International Journal of Pure and Applied Mathematics
Volume 66 No. 1 2011, 25-32

RISK OF ESTIMATORS FOR POINT SYMMETRY AND DOUBLE SYMMETRY MODELS FOR SQUARE CONTINGENCY TABLES

Kouji Tahata[1 §], Takahiro Inoue[2], Sadao Tomizawa[3]
[1,2,3]Department of Information Sciences
Faculty of Science and Technology
Tokyo University of Science
2641, Yamazaki, Noda City, Chiba, 278-8510, JAPAN
[1]e-mail: kouji_tahata@is.noda.tus.ac.jp
[3]e-mail: tomizawa@is.noda.tus.ac.jp

Abstract: For contingency tables, Bishop, Fienberg and Holland [1] gave the expected mean squared error (risk) of the maximum likelihood estimators (MLEs) of cell probabilities for the independence model, and Tahata, Miyamoto, Sasajima, and Tomizawa [4] gave the conditional risks of the MLEs for the symmetry and conditional symmetry models.

The present paper gives the risks of those for the point symmetry model and the double symmetry model for square contingency tables. The paper also compares the risks for these models.

AMS Subject Classification: 62H17
Key Words: double symmetry, mean squared error, point symmetry, risk, square tables

1. Introduction

For an $r \times r$ square contingency table with the same row and column classifications, let p_{ij} denote the probability that an observation will fall in the i-th

Received: October 4, 2010

§Correspondence author

row and j-th column of the table $(i = 1, \cdots, r; j = 1, \cdots, r)$. The symmetry (S) model is defined by

$$p_{ij} = p_{ji} \quad (i = 1, \cdots, r; j = 1, \cdots, r);$$

see Bowker [2], and Bishop, Fienberg and Holland [1, p. 282]. This indicates the symmetry of cell probabilities with respect to the main diagonal of the $r \times r$ table. Wall and Lienert [6] defined the point symmetry (PS) model by

$$p_{ij} = p_{i^* j^*} \quad (i = 1, \cdots, r; j = 1, \cdots, r),$$

where $i^* = r + 1 - i$ and $j^* = r + 1 - j$. This indicates the symmetry of cell probabilities with respect to the center point of the $r \times r$ table. Tomizawa [5] considered the double symmetry (DS) model by

$$p_{ij} = p_{ji} = p_{i^* j^*} = p_{j^* i^*} \quad (i = 1, \cdots, r; j = 1, \cdots, r).$$

This model indicates that both of S and PS hold.

Let x_{ij} be the observed frequency in cell (i, j) of the table with $N = \sum\sum x_{ij}$. Assume that $\{x_{ij}\}$ have a multinomial distribution. Let $T_{ij}^{(M)}$ be the maximum likelihood estimator (MLE) of the cell probability p_{ij} under model M.

The expected mean squared error (risk) for the MLEs of $\{p_{ij}\}$ for model M is

$$\begin{aligned} R_M &= \sum_{i=1}^{r}\sum_{j=1}^{r} E\left[\left(T_{ij}^{(M)} - p_{ij}\right)^2\right] \\ &= \sum_{i=1}^{r}\sum_{j=1}^{r} \left(\text{Var}\left(T_{ij}^{(M)}\right) + \text{Bias}\left(T_{ij}^{(M)}\right)\right), \end{aligned}$$

where

$$\begin{aligned} \text{Var}\left(T_{ij}^{(M)}\right) &= E\left[\left(T_{ij}^{(M)} - E\left(T_{ij}^{(M)}\right)\right)^2\right], \\ \text{Bias}\left(T_{ij}^{(M)}\right) &= \left[E\left(T_{ij}^{(M)}\right) - p_{ij}\right]^2. \end{aligned}$$

Bishop et al [1, p. 313] gave the risk of MLEs of cell probabilities for the independence model. Tahata, Miyamoto, Sasajima and Tomizawa [4] gave the conditional risks of MLEs of cell probabilities for the S model, the saturated (SA) model and McCullagh's [3] conditional symmetry model on condition that an observation falls in one of the off-diagonal cells of the square contingency table.

