REVSTAT – Statistical Journal
Volume 4, Number 2, June 2006, 153–161

DECOMPOSITIONS OF SYMMETRY MODEL INTO MARGINAL HOMOGENEITY AND DISTANCE SUBSYMMETRY IN SQUARE CONTINGENCY TABLES WITH ORDERED CATEGORIES

Authors: Sadao Tomizawa
– Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Noda City, Chiba, 278-8510, Japan
tomizawa@is.noda.tus.ac.jp

Nobuko Miyamoto
– Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Noda City, Chiba, 278-8510, Japan
miyamoto@is.noda.tus.ac.jp

Masami Ouchi
– Graduate School of Sciences and Technology, Tokyo University of Science, Noda City, Chiba, 278-8510, Japan

Received: July 2005 Revised: October 2005 Accepted: October 2005

Abstract:

- For square contingency tables with ordered categories, this paper proposes some distance subsymmetry models. The one model indicates that the cumulative probability that an observation will fall in row category i or below and column category $i+k$ ($k \geq 2$) or above, is equal to the probability that it falls in column category i or below and row category $i+k$ or above. This paper also gives the decomposition of the symmetry model into the marginal homogeneity model and some distance subsymmetry models. The father-son occupational mobility data in Britain and the women's unaided vision data in Britain are analyzed.

Key-Words:

- *decomposition; distance subsymmetry model; marginal homogeneity model; ordered category; square contingency table; symmetry model.*

AMS Subject Classification:

- 62H17.

1. INTRODUCTION

For an $r \times r$ square contingency table with ordered categories, let p_{ij} denote the probability that an observation will fall in the i^{th} row and j^{th} column of the table ($i = 1, 2, ..., r$; $j = 1, 2, ..., r$). The symmetry (S) model is defined by

$$p_{ij} = p_{ji} \qquad \text{for} \quad i = 1, 2, ..., r; \quad j = 1, 2, ..., r \ .$$

See Bishop, Fienberg, and Holland ([2], p. 282). This model indicates that the probability that an observation will fall in cell (i, j) of the table is equal to the probability that it falls in cell (j, i). Namely, this describes a structure of symmetry of the cell probabilities $\{p_{ij}\}$ with respect to the main diagonal of the table.

Let X_1 and X_2 denote the row and column variables, respectively. The marginal homogeneity (MH) model is defined by

$$\Pr(X_1 = i) \ = \ \Pr(X_2 = i) \qquad \text{for} \quad i = 1, 2, ..., r \ ,$$

namely

$$p_{i\cdot} = p_{\cdot i} \qquad \text{for} \quad i = 1, 2, ..., r \ ,$$

where $p_{i\cdot} = \sum_{t=1}^{r} p_{it}$ and $p_{\cdot i} = \sum_{s=1}^{r} p_{si}$ (Stuart, [8]). This indicates that the row marginal distribution is identical with the column marginal distribution.

Let

$$G_{ij} \ = \ \Pr\bigl(X_1 \le i, \, X_2 \ge j\bigr) = \sum_{s=1}^{i} \sum_{t=j}^{r} p_{st} \qquad \text{for} \quad i < j \ ,$$

and

$$G^*_{ij} \ = \ \Pr\bigl(X_1 \ge i, \, X_2 \le j\bigr) = \sum_{s=i}^{r} \sum_{t=1}^{j} p_{st} \qquad \text{for} \quad i > j \ .$$

Then the S model may be expressed as

$$G_{ij} = G^*_{ji} \qquad \text{for} \quad i < j \ . \tag{1.1}$$

The MH model may be expressed as

$$G_{i,i+1} = G^*_{i+1,i} \qquad \text{for} \quad i = 1, 2, ..., r-1 \ . \tag{1.2}$$

The S model implies the MH model. So, from (1.1) and (1.2), we are interested in decomposing (1.1) into (1.2) and the structure of

$$G_{ij} = G^*_{ji} \qquad \text{for} \quad j - i = 2, 3, ..., r-1; \quad i < j \ .$$

The purpose of this paper is to give the decompositions of the S model into some new models. The decompositions may be useful for seeing the reason for the poor fit when the S model fits the data poorly.

2. DECOMPOSITIONS OF SYMMETRY MODEL

This section proposes some new models based on $\{G_{ij}\}$ and based on $\{p_{ij}\}$, and gives the decompositions of the S model.

2.1. Distance Cumulative Subsymmetry Model

Consider a model defined by

$$G_{ij} = G^*_{ji} \qquad \text{for} \quad j-i = 2, 3, ..., r-1; \quad i < j \; , \tag{2.1}$$

which is equivalent to

$$p_{ij} = p_{ji} \qquad \text{for} \quad j-i = 2, 3, ..., r-1; \quad i < j \; .$$

This model indicates that the probability that an observation will fall in cell (i, j), which is one of cells such that the distance from the main diagonal is greater than or equal to 2, is equal to the probability that the observation falls in cell (j, i). We shall refer to (2.1) as the subsymmetry (SS) model.

