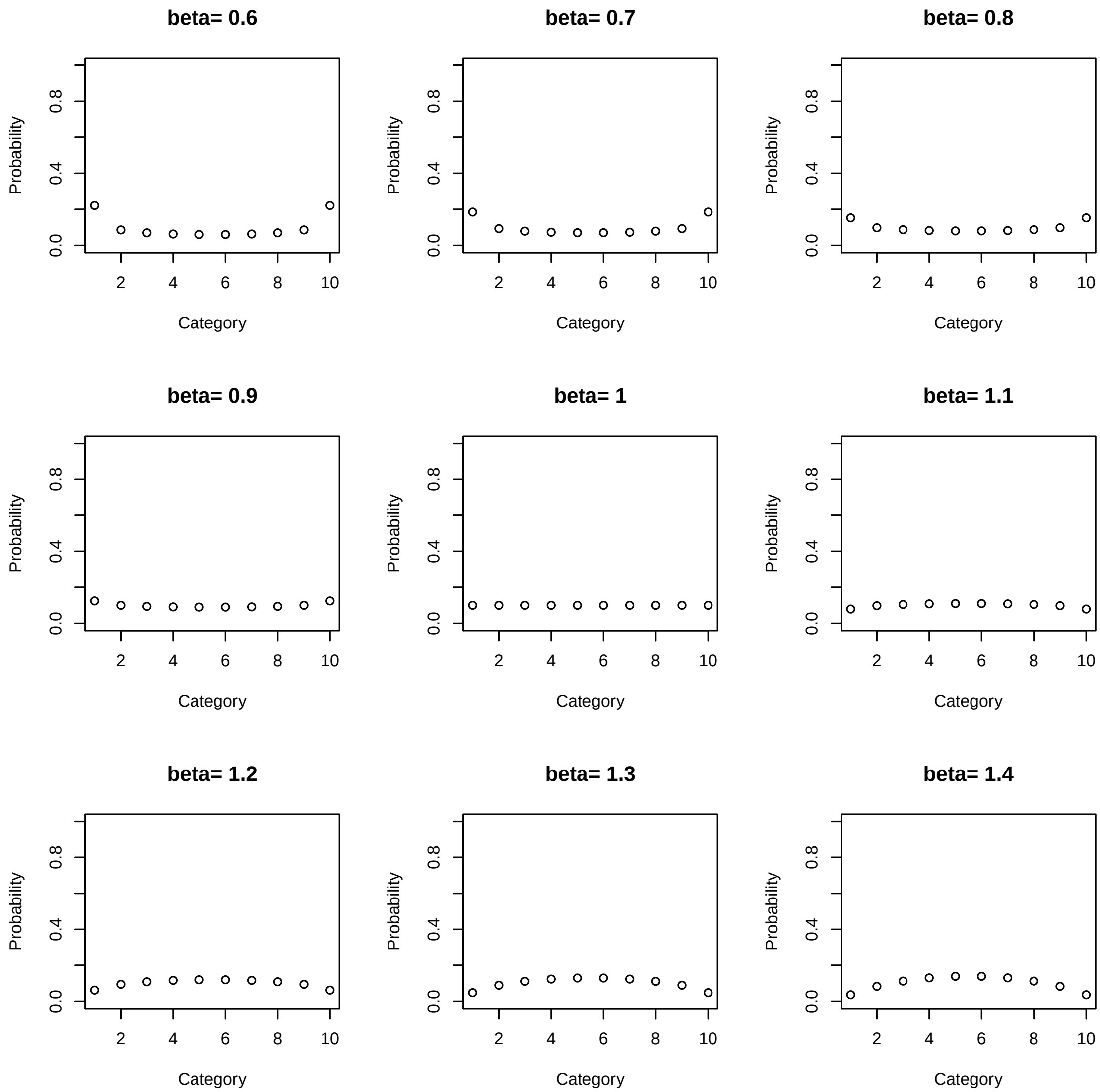

Figure A.6: Probabilities in simulation studies generated by alternative (4.5) for β range from 0.6 to 1.4 with step 0.1. Maximum probabilities are $p(1)$ and $p(10)$ when $\beta < 1$, and maximum probabilities are $p(5)$ and $p(6)$ when $\beta > 1$.

Appendix B

Estimators from Contingency Tables in Stratified Sampling

In this appendix, we discuss the properties of the estimators from contingency tables under stratified sampling. We will derive the expected values, variance, and covariance of the estimators.