The purpose of the present paper is to give the risks of MLEs of cell probabilities for the PS model and the DS model for square contingency tables.

2. Risks of MLEs for Models

Consider the $r \times r$ contingency table. The MLEs of cell probabilities $\{p_{ij}\}$ under the SA model are given by

$$T_{ij}^{(SA)} = \frac{x_{ij}}{N} \quad (i = 1, \cdots, r; j = 1, \cdots, r).$$

Thus the risk for the SA model is

$$R_{SA} = \frac{1}{N}\left(1 - \sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij}^2\right).$$

Also, the MLEs for $\{p_{ij}\}$ under the S model are given by

$$T_{ij}^{(S)} = \frac{x_{ij} + x_{ji}}{2N} \quad (i = 1, \cdots, r; j = 1, \cdots, r).$$

Thus the risk for the S model is

$$\begin{aligned} R_S &= \sum\sum_{i \neq j} \frac{1}{4}\left[\frac{1}{N}\left\{(p_{ij} + p_{ji}) - (p_{ij} + p_{ji})^2\right\} + (p_{ji} - p_{ij})^2\right] \\ &\quad + \sum_{i=1}^{r} \frac{1}{N} p_{ii}(1 - p_{ii}). \end{aligned}$$

2.1. Risk for Point Symmetry Model

Denote $\frac{r+1}{2}$ by c. The MLEs of cell probabilities $\{p_{ij}\}$ under the PS model are given by

$$T_{ij}^{(PS)} = \begin{cases} \dfrac{x_{ij} + x_{i^*j^*}}{2N} & (i,j) \neq (c,c), \\ \dfrac{x_{cc}}{N} & (i,j) = (c,c), \end{cases}$$

where cell (c,c) does not exist when r is even. We see

$$E\left(T_{ij}^{(PS)}\right) = \begin{cases} \dfrac{p_{ij} + p_{i^*j^*}}{2} & (i,j) \neq (c,c), \\ p_{cc} & (i,j) = (c,c). \end{cases}$$

When the PS model holds, $T_{ij}^{(PS)}$ is an unbiased estimator of p_{ij}. Also, by considering

$$\mathrm{Var}\left(T_{ij}^{(PS)}\right) = \frac{1}{4N^2}\left(\mathrm{Var}(x_{ij}) + \mathrm{Var}(x_{i^*j^*}) + 2\mathrm{Cov}(x_{ij}, x_{i^*j^*})\right)$$

for $(i,j) \neq (c,c)$, we find the variance term as

$$\mathrm{Var}\left(T_{ij}^{(PS)}\right) = \begin{cases} \dfrac{(p_{ij}+p_{i^*j^*}) - (p_{ij}+p_{i^*j^*})^2}{4N} & (i,j) \neq (c,c)\,, \\ \dfrac{p_{cc}\,(1-p_{cc})}{N} & (i,j) = (c,c)\,. \end{cases}$$

Also the bias term is

$$\mathrm{Bias}\left(T_{ij}^{(PS)}\right) = \begin{cases} \dfrac{(p_{i^*j^*} - p_{ij})^2}{4} & (i,j) \neq (c,c)\,, \\ 0 & (i,j) = (c,c)\,. \end{cases}$$

Thus we can obtain the risk for the PS model as

$$R_{PS} = \begin{cases} \displaystyle\sum\sum_{(i,j)\neq(c,c)} \frac{1}{4}\left[\frac{1}{N}\left\{(p_{ij}+p_{i^*j^*}) - (p_{ij}+p_{i^*j^*})^2\right\} + (p_{i^*j^*}-p_{ij})^2\right] \\ \displaystyle +\frac{1}{N}p_{cc}\,(1-p_{cc}) \\ \qquad\qquad\qquad\qquad (\text{when } r \text{ is odd})\,, \\ \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{r}\frac{1}{4}\left[\frac{1}{N}\left\{(p_{ij}+p_{i^*j^*}) - (p_{ij}+p_{i^*j^*})^2\right\} + (p_{i^*j^*}-p_{ij})^2\right] \\ \qquad\qquad\qquad\qquad (\text{when } r \text{ is even})\,. \end{cases}$$

