Empirical Analysis on Rao-Scott First Order Adjustment for Two Population Homogeneity test Based on Stratified Three-Stage Cluster Sampling with PPS

Sunyeong Heo†

Abstract

National-wide and/or large scale sample surveys generally use complex sample design. Traditional Pearson chi-square test is not appropriate for the categorical complex sample data. Rao-Scott suggested an adjustment method for Pearson chi-square test, which uses the average of eigenvalues of design matrix of cell probabilities. This study is to compare the efficiency of Rao-Scott first order adjusted test to Wald test for homogeneity between two populations using 2009 Gyeongnam regional education offices's customer satisfaction survey (2009 GREOCSS) data. The 2009 GREOCSS data were collected based on stratified three-stage cluster sampling with probability proportional to size. The empirical results show that the Rao-Scott adjusted test statistic using only the variances of cell probabilities is very close to the Wald test statistic, which uses the covariance matrix of cell probabilities, under the 2009 GREOCSS data based. However it is necessary to be cautious to use the Rao-Scott first order adjusted test statistic in the place of Wald test because its efficiency is decreasing as the relative variance of eigenvalues of the design matrix of cell probabilities is increasing, specially more when the number of degrees of freedom is small.

Key words: J Categorical Data, Complex Sample Design, Design Effect, Rao-Scott First Order Adjustment, Wald Test

1. Introduction

National-wide and/or large scale surveys generally use several sampling methods together to select a representative sample, such as stratification, clustering, unequal probability sampling, multi-stage or multi-phase design, multi-frame sampling, and so on. This type of sampling is called complex sampling or complex sample design. Complex sample data do not satisfy the assumption of being independent and identically distributed (iid) which is generally required in traditional statistical theory.

Traditional Pearson chi-square tests are most often used for the tests of independence, homogeneity, and goodness-of-fit of categorical data. The Pearson chi-square tests require the conditions that data are iid with multi-normal distribution and the expected frequencies are so large that Pearson chi-square test statistics have asymptotically chi-square distribution.

Therefore, when a sample data does not satisfy the iid condition and one uses Pearson chi-square test, the results of analyses can be severely distorted. Heo and Chung[1] showed the importance of analysis reflecting the sample design through empirical study comparing traditional Pearson chi-square test and Wald test for two sample homogeneity test based on complex sample design.

In the case of using the secondary data or some statistical software for analysis of categorical data, it is often only to be usable or known the estimates of cell probabilities and their variances but not usable or unknown the their covariance matrix. In the occasion, Wald test, an unbiased test, is not applicable method. Holt et al.[2], Rao and Scott[3-5] and Tomas and Rao[6] have suggested a adjustment method for Pearson test when one only knows the variances of cell probabilities but does not know their covariance matrix.

This study compares the efficiency of Rao-Scott first order adjusted test to Wald test for homogeneity between two populations using 2009 Gyeongnam regional education offices's customer satisfaction survey (2009 GREOCSS) data. The 2009 GREOCSS data were
collected based on stratified three-stage cluster sampling with probability proportional to size (pps) at the selection of primary sample units. In Chapter 2, it is reviewed the Rao-Scott first order adjusted test. In Chapter 3, empirical analysis is given for two population homogeneity using the 2009 GREOCSS data for both test: Rao-Scott first order adjusted test and Wald test. In Chapter 4, conclusions from numerical results are given.

2. Rao-Scott First Order Adjustment

Assume that two independent samples of size $n_1$ and $n_2$ are selected according to given sampling designs, and there are $K$ categories $C_1, C_2, \ldots, C_K$ with probability $p_{ij}$ of which an element belongs to $C_i$th category in $C_j$th population ($i = 1, 2; \ j = 1, 2, \ldots, K$). The null hypothesis for test of homogeneity is

$$H_0: p_1 = p_2 \ (= p)$$

where $p_i = (p_{i1}, \ldots, p_{i, K-1})^T$ with $\sum_{j=1}^{K-1} p_{ij} = 1$, and $p = (p_1, \ldots, p_{K-1})^T$ with $\sum_{i=1}^{K-1} p_i = 1$.

