

Improved Measure on Extended Marginal Homogeneity for Ordinal Square Contingency Tables

Kouji Yamamoto¹, Ryota Shinjo² & Sadao Tomizawa²

¹ Department of Medical Innovation, Osaka University Hospital, Yamadaoka, Suita, Osaka, Japan

² Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Yamazaki, Noda City, Chiba, Japan

Correspondence: Kouji Yamamoto, Department of Medical Innovation, Osaka University Hospital, Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: yamamoto-k@hp-crc.med.osaka-u.ac.jp

Received: October 7, 2012 Accepted: January 15, 2013 Online Published: January 22, 2013

doi:10.5539/ijsp.v2n1p108 URL: <http://dx.doi.org/10.5539/ijsp.v2n1p108>

Abstract

For square contingency tables with ordered categories, Yamamoto et al. (2007) considered a measure to represent the degree of departure from extended marginal homogeneity. It attains the maximum value when one of two symmetric cumulative probabilities is zero. The present paper proposes an improved measure so that the degree of departure from extended marginal homogeneity can attain the maximum value even when the cumulative probabilities are not zeros. An example is given.

Keywords: marginal homogeneity, measure, Patil-Taillie diversity index, Shannon entropy

1. Introduction

For the $R \times R$ square contingency table, let π_{ij} denote the probability that an observation will fall in cell (i, j) ($i = 1, \dots, R; j = 1, \dots, R$). The marginal homogeneity (MH) model is defined by

$$\pi_i = \pi_{\cdot i} \quad (i = 1, \dots, R),$$

where $\pi_{i\cdot} = \sum_{k=1}^R \pi_{ik}$ and $\pi_{\cdot i} = \sum_{k=1}^R \pi_{ki}$ (Stuart, 1955; Bishop et al., 1975, p. 294). Let

$$H_{1(i)} = \sum_{s=1}^i \sum_{t=i+1}^R \pi_{st}, \quad H_{2(i)} = \sum_{s=i+1}^R \sum_{t=1}^i \pi_{st},$$

for $i = 1, \dots, R-1$. This model may be expressed as

$$H_{1(i)} = H_{2(i)} \quad (i = 1, \dots, R-1).$$

This states that the cumulative probability that an observation will fall in row category i or below and column category $i+1$ or above is equal to the cumulative probability that the observation falls in column category i or below and row category $i+1$ or above for $i = 1, \dots, R-1$.

Tomizawa (1984, 1995) considered the extended marginal homogeneity (EMH) model which is expressed as

$$H_{1(i)} = \delta H_{2(i)} \quad (i = 1, \dots, R-1).$$

When $\delta = 1$, this is the MH model. Let

$$H_1 = \sum_{i=1}^{R-1} H_{1(i)}, \quad H_2 = \sum_{i=1}^{R-1} H_{2(i)}.$$

Assume that $\{H_{1(i)} + H_{2(i)} > 0\}$, $H_1 > 0$, and $H_2 > 0$. The EMH model may also be expressed as

$$Q_{1(i)} = Q_{2(i)} \quad (i = 1, \dots, R-1),$$

where

$$Q_{1(i)} = \frac{H_{1(i)}^*}{H_{1(i)}^* + H_{2(i)}^*}, \quad Q_{2(i)} = \frac{H_{2(i)}^*}{H_{1(i)}^* + H_{2(i)}^*},$$

$$H_{1(i)}^* = \frac{H_{1(i)}}{H_1}, \quad H_{2(i)}^* = \frac{H_{2(i)}}{H_2}.$$

This indicates that there is a structure of symmetry between $\{Q_{1(i)}, Q_{2(i)}\}$. Yamamoto et al. (2007) considered a measure to represent the degree of departure from EMH, using Patil and Taillie (1982) diversity index. The measure ranges between 0 and 1, and the degree of departure from EMH is maximum when $Q_{1(i)} = 0$ or $Q_{2(i)} = 0$ for all $i = 1, \dots, R-1$. [Note that for measures for other models, e.g., the symmetry model (Bowker, 1948) and the MH model, see (e.g., Tomizawa et al., 2001; Tahata et al., 2006; Tahata et al., 2009)].

However, for analyzing square contingency tables, all $Q_{1(i)}$ and $Q_{2(i)}$ ($i = 1, \dots, R-1$) are positive in many cases. Thus, then Yamamoto et al. (2007) measure cannot attain the maximum value. So, we are now interested in a measure to represent the degree of departure from EMH such that it can attain the maximum value even when each of $\{Q_{1(i)}\}$ and $\{Q_{2(i)}\}$ is not zero.