2.2. Risk for Double Symmetry Model

Define the sets of cells as follows:

$$\begin{aligned} D_1 &= \{(i,j)\,|i=1,\cdots,r; j=1,\cdots,r; i\neq j; j\neq i^*\}\,, \\ D_2 &= \{(i,j)\,|i=1,\cdots,r; i=j; (i,j)\neq(c,c)\}\,, \\ D_3 &= \{(i,j)\,|i=1,\cdots,r; j=i^*; (i,j)\neq(c,c)\}\,, \\ D_4 &= \{(i,j)\,|i=j=c\}\,. \end{aligned}$$

Note that D_4 is defined only when r is odd. The MLEs of $\{p_{ij}\}$ under the DS model are given by

$$T_{ij}^{(DS)} = \begin{cases} \dfrac{x_{ij}+x_{ji}+x_{i^*j^*}+x_{j^*i^*}}{4N} & (i,j)\in D_1, \\ \dfrac{x_{ii}+x_{i^*i^*}}{2N} & (i,j)\in D_2, \\ \dfrac{x_{ii^*}+x_{i^*i}}{2N} & (i,j)\in D_3, \\ \dfrac{x_{cc}}{N} & (i,j)\in D_4. \end{cases}$$

We see

$$E\left(T_{ij}^{(DS)}\right) = \begin{cases} \dfrac{p_{ij}+p_{ji}+p_{i^*j^*}+p_{j^*i^*}}{4} & (i,j)\in D_1, \\ \dfrac{p_{ii}+p_{i^*i^*}}{2} & (i,j)\in D_2, \\ \dfrac{p_{ii^*}+p_{i^*i}}{2} & (i,j)\in D_3, \\ p_{cc} & (i,j)\in D_4. \end{cases}$$

When the DS model holds, $T_{ij}^{(DS)}$ is an unbiased estimator of p_{ij}.

The variance term is expressed by

$$\mathrm{Var}\left(T_{ij}^{(DS)}\right) = \begin{cases} \dfrac{(p_{ij}+p_{ji}+p_{i^*j^*}+p_{j^*i^*}) - (p_{ij}+p_{ji}+p_{i^*j^*}+p_{j^*i^*})^2}{16N} & (i,j)\in D_1, \\ \dfrac{(p_{ii}+p_{i^*i^*}) - (p_{ii}+p_{i^*i^*})^2}{4N} & (i,j)\in D_2, \\ \dfrac{(p_{ii^*}+p_{i^*i}) - (p_{ii^*}+p_{i^*i})^2}{4N} & (i,j)\in D_3, \\ \dfrac{p_{cc}(1-p_{cc})}{N} & (i,j)\in D_4. \end{cases}$$

The bias term is

$$\mathrm{Bias}\left(T_{ij}^{(DS)}\right) = \begin{cases} \dfrac{(p_{ji}+p_{i^*j^*}+p_{j^*i^*}-3p_{ij})^2}{16} & (i,j)\in D_1, \\ \dfrac{(p_{i^*i^*}-p_{ii})^2}{4} & (i,j)\in D_2, \\ \dfrac{(p_{i^*i}-p_{ii^*})^2}{4} & (i,j)\in D_3, \\ 0 & (i,j)\in D_4. \end{cases}$$

Therefore the risk for the DS model is given as follows:

When r is odd,

$$\begin{aligned} R_{DS} &= \sum\sum_{(i,j)\in D_1} \frac{1}{16}\left[\frac{1}{N}\left(A_{ij}-A_{ij}^2\right) + (A_{ij}-4p_{ij})^2\right] \\ &+ \sum\sum_{(i,j)\in D_2} \frac{1}{4}\left[\frac{1}{N}\left(B_{ij}-B_{ij}^2\right) + (B_{ij}-2p_{ii})^2\right] \\ &+ \sum\sum_{(i,j)\in D_3} \frac{1}{4}\left[\frac{1}{N}\left(C_{ij}-C_{ij}^2\right) + (C_{ij}-2p_{ii^*})^2\right] \end{aligned}$$

