Power Analysis of the Rao-Scott First Order Adjustment to the Pearson Test for Homogeneity

Sunyeong Heo, Changwon National University
Department of Statistics, Changwon National University
Changwon, 641-773 Republic of Korea (syheo@sarim.changwon.ac.kr)

Abstract

In a secondary categorical data analysis of complex sample data, situations often arises when the full variance-covariance matrices of cell proportion estimates are not available but only their variances. In the case, researchers often use the Rao-Scott first order adjustments for inferences. In some circumstances, the type I error rate is greater than the nominal level when using the adjusted test. In addition, there is a concern about the loss of power. I have searched for a way to evaluate powers of the Rao-Scott first order adjustments to Pearson-type test statistics for the homogeneity test, and compared them with the powers of the Wald test. To evaluate power, I used Solomon and Stephens's three-moments approximation. The proposed methods are applied to 2003 NHIS (National Health Interview Survey) public use data.

Keywords: Complex sample; Homogeneity test; Rao-Scott first order adjustment; Wald test; Non-central chi-square distribution; 2003 NHIS (National Health Interview Survey.

1. Introduction

For data from a complex sample design, standard Pearson multinomial-based chi-square tests generally do not achieve nominal levels of type I error. See, e.g., Holt, Scott and Ewings (1980), Rao and Scott (1981, 1984, 1987) and Thomas and Rao (1987) for general background on the effect of a complex sample design on the Pearson chi-squared test. Holt, Scott and Ewings (1980) and Rao and Scott (1981) suggested misspecification-effect-based adjustments to the usual Pearson-type chi-square test statistics to give asymptotically valid homogeneity tests based on a complex sample design.

This paper investigates the relative efficiency of Rao-Scott first order adjusted test to the Wald test in terms of power. The proposed methods are applied to the family-level data from the 2003 National Health Interview Survey.

In section 2, I review the Wald test and Rao-Scott first order adjusted test for homogeneity. In addition, I drive powers for both tests. In section 3, I evaluate the powers for both tests; Wald test and Rao-Scott first order adjusted test for Pearson-type test, and examine the degree of loss of power from the latter test.

2. Homogeneity Test

Suppose we have two independent populations and take independent samples of size n_1 and n_2 from these populations; and suppose we are interested in a categorical variable with K mutually exclusive categories $C_1, C_2, \cdots, C_K$ and each observation in the ith sample is classified into one of these categories. Let n_{ik} be the observed cell frequencies for category C_k in the ith sample, and p_{ik} be the proportion of units for category C_k in the ith population ($\sum_{k=1}^{K} p_{ik} = 1$); and let $p_i = (p_{i1}, \cdots, p_{i,K-1})'$ be the vector of proportions in the first $K-1$ categories for the ith population. The hypothesis of homogeneity of the two populations is $H_0 : p_1 = p_2 (= p)$ versus $H_1 : p_1 \neq p_2$, where p is an unknown vector.

2.1 Wald Test

Let $\hat{p}_i = (\hat{p}_{i1}, \cdots, \hat{p}_{i,K-1})'$ be a consistent estimator of p_i based on a complex sample. When $n_i^{1/2}(\hat{p}_i - p_i)$ is asymptotically distributed to $N_{K-1}(0, V_i)$ as n_i increases and a consistent estimator $\hat{V}_i$ of V_i is available ($i = 1, 2$), Wald test statistics

$$X_W^2 = (\hat{p}_1 - \hat{p}_2)'(n_1^{-1}\hat{V}_1 + n_2^{-1}\hat{V}_2)(\hat{p}_1 - \hat{p}_2) \qquad (1)$$

is asymptotically distributed as c_{K-1}^2 under H_0 for sufficiently large n_i (Holt *et al.*, 1980).

The power of Wald test in (1) is

$$1 - b_W = \Pr\{X_W^2 > c_{K-1,a}^2 \mid D\}$$
$$\approx \Pr\{W_{K-1,d}^2 > c_{K-1,a}^2 \mid D\}$$

where $c_{K-1,a}^2$ is the upper $100a\%$ percentile of a central chi-square distribution on $K-1$ degrees of freedom; $W^2_{K-1,d}$ is a non-central chi-square random variable on $K-1$ degrees of freedom with non-centrality parameter $d = D'V_D^{-1}D/2$; and $D = p_1 - p_2$, $V_D = Var(\hat{p}_1 - \hat{p}_2)$. The distribution of $W^2_{K-1,d}$ is equal to the distribution of $\sum_{k=1}^{K-1} Z^2_{k,d}$ where $Z^2_{k,d}$ are independent non-central chi-square random variables $Z^2_{k,d}$ with one degree of freedom and non-centrality parameter $d_{Z,k} = m^2_{Z,k}/2$. The $m_{Z,k}$ is the k th element of $m_Z = V_D^{-1/2}(p_1 - p_2)$. Thus, the power of the Wald test is also expressed by

$$\begin{aligned} 1-b_W &= \Pr\{X^2_W > c^2_{K-1,a} \mid D\} \\ &\approx \Pr\{\textstyle\sum_{k=1}^{K-1} Z^2_{k,d} > c^2_{K-1,a} \mid D\}. \end{aligned} \quad (2)$$

