Analysis of Stock Prices using Symmetry Models for Multi-way Contingency Tables

Kiyotaka Iki[1]

1. Introduction

For analysis of contingency tables, we are interested in whether the two classificatory variables are independent of one another. When the independent variable does not hold, we may use Pearson's correlation coefficient to estimate the correlation between two variables. Also, it is important to interpret the data, and to propose models that fit the data well. Goodman (1979a) considered the uniform association model, and Agresti (1983a) considered the linear-by-linear association model. In addition, various extended association models have been proposed, for instance, Liu and Agresti (2005) and Ando et al. (2021).

In particular, we consider tables with the same row and column classifications. Such the table is called square contingency tables. For square contingency tables, the independence between the row and column is unlikely to hold because many observations fall in the main diagonal cells, which indicates that the value of the row category is the same as the value of the column category. Therefore, for the analysis of square contingency tables, instead of independence, we are interested in whether or not the row variable is symmetric with the column variable. As models of symmetry, the symmetry model (Bowker, 1948), the marginal homogeneity model (Stuart, 1955), the quasi-symmetry model (Caussinus, 1965), and the quasi-diagonal exponent symmetry model (Iki et al., 2014) have been proposed. Further, as models of asymmetry, the conditional symmetry model (McCullagh, 1978), the diagonals-parameter symmetry model (Goodman, 1979b), the linear diagonals-parameter symmetry model (Agresti, 1983b) and the log-normal type symmetry model (Iki and Tomizawa, 2018) have been proposed.

In particular, the r^T tables are constructed from a qualitative variable with r values observed $T\ (\geq 3)$ times. This table is called a multi-way contingency table. For the analysis of multi-way contingency tables, as models of symmetry, the symmetry model (Bhapkar and Darroch, 1990), the marginal cumulative logistic model (Yoshimoto et al., 2020) and the marginal continuation odds ratio model (Shinoda et al., 2020) have been proposed.

Many researchers have analyzed the stock prices of companies listed in the 1st section of the Tokyo Stock Exchange. In this study, we are interested in representing stock price data in a contingency table. Let x_1, x_2, x_3 and x_4 denote the closing stock prices at the 1st, 2nd, 3rd and 4th week in a month, respectively. In addition, x_0, x_{-1}... x_{-8} indicate the corresponding past closing stock prices. Let α and β denote the minimum and maximum closing stock prices in that period, respectively:

[1] E-mail: iki.kiyotaka@nihon-u.ac.jp

$$\alpha = \min\{x_{-8}, \ldots, x_4\}, \quad \beta = \max\{x_{-8}, \ldots, x_4\}.$$

Let X_i denote

$$X_i = \begin{cases} 1 \left(\frac{1}{3}\alpha + \frac{2}{3}\beta < x_i \leq \beta\right) \\ 2 \left(\frac{2}{3}\alpha + \frac{1}{3}\beta < x_i \leq \frac{1}{3}\alpha + \frac{2}{3}\beta\right) \\ 3 \left(\alpha < x_i \leq \frac{2}{3}\alpha + \frac{1}{3}\beta\right) \end{cases} \quad (i = 1, 2, 3, 4).$$

The variables $\{X_i\}$ are classified into three categories (levels); (1) high, (2) middle and (3) low. The data in Table 1 are constructed from the stock prices of the 2009 companies listed in the 1st section of the Tokyo Stock Exchange in August 2016. These data are a $3 \times 3 \times 3 \times 3$ (namely, the 3^4) contingency table. We are interested in investigating the types of symmetry structure that exists for the data in Table 1 and propose new models of symmetry with respect to time transition.

The new models are introduced in Section 2. Section 3 presents the goodness-of-fit test, and Section 4 provides an example. Finally, Section 5 concludes the paper.

Table 1

The stock prices of the 2009 companies listed in the 1st section of the Tokyo Stock Exchange in August 2016. The parenthesized values are the maximum likelihood estimates of the expected frequencies under the M3D model.

