Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/110104

Partial Quasi-Symmetry Model for Square Contingency Tables with Nominal Categories

Kiyotaka Iki[1,*] *and Sadao Tomizawa*[2]

[1]College of Economics, Nihon University, Chiyoda Ward, Tokyo, Japan
[2]Department of Information Science, School of Information Science, Meisei University, Hino City, Tokyo, Japan

Received: 2 Jul. 2023, Revised: 2 Sep. 2023, Accepted: 13 Oct. 2023
Published online: 1 Jan. 2024

Abstract: For square contingency tables with nominal categories, the quasi-symmetry model (Caussinus, 1965) was characterized in terms of the symmetry of odds ratios. This study proposes new models that partially indicate the structure of the symmetry of odds ratios. This study also decomposed the symmetry model using the proposed model. An analysis of the data representing changes in membership and attitudes toward the leading crowd was provided.

Keywords: Decomposition, quasi-symmetry, square contingency table, symmetry

1 Introduction

We consider an $r \times r$ square contingency table with the same row and column classifications. Let p_{ij} denote the probability that an observation falls in the ith row and jth column of the table ($i = 1,\ldots,r; j = 1,\ldots,r$). The symmetry (S) model is defined as

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

where $\psi_{ij} = \psi_{ji}$; see Bowker [1] and Bishop et al. [2]. This model indicates a structure of probability symmetry with respect to the main diagonal of the table. As an extension of the S model, Caussinus [3] proposed the quasi-symmetry (QS) model defined by

$$p_{ij} = \alpha_i \beta_j \psi_{ij} \quad (i = 1,\ldots,r; j = 1,\ldots,r),$$

where $\psi_{ij} = \psi_{ji}$; see also Bradley and Terry [4] and Tahata et al. [5]. A special case of this model obtained by setting $\{\alpha_i = \beta_i\}$ as the S model. Denote the odds ratio by

$$\theta_{(i<s;j<t)} = \frac{p_{ij}p_{st}}{p_{sj}p_{it}} \quad (i < s : j < t).$$

The QS model is expressed as

$$\theta_{(i<s;j<t)} = \theta_{(j<t;i<s)} \quad (i < s : j < t). \tag{1}$$

Therefore, the QS model is characterized in terms of the symmetry of odds ratios (although the S model indicates the symmetry of cell probabilities). From Equation (1), the QS model is further expressed as

$$p_{ij}p_{jk}p_{ki} = p_{ji}p_{kj}p_{ik} \quad (1 \le i < j < k \le r).$$

The marginal homogeneity (MH) model is defined as follows:

$$p_{i\cdot} = p_{\cdot i} \quad (i = 1,\ldots,r),$$

* Corresponding author e-mail: iki.kiyotaka@nihon-u.ac.jp

where $p_{i\cdot} = \sum_{t=1}^{r} p_{it}$ and $p_{\cdot i} = \sum_{s=1}^{r} p_{si}$ (Stuart [6]). This model indicates that the row marginal distribution is identical to the column marginal distribution. Caussinus [3] provided the theorem that the S model holds if and only if both the QS and MH models hold.

Tomizawa et al. [7] proposed a conditional difference asymmetry (CDAS) model defined as

$$p_{ij} = \begin{cases} e^{\Delta_{ij}} \psi_{ij} & (i < j), \\ \psi_{ij} & (i \geq j), \end{cases}$$

where $|\Delta_{ij}| = \Delta$ and $\psi_{ij} = \psi_{ji}$. This model indicates that the odds p_{ij}/p_{ji} $(i < j)$ are equal to e^{Δ} for some $i < j$ and $e^{-\Delta}$ for some $i < j$. Note that a special case of CDAS model obtained by setting $\Delta = 0$ is the S model. The CDAS model also indicates that $|p_{ij}^{c} - p_{ji}^{c}|$ for $i < j$ is constant, where $p_{ij}^{c} = p_{ij}/(p_{ij} + p_{ji})$.

Under the QS model, the structure $p_{ij}p_{jk}p_{ki} = p_{ji}p_{kj}p_{ik}$ holds for any $1 \leq i < j < k \leq r$. We then proposed a model in which the structure holds for some $i < j < k$. Furthermore, if the S model holds, the CDAS model holds; however, the converse does not necessarily hold. In addition to the structure of the CDAS model, we are interested in what kind of structure the S model holds. We propose a model that satisfies these constraints.

Section 2 proposes the models and describes the properties of the new models, includes the decompositions using the proposed models, and shows the maximum likelihood estimates of expected frequencies under the proposed models. Section 3 applies the proposed model to the data in table representing changes in membership and attitude for the "Leading Crowd". Finally, Section 4 provides the conclusions.

