Journal of Statistical Theory and Applications (2022) 21:21–30
https://doi.org/10.1007/s44199-022-00039-z

RESEARCH ARTICLE

Asymmetry Models Based on Non-integer Scores for Square Contingency Tables

Shuji Ando[1]

Received: 25 October 2021 / Accepted: 14 January 2022 / Published online: 25 January 2022

Abstract

Square contingency tables with ordinal classifications are used in many disciplines that include but are not limited to data science, engineering, and medical research. This study proposes two original asymmetry models based on non-integer scores for the analysis of square contingency tables. The ordinal quasi-symmetry model applies to data sets that can be assigned to known ordered scores for all categories. When we assign the equally spaced score for categories, the ordinal quasi-symmetry model is equivalent to the linear diagonals-symmetry model. The ordinal quasi-symmetry model, however, is not applicable to data sets that cannot be assigned the known ordered scores for all categories. This study addresses this issue. The proposed models apply to data sets that: (i) can be assigned the known ordered scores for all except one category and (ii) cannot be assigned the known ordered scores for all categories. These two models provide a better fit than existing models for real-world data.

Keywords Equally spaced score · Midpoint score · Open-ended categories · Power parameter score · Symmetry

Mathematics Subject Classification 62H17

Abbreviations

OQS	Ordinal quasi-symmetry
LDPS	Linear diagonals-parameter symmetry
S	Symmetry
RQS	Ridit score type quasi-symmetry
OEAS	Open-ended category type asymmetry
PPAS	Power parameter type asymmetry

✉ Shuji Ando
shuji.ando@rs.tus.ac.jp

[1] Department of Information and Computer Technology, Tokyo University of Science, 6-3-1 Niijuku, Katsushika-ku, Tokyo 1258585, Japan

MLE Maximum likelihood estimate
AIC Akaike information criterion

1 Introduction

Consider $R \times R$ square contingency tables with the same row and column ordinal classifications. Square contingency tables are used in many disciplines that include data science, engineering, and medical research, see, for example, Agresti [1].

Let s_k be the ordered score of category k for all $k = 1, \ldots, R$, where $s_1 < \cdots < s_R$. We consider the data sets that: (i) can be assigned the known ordered scores for all categories, (ii) can be assigned the known ordered scores for all except one category, and (iii) cannot be assigned the known ordered scores for all categories.

Typical examples of types (i) and (ii) are categorical variables set as intervals based on a continuous variable. When there is clear information about category intervals, it is recommended that ordered scores be assigned as midpoint intervals (midpoint scores) instead of equally spaced scores (see Graubard and Korn [2] and Senn [3]).

For the analysis of the data sets of type (i), the ordinal quasi-symmetry (OQS) model proposed by Agresti [1] is often used. The OQS model indicates the asymmetric structure of the cell probabilities with respect to the main-diagonals cell of the table. The OQS model assumes that the row category k and column category k are assigned the same known scores s_k for all $k = 1, \ldots, R$. This assumption is natural for square contingency tables with the same row and column ordinal classifications.

We consider the data set in Table 1, that presents the cross-classification of 1995 income data for espoused couples in Japan. Individual and spouse incomes are categorized as "less than 70", "from 70 to less than 150", "from 150 to less than 450", and "450 or more". We assign 35, 110, and 300 as the ordered scores for the first, second, and third categories, respectively. However, as the fourth category is unbounded above, we could not assign the known ordered score.

Gautam [4] suggests that the ordered scores for a data set with an open-ended category should be assigned as follows: the scores s_1 to s_{R-1} are midpoint scores, and the score s_R is unknown. Therefore, the score s_R can be expressed as $s_R = w_0 + w$; where w_0 is the smallest value of the interval for the open-ended category, and $w \geq 0$ is unknown. Gautam [4] assumes that the row category k and column category k are

Table 1 Cross-classification of 1995 income data (in units of 10,000 yen) for income of individuals (male) and their spouses (female) in Japan are derived from the Social Science Japan Data Archive (available at https://nesstar.iss.u-tokyo.ac.jp/webview/)

Individual income	Spouse income				Total
	(1)	(2)	(3)	(4)	
Less than 70 (1)	9	2	5	2	18
From 70 to less than 150 (2)	13	7	2	1	23
From 150 to less than 450 (3)	175	61	70	4	310
450 or more (4)	298	88	82	66	534
Total	495	158	159	73	885

assigned the same scores s_k for all $k = 1, \ldots, R$. Additionally, as the ordered scores for a data set with an open-ended category, Aktas and Wu [5] introduces the standardized z-scores for the row and column categories, and proposes a model that indicates the asymmetric structure depending on the standardized z-scores. However, the standardized z-scores assumes that the row category k and column category k are assigned the different scores s_k for all $k = 1, \ldots, R$. Therefore, for the analysis of the data sets of type (ii), we are interested in considering a model the asymmetric structure of the cell probabilities depending on the score proposed by Gautam [4].