Assume also that for an estimator $\hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{i, K-1})^T$ of $p_i$ based on a given sample design, $n_i^{-1/2}(\hat{p}_i - p_i)$ follows asymptotically normal distribution $N_{K-1}(0, V_i)$. In addition assume that $v_{ij} = \text{Var}(p_{ij})$ is the $j$th diagonal element of the covariance matrix $V_i$ of $p_i$ and its consistent estimator $\hat{v}_{ij} = \widehat{\text{Var}}(\hat{p}_{ij})$ is available. Under these assumptions, the test statistic applying $\hat{p}_i$ to the traditional Pearson chi-square test statistic for $H_0$ is

$$Q^2 \approx \sum_{j=1}^{K-1} \hat{v}_{ij}(\hat{p}_{ij} - p_{ij})^2 \bigg/ \hat{P}$$

$$= (\hat{p}_1 - \hat{p}_2)^T \left( \frac{\hat{P}}{n_1} + \frac{\hat{P}}{n_2} \right)^{-1} (\hat{p}_1 - \hat{p}_2)$$

where $\hat{p}_i = (n_1 \hat{p}_{i1} + n_2 \hat{p}_{i2}) / n$ and $\hat{P} = \text{diag}(\hat{p}) - \hat{p} \hat{p}^T$ with $\hat{p} = (\hat{p}_{i1}, \ldots, \hat{p}_{K-1})^T$. The test statistics $Q^2$'s asymptotic distribution is

$$Q^2_i \approx \sum_{j=1}^{K-1} \delta_j Z^2_j$$

where $Z_j \sim N(0, 1)$. The $\delta_j$ ($j = 1, 2, \ldots, K-1$) are eigenvalues of $\tilde{n} V^{1/2} P^{-1} V^{1/2}$, $V = \text{Var}(\hat{p}_1 + \hat{p}_2)$, $P = \text{diag}(\hat{p}) - \hat{p} \hat{p}^T$, and $\tilde{n} = (n_1 + n_2) / n$.

The mean and variance of $\sum_{i=1}^{K-1} \delta_i Z^2_i$ are

$$E\left[ \sum_{i=1}^{K-1} \delta_i Z^2_i \right] = (K-1) \bar{\delta}$$

and

$$\text{Var}\left[ \sum_{i=1}^{K-1} \delta_i Z^2_i \right] = 2 \sum_{i=1}^{K-1} (\delta_i - \bar{\delta})^2 + 2 (K-1) \bar{\delta}^2.$$

Therefore, the power of $Q^2$ is affected by the size and dispersion of $\delta$s.

Based on the fact that $\bar{\delta}$ can be estimated by only using the $j$th diagonal elements of $\hat{V}_i$, $v_{ij} = \hat{\text{Var}}(p_{ij})$, Holt et al. [2], Rao and Scott [3-5], and Tomas and Rao [6] have suggested adjustments of Pearson test statistic $Q^2$. The first order adjusted test statistic is

$$Q^2 = Q^2 / \bar{\delta}.$$
$Q^2_{W}$ can be larger than the nominal level of type I error. 

The size of increment of the variance depends on the size of relative variance of $\hat{\delta}_j, s^2_j/\hat{\delta}^2$, and the amount of inflation of type I error also depends on it. 

Wald test statistic for test of homogeneity of two populations is

$$Q^2_W = (\hat{p}_1 - \hat{p}_2)^T \left( \frac{\hat{V}_1}{n_1} + \frac{\hat{V}_2}{n_2} \right)^{-1} (\hat{p}_1 - \hat{p}_2).$$

The asymptotic distribution of $Q^2_W$ is chi-square distribution with $K-1$ degrees of freedom under $H_0$.

3. Empirical Analysis

Empirical analysis was conducted by applying to the 2009 Gyeongnam regional education office's customer satisfaction survey (2009 GREOCSS) data. The 2009 GREOCSS was conducted to measure the level of satisfaction of students, parents and teachers about educational service offered by 20 Gyengnam regional education offices.

The 2009 GREOCSS data were collected by stratified three-stage cluster sampling with probability proportional to size. Gyeongnam was first stratified into 20 regions (10 cities and 10 counties) and the types of schools. For each stratum, sample schools were selected by probability proportional to the number of classes. On the second stage of sampling, one class per grade was randomly selected within each sample school. On the final stage, sample students within each sample class were randomly surveyed. Parents of sample students were surveyed, and about 10 teachers per sampled school were surveyed. One can refer to Heo and Chang[7] and Chung et al.[8] for details of the sample design of the 2009 GREOCSS. In the 2009 GREOCSS, the level of satisfaction of students and parents were measured for 6 interested domains and the teachers's satisfaction level for 5 domains. Each item in a domain was measured by 5-point scale. In this study, one item selected per each interested domain for analysis.