Consider a multinomial data with K categories and n observations. Suppose the hypothesis of interest is the same as (2.2)

$$H_0 : p(k) = p_0(k), \text{ for } k = 1, \cdots, K.$$

Define a series of uncorrelated random variables as follows

$$y_j(k) = \begin{cases} 1 & \text{if unit } j \text{ in cell } k \text{ is selected,} \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } j = 1, \cdots, n.$$

Let w_j denote the sampling weight of the jth observation, for $j = 1, \cdots, n$. The estimated proportion for kth category is

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

a weighted average of the observations. It can be proved that $\hat{p}(k)$ is a consistent estimator of the unknown parameter $p(k)$. We can show that

$$\begin{aligned} \mathrm{E}(\hat{p}(k)) &= \mathrm{E}\left(\frac{\sum_{j=1}^{n} w_j y_j(k)}{\sum_{j=1}^{n} w_j}\right) = \frac{\sum_{j=1}^{n} w_j \mathrm{E}[y_j(k)]}{\sum_{j=1}^{n} w_j} \\ &= \frac{1}{\sum_{j=1}^{n} w_j}\left(\sum_{j=1}^{n} w_j p(k)\right) \\ &= p(k) \end{aligned}$$

The variance of $\hat{p}(k)$ can be derived as follows,

$$\begin{aligned} \mathrm{var}(\hat{p}(k)) &= \mathrm{var}\left(\frac{\sum_{j=1}^{n} w_j y_j(k)}{\sum_{j=1}^{n} w_j}\right) = \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2}\mathrm{var}\left(\sum_{j=1}^{n} w_j y_j(k)\right) \\ &= \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2}\left[\sum_{j=1}^{n} w_j^2 \mathrm{var}(y_j(k))\right] \\ &= \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2}\left[\sum_{j=1}^{n} w_j^2 p(k)(1-p(k))\right] \\ &= \frac{\sum_{j=1}^{n} w_j^2}{\left(\sum_{j=1}^{n} w_j\right)^2} p(k)(1-p(k)). \end{aligned} \tag{B.2}$$

The reciprocal of the coefficient of equation (B.2) is Kish's approximate formula for computing effective sample size, which was originally defined in Kish (1965). Denote the Kish's effective sample size as

$$\tilde{n} = \frac{\left(\sum_{j=1}^{n} w_j\right)^2}{\sum_{j=1}^{n} w_j^2}, \tag{B.3}$$

The variance of the weighted estimated proportion is

$$\mathrm{var}(\hat{p}(k)) = \frac{\sum_{j=1}^{n} w_j^2}{\left(\sum_{j=1}^{n} w_j\right)^2} p(k)(1-p(k)) = \frac{p(k)(1-p(k))}{\tilde{n}}. \tag{B.4}$$

In stratified sampling, Kish's effective sample size $\tilde{n}$ is usually different from the observed sample size n, so the variances of the weighted and unweighted estimated proportions are usually different. In SRS, it is trivial that $\tilde{n} = n$ and thus, $\mathrm{var}(\hat{p}(k)) = \mathrm{var}(\tilde{p}(k))$, for $k = 1, \cdots, K$.

Next, we derive the covariance between $\hat{p}(k)$ and $\hat{p}(l)$, for all $k \neq l$ and $k, l = 1, \cdots, K$.

$$\begin{aligned}
& \mathrm{cov}(\hat{p}(k), \hat{p}(l)) \\
= \; & \mathrm{cov}\left(\frac{\sum_{j=1}^{n} w_j y_j(k)}{\sum_{j=1}^{n} w_j}, \frac{\sum_{j=1}^{n} w_j y_j(l)}{\sum_{j=1}^{n} w_j}\right) \\
= \; & \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2} \mathrm{cov}\left(\sum_{j=1}^{n} w_j y_j(k), \sum_{j=1}^{n} w_j y_j(l)\right) \\
= \; & \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2} [\mathrm{cov}(w_1 y_1(k), w_1 y_1(l)) + \cdots + \mathrm{cov}(w_1 y_1(k), w_n y_n(l)) \\
& + \mathrm{cov}(w_2 y_2(k), w_1 y_1(l)) + \cdots + \mathrm{cov}(w_2 y_2(k), w_n y_n(l)) \\
& + \cdots \\
& + \mathrm{cov}(w_n y_n(k), w_1 y_1(l)) + \cdots + w_n \mathrm{cov}(y_n(k), w_n y_n(l))]
\end{aligned}$$

Notice that observations are uncorrelated, so their covariance is 0. Hence,

$$\mathrm{cov}(\hat{p}(k), \hat{p}(l)) = \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2} \left[\sum_{j=1}^{n} w_j^2 \mathrm{cov}(y_j(k), y_j(l))\right].$$