2.2 Rao-Scott First Order Correction to Pearson Test

2.2.1 Rao-Scott first order adjusted test statistics

In the cases of which the estimates of V_i are not available, researchers often use a Pearson-type chi-squared test statistics for testing $H_0 : p_1 = p_2 (= p)$, which is

$$X^2_P = \tilde{n}(\hat{p}_1 - \hat{p}_2)' \hat{P}_0^{-1} (\hat{p}_1 - \hat{p}_2) \quad (3)$$

where $\hat{P}_0 = diag(\hat{p}_0) - \hat{p}_0 \hat{p}_0'$, $\hat{p}_0 = (\hat{p}_{01}, \cdots, \hat{p}_{0,K-1})'$ and $\hat{p}_{0k} = (n_1 \hat{p}_{1k} + n_2 \hat{p}_{2k})/n$; and $\tilde{n} = n_1 n_2 / n$, $n = n_1 + n_2$.

There are often the cases in which the estimated cell variances $\hat{v}_{ikk} = \hat{V}ar(\hat{p}_{ik})$ are available. Rao and Scott (1980) noted that under general design conditions and H_0, the test statistics X^2_P is distributed asymptotically as a weighted sum, $\sum_{k=1}^{K-1} l_k Z^2_k$, of independent chi-square random variable Z^2_k with one degree of freedom. The weights l_k are the singular values of $(n_2 D_1 + n_1 D_2)/n$ where $D_i = P_0^{-1} V_i$ is the misspecification effect matrix for the i th population; and $P_0 = diag(p) - pp'$. They called D_i the design effect matrix and used it to develop a first-order adjustment for X^2_P. With consistency with Skinner *et al.* (1989, p. 28) I use a term of misspecification effect matrix for D_i, instead. Rao and Scott (1980) suggested the simple correction of the test statistic X^2_P such as

$$X^2_C = X^2_P / \bar{\hat{l}} \quad (4)$$

where $\bar{\hat{l}}$ is the average of $\hat{l}_1, \hat{l}_2, \cdots, \hat{l}_{K-1}$, and the estimates $\hat{l}_k$ of l_k are obtained by $V_i = \hat{V}_i$ and $P_0 = \hat{P}_0$. The $\bar{\hat{l}}$ can be obtained by only using cell variances $\hat{v}_{ikk}$ (Rao and Scott, 1980).

Through empirical analyses, Holt *et al.* (1980) showed that X^2_C is a very good approximation to the nominal type I error rate. They also noted that the inflation of the type I error is getting bigger the variability of l_k's or K increase.

2.2.2 Power of Rao-Scott first order correction

Under a complex sample design and for a design based consistent estimator $\hat{p}_i$ of p_i, assume the following conditions.

(C.1) The asymptotic distribution of $\left\{n_1^{1/2}(\hat{p}_1 - p_1), n_2^{1/2}(\hat{p}_2 - p_2)\right\}'$ is $N_{2(K-1)}(0, V_p)$ where $V_p = blockdiag(V_1, V_2)$.

(C.2) There exists a consistent estimator $\hat{V}_i$ of V_i $(i = 1, 2)$.

Corollary 1: Assume conditions (C.1) and (C.2). For any non-zero $D = p_1 - p_2$, the Pearson-type test statistic X^2_P in expression (3) is asymptotically distributed as $\sum_{k=1}^{K-1} l_k Y^2_{k,d}$ where $Y^2_{k,d}$ are independent non-central chi-square random variable on one degree of freedom and non-centrality parameter $d_{Y,k} = m^2_{Y,k}/2$. The $m_{Y,k}$ is the k th element of $m_Y = Q'\{V_D^{-1/2}(p_1 - p_2)\}$ where Q is an orthogonal matrix such as $\tilde{n}\{(V_D^{1/2})' P_0^{-1} (V_D^{1/2})\} = Q \Lambda Q'$; and $\Lambda = diag(l_1, l_2, \cdots, l_{K-1})$ with $l_1 \ge l_2 \ge \cdots \ge l_{K-1}$.

