On measure of departure from uniform association based on concordant and discordant pairs in contingency tables

Tahata, Kouji
Tokyo University of Science, Department of Information Sciences
2641, Yamazaki
Noda City (2788510), Japan
E-mail: kouji_tahata@is.noda.tus.ac.jp

Tomizawa, Sadao
Tokyo University of Science, Department of Information Sciences
2641, Yamazaki
Noda City (2788510), Japan
E-mail: tomizawa@is.noda.tus.ac.jp

For a contingency table with ordered categories, Goodman (1979) considered the uniform association model that is the extension of null association (independence) model. This model indicates that the local odds ratios for adjacent rows and adjacent columns have common value. Also, the uniform association model may be expressed as the structure that the conditional probability that a randomly selected adjacent pair of observations is in (i, j) and in $(i+1, j+1)$ given that a randomly selected adjacent pair is concordant, is equal to the conditional probability that a randomly selected adjacent pair of observations is in $(i, j+1)$ and in $(i+1, j)$ given that a randomly selected adjacent pair is discordant.

When the uniform association model does not hold, we are interested in measuring the degree of departure from the equality of these conditional probabilities, which is also the degree of departure from the uniform association model. The purpose of present paper is to propose a measure, which is a generalization of the measure proposed by Tahata and Tomizawa (2011), to represent the degree of departure from uniform association.

Measure for uniform association

For an $I \times J$ cross-classification table with ordered categories, let p_{ij} denote the probability that an observation will fall in the ith row and jth column of the table ($i = 1, \ldots, I; j = 1, \ldots, J$). Let

$$C^* = \sum_{s=1}^{I-1}\sum_{t=1}^{J-1} p_{st}p_{s+1,t+1} \quad \text{and} \quad D^* = \sum_{s=1}^{I-1}\sum_{t=1}^{J-1} p_{s,t+1}p_{s+1,t}.$$

Thus, $2C^*$ ($2D^*$) indicates the probability of concordance (discordance) for a randomly selected adjacent pair of observations. Also let

$$c_{ij} = \frac{p_{ij}p_{i+1,j+1}}{C^*} \quad \text{and} \quad d_{ij} = \frac{p_{i,j+1}p_{i+1,j}}{D^*}$$

for $i = 1, \ldots, I-1; j = 1, \ldots, J-1$. Thus, c_{ij} indicates the conditional probability that a randomly selected adjacent pair of observations is in (i, j) and in $(i+1, j+1)$ given that a randomly selected adjacent pair is concordant, and d_{ij} indicates the conditional probability that a randomly selected adjacent pair of observations is in $(i, j+1)$ and in $(i+1, j)$ given that a randomly selected adjacent pair is discordant. Using $\{c_{ij}\}$ and $\{d_{ij}\}$, the uniform association model (Goodman, 1979) may be expressed as

$$c_{ij} = d_{ij} \quad \text{for} \quad i = 1, \ldots, I-1; j = 1, \ldots, J-1.$$

Assuming that $C^* \neq 0$, $D^* \neq 0$ and $\{c_{ij} + d_{ij} \neq 0\}$, consider a measure defined by, for $\lambda > -1$

$$\Phi^{(\lambda)} = \frac{\lambda(\lambda+1)}{2(2^\lambda - 1)} \left[I^{(\lambda)} \left(\{c_{ij}\}; \left\{ \frac{c_{ij}+d_{ij}}{2} \right\} \right) + I^{(\lambda)} \left(\{d_{ij}\}; \left\{ \frac{c_{ij}+d_{ij}}{2} \right\} \right) \right],$$

where

$$I^{(\lambda)} \left(\{a_{ij}\}; \{b_{ij}\} \right) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} a_{ij} \left[\left(\frac{a_{ij}}{b_{ij}} \right)^\lambda - 1 \right]$$

and the value at $\lambda = 0$ is taken to be the limit as $\lambda \to 0$. Note that $I^{(\lambda)}(\{a_{ij}\}; \{b_{ij}\})$ is the power-divergence between two arbitrary probability distributions $\{a_{ij}\}$ and $\{b_{ij}\}$, and the value of λ is a real value that is chosen by the user. For more details of the power-divergence, see Read and Cressie (1988, p.15). When $\lambda = 0$, the measure $\Phi^{(0)}$ is the same as the measure proposed by Tahata and Tomizawa (2011).

