Brazilian Journal of Probability and Statistics (2006), 20, pp. 191–198.

Decomposing independence model using collapsed tables for contingency tables having ordered categories

Sadao Tomizawa and **Masaya Sakurai**
Tokyo University of Science

Abstract: For two-way contingency tables with ordered categories, it is shown that the independence model holds if and only if the Goodman-Kruskal gamma measure equals zero and the independence model holds for the collapsed tables which are obtained by combining a pair of adjacent rows and by combining a pair of adjacent columns. Using this decomposition, examples are analyzed.

Key words: Concordance, decomposition, discordance, gamma measure, independence.

1 Introduction

For an $r \times c$ contingency table with ordered categories, let X and Y denote the row and column variables, respectively, and let $\mathrm{P}(X = i, Y = j) = p_{ij}$ for $i = 1, \cdots, r$ and $j = 1, \cdots, c$. The independence model is defined by

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

where

$$p_{i.} = \sum_{t=1}^{c} p_{it}, \quad p_{.j} = \sum_{s=1}^{r} p_{sj};$$

see Bishop, Fienberg and Holland (1975, p.14). Let $\theta_{(i<j;s<t)}$ denote the odds ratio defined for rows i and j $(> i)$ and columns s and t $(> s)$. Thus

$$\theta_{(i<j;s<t)} = \frac{p_{is}p_{jt}}{p_{js}p_{it}} \quad (1 \le i < j \le r; 1 \le s < t \le c).$$

The independence model may be expressed as

$$\theta_{(i<j;s<t)} = 1 \quad (1 \le i < j \le r; 1 \le s < t \le c).$$

Let P_C and P_D denote the probability of concordance for a randomly selected pair of observations and the probability of discordance for the pair, respectively. Thus

$$P_C = 2\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}\sum_{s=1}^{c-1}\sum_{t=s+1}^{c} p_{is}p_{jt}, \quad P_D = 2\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}\sum_{s=1}^{c-1}\sum_{t=s+1}^{c} p_{js}p_{it}.$$

The gamma measure, proposed by Goodman and Kruskal (1954), is defined by

$$\gamma = \frac{P_C - P_D}{P_C + P_D};$$

see Agresti (1984, p.160). If the independence model holds, then $\gamma = 0$, namely, $P_C = P_D$; but the converse does not hold. We are now interested in what structure between X and Y is necessary for obtaining the independence, in addition to $P_C = P_D$.

Table 1a is the artificial 3×3 table of the probabilities $\{p_{ij}\}$, and Table 1b is the 2×3 table obtained by combining rows 1 and 2. It is easily seen that the independence model does not hold for Table 1a but the independence model holds for Table 1b. Generally, if the independence model holds for the original $r \times c$ table, then the independence model holds for the collapsed $s \times t$ table (where $s \leq r$ and $t \leq c$) obtained by combining some rows and/or some columns. However, the converse does not hold, for example, as Table 1. We are now interested in what structure between X and Y is necessary for obtaining the independence for the original $r \times c$ table when the independence model holds for the collapsed $s \times t$ table.

Table 1 *(a) Artificial* 3×3 *table of the probabilities and (b) collapsed* 2×3 *table obtained by combining rows 1 and 2.*

(a) 3×3 table

$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$
$\frac{3}{15}$	$\frac{2}{15}$	$\frac{1}{15}$
$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

(b) Collapsed 2×3 table

$\frac{4}{15}$	$\frac{4}{15}$	$\frac{4}{15}$
$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

The purpose of this paper is to give a decomposition of the independence model for the original $r \times c$ table.

2 Decomposition of independence model

Consider the $r \times c$ contingency table. For $a = 1, \cdots, r-1$ and $b = 1, \cdots, c-1$, collapse the $r \times c$ table into the $(r-1) \times c$ table $\Delta_{R(a)}$ obtained by combining the

ath and $(a+1)$th row categories and into the $r \times (c-1)$ table $\Delta_{C(b)}$ obtained by combining the bth and $(b+1)$th column categories, respectively. Then, we obtain the following theorem.