Corollary 2: Under conditions (C.1) and (C.2), $X^2_C = X^2_P / \bar{\hat{l}}$ is distributed asymptotically as $\sum_{k=1}^{K-1} \left(l_k / \bar{l} \right) Y^2_{k,d}$.

Corollary 2 is driven directly from Corollary 1. The $\bar{l}$ is the average of $l_1, l_2, \cdots, l_{K-1}$. From the Corollary 2, the power of X^2_C in expression (4) is obtained by

$$1-b_C=\Pr\{X_C^2>c^2_{K-1,a}\mid D\}$$
$$\approx \Pr\{\sum_{k=1}^{K-1}(l_k/\bar{l})Y^2_{k,d}>c^2_{K-1,a}\mid D\}. \quad (5)$$

3. Power Evaluation

For power evaluation I applied to the public use 2003 National Health Interview Survey (NHIS) data.

3.1 2003 NHIS Data

The National Health Interview Survey (NHIS) is a national level face-to-face survey carried out in all 50 states to get the general health information of the non-institutionalized population in the United States. The sample is selected according to multistage sample design with unequal selection probability.

I selected four variables from the 2003 NHIS family-level data file. The table 1 shows variable names with meanings, responded family sizes and the number of categories. For the variables with small cell proportions, I collapsed some categories. The numbers inside of parentheses are the number of categories before collapsing. In addition, I selected two regions, Northeast and Midwest, and considered them as independent populations. The sample family sizes of Northeast (NE) and Midwest (MW) are 6,546 and 8,144, respectively. For a given variable, the hypothesis in which we are interested is $H_0: P(C_k\mid NE)-P(C_k\mid MW)=0$ for all k.

For simulation purposes, I reordered the categories by the descending order of the estimated cell proportions.

3.2 Evaluation Power

For each variable, I estimated the survey based variance estimate $\hat{V}ar(\hat{p}_i)$ using linearization method and considered them as the true variance. In addition, for a given value l $(0<l\le\sqrt{K-1})$ I generated 100 vectors of $D=p_{NE}-p_{MW}$ from $Uniform(0,1)$ such as $\|p_{NE}-p_{MW}\|=l$ and $|p_{NE}-p_{MW}|\le 1$ where $\|a\|$ is the norm of vector a.

For each vector $D=p_{NE}-p_{MW}$, I evaluated asymptotic powers from expression (2) for Wald test and from (5) for Rao-Scott adjusted test. In this step, we need to know the distribution of $\sum_{k=1}^{K-1}Z^2_{k,d}$ in (2)

Table 1. Four variables names, numbers of categories, responded family sizes and meanings from the 2003 NHIS family-level file.

Variable	No. of Cat.	Sample Size	Explanaton
Lng_intv	4	$n_{NE}=6542$ $n_{MW}=8140$	Interview Language
Fspedct	4(6)	$n_{NE}=2141$ $n_{MW}=2740$	No. children in fam rec Spec Ed/EIS
Fwrklwct	6(8)	$n_{NE}=4699$ $n_{MW}=6126$	No. of fam members work full time last week
fhicost	8	$n_{NE}=6546$ $n_{MW}=8144$	Cost of fam med/dental care past 12 mo.

* Data Source: National Center for Health Statistics (2003)

Table 2. Means, standard deviations and coefficient of variation of $\hat{l}_k$.

Variable	$\bar{\hat{l}}$	sd_l	$sd_l/\bar{\hat{l}}$
Lng_intv	1.255	0.561	0.447
fspedct	1.084	0.115	0.110
Fwrklwct	1.375	0.556	0.405
fhicost	1.871	1.210	0.647

* Data Source: National Center for Health Statistics (2003)

and $\sum_{k=1}^{K-1}\left(l_k/\bar{l}\right)Y^2_{k,d}$ in (5), and I used Solomon-Stephens's approximation method (1977). As the result, for a given l and test statistics I had 100 powers and I took their average.

Table 2 shows averages, standard deviations and coefficients of variation of $\hat{l}_k$ for four variables. Figure 1 to Figure 4 show the power curves of four variables for Wald test and Rao-Scott adjusted test In the figures, x-axis is norms of $D=p_{NE}-p_{MW}$ and y-axis is the average of 100 powers. From the graphs along with Table 2, we can see that the larger coefficient of variations of $\hat{l}_k$ are the more we lose power.

Acknowledgements

The view expressed in this paper are those of the author and do not necessarily represent the policies of the National Center for Health Statistics.

