Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 1, pp. 137～144

New Wald Test Compared with Chen and Fienberg's for Testing Independence in Incomplete Contingency Tables

Shin-Soo Kang[1]

Abstract

In $I \times J$ incomplete contingency tables, the test of independence proposed by Chen and Fienberg(1974) uses $I \times J-1$ instead of $(I-1)(J-1)$ degrees of freedom without providing much of an increase in the value of the test statistic. For these reasons, Chen and Fienberg tests are expected to have less power. New Wald test statistic related to the part of Chen and Fienberg test statistic is proposed using delta method. These two tests are compared through Monte Carlo studies.

Keywords : Complete Case Analysis, MLE, Wald Statistic

1. Introduction

In the analysis of contingency tables, it may happen that some observations are fully classified or others are partially classified. These tables are called 'incomplete contingency table'.

Blumenthal (1968) considered two way contingency tables for multinomial samples where the column classification is missing. Chen and Fienberg (1974) used an iterative procedure for computing maximum likelihood estimates and developed Pearson and likelihood ratio tests of independence for two-way tables for which either the row or column classification could be missing for some cases. As in Chen and Fienberg (1974), Hocking and Oxspring (1974) consider three independent multinomial distributions corresponding to the set of fully cross-classified counts and the two sets of partially classified counts, where either the row classification or the column classification is missing.

[1] Professor, Department of Information Statistics, Kwandong University, Kangnung, 210-701, Korea
E-mail: sskang@kd.ac.kr

An alternative approach involves constructing a complete table, in which all cases are completely classified, by imputing information for the missing row or column classification. Multiple imputation, proposed by Rubin (1978), provides a way to take advantage of common tests of independence for completely classified tables.

Little and Rubin (2002) in Section 1.3 discussed three general mechanisms for missing data: missing completely at random(MCAR), missing at random(MAR), and not missing at random(NMAR). Let X_1 and X_2 denote categorical variables for a two-way incomplete contingency table. If the missing probability of X_i does not depend on either the value of the other variable or the value of X_i, then it is MCAR. If the missing probability of X_i depends on the value of the other variable but not on the value of X_i, then it is MAR. If the missing probability of X_i depends on its value, then it is NMAR.

The proposed new Wald tests of independence in Section 4 and the tests derived by Chen and Fienberg(1974) are examined through Monte Carlo studies in Section 5 considering both type I error level and power.

2. Notation and MLE under independence model

Consider an $I \times J$ contingency table where the row factor X_1 has I categories and the column factor X_2 has J categories. Assume simple random sampling with replacement. In a complete table, where the row and column categories are observed for every case in the sample, the counts have a multinomial distribution with sample size N and probability vector θ, where $\theta = (\theta_{11}, \theta_{12}, \cdots \theta_{1J}, \theta_{21}, \cdots, \cdots, \theta_{IJ})$. Let n_{ij} denote the count for the cell (i, j), and let θ_{ij}, an element of θ, denote the population proportion for the cell (i, j).

When information on either the row or column classification is missing, we can construct a table of counts for the completely classified cases where x_{ij} denotes the number of cases observed in the (i, j) cell. We can also construct one-way tables of counts for partially classified cases. Let x_{im} denote the number of cases in the i^{th} row category, $i = 1, 2, \cdots, I$, where the column category is unknown, and let x_{mj} denote the number of cases in the j^{th} column category, $j = 1, 2, \cdots, J$, where the row category is unknown. Then, x_{im} and x_{mj} are marginally observed counts on a single variable. Let x_{mm} denote the number of cases where both the row and column categories are missing. The total sample size is

$$\begin{aligned} N &= \sum_{ij} x_{ij} + \sum_i x_{im} + \sum_j x_{mj} + x_{mm} \\ &= n_{cc} + x_{+m} + x_{m+} + x_{mm}. \end{aligned}$$

The likelihood function for the observed counts assuming MCAR or MAR is proportional to

$$\left[\prod_{i=1}^{I}\prod_{j=1}^{J} \theta_{ij}^{\,x_{ij}}\right]\left[\prod_{i=1}^{I} \theta_{i+}^{\,x_{im}}\right]\left[\prod_{j=1}^{J} \theta_{+j}^{\,x_{mj}}\right].$$

The MLE's of $\{\theta_{ij}\}$ under independence model proposed by Chen and Fienberg(1974) are $\hat{\theta}_{ij} = \left(\frac{x_{i+} + x_{im}}{n_{cc} + x_{+m}}\right)\left(\frac{x_{+j} + x_{mj}}{n_{cc} + x_{m+}}\right)$, $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$.