For two population homogeneity test, Gyeongnam was divided into two subpopulation; one is consisted of 10 cities and the other is 10 counties. The first subpopulation is called city and the second is called county from the subsequence. The test was conducted for homogeneity between city and county.

Fig. 1–Fig. 3 show that three test statistics for homogeneity between city and county for students,
Empirical Analysis on Rao-Scott First Order Adjustment for Two Population... 211

parents and teachers. Wald is the value of Wald test statistic \( Q_w \), Rao-Scott first order adjusted \( Q_s^1 \), and Pearson \( Q_p^2 \).

Fig. 1–Fig. 3 show that Rao-Scott first order adjusted test statistics give values very close to Wald test statistics, but the values of Pearson test statistic are much larger than both statistics. It is more clear when they are compared with numerical results in Table 1–Table 3.

Table 1–Table 3 give the numerical results of Wald test and Rao-Shao first order adjusted test, and summary of eigenvalues of design effect matrix. As reviewed in chapter 2, the variance of asymptotic distribution of Rao-Scott first order adjusted test statistic \( Q_s^1 \) is \( 1 + s_p^2/\hat{\delta}^2 \) times greater than the variance of chi-square distribution with \( K-1 \) degrees of freedom, and hence it has longer tail than \( \chi^2_{K-1} \). From Table 1. and Table 2. which have the same number of categories, the ratio of p-value of \( Q_w^2 \) to p-value of \( Q_s^2 \) is the largest when \( 1 + s_p^2/\hat{\delta}^2 \) is 1.799, the largest value. And the ratio is the smallest when \( 1 + s_p^2/\hat{\delta}^2 \) is 1.295, the second smallest value, but at which the ratio of test statistic \( Q_w^2 \) to \( Q_s^2 \) is the smallest. From Table 1–Table 3 the ratio of two

Table 1. Values of Wald and Rao-Scott adjusted test statistics and summary of eigenvalues of design matrix for two population homogeneity test between city and county for student

<table>
<thead>
<tr>
<th>Domain No. of Cat. ((K))</th>
<th>Test statistics</th>
<th>p-value</th>
<th>Test statistics</th>
<th>p-value</th>
<th>Design effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( Q_w^2 )</td>
<td>( Q_s^1 )</td>
<td>( \hat{\delta} )</td>
<td>( s_p )</td>
<td>( 1 + s_p^2/\hat{\delta}^2 )</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>11.01</td>
<td>15.35</td>
<td>0.0040</td>
<td>2.894</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>8.08</td>
<td>12.54</td>
<td>0.0040</td>
<td>1.770</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>7.88</td>
<td>11.86</td>
<td>0.0184</td>
<td>1.547</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>10.84</td>
<td>17.08</td>
<td>0.0019</td>
<td>2.234</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5.50</td>
<td>5.50</td>
<td>0.2397</td>
<td>1.847</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>13.78</td>
<td>21.24</td>
<td>0.0003</td>
<td>2.234</td>
</tr>
</tbody>
</table>

Table 2. Values of Wald and Rao-Scott adjusted test statistics and summary of eigenvalues of design matrix for two population homogeneity test between city and county for parents

<table>
<thead>
<tr>
<th>Domain No. of Cat. ((K))</th>
<th>Test statistics</th>
<th>p-value</th>
<th>Test statistics</th>
<th>p-value</th>
<th>Design effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( Q_w^2 )</td>
<td>( Q_s^1 )</td>
<td>( \hat{\delta} )</td>
<td>( s_p )</td>
<td>( 1 + s_p^2/\hat{\delta}^2 )</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>7.17</td>
<td>9.14</td>
<td>0.0577</td>
<td>1.914</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5.69</td>
<td>9.99</td>
<td>0.0406</td>
<td>2.255</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>20.95</td>
<td>33.46</td>
<td>0.0000</td>
<td>1.875</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>17.89</td>
<td>24.29</td>
<td>0.0001</td>
<td>1.903</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5.50</td>
<td>11.47</td>
<td>0.0218</td>
<td>1.298</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>12.07</td>
<td>21.68</td>
<td>0.0002</td>
<td>2.022</td>
</tr>
</tbody>
</table>