We further consider the data set in Table 2 obtained from Agresti [1], that present the cross-classification of occupational status categories for father and son dyads in Britain. In this case, the scores s_1 to s_R are treated as unknown because it was difficult to assign an ordered known score to any category.

For the analysis of the data sets of type (iii), the ridit score type quasi-symmetry (RQS) model proposed by Iki et al. [6] is often used. The RQS model indicates the asymmetric structure of the cell probabilities depending on the ridit scores. The ridit scores is the unknown ordered scores, and is defined the average of row and column marginal ridits. Thus, the RQS model also assumes that the row category k and column category k are assigned the same unknown scores s_k for all $k = 1, \ldots, R$.

Bagheban and Zayeri [7] proposes a power parameter score. We modify the power parameter score so that it can be treated as the unknown ordered score, although Bagheban and Zayeri [7] is treated the power parameter score as the known ordered scores. Thus, for the analysis of the data sets of type (iii), we are interested in proposing a model the asymmetric structure of the cell probabilities depending on the power parameter score.

This study proposes two original asymmetry models based on non-integer scores for square contingency tables with the same row and column ordinal classifications. One of the proposed models is useful for the data set with open-ended categories–namely type (ii). The other is applicable to a data set that cannot be assigned the known ordered scores for all categories–namely type (iii).

Table 2 Cross-classification of occupational status categories for father and son dyads in Britain, obtained from Agresti [1]

Father status	Son status					Total
	(1)	(2)	(3)	(4)	(5)	
Highest (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
Lowest (5)	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

The remainder of this paper is organized as follows. Sect. 2 proposes two original models based on non-integer scores. Sect. 3 demonstrates the utility of these proposed models as applied to the real-world data presented in Tables 1 and 2. We conclude the paper in Sect. 4.

2 Proposed Models

2.1 Existing Models Based on Non-integer Scores

Let p_{ij} denote the probability that an observation will fall in the (i, j)th cell of the table $(i = 1, \dots, R; j = 1, \dots, R)$.

This study focuses on a model having the following formula:

$$p_{ij} = \delta^{s_j - s_i} p_{ji} \quad (i < j).$$

Note that s_k is the ordered score of category k for all $k = 1, \dots, R$, where $s_1 < \cdots < s_R$. This model can represent various models depending on how s_k is set.

As the model based on integer scores (i.e., $s_k = k$ for all $k = 1, \dots, R$), the linear diagonals-parameter symmetry (LDPS) model proposed by Agresti [8] was defined as

$$p_{ij} = \delta^{j-i} p_{ji} \quad (i < j).$$

We introduce existing models based on non-integer scores (i.e., $s_k \neq k$ for all $k = 1, \dots, R$). We can assign the known ordered scores s_k to the category k for all $k = 1, \dots, R$. The ordinal quasi-symmetry (OQS) model proposed by Agresti [1] was defined as

$$p_{ij} = \delta^{s_j - s_i} p_{ji} \quad (i < j).$$

Note that the OQS model corresponds to the type (i) data set. The OQS model with equally spaced scores ($s_k = s_1 + (k-1)d$ for $k = 1, \dots, R$) is equivalent to the LDPS model. The OQS model with $\delta = 1$ is also identical to the symmetry (S) model proposed by Bowker [9]. Kateri and Agresti [10] considered the OQS model based on f-divergence, also see Saigusa et al. [11].