3. Chen and Fienberg Test

The test proposed by Chen and Fienberg(1974) consists of three parts, one is essentially a test for independence for the fully observed data and the other two correspond to tests for row and column margins, respectively:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(x_{ij} - \hat{\alpha}_{ij})^2}{\hat{\alpha}_{ij}} + \sum_{i=1}^{I} \frac{(x_{im} - \hat{\beta}_i)^2}{\hat{\beta}_i} + \sum_{j=1}^{J} \frac{(x_{mj} - \hat{\sigma}_j)^2}{\hat{\sigma}_j}, \tag{1}$$

where $\hat{\alpha}_{ij}$, $\hat{\beta}_i$ and $\hat{\sigma}_j$ are the expected values under independence model assuming MCAR such as

$$\begin{aligned} \hat{\alpha}_{ij} &= n_{cc}\left(\frac{x_{i+} + x_{im}}{n_{cc} + x_{+m}}\right)\left(\frac{x_{+j} + x_{mj}}{n_{cc} + x_{m+}}\right), \\ \hat{\beta}_i &= x_{+m}\left(\frac{x_{i+} + x_{im}}{n_{cc} + x_{+m}}\right), \\ \hat{\sigma}_j &= x_{m+}\left(\frac{x_{+j} + x_{mj}}{n_{cc} + x_{m+}}\right). \end{aligned}$$

If the missing mechanism does not satisfy missing completely at random(MCAR) criterion, there is no explicit form for the expected values in (1). We require iteration procedure to get the expected values.

The last two parts in (1), for partially classified counts, do not provide much information for the association between two categorical variables. Those two parts, for partially classified counts, are close to central Chi-square distributions even though the null hypothesis of independence is not true. The test statistic in (1) uses $I\times J-1$ instead of $(I-1)(J-1)$ degrees of freedom without providing much of an increase in the value of the test statistic. For these reasons, Chen and Fienberg tests are expected to have less power.

4. New Wald test

Let $C_0=(x_{11},\cdots,x_{1J},x_{21},\cdots,x_{2J},\cdots x_{IJ},x_{m1},\cdots,x_{mJ},x_{1m},\cdots x_{Im})'$. Conditional on the value of x_{mm}, C_0 has a multinomial distribution with sample size $n=N-x_{mm}$ and probabilities

$$\pi=(\pi_{11},\cdots,\pi_{1J},\pi_{21},\cdots,\pi_{2J},\cdots\pi_{IJ},\pi_{m1},\cdots,\pi_{mJ},\pi_{1m},\cdots\pi_{Im})'.$$

The expected counts for the fully classified counts are $n_{cc}\hat{\theta}_{ij}$. The differences between the fully classified counts and the expected counts under independence model are

$$a_{ij}=x_{ij}-n_{cc}\left(\frac{x_{i+}+x_{im}}{n_{cc}+x_{+m}}\right)\left(\frac{x_{+j}+x_{mj}}{n_{cc}+x_{m+}}\right).$$

The variance-covariance matrix of C_0 is $Var(C_0)=n(\Delta_\pi-\pi\pi')\equiv V$, where Δ_π is a diagonal matrix with the elements of π on the main diagonal. For $I\times J$ tables, let $A'=(a_{11},\cdots,a_{1J},a_{21},\cdots,a_{2J},\cdots a_{IJ})'$, then $Var(A)=DVD'\equiv\Sigma_a$, where D is the matrix of the first partial derivatives of A with respect to x's in C_0 as follows:

$$D_{p\times q}=\begin{pmatrix} \frac{\partial a_{11}}{\partial x_{11}} & \cdots & \frac{\partial a_{11}}{\partial x_{IJ}} & \frac{\partial a_{11}}{\partial x_{m1}} & \cdots & \frac{\partial a_{11}}{\partial x_{mJ}} & \frac{\partial a_{11}}{\partial x_{1m}} & \cdots & \frac{\partial a_{11}}{\partial x_{Im}} \\ \frac{\partial a_{12}}{\partial x_{11}} & \cdots & \frac{\partial a_{12}}{\partial x_{IJ}} & \frac{\partial a_{12}}{\partial x_{m1}} & \cdots & \frac{\partial a_{12}}{\partial x_{mJ}} & \frac{\partial a_{12}}{\partial x_{1m}} & \cdots & \frac{\partial a_{12}}{\partial x_{Im}} \\ \vdots & \vdots & & & & & & & \vdots \\ \frac{\partial a_{IJ}}{\partial x_{11}} & \cdots & \frac{\partial a_{IJ}}{\partial x_{IJ}} & \frac{\partial a_{IJ}}{\partial x_{m1}} & \cdots & \frac{\partial a_{IJ}}{\partial x_{mJ}} & \frac{\partial a_{IJ}}{\partial x_{1m}} & \cdots & \frac{\partial a_{IJ}}{\partial x_{Im}} \end{pmatrix},$$

where $p=I\times J$ and $q=I\times J+I+J$.