For square contingency tables with ordered categories, the present paper proposes such a measure on EMH when all cumulative probabilities are positive.

2. New Measure

Let

$$E_i = \frac{H_{1(i)}^* + H_{2(i)}^*}{2} \quad (i = 1, \dots, R-1).$$

For a specified d with $0.5 < d \leq 1$ and $1-d \leq Q_{1(i)} \leq d$ ($i = 1, \dots, R-1$), define the new measure as, for $\lambda (> -1)$ fixed,

$$\Omega = \frac{1}{K} \left(1 - \frac{\lambda 2^\lambda}{2^\lambda - 1} \sum_{i=1}^{R-1} E_i W_i \right),$$

where

$$K = 1 - \frac{\lambda 2^\lambda}{2^\lambda - 1} L,$$

$$L = \frac{1}{\lambda} (1 - d^{\lambda+1} - (1-d)^{\lambda+1}),$$

$$W_i = \frac{1}{\lambda} (1 - Q_{1(i)}^{\lambda+1} - Q_{2(i)}^{\lambda+1}),$$

and the value at $\lambda = 0$ is taken to be continuous limit as $\lambda \rightarrow 0$. Thus, when $\lambda = 0$,

$$\Omega = \frac{1}{K} \left(1 - \frac{1}{\log 2} \sum_{i=1}^{R-1} E_i W_i \right),$$

where

$$K = 1 - \frac{1}{\log 2} L,$$

$$L = -d \log d - (1-d) \log(1-d),$$

$$W_i = -Q_{1(i)} \log Q_{1(i)} - Q_{2(i)} \log Q_{2(i)}.$$

Note that W_i is Patil-Taillie diversity index including Shannon entropy (when $\lambda = 0$). A value of d is chosen by the user such that $1-d \leq Q_{1(i)} \leq d$ for any $i = 1, \dots, R-1$. When $d = 1$, the measure Ω is identical to Yamamoto et al. (2007) measure. [Although the detail is omitted, note that Ω can also be expressed by using the power-divergence.]

Then, we can obtain the following theorem:

Theorem 1 For each λ and a fixed d ,

- (i) $0 \leq \Omega \leq 1$,
- (ii) $\Omega = 0$ if and only if the EMH model holds,

(iii) $\Omega = 1$ if and only if the degree of departure from EMH is the largest in the sense that $Q_{1(i)} = d$ or $Q_{2(i)} = d$ for all $i = 1, \dots, R-1$.

Proof. When $d = 1$, for each λ , the minimum value of W_i is 0 when $Q_{1(i)} = 0$ or $Q_{2(i)} = 0$ for all $i = 1, \dots, R-1$, and the maximum value of it is $(2^\lambda - 1)/(\lambda 2^\lambda)$ (if $\lambda \neq 0$) or $\log 2$ (if $\lambda = 0$), when $Q_{1(i)} = Q_{2(i)} = 1/2$ for all $i = 1, \dots, R-1$. When $d \neq 1$, the minimum value of it is L , which is not equal to 0, and the maximum value of it is the same as $d = 1$. Thus, the measure Ω lies between 0 and 1. So the proof is completed.

We note that the measure Ω is the modified measure of Yamamoto et al. (2007) by using a coefficient $1/K$.

Consider the artificial 4×4 table data in Table 1a on cell probabilities $\{p_{ij}\}$. Then, we see the degree of departure from EMH by using the existing measure Ω with $d = 1$ (i.e., Yamamoto et al. measure) and the measure Ω with $d < 1$ (in this case we set $d = 0.9$). We see from Table 1b that the true value of Ω with $d = 1$ is 0.531 (when $\lambda = 0$), and that of Ω with $d = 0.9$ is 1 (when $\lambda = 0$). Thus, we can see that the new measure Ω with $d < 1$ attains the maximum value 1, though all cumulative probabilities are positive.