Let X and Y denote the row and column variables, $p_{k\cdot} = \sum_{l=1}^{R} p_{kl}$ and $p_{\cdot k} = \sum_{l=1}^{R} p_{lk}$ for the marginal probabilities for $k = 1, \dots, R$. and $F_k^X = \sum_{l=1}^{k} p_{l\cdot}$ and $F_k^Y = \sum_{l=1}^{k} p_{\cdot l}$ for the marginal distribution functions for $k = 1, \dots, R$, where $F_R^X = 1$ and $F_R^Y = 1$. Then the marginal ridits are defined as,

$$r_k^X = \sum_{l=1}^{k-1} p_{l\cdot} + \frac{p_{k\cdot}}{2} \quad \text{and} \quad r_k^Y = \sum_{l=1}^{k-1} p_{\cdot l} + \frac{p_{\cdot k}}{2} \quad (k = 1, \dots, R),$$

see Bross [12].

When we cannot assign the known ordered scores s_k to the category k for all $k = 1, \dots, R$, we adopt the RQS model proposed by Iki et al. [6]:

$$p_{ij} = \delta^{s_j - s_i} p_{ji} \quad (i < j),$$

where $s_k = (r_k^X + r_k^Y)/2$ for all $k = 1, \ldots, R$. Note that $\{s_k\}$ in the RQS model are unspecified (i.e., the unknown ordered scores), and the RQS model corresponds to the data set of type (iii).

We highlight that a model corresponding to the data set of type (ii) does not exist in a similar form to the OQS and RQS models.

2.2 Proposed Models Based on Non-integer Scores

We propose two original models based on non-integer scores corresponding to the data set of types (ii) and (iii). First, we propose an original model corresponding to the data set of type (ii), defined as

$$p_{ij} = \delta^{s_j - s_i} p_{ji} \quad (i < j),$$

where s_1 to s_{R-1} are known, and s_R is unknown. Therefore, s_1 to s_{R-1} are assigned to known ordered scores (e.g., midpoint scores), s_R is defined as $s_R = w_0 + w$, where w_0 is the smallest value of the interval for the open-ended category, and $w\ (\geq 0)$ is unspecified. We refer this model as the open-ended category type asymmetry (OEAS) model.

Second, we consider a model corresponding to the data set of type (iii). Bagheban and Zayeri [7] consider a power parameter score as follows:

$$k^a \quad (k = 1, \ldots, R),$$

where $a > 0$. The power parameter score has the following properties:

(1) if $a < 1$ then the difference in scores between category $k + 1$ and k decreases as k increases;
(2) if $a > 1$ then the difference in scores between category $k + 1$ and k increases as k increases;
(3) if $a = 1$ then the power parameter score is equivalent to the equally spaced score.

Bagheban and Zayeri [7] treated a as known but did not discuss how to select the optimal value of a. In the OQS model, Ando [13] used the power parameter score as the known ordered scores, selected the optimal value of a by a grid search. In contrast, we propose the following original model treating a as unknown:

$$p_{ij} = \delta^{s_j - s_i} p_{ji} \quad (i < j),$$

where $s_k = k^a$ for $k = 1, \ldots, R$, and $a\ (> 0)$ are unknown. We refer to this model as the power parameter type asymmetry (PPAS) model. The PPAS model with $a = 1$ is identical to the LDPS model.

Under the OEAS and PPAS models, the following properties hold:

(1) if $\delta > 1$ then $F_k^X > F_k^Y$ for all $k = 1, \ldots, R-1$ because $p_{ij} > p_{ji}$ for all $i < j$;
(2) if $\delta < 1$ then $F_k^X < F_k^Y$ for all $k = 1, \ldots, R-1$ because $p_{ij} < p_{ji}$ for all $i < j$;
(3) if $\delta = 1$ then the S model holds because $p_{ij} = p_{ji}$ for all $i < j$.

For properties (1) and (2), the parameter δ in the OEAS or PPAS models infers whether X is stochastically greater than Y or vice versa.

2.3 Goodness-of-Fit Test

Let n_{ij} denote the observed frequency in the (i, j)th cell of the table $(i, j = 1, \ldots, R)$. Assume that a multinomial distribution applies to the $R \times R$ table. The maximum likelihood estimates of expected frequencies under the model can be obtained using the Newton–Raphson method in the log-likelihood equation.

Each model can be tested for goodness-of-fit by, the likelihood ratio and chi-square statistic (denoted by G^2) with the corresponding degrees of freedom. The test statistic G^2 of model M is given as

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

where $\hat{m}_{ij}$ is the maximum likelihood estimate (MLE) of the expected frequency m_{ij} under model M.