Table 3. Values of Wald and Rao-Scott adjusted test statistics and summary of eigenvalues of design matrix for two population homogeneity test between city and county for teachers

<table>
<thead>
<tr>
<th>Domain No. of Cat. ((K))</th>
<th>Test statistics</th>
<th>p-value</th>
<th>Test statistics</th>
<th>p-value</th>
<th>Design effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( Q_w^2 )</td>
<td>( Q_s^1 )</td>
<td>( \hat{\delta} )</td>
<td>( s_p )</td>
<td>( 1 + s_p^2/\hat{\delta}^2 )</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>8.92</td>
<td>16.23</td>
<td>0.0010</td>
<td>2.329</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>10.74</td>
<td>20.28</td>
<td>0.0001</td>
<td>2.351</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8.81</td>
<td>13.89</td>
<td>0.0077</td>
<td>2.858</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>9.32</td>
<td>13.60</td>
<td>0.0035</td>
<td>2.776</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>7.44</td>
<td>14.29</td>
<td>0.0064</td>
<td>2.957</td>
</tr>
</tbody>
</table>

p-values is the largest for $K = 4$ and $1 + \frac{s^2}{\hat{\sigma}^2} = 1.724$. From the Tables, it is shown that the efficiency of Rao-Scott first adjusted test to Wald test depends on both the number of categories and the relative variance of eigenvalues of design matrix of cell probabilities. In general, the p-value of Rao-scott first order adjusted test becomes smaller as the number of categories become smaller and the relative variances become larger.

On the other hand, it needs to be careful before making decision based on the p-value of $\chi^2$ which is near to the nominal level of significance $\alpha$. For example, the domain 2 and 3 of students (Table 1) has $1 + \frac{s^2}{\hat{\sigma}^2}$ equal to 1.499 and 1.421 respectively, and these values are not that large comparing to others. Their corresponding p-values of $Q_n^\alpha$ are 0.0136 and 0.0184 respectively, and one can make decision of reject $H_0$ under $\alpha = 0.05$. However, the p-values of $Q_n^\alpha$ on these two domains are 0.0887 and 0.0961, and the null hypotheses are not rejected under $\alpha = 0.05$. It is similar at the domain 2 of parents (Table 2).

4. Conclusions

Today, it is in general using complex sample design. The categorical complex sample data generally do not satisfy the iid condition required for the traditional Pearson chi-square tests. If one uses the Pearson chi-square test for complex sample data, the results of analysis can be severely distorted. For test of homogeneity, it is needed to know full covariance matrix of cell probabilities. However, when one uses the secondary data or some statistical softwares for analysis of categorical data, it is often the case only to be usable or known the estimates of cell probabilities and their variances, but not usable or unknown the their covariance matrix.

Holt et al.[2], Rao and Scott[3-5] and Tomas and Rao[6] have suggested an adjustment method of Pearson test for the situation, which is generally called Rao-Scott first order adjustment. The adjusted test statistic corrects Pearson test statistic using the average of eigenvalues of design matrix of cell probabilities, and the average of eigenvalues can be calculated by only using the variances of cell probabilities. The asymptotic distribution of the adjusted test statistic has the same mean as $\chi^2_{K-1}$, but has the variance $1 + \frac{s^2}{\hat{\sigma}^2}$ times grater than the variance of $\chi^2_{K-1}$.

In this study, the efficiency of Rao-Scott first order adjusted test to Wald test for two population homogeneity is examined based on 2009 Gyeongnam regional education offices's customer satisfaction survey (2009 GREOCSS) data. The 2009 GREOCSS data was collected based on stratified three-stage cluster sampling with probability proportional to size (pps) at the selection of primary sample units. The numerical analysis shows that the p-value of Rao-scott first order adjusted test generally becomes smaller as the number of categories becomes smaller and the relative variance larger. However, it needs to be careful before making final decision based on the p-value of $\chi^2$ even though the relative variance is not big if its p-value is a little smaller than the nominal level of significance $\alpha$.

Acknowledgements

This research is financially supported by Changwon National University in 2013 ~ 2014.

References