AN APPROXIMATE RAO-SCOTT MODIFICATION FACTOR IN TWO-WAY TABLES WITH ONLY KNOWN MARGINAL *DEFF*s

Hee-Choon Shin, National Opinion Research Center
1155 E. 60th St., Chicago, IL 60637

Key words: Chi-Squared Tests, Complex Survey Data

1. INTRODUCTION

Consider the problem of testing for independence in a two-way table with I rows and J columns. The hypothesis of interest is

$$H_0: p_{ij} = p_{i+}p_{+j}, \text{ for } i = 1, 2, ..., I; j = 1, 2, ..., J,$$

where p_{ij} = the population proportion in (i, j)th cell, $p_{i+} = \Sigma_{j=1}^{J} p_{ij}$, and $p_{+j} = \Sigma_{i=1}^{I} p_{ij}$. The conventional Pearson chi-squared statistic for testing H_0 is

$$X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(\hat{p}_{ij}-\hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}, \quad (1)$$

where $\hat{p}_{ij}$ is an unbiased estimator of p_{ij}. However, complex sample designs usually invalidate the direct application of the Pearson chi-squared statistics to test independence in a cross-classified table.

Let d_i and d_j be the estimated deffs of $\hat{p}_{i+}$ and $\hat{p}_{+j}$, respectively, and d_{ij} be the estimated deffs of $\hat{p}_{ij}$. A modified statistic, X_m^2, proposed by Rao and Scott (1981, 1984), is given by

$$X_m^2 = X^2/\hat{\delta}_{\cdot},$$

where $\hat{\delta}_{\cdot}$ =

$$\frac{1}{(I-1)(j-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(1-\hat{p}_{i+}\hat{p}_{+j})d_{ij} - \sum_{i=1}^{I}(1-\hat{p}_{i+})d_i - \sum_{j=1}^{J}(1-\hat{p}_{+j})d_j. \quad (2)$$

2. STATEMENT OF THE PROBLEM

As we see in (2), the Rao-Scott chi-squared statistic can be obtained with the complete information on the estimated cell proportions $\hat{p}_{ij}$ and their estimated deffs d_{ij}, and the estimated deffs d_i and d_j of the marginal proportions $\hat{p}_{i+}$ and $\hat{p}_{+j}$ respectively.

Table 1 shows the estimated popupation proportions of drinking levels by age group of 1081 women in 1991, and Table 2 shows corresponding cell and marginal deffs. Our data come from a longitudinal study of drinking behavior among U.S. women living in non-institutional settings (The National Longitudinal Study of Women's Drinking, Sharon C. Wilsnack & Richard W. Wilsnack, Co-principal investigators, funded by the National Institute on Alcohol Abuse and Alcoholism). The conventional Pearson chi-squared statistic, X^2, for testing independence between drinking level and age is 124.692 with 12 degrees of freedom. The Rao-Scott chi-squared statistic, X_m^2, is 103.051 with the modification factor, $\hat{\delta}_{\cdot}$, of 1.210.

However, the information on estimated deffs may not be complete in a published report. Typically, some marginal deffs are reported, if any, to measure the impact of survey design. The exact Rao-Scott chi-squared statistic can not be computed under this situation. This paper will examine some reasonable substitutes to the Rao-Scott chi-squared statistic to test independence of two-way cross-classification tables, and suggests a new approximate Rao-Scott modification factor under a lack of complete information on cell deffs.

3. SUBSTITUTES TO RAO-SCOTT STATISTIC

3.1. Known Cell Deffs

The first immediate substitute to the Rao-Scott modification factor is the mean eigenvalue, $\hat{\lambda}_{\cdot}$, for goodness-of-fit problem:

$$\hat{\lambda}_{\cdot} = \frac{1}{(IJ-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(1-\hat{p}_{i+}\hat{p}_{+j})d_{ij}. \quad (3)$$

The second substitute is the arithmetic mean, $\hat{d}_{\cdot}$, of cell deffs (Fellegi, 1980):

$$\hat{d}_{\cdot} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}d_{ij}. \quad (4)$$

If $\hat{\lambda}_{\cdot}$ and $\hat{d}_{\cdot}$ are given, $\hat{\delta}_{\cdot}$ is usually available or the calculation of $\hat{\delta}_{\cdot}$ is also possible unless marginal deffs are missing. Accordingly, the first two substitutes are not necessary and $\hat{\delta}_{\cdot}$ should be preferred to $\hat{\lambda}_{\cdot}$ and $\hat{d}_{\cdot}$ as long as the computing is realistically possible.

3.2. Known Marginal Deffs and Unknown Cell Deffs

Now consider a situation where d_i and d_j are known but d_{ij} are not. This situation is quite a common one in practice, particularly in published reports.