Theorem 2.1 *For arbitrary a and b $(a = 1, \cdots, r-1; b = 1, \cdots, c-1)$, the independence model holds for the $r \times c$ table if and only if P_C equals P_D for the $r \times c$ table and the independence models hold for both tables $\Delta_{R(a)}$ and $\Delta_{C(b)}$.*

Proof. It is easily seen that for arbitrary a and b, if the independence model holds for the $r \times c$ table, then $P_C = P_D$ and the independence model holds for both tables $\Delta_{R(a)}$ and $\Delta_{C(b)}$. Therefore we shall prove that if $P_C = P_D$ and the independence models hold for both tables $\Delta_{R(a)}$ and $\Delta_{C(b)}$, then the independence model holds for the $r \times c$ table. Consider the case of $a = 1$ and $b = 1$. For the $r \times c$ table, we obtain

$$\begin{aligned}\frac{1}{2}P_C &= \sum_{i=1}^{r-1}\sum_{j=i+1}^{r}\sum_{s=1}^{c-1}\sum_{t=s+1}^{c} p_{is}p_{jt} \\ &= \sum_{s=1}^{c-1}\sum_{t=s+1}^{c}\left[p_{1s}p_{2t} + (p_{1s}+p_{2s})\sum_{j=3}^{r} p_{jt} + \sum_{i=3}^{r-1}\sum_{j=i+1}^{r} p_{is}p_{jt}\right]. \quad (2.1)\end{aligned}$$

From the assumption that the independence model holds for the table $\Delta_{R(1)}$, the equation (2.1) can be expressed as

$$\begin{aligned}\frac{1}{2}P_C &= \sum_{s=1}^{c-1}\sum_{t=s+1}^{c}\left[p_{1s}p_{2t} + (p_{1t}+p_{2t})\sum_{j=3}^{r} p_{js} + \sum_{i=3}^{r-1}\sum_{j=i+1}^{r} p_{js}p_{it}\right] \\ &= \sum_{s=1}^{c-1}\sum_{t=s+1}^{c}\left[p_{1s}p_{2t} + \left(p_{1t}\sum_{j=2}^{r} p_{js} - p_{1t}p_{2s}\right) + \sum_{i=2}^{r-1}\sum_{j=i+1}^{r} p_{js}p_{it}\right] \\ &= \sum_{s=1}^{c-1}\sum_{t=s+1}^{c}(p_{1s}p_{2t} - p_{1t}p_{2s}) + \sum_{i=1}^{r-1}\sum_{j=i+1}^{r}\sum_{s=1}^{c-1}\sum_{t=s+1}^{c} p_{js}p_{it} \\ &= \sum_{s=1}^{c-1}\sum_{t=s+1}^{c}(p_{1s}p_{2t} - p_{1t}p_{2s}) + \frac{1}{2}P_D.\end{aligned}$$

From the assumption that $P_C = P_D$, we obtain

$$\sum_{s=1}^{c-1}\sum_{t=s+1}^{c} p_{1t}p_{2s} = \sum_{s=1}^{c-1}\sum_{t=s+1}^{c} p_{1s}p_{2t}.$$

Hence, we see

$$\begin{aligned} & p_{21}p_{12} + (p_{21} + p_{22}) \sum_{t=3}^{c} p_{1t} + \sum_{s=3}^{c-1} \sum_{t=s+1}^{c} p_{2s}p_{1t} \\ = \quad & p_{11}p_{22} + (p_{11} + p_{12}) \sum_{t=3}^{c} p_{2t} + \sum_{s=3}^{c-1} \sum_{t=s+1}^{c} p_{1s}p_{2t}. \end{aligned} \tag{2.2}$$

From the assumption that the independence model holds for the table $\Delta_{C(1)}$, the left term of equation (2.2) can be expressed as

$$p_{21}p_{12} + (p_{11} + p_{12}) \sum_{t=3}^{c} p_{2t} + \sum_{s=3}^{c-1} \sum_{t=s+1}^{c} p_{1s}p_{2t}.$$

Therefore we see $p_{11}p_{22} = p_{21}p_{12}$. Hence, from the assumption that the independence models hold for both tables $\Delta_{R(1)}$ and $\Delta_{C(1)}$, we see

$$\theta_{(i<j;s<t)} = 1 \quad (1 \le i < j \le r; 1 \le s < t \le c).$$

Namely, the independence model holds for the $r \times c$ table. In the similar way, for the other cases of a and b, the independence model holds for the $r \times c$ table. The proof is completed.

3 Examples

Example 1. The data in Table 2, taken from Fienberg (1980, p.20), present the relationship between aptitude (as measured at an earlier data by a scholastic aptitude test) and occupation.