Similarly, since the jth observation can not fall into two different categories simultaneously, $\mathrm{E}\left(y_j(k)y_j(l)\right) = 0$ holds. As a result,

$$\begin{aligned}\mathrm{cov}(y_j(k), y_j(l)) &= \mathrm{E}\left[y_j(k)y_j(l)\right] - \mathrm{E}\left[y_j(k)\right]\mathrm{E}\left[y_j(l)\right] \\ &= -\mathrm{E}\left[y_j(k)\right]\mathrm{E}\left[y_j(l)\right] \\ &= -p(k)p(l).\end{aligned}$$

Consequently, the covariance between $\hat{p}(k)$ and $\hat{p}(l)$ is

$$\begin{aligned}\mathrm{cov}(\hat{p}(k), \hat{p}(l)) &= \frac{1}{\left(\sum_{j=1}^{n} w_j\right)^2}\left[-\sum_{j=1}^{n} w_j^2 p(k)p(l)\right] \\ &= \frac{\sum_{j=1}^{n} w_j^2}{\left(\sum_{j=1}^{n} w_j\right)^2}(-p(k)p(l)) \\ &= -\frac{p(k)p(l)}{\tilde{n}}, \text{ for } k \neq l,\end{aligned}$$

where $\tilde{n}$ is Kish's effective sample size defined in (B.3).

Let

$$\mathbf{V} = \begin{pmatrix} p(1)(1-p(1)) & -p(1)p(2) & \cdots & -p(1)p(K-1) \\ -p(2)p(1) & p(2)(1-p(2)) & \cdots & -p(2)p(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ -p(K-1)p(1) & -p(K-1)p(2) & \cdots & p(K-1)(1-p(K-1)) \end{pmatrix} \tag{B.5}$$

be a $(K-1) \times (K-1)$ matrix. According to the variance and covariance of the estimated proportions of multinomial data in complex surveys, $\mathbf{V}/\tilde{n}$ is the covariance matrix of $\hat{\mathbf{p}}$. If we denote $\hat{\mathbf{V}}$ to be the estimator of $\mathbf{V}$, then $\hat{\mathbf{V}}/\tilde{n}$ is the estimator of the covariance matrix $\hat{\mathbf{p}}$.

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.

Brier, S. S. (1978). Discrete data models with random effects (Technical Paper). Minneapolis, MN: University of Minnesota, School of Statistics.

Christensen, R. (1997). Log-linear models and logistic regression (2nd ed.). New York: Springer.

Christensen, R. (2011). Plane answers to complex questions: The theory of linear models(4th ed.). New York: Springer.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, 46, 440-464.

Dodge, R. W., & Lentzner, H. R. (1980). Crime and seasonality. Bureau of Justice Statistics.

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.

Eubank, R. L., & Spiegelman, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. Journal of the American Statistical Association, 85, 387-392.

Euler, L. (1743). De integratione aequationum differentialium altiorum gradduum. Miscellanea Berolinensia, 7, 193-242.

Fan, J., & Huang, L.-S. (2001). Goodness-of-fit tests for parametric regression models. Journal of the American Statistical Association, 96, 640-652.

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.

Feller, W. (1968). An introduction to probability theory and its applications, vol. 1 (3rd ed.). New York: Wiley.

Fourier, J. B. J. (1822). Theorie analytique de la chaleur. Paris: Chez Firmin Didot, pre et fils.

Fourier, J. B. J., & Freeman, A. (1878). The analytical theory of heat. The University Press.

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.

Inglot, T., Jurlewicz, T., & Ledwina, T. (1990). On neyman-type smooth tests of fit. Statistics, 21, 549-568.

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

Kish, L. (1965). Survey sampling. New York: Wiely.

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

Ledwina, T. (1994). Data-driven version of neyman's smooth test of fit. American Statistical Association, 89, 1000-1005.

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.

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

Lu, Y., & Lohr, S. L. (2010). Gross flow estimation in dual frame surveys. Survey

Methodology, 36, 13-22.

Nair, V. (1987). Chi-squared-type tests for ordered alternatives in contingency tables. Journal of the American Statistical Association, 82, 283-291.

Nair, V. (1988). Testing in industrial experiments with ordered categorical data. Technometrics, 28, 283-311.

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`

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5, 50, 157-175.

Rahman, M. (2011). Applications of fourier transforms to generalized functions. WIT Press.

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.

Rao, J. N. K., & Thomas, D. R. (1988). The analysis of cross-classified categorical data from complex sample surveys. Sociological Methodology, 213-269.

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.

Shibata, R. (1976). Selection of the order of an autoregressive model by akaike's information criterion. Biometrika, 63, 117-126.

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

Thomas, D. R., & Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. Journal of the American Statistical Association, 82, 630-636.

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.

U.S. Census Bureau. (2002). Current population survey: Design and methodology (Technical Paper No. 63RV). Washington, DC: U. S. Census Bureau. Retrieved from `www.census.gov/prod/2002pubs/tp63rv`

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.