The number of degrees of freedom for both the OEAS and PPAS models are $(R^2 - R - 4)/2$. Note that the number of degrees of freedom for the OEAS and PPAS models is one less than that of the LDPS, OQS, and RQS models, and two less than the S model.

Applied economists often use the Akaike information criterion (AIC) as a quick method for choosing the best-fitting model among alternatives. The AIC is defined as

$$\text{AIC} = -2(\text{the maximum log likelihood}) + 2(\text{the number of parameters}),$$

for each model, see Akaike [14]. This criterion recommends a model with minimum AIC as the best-fitting model. When two models are compared, only the difference between AICs is required. It is therefore possible to ignore a common constant AIC, and use a modified AIC defined as

$$\text{AIC}^+ = G^2 - 2(\text{the number of degrees of freedom}).$$

Thus, the model with the minimum AIC^+ (i.e., the minimum AIC) is the best-fitting model among the applied models.

Table 3 The maximum likelihood estimates of expected frequencies under the open-ended category type asymmetry model applied to the data set in Table 1 are shown in parentheses in the second line

Individual income	Spouse income				Total
	(1)	(2)	(3)	(4)	
Less than 70 (1)	9	2	5	2	18
	(9.000)	(3.943)	(4.588)	(0.581)	
From 70 to less than 150 (2)	13	7	2	1	23
	(11.057)	(7.000)	(4.306)	(0.482)	
From 150 to less than 450 (3)	175	61	70	4	310
	(175.412)	(58.694)	(70.000)	(5.938)	
450 or more (4)	298	88	82	66	534
	(299.419)	(88.518)	(80.062)	(66.000)	
Total	495	158	159	73	885

Table 4 Values of the likelihood ratio, chi-square statistic (G^2) and the modified Akaike information criterion (AIC^+), for each model applied to the data are shown in Table 1

Models	Degrees of freedom	G^2	AIC^+
S	6	873.592*	861.592
LDPS	5	13.575*	3.575
OEAS†	4	6.467	−1.533

Note: The symbol ∗ implies significance at the 5% level, and † indicates a proposed model

3 Application to Real-World Data

3.1 Application to Income Data

We apply the S, LDPS, and OEAS models to the data set in Table 1. The ordered scores s_1, s_2, and s_3 of the OEAS model are assigned as 35, 110, and 300 respectively, and s_4 is assigned $450 + w$ ($w \geq 0$).

Table 3 shows the MLEs of the expected frequencies under the OEAS model. The goodness-of-fit results in Table 4 reveal that (1) the S and LDPS models fit poorly, (2) the OEAS model fits well, and (3) the OEAS model is significantly better compared to the LDPS model.

Under the OEAS model, the MLEs of δ and w are $\hat{\delta} = 0.986$ and $\hat{w} = 39.203$ respectively. Thus, the MLE of s_4 is $\hat{s}_4 = 489.203$. Since $s_2 - s_1 = 75$, $s_3 - s_2 = 190$, and $s_4 - s_3 = 189.203$, the $s_{k+1} - s_k$ for $k = 1, \ldots, R-1$ are unlikely to be constant. Since $\hat{\delta} < 1$, we infer that the male individuals' incomes tend to be higher than that of their female spouses' incomes.

3.2 Application to Occupational Status Data

We apply the S, LDPS, and PPAS models to the data set in Table 2.

Table 5 The maximum likelihood estimates of expected frequencies under the power parameter type asymmetry model applied to the data set in Table 2 are shown in parentheses in the second line

Father status	Son status					Total
	(1)	(2)	(3)	(4)	(5)	
Highest (1)	50	45	8	18	8	129
	(50.000)	(36.520)	(9.571)	(16.760)	(6.586)	
(2)	28	174	84	154	55	495
	(36.480)	(174.000)	(81.559)	(159.136)	(58.050)	
(3)	11	78	110	223	96	518
	(9.429)	(80.441)	(110.000)	(212.172)	(99.983)	
(4)	14	150	185	714	447	1510
	(15.240)	(144.864)	(195.828)	(714.000)	(441.549)	
Lowest (5)	3	42	72	320	411	848
	(4.414)	(38.950)	(68.017)	(325.451)	(411.000)	
Total	106	489	459	1429	1017	3500

Table 5 shows the MLEs of expected frequencies under the PPAS model. The results in Table 6 reveal that (1) the S and LDPS models fit poorly, (2) the RQS and PPAS models fit well, and (3) the PPAS model is preferred over the RQS model when values of AIC^+ are compared.