Table 1. (a) An artificial 4×4 table data on cell probabilities $\{p_{ij}\}$, and (b) the values of measure Ω with $d = 1$ (existing measure) and Ω with $d = 0.9$ (new measure) applied to Table 1a

(a) Artificial data				
	(1)	(2)	(3)	(4)
(1)	0.2	0.00025	0.00025	0.0005
(2)	0.003	0.2	0.089	0.00025
(3)	0.003	0.001	0.2	0.00825
(4)	0.003	0.003	0.075	0.2135

(b) Value of the existing measure and new measure

Existing measure	New measure
0.531	1

3. Asymptotic Variance for Estimated Measure

Let n_{ij} denote the observed frequency in cell (i, j) ($i = 1, \dots, R; j = 1, \dots, R$). Assuming a multinomial distribution, the estimated measure $\hat{\Omega}$ is given by Ω with $\{\pi_{ij}\}$ replaced by $\{\hat{\pi}_{ij}\}$, where $\hat{\pi}_{ij} = n_{ij}/n$ and $n = \sum \sum n_{ij}$. Using the delta method, $\hat{\Omega}$ has asymptotically (as $n \rightarrow \infty$) a normal distribution with mean Ω and variance

$$\sigma^2 = \frac{1}{nK^2} \sum_{k=1}^{R-1} \sum_{l=k+1}^R [\pi_{kl}(v_{1(kl)})^2 + \pi_{lk}(v_{2(kl)})^2],$$

where for $\lambda \neq 0$,

$$v_{s(kl)} = \frac{2^\lambda}{2(2^\lambda - 1)H_s} \left[\sum_{i=k}^{l-1} \tau_{s(i)} - (l-k) \sum_{i=1}^{R-1} H_{s(i)}^* \tau_{s(i)} \right] \quad (s = 1, 2),$$

with

$$\begin{aligned} \tau_{1(i)} &= (Q_{1(i)})^\lambda + \lambda \{ (Q_{1(i)})^\lambda - (Q_{2(i)})^\lambda \} Q_{2(i)}, \\ \tau_{2(i)} &= (Q_{2(i)})^\lambda + \lambda \{ (Q_{2(i)})^\lambda - (Q_{1(i)})^\lambda \} Q_{1(i)}, \end{aligned}$$

and for $\lambda = 0$,

$$v_{s(kl)} = \frac{1}{2H_s(\log 2)} \left[\sum_{i=k}^{l-1} \log Q_{s(i)} - (l-k) \sum_{i=1}^{R-1} H_{s(i)}^* \log Q_{s(i)} \right] \quad (s = 1, 2).$$

Let $\hat{\sigma}^2$ denote σ^2 with $\{\pi_{ij}\}$ replaced by $\{\hat{\pi}_{ij}\}$. Using these, the approximate confidence interval for the measure Ω is obtained as follows:

$$\hat{\Omega} \pm Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}},$$

where $Z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile of the standard normal distribution.

4. An Example

Consider the data in Table 2, taken from Hattori et al. (2002, p. 244). These data describe the cross-classification of father's and son's occupational status categories in Japan which were examined in 1955 and in 1975.

Table 2. Occupational status for Japanese father-son pairs (from Hattori et al., 2002, p. 244)

(a) Examined in 1955					
Father's status	Son's status				Total
	(1)	(2)	(3)	(4)	
(1)	59	41	18	13	131
(2)	45	136	70	27	278
(3)	25	75	236	43	379
(4)	62	131	212	686	1091
Total	191	383	536	769	1879

(b) Examined in 1975					
Father's status	Son's status				Total
	(1)	(2)	(3)	(4)	
(1)	127	101	54	12	294
(2)	86	207	125	13	431
(3)	78	124	310	24	536
(4)	109	206	437	325	1077
Total	400	638	926	374	2338

Note: (1) is Upper White-collar; (2) Lower White-collar; (3) Blue-collar and (4) Farming.

It seems natural to assume that all cumulative probabilities are positive because any observations can fall in all cells of the table. Therefore, it may not be appropriate to use the measure Ω with $d = 1$ because there is not a structure of cumulative probabilities such that Ω with $d = 1$ attains the maximum value 1. So we should use Ω with $d < 1$ (for example, $d = 0.99$) so that the measure can attain the maximum value 1.

Since the confidence intervals for Ω with $d = 0.99$ applied to the data in each of Tables 2a and 2b, do not include zero for all λ (see Table 3), these would indicate that there is not a structure of EMH in neither of tables.