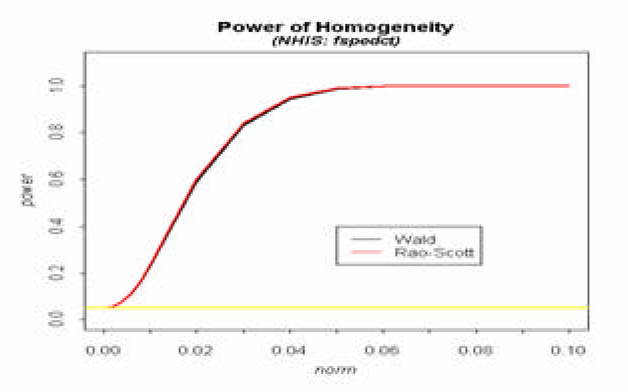

Figure 1. Power of Wald test and Rao-Scott adjusted test for variable fspedct ($a = 0.05$).

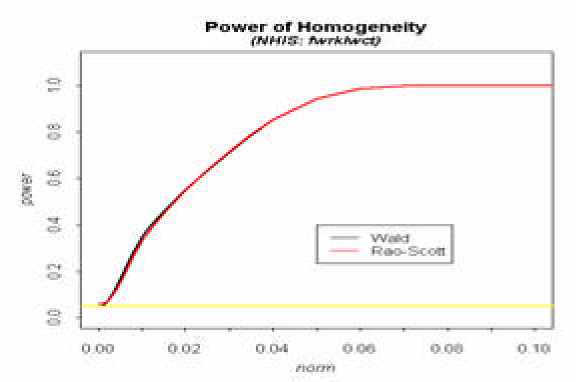

Figure 2. Power of Wald test and Rao-Scott adjusted test for variable fwrktwct ($a = 0.05$).

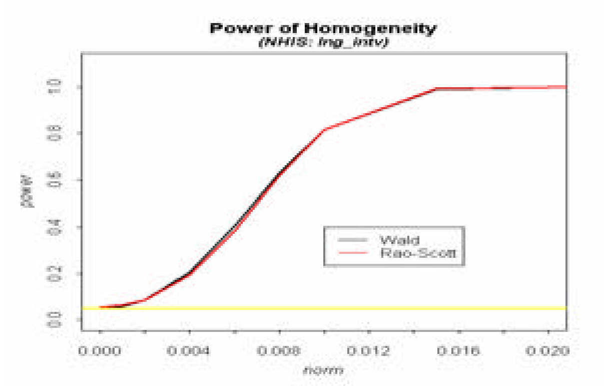

Figure 3. Power of Wald test and Rao-Scott adjusted test for variable lng_intv ($a = 0.05$).

* Data Source: National Center for Health Statistics (2003)

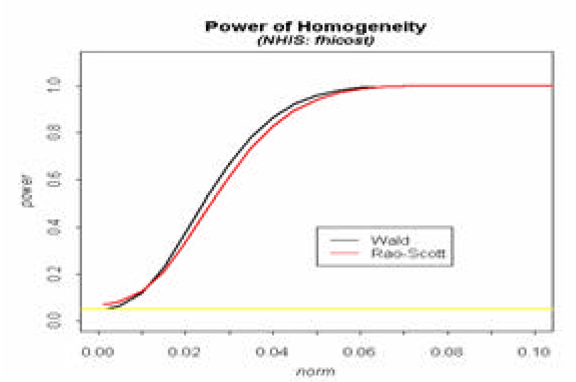

Figure 4. Power of Wald test and Rao-Scott adjusted test for variable fhicost($a = 0.05$).

References

Holt, D., Scott, A. J., and Ewings, P. D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A*, **143**, 302-320.

National Center for Health Statistics (2004). Data File Documentation, National Health Interview Survey, 2003 (machine readable data file and documentation). National Center for Health Statistics, Center for Disease Control and Prevention, Hyattsville, Maryland.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit the independence in two-way tables. *Journal of the American Statistical Association* **76**, 221-230.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared test for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics* **12**, 46-60.

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

Scott, A. J. and Rao, J. N. K. (1981). Chi-squared tests for contingency tables with proportions estimated from survey data. In D. Krewski, R. Platek, and J. N. K. Rao (eds.), *Current Topics in Survey Sampling, pp. 247-265*. New York: Academic Press.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons Ltd.

Solomon, H and Stephens, M.A. (1977). Distribution of a Sum of Weighted Chi-Square Variables. *Journal of the American Statistical Association*, Vol. 72, 881-885.

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