Table 6 Values of the likelihood ratio chi-square statistic (G^2) and the modified Akaike information criterion (AIC^+), for each model applied to the data are shown in Table 2

Models	Degrees of freedom	G^2	AIC^+
S	10	37.464^*	17.464
LDPS	9	17.126^*	−0.874
RQS	9	12.669	−5.331
PPAS†	8	8.080	−7.920

Note: The symbol $*$ implies significance at the 5% level, and † indicates a proposed model

Under the PPAS model, the MLEs of δ and a are $\hat{\delta} = 1.000013$ and $\hat{a} = 6.441$ respectively. As $\hat{a} > 1$, the difference in scores between $k+1$ and k increases as k increases. We provide evidence that, $s_2 - s_1 = 85.891$, $s_3 - s_2 = 1096.712$, $s_4 - s_3 = 6366.527$, and $s_5 - s_4 = 24230.730$. As $\hat{\delta} > 1$, we infer that the occupational statuses of fathers tend to be higher than those of their sons.

4 Conclusion

This study introduced three types of data set, namely those that: (i) can be assigned the known ordered scores for all categories, (ii) can be assigned the known ordered scores for all except one category, and (iii) cannot be assigned the known ordered scores for all categories. This study proposed two original asymmetry models based

on non-integer scores corresponding to data sets of types (ii) and (iii). The proposed models are simple asymmetry models, and therefore easier to apply and interpret. The findings demonstrate that the proposed models are applicable to real-world data.

Acknowledgements The author would like to thank the anonymous reviewers and the editors for their comments and suggestions to improve this paper. The data set in Table 1 used in this analysis–namely Social Stratification and Mobility (SSM95A)–was provided by the Social Science Japan Data Archive, Center for Social Research and Data Archives, Institute of Social Science, and the University of Tokyo.

Funding The authors have solely funded the research by themselves.

Data Availability The data set of Table 1 is available at https://nesstar.iss.u-tokyo.ac.jp/webview/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval Not applicable.

Consent for Publication Not applicable.

References

1. Agresti, A.: Categorical data analysis, 2nd edn. Wiley, New York (2002)
2. Graubard, B.I., Korn, E.L.: Choice of column scores for testing independence in ordered $2 \times k$ contingency tables. Biometrics **43**, 471–476 (1987)
3. Senn, S.: Drawbacks to noninteger scoring for ordered categorical data. Biometrics **63**, 296–299 (2007)
4. Gautam, S.: Test for linear trend in $2 \times k$ ordered tables with open-ended categories. Biometrics **53**, 1163–1169 (1997)
5. Aktas, S., Wu, S.: Marginal homogeneity model for ordered categories with open ends in square contingency tables. REVSTAT-Statistical Journal **13**(3), 233–243 (2015)
6. Iki, K., Tahata, K., Tomizawa, S.: Ridit score type quasi-symmetry and decomposition of symmetry for square contingency tables with ordered categories. Austrian Journal of Statistics **38**, 183–192 (2009)
7. Bagheban, A.A., Zayeri, F.: A generalization of the uniform association model for assessing rater agreement in ordinal scales. Journal of Applied Statistics **37**, 1265–1273 (2010)
8. Agresti, A.: A simple diagonals-parameter symmetry and quasi-symmetry model. Statistics & Probability Letters **1**, 313–316 (1983)
9. Bowker, A.H.: A test for symmetry in contingency tables. Journal of the American Statistical Association **43**, 572–574 (1948)
10. Kateri, M., Agresti, A.: A class of ordinal quasi-symmetry models for square contingency tables. Statistics & Probability Letters **77**(6), 598–603 (2007)

11. Saigusa, Y., Tahata, K., Tomizawa, S.: Orthogonal decomposition of symmetry model using the ordinal quasi-symmetry model based on f-divergence for square contingency tables. Statistics & Probability Letters **101**, 33–37 (2015)
12. Bross, I.D.J.: How to use ridit analysis. Biometrics **14**, 18–38 (1958)
13. Ando, S.: Asymmetry models based on ordered score and separations of symmetry model for square contingency tables. Biometrical Letters **58**(1), 27–39 (2021)
14. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control AC**19**, 716–723 (1983)