$$+\frac{1}{N}p_{cc}\left(1-p_{cc}\right),$$

and when r is even,

$$\begin{aligned} R_{DS} &= \sum\sum_{(i,j)\in D_1}\frac{1}{16}\left[\frac{1}{N}\left(A_{ij}-A_{ij}^2\right)+\left(A_{ij}-4p_{ij}\right)^2\right] \\ &+\sum\sum_{(i,j)\in D_2}\frac{1}{4}\left[\frac{1}{N}\left(B_{ij}-B_{ij}^2\right)+\left(B_{ij}-2p_{ii}\right)^2\right] \\ &+\sum\sum_{(i,j)\in D_3}\frac{1}{4}\left[\frac{1}{N}\left(C_{ij}-C_{ij}^2\right)+\left(C_{ij}-2p_{ii^*}\right)^2\right], \end{aligned}$$

where $A_{ij}=p_{ij}+p_{ji}+p_{i^*j^*}+p_{j^*i^*}$, $B_{ij}=p_{ii}+p_{i^*i^*}$, and $C_{ij}=p_{ii^*}+p_{i^*i}$.

3. Comparisons between Risks

We can obtain the following theorems from comparing the risks for the SA, S, PS, and DS models.

Theorem 1. *When the PS model holds,*

$$R_{SA}-R_{PS}=\begin{cases}\dfrac{1}{2N}\sum\sum_{(i,j)\neq(c,c)}p_{ij} & (\text{when } r \text{ is odd}), \\ \dfrac{1}{2N} & (\text{when } r \text{ is even}).\end{cases}$$

Theorem 2. *When the DS model holds,*

$$R_{SA}-R_{DS}=\frac{3}{4N}\sum\sum_{(i,j)\in D_1}p_{ij}+\frac{1}{2N}\sum\sum_{(i,j)\in D_2}p_{ii}+\frac{1}{2N}\sum\sum_{(i,j)\in D_3}p_{ii^*}.$$

Theorem 3. *When the DS model holds,*

$$R_{S}-R_{DS}=\frac{1}{4N}\sum\sum_{(i,j)\in D_1}p_{ij}+\frac{1}{2N}\sum\sum_{(i,j)\in D_2}p_{ii}.$$

Theorem 4. *When the DS model holds,*

$$R_{PS}-R_{DS}=\frac{1}{4N}\sum\sum_{(i,j)\in D_1}p_{ij}.$$

We see that

(1) from Theorem 1, when the PS model holds, $R_{PS}<R_{SA}$,

(2) from Theorem 2, when the DS model holds, $R_{DS} < R_{\mathrm{SA}}$*,*

(3) from Theorem 3, when the DS model holds, $R_{DS} < R_{S}$*,*

(4) from Theorem 4, when the DS model holds, $R_{DS} < R_{PS}$*.*

Thus we see that the overall variability for the estimators based on the PS model is smaller than that based on the SA model when the PS model is correct. Also the overall variability based on the DS model is smaller than that based on the SA, S, and PS models when DS model is correct.

4. Concluding Remark

By comparing the risks between the models in terms of Theorems 1, 2, 3, and 4, we find that when the simpler model is correct, the overall variability for the estimators based on the simpler model is smaller than the estimator based on the more complicated model.

Acknowledgments

We would like to thank the referees.

References

[1] Y.M.M. Bishop, S.E. Fienberg, P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Massachusetts (1975).

[2] A.H. Bowker, A test for symmetry in contingency tables, *Journal of the American Statistical Association*, **43** (1948), 572-574.

[3] P. McCullagh, A class of parametric models for the analysis of square contingency tables with ordered categories, *Biometrika*, **65** (1978), 413-418.

[4] K. Tahata, N. Miyamoto, T. Sasajima, S. Tomizawa, Expected mean squared error of estimators for symmetry and asymmetry models for contingency tables, *International Journal of Pure and Applied Mathematics*, **46**, No. 4 (2008), 389-401.

[5] S. Tomizawa, Double symmetry model and its decomposition in a square contingency table, *Journal of the Japan Statistical Society*, **15** (1985), 17-23.

[6] K.D. Wall, G.A. Lienert, A test for point-symmetry in J-dimensional contingency-cubes, *Biometrical Journal*, **18** (1976), 259-264.