Next, for fixed k $(k = 2, 3, ..., r-1)$, consider a model defined by

$$G_{i,i+k} = G^*_{i+k,i} \qquad \text{for} \quad i = 1, 2, ..., r-k \; . \tag{2.2}$$

This model indicates that the cumulative probability that an observation will fall in row category i or below and column category $i+k$ or above, is equal to the cumulative probability that the observation falls in column category i or below and row category $i+k$ or above. We shall refer to (2.2) as the model of the distance cumulative subsymmetry with the difference k between the diagonal containing the cutpoint [i and $i+k$] and the main diagonal (denoted by the DCS-k model).

2.2. Distance Subsymmetry Model

For fixed k $(k = 1, 2, ..., r-1)$, consider a model defined by

$$p_{ij} = p_{ji} \qquad \text{for} \quad j-i = k; \quad i < j \; . \tag{2.3}$$

This model indicates that the probability that an observation will fall in cell (i, j) with the distance k from the main diagonal, is equal to the probability that the observation falls in cell (j, i) with the same distance k. We shall refer to (2.3) as the distance subsymmetry with distance k (DS-k) model. We obtain the following theorem.

Theorem 2.1. *The following four statements are equivalent:*

(1) *the S model holds,*

(2) *the MH and SS models hold,*

(3) *the MH and* $\{$*DCS-*$k\}$ $(k = 2, 3, ..., r-1)$ *models hold,*

(4) *all the* $\{$*DS-*$k\}$ $(k = 1, 2, ..., r-1)$ *models hold.*

2.3. Goodness-of-Fit Test

Assume that a multinomial distribution is applied to the $r \times r$ table. The maximum likelihood estimates (MLEs) of expected frequencies under the S, SS and DS-k models are obtained in the closed-forms. The MLEs of them under the MH and DCS-k models could not be obtained in the closed-forms, however, they could be obtained using the Newton–Raphson methods in the log-likelihood equations.

The likelihood ratio statistic for testing the goodness-of-fit of the model is

$$G^2 = 2 \sum_{i=1}^{r} \sum_{j=1}^{r} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right),$$

with the corresponding degrees of freedom (df), where n_{ij} is the observed frequency in cell (i, j), and $\hat{m}_{ij}$ is the MLE of expected frequency m_{ij} under the model. The numbers of df for the MH and SS models are $r-1$ and $(r-1)(r-2)/2$. Also the numbers of df for the DCS-k model $(k = 2, 3, ..., r-1)$ are $r-k$, and those for the DS-k model $(k = 1, 2, ..., r-1)$ are $r-k$.

3. EXAMPLES

We shall analyze the data in Tables 1 and 2.

3.1. Analysis of Table 1

Consider the data in Table 1, taken directly from Goodman [5]. These data relate the father's and his son's occupational status category in Britain. These data have been analyzed by some statisticians, including Agresti ([1], p. 206) and Tomizawa [9].

Table 1: The father's and son's occupational mobility data in Britain; from Goodman [5].

Father's status	**Son's status**					**Total**
	(1)	(2)	(3)	(4)	(5)	
(1)	50	45	8	18	8	129
(2)	28	174	84	154	55	495
(3)	11	78	110	223	96	518
(4)	14	150	185	714	447	1510
(5)	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

Table 2: Unaided distance vision of 7477 women aged 30-39 employed in Royal Ordnance factories in Britain from 1943 to 1946; from Stuart [8].

Right eye grade	**Left eye grade**				**Total**
	Best (1)	Second (2)	Third (3)	Worst (4)	
Best (1)	1520	266	124	66	1976
Second (2)	234	1512	432	78	2256
Third (3)	117	362	1772	205	2456
Worst (4)	36	82	179	492	789
Total	1907	2222	2507	841	7477

Table 3 presents the likelihood ratio chi-square values G^2 for the models applied to these data. The S model fits these data poorly, yielding $G^2 = 37.46$ with 10 df (Table 3). By using the decompositions of the S model, we shall consider the reason why the S model fits these data poorly.

The MH model fits these data poorly, however, the SS model fits these data well (Table 3). Therefore we can see from Theorem 2.1 that the poor fit of the S model is caused by the poor fit of the MH model (rather than the SS model).

Moreover, all the DCS-k ($k = 2, 3, 4$) models fit the data in Table 1 well. Therefore we can also see from Theorem 2.1 that the poor fit of the S model is caused by the poor fit of the MH model (rather than the DCS-k ($k = 2, 3, 4$) models). The DCS-k ($k = 2, 3, 4$) models provide that the probability that the occupational status category of the father in a pair is k or above higher than that of his son, is estimated to equal the probability that the status category of the son is k or above higher than that of his father.