Table 3. When $d = 0.99$, the estimate of Ω , estimated approximate standard error (S.E.) for $\hat{\Omega}$, and approximate 95% confidence interval (C.I.) for Ω , applied to Tables 2a and 2b

	λ	$\hat{\Omega}$	S.E.	C.I.
For Table 2a	-0.5	0.023	0.007	(0.010, 0.036)
	0.0	0.033	0.009	(0.014, 0.051)
	0.5	0.039	0.011	(0.018, 0.061)
	1.0	0.043	0.012	(0.019, 0.067)
	1.5	0.044	0.012	(0.020, 0.068)
	2.0	0.043	0.012	(0.019, 0.067)
	2.5	0.041	0.012	(0.018, 0.063)
For Table 2b	-0.5	0.105	0.012	(0.080, 0.129)
	0.0	0.141	0.016	(0.110, 0.172)
	0.5	0.165	0.017	(0.131, 0.199)
	1.0	0.177	0.018	(0.141, 0.213)
	1.5	0.180	0.018	(0.144, 0.216)
	2.0	0.177	0.018	(0.141, 0.213)
	2.5	0.170	0.018	(0.135, 0.205)

Moreover, we compare the degree of departure from EMH in Tables 2a and 2b using the confidence intervals for Ω . For any λ , the values in the confidence interval for Ω applied to the data in Table 2b are greater than those

applied to the data in Table 2a. In addition, the values in the confidence interval do not overlap for Table 2a and for Table 2b. Thus, the degree of departure from EMH is greater for Table 2b than for Table 2a.

5. Concluding Remarks

We have proposed Ω which is an improvement of Yamamoto et al. (2007) measure (i.e., Ω with $d = 1$) to represent the degree of departure from EMH. For analyzing the data of square table such that all cumulative probabilities are positive, it may not be adequate to use the measure Ω with $d = 1$ because then the measure cannot attain the maximum value 1. For such data, it would be natural to use the measure Ω with $d < 1$ because then the measure can attain maximum value 1 even when all cumulative probabilities are positive.

The analyst may also be interested in how the value of d is determined. However it seems difficult to discuss this. The measure Ω depends on the value of a fixed d . Also, the value of Ω increases as the value of d decreases. But when we compare several tables, the result of comparisons is invariant without depending on the value of d . For analyzing a square table data, we note that if $1 - d \leq Q_{1(i)} \leq d$ is not satisfied for all $i = 1, \dots, R - 1$, the measure Ω cannot be used for the given data. Thus, the analyst must set the value of d carefully, so as to satisfy the condition $1 - d \leq Q_{1(i)} \leq d$ for all $i = 1, \dots, R - 1$. Therefore we recommend a value being close to 1 (for example, $d = 0.99$) as the value of d .

Acknowledgments

The authors would like to thank the editor and two anonymous reviewers for the helpful comments and suggestions.

References

- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: The MIT Press.
- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43, 572-574. <http://dx.doi.org/10.1080/01621459.1948.10483284>
- Hattori, T., Funatsu, T., & Torii, T. (2002). *Ajia Chukanso no Seisei to Tokushitsu (The Emergence and Features of the Asian Middle Classes)*. The Institute of Developing Economies, Chiba, Japan (in Japanese).
- Patil, G. P., & Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77, 548-561. <http://dx.doi.org/10.1080/01621459.1982.10477845>
- Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412-416. <http://dx.doi.org/10.1093/biomet/42.3-4.412>
- Tahata, K., Iwashita, T., & Tomizawa, S. (2006). Measure of departure from symmetry of cumulative marginal probabilities for square contingency tables with ordered categories. *SUT Journal of Mathematics*, 42, 7-29. Retrieved from http://www3.ma.kagu.tus.ac.jp/sutjmath/_userdata/42-1/02-tomizawa.pdf
- Tahata, K., Yamamoto, K., Yamada, A., & Tomizawa, S. (2009). Generalized measures of departure from symmetry for square contingency tables. *Behaviormetrika*, 36, 75-86. <http://dx.doi.org/10.2333/bhmk.36.75>
- Tomizawa, S. (1984). Three kinds of decompositions for the conditional symmetry model in a square contingency table. *Journal of the Japan Statistical Society*, 14, 35-42. Retrieved from <http://www.jss.gr.jp/ja/journal/jss1984.html>
- Tomizawa, S. (1995). A generalization of the marginal homogeneity model for square contingency tables with ordered categories. *Journal of Educational and Behavioral Statistics*, 20, 349-360. <http://dx.doi.org/10.3102/10769986020004349>
- Tomizawa, S., Miyamoto, N., & Hatanaka, Y. (2001). Measure of asymmetry for square contingency tables having ordered categories. *Australian and New Zealand Journal of Statistics*, 43, 335-349. <http://dx.doi.org/10.1111/1467-842X.00180>
- Yamamoto, K., Furuya, Y., & Tomizawa, S. (2007). Measure of departure from extended marginal homogeneity for square contingency tables with ordered categories. *REVSTAT: Statistical Journal*, 5, 269-283. Retrieved from <http://www.ine.pt/revstat/pdf/rs070303.pdf>