Let $R_i = \frac{r_i}{r_+} = \left(\frac{x_{i+} + x_{im}}{n_{cc} + x_{+m}}\right)$ and $C_j = \frac{c_j}{c_+} = \left(\frac{x_{+j} + x_{mj}}{n_{cc} + x_{m+}}\right)$, then the elements of D matrix are

$$\frac{\partial a_{ij}}{\partial x_{cd}} = \begin{cases} 1 - R_i C_j - n_{cc}\left(C_j \frac{r_+ - r_i}{r_+^2} + R_i \frac{c_+ - c_j}{c_+^2}\right), & c = i, d = j \\ -R_i C_j - n_{cc}\left(C_j \frac{r_+ - r_i}{r_+^2} - R_i \frac{c_j}{c_+^2}\right), & c = i, d \neq j, d \neq m \\ -R_i C_j - n_{cc}\left(C_j \frac{-r_i}{r_+^2} + R_i \frac{c_+ - c_j}{c_+^2}\right), & c \neq i, d = j, c \neq m \\ -n_{cc}\left(R_i \frac{c_+ - c_j}{c_+^2}\right), & c = m, d = j \\ -n_{cc}\left(R_i - \frac{c_j}{c_+^2}\right), & c = m, d \neq j \\ -n_{cc}\left(C_j \frac{r_+ - r_i}{r_+^2}\right), & c = i, d = m \\ -n_{cc}\left(C_j - \frac{r_i}{r_+^2}\right), & c \neq i, d = m \\ -R_i C_j - n_{cc}\left(C_j - \frac{r_i}{r_+^2} + R_i - \frac{c_j}{c_+^2}\right), & c \neq i,\ d \neq j, c \neq m, d \neq m. \end{cases}$$

The hypothesis of independence for $I \times J$ tables is $\begin{cases} H_0: E(A) = \underset{\sim}{0} \\ H_1: E(A) \neq \underset{\sim}{0}. \end{cases}$ The Wald statistic to test of independence is $A' \,\widehat{\Sigma}_a^{-} A$ and this test statistic has an asymptotic χ^2 distribution with $I \times J - 1$ degree of freedom. π is evaluated using sample proportion to estimate Σ_a

5. Simulation Study

The proposed test in Section 4 is compared with the Chen and Fienberg(1974) test through Monte Carlo simulations. For each combination of sample size and level of missingness, Type I error levels are estimated from 1,000 simulated tables under the independence assumption. Power levels are examined by simulating 1000 tables under an alternative to independence.

5.1 Type I error levels

All of the 2×2 incomplete contingency tables to check Type I error levels were generated with equal cell probabilities and data missing completely at random. Four combinations of sample size and level of missing data were considered and 1000 tables were generated for each combination. X_1 and X_2 were independently generated from Bernouli(0.5) random variables. There are two levels N=200 and N=400 for the total sample size N. MX_i is a missing indicator variable independent of X_i. If $MX_i = 1$, the corresponding variable X_i is missing. The four combinations of factors are summarized in Table 1. The percentages of cases with missing information on at least one variable are expected to be 19%, 36%, 51%, and 91% for combination 1, 2, 3, and 4, respectively.

Table 1: Combination of factors generated to check Type I error levels

		$MX_i \sim Ber(p)$
Combination	N	p
1	200	0.1
2	200	0.2
3	200	0.3
4	400	0.7

Table 2 shows the numbers of tables for which the independence null hypothesis was falsely rejected out of 1000 tables for three nominal Type I error levels. The results of two methods seem to have appropriate Type I error levels but the new method has more inflated Type I error levels than Chen and Fienberg test in all combinations.

Table 2: Comparison of Type I error levels

	Chen & Fienberg			New test		
α Combi.	1%	5%	10%	1%	5%	10%
1	12	49	101	15	59	119
2	11	47	100	13	54	111
3	11	60	103	18	66	111
4	7	45	91	15	62	109

5.2 Power Study

There are 4 alternatives to independence for 2×2 tables with equal probability margins to check power. The generated multinomial variables have the cell probability such that

$$(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = (0.2, 0.3, 0.3, 0.2)$$

for all combinations in Table 1.

The numbers in Table 3 indicate the number of tables out of 1000 for which the independence null hypothesis was rejected under the given α levels among 1000 tables in each combination.

Table 3 shows new method has more power than Chen and Fienberg test. As the proportion of missing cases increases from 19% to 91%, the power decreases as expected.

Table 3: Power Comparison

	Chen & Fienberg			New test		
Combi. \ α	1%	5%	10%	1%	5%	10%
1	345	573	687	380	604	698
2	224	454	585	249	483	608
3	160	351	463	185	370	488
4	29	134	211	58	151	235

6. Conclusion

Chen and Fienberg's test is more conservative and has less power in most cases than new Wald test. If there are lots of missing values, we can expect that new test has more appropriate Type I error level and more power.

If the missing mechanism does not satisfy missing completely at random(MCAR) criterion, Chen and Fienberg test statistic has no explicit formula. The Wald test proposed in Section 4 also works under MAR missing mechanism.

References

1. Blumenthal, S. (1968). Multinomial sampling with Partially Categorized Data, *Journal of the American Statistical Association*, 63, 542-551.
2. Chen, T. T., and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data, *Biometrics,* 32, 133-144.
3. Hocking, R. R., and Oxspring, H. H. (1974). The Analysis of Partially Categorized contingency Data, *Biometrics,* 30, 469-483.
4. Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data*, J. Wiley and Sons, New York.
5. Rubin, D. B. (1978). Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, *Proceedings of the Survey Research Methods Section, American Statistical Association,* 1978, 20-34.

[received date : Sep. 2004, accepted date : Dec. 2004]