The immediate substitute uses the smaller or minimum value, $\hat{d}_{\cdot}^{min}$, of average deffs for rows and column marginals (Holt, Scott, and Ewings, 1980):

$$\hat{d}_{\cdot}^{min} = \min\left(\sum_{i=1}^{I} d_i/I, \sum_{j=1}^{J} d_j/J\right). \quad (5)$$

This has been considered as the best subsitute. We propose a better approximate Rao-Scott modification factor, $\hat{\delta}_{\cdot}^{min}$. Since d_{ij} are not available, each d_{ij} is replaced by the smaller value of d_i and d_j, i.e., $\min(d_i, d_j)$. Using (2), the new approximate Rao-Scott modification factor, $\hat{\delta}_{\cdot}^{min}$, is

$$\hat{\delta}^{min}_{\cdot} = \frac{1}{(I-1)(J-1)} \cdot \left[\sum_{i=1}^{I}\sum_{j=1}^{J} ((1-\hat{p}_{i+}\hat{p}_{+j}) \min[d_i, d_j]) - \sum_{i=1}^{I}(1-\hat{p}_{i+})d_i - \sum_{j=1}^{J}(1-\hat{p}_{+j})d_j \right]. \quad (6)$$

4. EMPIRICAL RESULTS

For a two-way table with known marginal deffs but with unknown cell deffs, we have suggested a new approximation, $\hat{\delta}^{min}_{\cdot}$, for the Rao-Scott modification factor. In this section, we examine the performance of each substitute ($\hat{\lambda}_{\cdot}$, $\hat{d}_{\cdot}$, $\hat{d}^{min}$, and $\hat{\delta}^{min}_{\cdot}$), and show $\hat{\delta}^{min}_{\cdot}$ is a better approximation of Rao-Scott modification factor under a lack of complete information on deffs specified above, as compared to other substitutes.

Table 3 shows the sizes of each modification factor and their differences from the Rao-Scott factor. The mean eigenvalue for goodness-of-fit ($\hat{\lambda}_{\cdot}$) is 1.415, which is larger than the Rao-Scott modification factor, $\hat{\delta}_{\cdot}$, (1.210). The mean deff ($\hat{d}_{\cdot}$), or arithmetic mean value of cell deffs, is 1.438. Smaller average marginal deff ($\hat{d}^{min}$) is 1.487. Our proposed approximated Rao-Scott factor ($\hat{\delta}^{min}_{\cdot}$) is 1.188, which is quite close to the true Rao-Scott modification factor, $\hat{\delta}_{\cdot}$, as compared to other substitutes.

REFERENCES

Fellegi, I.P. (1980), "Approximate tests of independence and goodness of fit based on stratified multistage samples." *Journal of the American Statistical Association*, 75, 261-268.

Holt, D., Scott, A.J., and Ewings, P.O. (1980), "Chi-squared tests with survey data." *Journal of the Royal Statistical Society*, Ser. A, 143, 303-320.

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 and independence in two-way tables." *Journal of the American Statistical Association*, 76, 221-230.

—— (1984), "On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data." *The Annals of Statistics*, 12, 46-60.

TABLES

Table 1. Estimated population proportions of women's drinking levels by age group (n=1,081)

Drinking level	Age groups 21-34	35-49	50-64	65-	Total
(1)	.0779	.1086	.0949	.1403	.4217
(2)	.0263	.0175	.0085	.0019	.0542
(3)	.1363	.1314	.0706	.0442	.3825
(4)	.0484	.0349	.0225	.0089	.1147
(5)	.0134	.0080	.0032	.0023	.0269
Total	.3023	.3004	.1997	.976	1

Note; Drinking level: (1) Abstainers, (2) Temporary abstainers, (3) Lighter drinkers, (4) Moderate drinkers, (5) Heavier drinkers.

Table 2. Cell and marginal deffs of women's drinking levels by age group (n=1,081).

Drinking level	Age groups 21-34	35-49	50-64	65-	Total
(1)	.9755	2.1746	1.9773	3.3916	2.2137
(2)	1.2210	1.7189	1.6753	.4994	1.5298
(3)	.9897	2.6144	2.0835	1.5397	1.5261
(4)	.8089	.9375	.9850	2.1393	1.1926
(5)	1.0299	.6968	.7015	.6003	.9704
Total	1.3136	2.5720	2.1833	2.9280	

Note; Drinking level: (1) Abstainers, (2) Temporary abstainers, (3) Lighter drinkers, (4) Moderate drinkers, (5) Heavier drinkers.

Table 3. Sizes of substitutive modification factors.

Modification factor	Size
Mean eigenvalue($\hat{\lambda}_{\cdot}$)	1.415
Mean deff($\hat{d}_{\cdot}$)	1.438
Smaller average marginal deff($\hat{d}^{min}$)	1.487
Approximated Rao-Scott($\hat{\delta}^{min}_{\cdot}$)	1.188
Rao-Scott($\hat{\delta}_{\cdot}$)	1.120