Table 3: Likelihood ratio chi-square values for models applied to the data in Table 1.

Applied models	Degrees of freedom	Likelihood ratio chi-square
S	10	37.46*
MH	4	32.80*
SS	6	8.58
DCS-2	3	6.89
DCS-3	2	4.29
DCS-4	1	2.36
DS-1	4	28.89*
DS-2	3	3.97
DS-3	2	2.25
DS-4	1	2.36

* means significant at the 0.05 level.

In addition, the DS-k ($k=2,3,4$) models fit these data well, but the DS-1 model fits these data poorly. Therefore we can see from Theorem 2.1 that the poor fit of the S model is caused by the poor fit of the DS-1 model (rather than the DS-k ($k = 2,3,4$) models). The DS-k ($k = 2,3,4$) models provide that the probability that the occupational status of the father in a pair is k categories higher than that of his son, is estimated to equal the probability that the status of the son is k categories higher than that of his father.

3.2. Analysis of Table 2

Consider the data in Table 2, taken directly from Stuart [8]. These data are constructed from the unaided distance vision of 7477 women aged 30–39 employed in Royal Ordnance factories in Britain from 1943 to 1946. These data have been analyzed by many statisticians, including Caussinus [3], Bishop *et al.* ([2], p. 284), McCullagh [6], Goodman [4], Tomizawa [10], and Miyamoto, Ohtsuka and Tomizawa [7].

From Table 4, we see that the S model fits these data poorly, yielding $G^2 = 19.25$ with 6 df. By using the decompositions of the S model, we shall consider the reason why the S model fits these data poorly.

Both the MH and SS models, being the decomposed models of the S model, fit these data poorly. So, in order to analyze these data in more details, we shall apply Theorem 2.1. The DCS-2 model fits these data well, however, the DCS-3 model fits them poorly (Table 4). Therefore we can see from Theorem 2.1 that the poor fit of the S model is caused by the poor fits of the MH and DCS-3 models (rather than the DCS-2 model). The DCS-2 model provides that the probability that a woman's right eye is 2 or 3 grades better than her left eye is estimated to equal the probability that the woman's left eye is 2 or 3 grades better than her right eye.

Table 4: Likelihood ratio chi-square values for models applied to the data in Table 2.

Applied models	**Degrees of freedom**	**Likelihood ratio chi-square**
S	6	19.25*
MH	3	11.99*
SS	3	9.26*
DCS-2	2	5.00
DCS-3	1	8.96*
DS-1	3	9.99*
DS-2	2	0.30
DS-3	1	8.96*

* means significant at the 0.05 level.

The DS-2 model fits these data very well, however, the DS-1 and DS-3 models fit them poorly (Table 4). Therefore we can see from Theorem 2.1 that the poor fit of the S model is caused by the poor fits of the DS-1 and DS-3 models (rather than the DS-2 model). The DS-2 model provides that the probability that a woman's right eye is 2 grades better than her left eye is estimated to equal the probability that the woman's left eye is 2 grades better than her right eye.

Therefore, these indicate that there are the structures of subsymmetry (not the complete symmetry) in these data.

4. CONCLUDING REMARKS

Theorem 2.1 gives the decompositions of the S model into some distance subsymmetry models including the MH model. These decompositions would be useful for seeing which structures of distance subsymmetry are lacking when the S model does not hold for analyzing the data.

Finally we note that Caussinus [3] gave the decomposition of the S model into the quasi-symmetry model, which indicates the symmetry of odds-ratios, and the MH model. Caussinus's decomposition would be useful for seeing which of the structure of symmetry of odds-ratios and the structure of marginal homogeneity is lacking when the S model does not hold for analyzing the data (although Caussinus's decomposition could not see which structures of some distance subsymmetry are lacking).

ACKNOWLEDGMENTS

The authors thank the referee and the associate editor for their useful comments.

REFERENCES

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, John Wiley & Sons, New York.

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

[3] Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, **29**, 77–182.

[4] Goodman, L.A. (1979). Multiplicative models for square contingency tables with ordered categories, *Biometrika*, **66**, 413–418.

[5] Goodman, L.A. (1981). Association models and the bivariate normal for contingency tables with ordered categories, *Biometrika*, **68**, 347–355.

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

[7] Miyamoto, N.; Ohtsuka, W. and Tomizawa, S. (2004). Linear diagonals-parameter symmetry and quasi-symmetry models for cumulative probabilities in square contingency tables with ordered categories, *Biometrical Journal*, **46**, 664–674.

[8] Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika*, **42**, 412–416.

[9] Tomizawa, S. (1990). Another linear diagonals-parameter symmetry model for square contingency tables with ordered categories, *South African Statistical Journal*, **24**, 117–125.

[10] Tomizawa, S. (1993). Diagonals-parameter symmetry model for cumulative probabilities in square contingency tables with ordered categories, *Biometrics*, **49**, 883–887.