X_1	X_3	1			2			3		
	$X_2 \backslash X_4$	1	2	3	1	2	3	1	2	3
1	1	152	29	0	21	63	3	1	2	2
		(152.0)	(28.4)	(0.0)	(21.0)	(61.6)	(2.9)	(1.0)	(2.0)	(2.0)
	2	3	1	0	4	15	5	1	4	15
		(3.0)	(1.0)	(0.0)	(4.0)	(14.7)	(4.9)	(1.0)	(3.9)	(14.7)
	3	1	0	0	0	1	0	0	1	7
		(1.0)	(0.0)	(0.0)	(0.0)	(1.0)	(0.0)	(0.0)	(1.0)	(6.8)
2	1	141	35	0	25	84	8	0	4	11
		(144.2)	(35.0)	(0.0)	(25.6)	(84.0)	(7.8)	(0.0)	(4.0)	(10.8)
	2	39	14	0	36	220	41	1	40	169
		(39.9)	(14.0)	(0.0)	(36.8)	(220.0)	(40.1)	(1.0)	(40.0)	(165.4)
	3	0	0	0	2	3	2	0	6	64
		(0.0)	(0.0)	(0.0)	(2.0)	(3.0)	(2.0)	(0.0)	(6.0)	(62.6)
3	1	13	1	0	2	9	3	0	1	2
		(13.3)	(1.0)	(0.0)	(2.0)	(9.2)	(3.0)	(0.0)	(1.0)	(2.0)
	2	11	7	0	9	61	29	2	24	150
		(11.2)	(7.2)	(0.0)	(9.2)	(62.4)	(29.0)	(2.0)	(24.5)	(150.0)
	3	3	0	0	3	7	11	2	25	358
		(3.1)	(0.0)	(0.0)	(3.1)	(7.2)	(11.0)	(2.0)	(25.6)	(358.0)

2. Models

In this section, for a $3\times3\times3\times3$ contingency table with ordered categories, we propose seven types of models for symmetry.

Let X_1, X_2, X_3 and X_4 denote the first, second, third, and fourth variables, respectively, and have time ordering. Also, let Pr $(X_1=i, X_2=j, X_3=k, X_4=l)=p_{ijkl}$ $(1\leq i, j, k, l\leq 3)$. We propose a model defined by

$$p_{ijkl}=p_{lkji} \quad (i\leq j\leq k\leq l; i<l).$$

This indicates that the probability of an observation falling in the (i, j, k, l)th cell is equal to the probability of the observation falling in the (l, k, j, i)th cell, under the condition that i, j, k and l are increase monotonically. This describes a structure of symmetry of the probability with respect to time transition only in the cell where the value of the category increases or decreases monotonically. We refer to this model as the M1 model.

Next, we propose a model defined by

$$\sum\sum\sum\sum_{\substack{i\leq j\leq k\leq l\\ l-i=t}} p_{ijkl}=\sum\sum\sum\sum_{\substack{i\geq j\geq k\geq l\\ i-l=t}} p_{ijkl} \quad (t=1, 2).$$

This indicates that the probability of the difference between the value of the fourth variable and the first variable, namely $l-i$, is t on the condition that i, j, k and l are increase monotonically, and is equal to the probability of the difference between the values of the first and fourth variables, namely $i-l$, being t $(t=1, 2)$. We refer to this model as the M2 model, and note that the M1 model implies the M2 model.

Moreover, we propose a model defined by

$$\sum\sum\sum\sum_{\substack{i\leq j\leq k\leq l\\ i<l}} p_{ijkl}=\sum\sum\sum\sum_{\substack{i\geq j\geq k\geq l\\ i>l}} p_{ijkl}.$$

This indicates that the probability of the difference between the value of the fourth variable and the first variable is positive when i, j, k and l are increase monotonically, and is equal to the probability of the difference between the value of the fourth variable and the first variable being negative. We refer to this model as the M3 model. The M2 model implies the M3 model.

We further propose a model defined by

$$p_{ijkl}=p_{lkji} \quad (i<l).$$

This indicates that the probability that an observation falling in the (i, j, k, l)th cell is equal to the probability of the observation falling in the (l, k, j, i)th cell on the condition that the difference between $l-i$ is positive. This describes a structure of symmetry of the probability with respect to the time transition only in the cell where the difference between $l-i$ is positive. We refer to this model as the M1D model. The M1D model implies the M1 model.