2 Models

We consider an $r \times r$ square contingency table. We propose a model defined as

$$\prod_{1 \leq i < j < k \leq r} \prod \left(p_{ij}p_{jk}p_{ki} - p_{ji}p_{kj}p_{ik} \right) = 0.$$

This model indicates that, for at least one triple l, m and n $(1 \leq l < m < n \leq r)$, the structure $p_{lm}p_{mn}p_{nl} - p_{ml}p_{nm}p_{ln} = 0$ holds. This model is referred to as the partial quasi-symmetry (PQS) model.

Additionally, for fixed i, j and k $(1 \leq i < j < k \leq r)$, we propose a model defined by

$$p_{ij}p_{jk}p_{ki} - p_{ji}p_{kj}p_{ik} = 0.$$

This model is denoted as PQS(i, j, k). When the number of categories in the square contingency table is r, the models can be defined as $r!/((r-3)!3!)$ types.

We obtain the following theorem.

Theorem 2.1. The S model holds if and only if both the CDAS and PQS models hold.

Proof. If the S model holds, then the CDAS and PQS models hold. Assuming that the CDAS and PQS models hold, we demonstrate that the S model holds. From the PQS model, for some i, j and k $(1 \leq i < j < k \leq r)$, we observe that

$$\frac{p_{ij}p_{jk}p_{ki}}{p_{ji}p_{kj}p_{ik}} = 1. \tag{2}$$

Additionally, from the CDAS model, for any i, j and k $(1 \leq i < j < k \leq r)$, we have

$$\frac{p_{ij}p_{jk}p_{ki}}{p_{ji}p_{kj}p_{ik}} = \frac{e^{\Delta_{ij}}e^{\Delta_{jk}}}{e^{\Delta_{ik}}}. \tag{3}$$

Since $|\Delta_{ij}| = \Delta$ $(i < j)$, the right-hand side of Equation (3) is expressed as $e^{3\Delta}$, e^{Δ}, $e^{-\Delta}$ or $e^{-3\Delta}$. Therefore, from Equations (2) and (3), we obtain $\Delta = 0$. That is, the S model holds. The proof is completed.

Additionally, we obtain the following theorem.

Theorem 2.2. For fixed i, j and k $(1 \leq i < j < k \leq r)$, the S model holds, if and only if both the CDAS and PQS(i, j, k) models hold.

The proof of Theorem 2.2 is omitted because it is similar to Theorem 2.1.

For an $r \times r$ contingency table, let n_{ij} denote the observed frequency in the ith row and jth column of the table, where $n = \sum\sum n_{ij}$ and let m_{ij} denote the corresponding expected frequency $(i = 1, \ldots, r; j = 1, \ldots, r)$. We assumed that the

observed frequencies have a multinomial distribution. Let $G^2(M)$ denote the likelihood ratio chi-squared statistic, defined by

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

where $\hat{m}_{ij}$ is the maximum likelihood estimate of the expected frequency m_{ij} in model M. Under model M, this statistics has an asymptotically central chi-squared distribution with the corresponding degrees of freedom. The numbers of degrees of freedom for the S, QS, MH, CDAS, and PQS models are $r(r-1)/2$, $(r-1)(r-2)/2$, $r-1$, $(r-2)(r+1)/2$, and 1, respectively. We considered the maximum likelihood estimates of the expected frequencies $\{m_{ij}\}$ under the PQS model in the log-likelihood equation. For the PQS model, we must maximize the Lagrangian

$$L = \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}\log p_{ij} - \lambda\left(\sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij} - 1\right) - \psi\left(\prod\prod\prod_{1\le i<j<k\le r}\left(\frac{p_{ij}p_{jk}p_{ki}}{p_{ji}p_{kj}p_{ik}} - 1\right)\right),$$

with respect to $\{p_{ij}\}$, λ and ψ. By setting the partial derivations of L equal to zero using the Newton-Raphson method, we can obtain the maximum likelihood estimates of $\{m_{ij}\}$ under the PQS model.

Table 1

Membership and attitude toward the Leading Crowd for a sample of schoolgirls: from Coleman [8]. The upper and lower parenthesized values are the maximum likelihood estimates of the expected frequencies in the PQS and PQS$(1,2,4)$ models, respectively.

(M,A) for first interview	(M,A) for second interview				Total
	Yes, P	Yes, N	No, P	No, N	
Yes, P	484	93	107	32	716
	(484.00)	(91.35)	(107.00)	(33.65)	
	(484.00)	(91.35)	(107.00)	(33.65)	
Yes, N	112	110	30	46	298
	(113.65)	(110.00)	(30.00)	(44.35)	
	(113.65)	(110.00)	(30.00)	(44.35)	
No, P	129	40	768	321	1258
	(129.00)	(40.00)	(768.00)	(321.00)	
	(129.00)	(40.00)	(768.00)	(321.00)	
No, N	74	75	303	536	988
	(72.35)	(76.65)	(303.00)	(536.00)	
	(72.35)	(76.65)	(303.00)	(536.00)	
Total	799	318	1208	935	3260

M, membership; A, attitude; P, positive; N, negative

3 Example

We consider the data in Table 1, taken from Coleman [8]. A sample of schoolgirls was interviewed twice, several months apart, and asked about their self-perceived membership in the "Leading Crowd" and whether they sometimes needed to go against their principles to belong to that group. Thus, there are two binary response variables, which we refer to as membership and attitude, measured at two interview for each subject. Table 1 labels the categories for attitude as (positive, negative), where "positive" refers to disagreement with the statement that one must go against her principles. For details of the data in Table 1, see Agresti [9].