From Table 4 we see that (1) the independence model for the original table does not hold, (2) the probability of concordance, P_C, equals the probability of discordance, P_D, but (3) the independence model does not hold for the collapsed $\Delta_{R(1)}$ table and for the collapsed $\Delta_{C(1)}$ table. Therefore, from Theorem 2.1 we see that the poor fit of the independence model for the original table is caused by the influence of the lack of structure of independence for the collapsed $\Delta_{R(1)}$ and $\Delta_{C(1)}$ tables (rather than the lack of structure that P_C equals P_D).

Example 2. Consider the artificial data in Table 3. From Table 4 we see that (1) the independence model for the original table does not hold, (2) P_C does not equal P_D, but (3) the independence model holds for the collapsed $\Delta_{R(1)}$ table and for the collapsed $\Delta_{C(1)}$ table. Therefore, from Theorem 2.1 we see that the poor fit of the independence model for the original table is caused by the influence of the lack of structure that P_C equals P_D (rather than the lack of structure of independence for the collapsed $\Delta_{R(1)}$ and $\Delta_{C(1)}$ tables).

Table 2 *(a) Cross-classification of subjects according to the aptitude and the occupation; from Fienberg (1980, p. 20), (b) collapsed* $\Delta_{R(1)}$ *table, and (c) collapsed* $\Delta_{C(1)}$ *table.*

(a) Original table

	Occupational level				
Aptitude	O1	O2	O3	O4	Totals
(low) A1	122	30	20	472	644
A2	226	51	66	704	1047
A3	306	115	96	1072	1589
A4	130	59	38	501	728
(high) A5	50	31	15	249	345
Totals	834	286	235	2998	4353

(b) $\Delta_{R(1)}$ table

348	81	86	1176
306	115	96	1072
130	59	38	501
50	31	15	249

(c) $\Delta_{C(1)}$ table

152	20	472
277	66	704
421	96	1072
189	38	501
81	15	249

Table 3 *(a) Artificial data of* 4×4 *table, (b) collapsed* $\Delta_{R(1)}$ *table, and (c) collapsed* $\Delta_{C(1)}$ *table.*

(a) Original table

	(1)	(2)	(3)	(4)	Totals
(1)	2	4	6	8	20
(2)	8	6	4	2	20
(3)	6	4	2	2	14
(4)	8	1	2	1	12
Totals	24	15	14	13	66

(b) $\Delta_{R(1)}$ table

10	10	10	10
6	4	2	2
8	1	2	1

(c) $\Delta_{C(1)}$ table

6	6	8
14	4	2
10	2	2
9	2	1

Table 4 *Likelihood ratio chi-square values $G^2(M_s)$ and Pearson's chi-square values $X^2(M_s)$ for testing goodness-of-fit of models M_s $(s = 1, 2, 3, 4)$ applied to Tables 2 and 3, where M_1 : independence (I) model for original table, $M_2 : P_C = P_D$ for original table, M_3 : I model for $\Delta_{R(1)}$ table, and M_4 : I model for $\Delta_{C(1)}$ table.*

	Table 2			Table 3		
Models	Degrees of freedom	$G^2(M_s)$	$X^2(M_s)$	Degrees of freedom	$G^2(M_s)$	$X^2(M_s)$
M_1	12	37.41*	35.80*	9	16.92**	16.23**
M_2	1	0.003	0.003	1	11.07*	10.25*
M_3	9	25.31*	24.64*	6	8.41	8.28
M_4	8	16.54*	15.47*	6	11.45	11.45

* means the significant at the 0.05 level
** means that the value is almost on the 5 percent point

4 Concluding remarks

As shown in Examples in Section 3, when the independence model fits the data poorly, Theorem 2.1 may be useful for seeing the reason for the poor fit, namely, which of the lack of structure that the probability of concordance, P_C, equals the probability of discordance, P_D, and the lack of structure of independence for collapsed tables, influences stronger.

Acknowledgements

The authors would like to thank Dr. N. Miyamoto, Tokyo University of Science, for many helpful comments. They also thank a referee for the helpful comments.

(Received May, 2006. Accepted June, 2006.)

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York, Wiley.

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

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd edition. Cambridge, Massachusetts, The MIT press.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association* **49**, 732-764.

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