Next, we propose a model defined by

$$\sum\sum\sum\sum_{l-i=t} p_{ijkl}=\sum\sum\sum\sum_{i-l=t} p_{ijkl} \quad (t=1, 2).$$

This indicates that the probability of the difference between the value of the fourth variable and the first variable is t, and is equal to the probability of the difference between the values of the first and fourth variables being

t $(t=1,2)$. We refer to this model as the M2D model. It should be noted that the M1D model implies the M2D model.

Moreover, we propose a model defined by

$$\sum\sum_{i<l}\sum\sum p_{ijkl}=\sum\sum_{i>l}\sum\sum p_{ijkl}.$$

This indicates that the probability of the difference between the value of the fourth variable and the first variable is positive, and is equal to the probability of the difference between the value of the fourth variable and the first variable being negative. We refer to this model as the M3D model. It should be noted that the M2D model implies the M3D model.

Finally,

$$p_{ijkl}=p_{lkji} \quad (1\leq i,j,k,l\leq 3).$$

This indicates that the probability of an observation falling in the (i,j,k,l)th cell is equal to the probability of the observation falling in the (l,k,j,i)th cell. This describes the structure of symmetry of the probability with respect to the time transition. We refer to this model as the M1DD model. We note that the M1DD model implies the M1D model.

Figure 1 shows the relationships among the models. In the figure, $A\rightarrow B$ indicates that model A implies model B.

Figure 1: Relationships among models.

$$\begin{tikzcd}
M1DD \arrow[d] & & \\
M1D \arrow[r] \arrow[d] & M2D \arrow[r] & M3D \\
M1 \arrow[r] & M2 \arrow[r] & M3
\end{tikzcd}$$

3. Goodness-of-fit Test

For a $3\times3\times3\times3$ contingency table, let n_{ijkl} denote the observed frequency in the (i,j,k,l)th cell of the table, where $n=\sum\sum\sum\sum n_{ijkl}$ and let m_{ijkl} denote the corresponding expected frequency $(1\leq i,j,k,l\leq 3)$. We assume that the observed frequencies have a multinomial distribution. The maximum likelihood estimates of the expected frequencies under each model are expressed in closed-forms. Each model can be tested for the goodness-of-fit using, for example, the likelihood ratio chi-squared statistic (denoted by G^2) with the corresponding degrees of freedom. The test statistic G^2 for model M is given by

$$G^2(M)=2\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3}n_{ijkl}\log\left(\frac{n_{ijkl}}{\hat{m}_{ijkl}}\right),$$

where $\hat{m}_{ijkl}$ is the maximum likelihood estimate of the expected frequency m_{ijkl} under model M. Under model M, these statistics have an asymptotic central chi-squared distribution with corresponding degrees of freedom. The number of degrees of freedom for the M1DD, M1D, M1, M2D, M2, M3D and M3 are 36, 27, 12, 2, 2, 1 and 1, respectively.

Akaike (1974) information criterion (AIC) is used to choose the preferable model among different models that include non-nested models. For details see Konishi and Kitagawa (2008). Because only the difference between AIC's is required when two models are compared, it is possible to ignore a common constant of AIC. We may use a modified AIC defined by

$$\mathrm{AIC}^{+} = G^2 - 2(\text{number of df}).$$

Thus, for the given data, the model with the minimum AIC^{+} is the preferable model. This criterion will be used in the next section.

4. An Example

We consider the data in Table 1 again. Table 2 shows that the M1DD, M1D and M1 models fit the data poorly, whereas the M2D, M2, M3D and M3 models fit the data well. Because these models include non-nested models, we use AIC^{+} to choose the preferable model. Because the M3D model has a minimum AIC^{+} value, the M3D model is the most preferred model.

Table 2

Values of the likelihood ratio chi-squared statistic for the models applied to Table 1.