We are interested in whether there is a structure of symmetry in membership and attitude between the first and second interviews for the data in Table 1. As shown in Table 2, the S and QS models do not fit the data well. Therefore, in these data, there is no structure for the symmetry of cell probabilities and odds ratios. Because the MH model does not fit these data well, it is not possible to know whether the poor fit of the S model is due to the poor fit of the QS or MH model using Caussinus' [3] theorem.

The PQS model fits these data well, whereas the CDAS model does not fit these data well. From Theorem 2.1, we observe that the poor fit of the S model is caused by the lack of structure of the CDAS model rather than the PQS model.

The PQS(i,j,k) models for $(i,j,k)=(1,2,3),(1,2,4),(2,3,4)$ fit these data well, whereas the PQS$(1,3,4)$ model does not fit the data well. Therefore, the probability that the students' membership and attitude change from (yes, positive) to (no, positive), from (no, positive) to (no, negative), and from (no, negative) to (yes, positive) is not equal to the probability that it changes from (no, positive) to (yes, positive), from (no, negative) to (no, positive), and (yes, positive) to (no, negative) between the first and second interviews.

Table 2

The likelihood ratio chi-squared values G^2 for the models applied in Table 1.

Applied models	Degrees of freedom	G^2	p-value
S	6	29.90*	< 0.001
QS	3	8.81*	0.032
MH	3	21.09*	< 0.001
CDAS	5	15.48*	0.008
PQS	1	0.27	0.604
PQS(1,2,3)	1	0.87	0.352
PQS(1,2,4)	1	0.27	0.604
PQS(1,3,4)	1	7.61*	0.006
PQS(2,3,4)	1	0.67	0.413

* means significant at 0.05 level.

4 Conclusions

In this study, we proposed the PQS and PQS(i,j,k) models for any fixed $1 \le i < j < k \le r$ for square contingency tables with the same row and column classifications. For analyzing the data in square contingency table, we note that the PQS model is always identical to the best fitting PQS(i,j,k) among the PQS(i,j,k) model all $1 \le i < j < k \le r$. Namely, under the PQS model, the maximum likelihood estimates of the expected frequency are equal to those of the best-fitting PQS(i,j,k) model. For the data in Table 1, the best-fitting model based on the likelihood ratio chi-squared statistic was the PQS(1,2,4) model. Indeed, the maximum likelihood estimates of expected frequency under the PQS and PQS(1,2,4) are equal from the data in Table 1. The likelihood ratio chi-squared statistics for the two models ware also equal.

It is suitable to use the PQS and $\{\text{PQS}(i,j,k)\}$ models for analyzing square tables with nominal categories; however, it would not be suitable to use this model for tables with ordered categories when one wants to use information about category ordering. This is because these models are invariant under the same arbitrary permutation of row and column categories.

The decomposition of the S model into the CDAS and PQS models, given by Theorem 2.1, is useful for determining the reason for its poor fit when the S model fits the data poorly. Indeed, for the data in Table 1, the poor fit of the S model is caused by the poor fit of the CDAS model rather than that of the PQS model.

Acknowledgement

The authors would like to thank the referee for many comments and suggestions.

References

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

[2] Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, 282, (1975).

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

[4] R.A. Bradley and M.E. Terry, Rank analysis of incomplete block designs I: The method of paired comparisons, *Biometrika*, **39**, 324-345 (1952).

[5] K. Tahata, N. Miyamoto and S. Tomizawa, Measure of departure from quasi-symmetry and Bradley-Terry models for square contingency tables with nominal categories, *Journal of the Korean Statistical Society*, **33**, 129-147 (2004).
[6] A. Stuart, A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika*, **42**, 412-416 (1955).
[7] S. Tomizawa, N. Miyamoto and R. Funato, Conditional difference asymmetry model for square contingency tables with nominal categories, *Journal of Applied Statistics*, **31**, 271-277 (2004).
[8] J.S. Coleman, *Introduction to Mathematical Sociology*, Free Press of Glencoe, London, 168, (1964).
[9] A. Agresti, *Categorical Data Analysis (3rd ed)*, Wiley and Sons, New Jersey, 530, (2013).