Applied models	Degrees of freedom	Likelihood ratio chi-square statistic	AIC^{+}
M1DD	36	693.4*	621.4
M1D	27	428.1*	374.1
M1	12	337.6*	313.6
M2D	2	3.5	−0.5
M2	2	8.6	4.6
M3D	1	0.4	−1.6(min)
M3	1	5.2	3.2

* significant at the 0.05 level.
(min) means a minimum AIC^{+}

Under the M3D model, the probability that closing stock prices in the 1st week in August 2016 is better than that in the 4th week is equal to the probability that closing stock prices in the 1st week in August 2016 is worse than that in the 4th week. Since $\sum\sum\sum\sum_{i>l} \hat{m}_{ijkl} = 433.5$, it is evident that the stock price of 21% (= 433.5/2009) of the companies listed in the 1st section of the Tokyo Stock Exchange rose during that period. In addition, because $\sum\sum\sum\sum_{i<l} \hat{m}_{ijkl} = 433.5$, it can be observed that the stock price of 21% of the companies fell during the same period. Therefore, the percentage of companies whose stock prices have fallen and the percentage of companies whose stock prices have risen are considered to be approximately the same. It can also be seen that the stock price of 57% (=(2009−867)/2009) of the companies does not fluctuate during this period.

5. Concluding Remarks

In Section 1, we define α and β as a function of the values of x_{-8}, ..., x_4, which represent 13 weeks (about 3 months) of the stock prices. However, the same could be applied to a different number of weeks or months as well.

The models proposed in this paper may be applicable not only to stock price data but also to data from other fields.

References

Agresti, A (1983a) “A survey of strategies for modeling cross-classifications having ordinal variables,” *Journal of the American Statistical Association*, Vol.**78**, pp.184-198.

——— (1983b) “A simple diagonals-parameter symmetry and quasi-symmetry model,” *Statistics and Probability Letters*, Vol.**1**, pp.313-316.

Akaike, H (1974) “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, Vol.**AC-19**, pp.716-723.

Ando, S., K. Tahata and S. Tomizawa (2021) “Extended diagonal uniform association symmetry model for square contingency tables with ordinal categories,” *Statistics and Probability Letters*, Vol.**169**, pp.1-4.

Bhapkar, V. P. and J. N. Darroch (1990) “Marginal symmetry and quasi symmetry of general order,” *Journal of Multivariate Analysis*, Vol.**34**, pp.173-184.

Bowker, A. H (1948) “A test for symmetry in contingency tables,” *Journal of the American Statistical Association*, Vol.**43**, pp.572-574.

Caussinus, H (1965) “Contribution à l’analyse statistique des tableaux de corrélation,” *Annales de la Faculté des Sciences de l’Université de Toulouse*, Vol.**29**, pp.77-182.

Goodman, L. A (1979a) “Simple models for the analysis of association in cross-classifications having ordered categories,” *Journal of the American Statistical Association*, Vol.**74**, pp.537-552.

——— (1979b) “Multiplicative models for square contingency tables with ordered categories,” *Biometrika*, Vol.**66**, pp.413-418.

Iki, K., K. Yamamoto and S. Tomizawa (2014) “Quasi-diagonal exponent symmetry model for square contingency tables with ordered categories,” *Statistics and Probability Letters*, Vol.**92**, pp.33-38.

——— (2018) “Log-normal distribution type symmetry model for square contingency tables with ordered categories,” *Austrian Journal of Statistics*, Vol.**47**, pp.39-48.

Konishi, S. and G. Kitagawa (2008) *Information Criteria and Statistical Modeling*, New York: Springer.

Liu, I., and A. Agresti (2005) “The analysis of ordered categorical data: An overview and a survey of recent developments,” *Sociedad Española de Estadística e Investigación Operativa Test*, Vol.**14**, pp.1-73.

McCullagh, P (1978) “A class of parametric models for the analysis of square contingency tables with ordered categories,” *Biometrika*, Vol.**65**, pp.413-418.

Shinoda, S., K. Tahata, K. Yamamoto and S. Tomizawa (2020) “Marginal continuation odds ratio model and decomposition of marginal homogeneity model for multi-way contingency tables,” *Sankhya: The Indian Journal of Statistics, Ser. B*, published online.

Stuart, A (1955) “A test for homogeneity of the marginal distributions in a two-way classification,” *Biometrika*, Vol.**42**, pp.412-416.

Yoshimoto, T., K. Tahata and S. Tomizawa (2020) “Marginal cumulative logistic model of general order for multi-way contingency tables,” *Austrian Journal of Statistics*, Vol.**49**, pp.53-61.