Let $c^*_{ij} = c_{ij}/(c_{ij} + d_{ij})$ and $d^*_{ij} = d_{ij}/(c_{ij} + d_{ij})$ for $i = 1, \ldots, I-1; j = 1, \ldots, J-1$. Then the measure $\Phi^{(\lambda)}$ may be expressed as, for $\lambda > -1$

$$\Phi^{(\lambda)} = 1 - \frac{\lambda 2^\lambda}{2(2^\lambda - 1)} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} (c_{ij} + d_{ij}) H^{(\lambda)}_{ij} \left(\{c^*_{ij}, d^*_{ij}\} \right),$$

where

$$H^{(\lambda)}_{ij} \left(\{c^*_{ij}, d^*_{ij}\} \right) = \frac{1}{\lambda} \left(1 - (c^*_{ij})^{\lambda+1} - (d^*_{ij})^{\lambda+1} \right)$$

and the value at $\lambda = 0$ is taken to be the limit as $\lambda \to 0$. Note that $H^{(\lambda)}_{ij}(\{c^*_{ij}, d^*_{ij}\})$ is the diversity index of degree λ (Patil and Taillie, 1982). We point out that for each λ, the minimum value of $H^{(\lambda)}_{ij}(\{c^*_{ij}, d^*_{ij}\})$ is 0 when $c^*_{ij} = 0$ (then $d^*_{ij} = 1$) or $d^*_{ij} = 0$ (then $c^*_{ij} = 1$), and the maximum value of it is $(2^\lambda - 1)/(\lambda 2^\lambda)$ for $\lambda \neq 0$ and $\log 2$ for $\lambda = 0$ when $c^*_{ij} = d^*_{ij}$. Therefore $\Phi^{(\lambda)}$ must lie between 0 and 1. Also, for each λ (> -1), (i) c_{ij} equals d_{ij} for all $i = 1, \ldots, I-1; j = 1, \ldots, J-1$ (i.e., there is the structure of uniform association in the $I \times J$ table) if and only if $\Phi^{(\lambda)}$ equals zero, and (ii) the degree of departure from uniform association is the largest, in the sense that $c_{ij} = 0$ (then $d_{ij} \neq 0$) or $d_{ij} = 0$ (then $c_{ij} \neq 0$) for all $i = 1, \ldots, I-1; j = 1, \ldots, J-1$, if and only if $\Phi^{(\lambda)}$ equals one. Consider $(I-1)(J-1)$ sub-tables constructed from adjacent rows and adjacent columns. When $\Phi^{(\lambda)} = 0$, the probability of concordance equals that of discordance for all sub-tables. On the other hand, when $\Phi^{(\lambda)} = 1$, the probability of concordance equals zero for some sub-tables and that of discordance equals zero for the others. Namely, the maximum departure from uniform association indicates that there is no concordant pair (or there is no discordant pair) for each sub-table. Since the structure of uniform association is equivalent to $c_{ij} = d_{ij}$ for $i = 1, \ldots, I-1; j = 1, \ldots, J-1$, this definition of maximum departure from uniform association would be natural.

Approximate confidence interval for measure

Let n_{ij} denote the observed frequency in the ith row and jth column of a table ($i = 1, \ldots, I; j = 1, \ldots, J$). Assume that a multinomial distribution applies to the $I \times J$ table. We consider an approximate standard error and large-sample confidence interval for $\Phi^{(\lambda)}$ using the delta method, of which description is given by, e.g., Bishop, Fienberg and Holland (1975, Sec. 14.6). The sample version of $\Phi^{(\lambda)}$ (denoted by $\hat{\Phi}^{(\lambda)}$), is given by $\Phi^{(\lambda)}$ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$, where $\hat{p}_{ij} = n_{ij}/n$ and $n = \sum\sum n_{ij}$. Using the delta method, $\sqrt{n}(\hat{\Phi}^{(\lambda)} - \Phi^{(\lambda)})$ has asymptotically (as $n \to \infty$) a normal distribution with mean zero and variance

$$\sigma^2 = \sum_{s=1}^{I} \sum_{t=1}^{J} \left(\frac{\partial \Phi^{(\lambda)}}{\partial p_{st}} \right)^2 p_{st} - \left(\sum_{s=1}^{I} \sum_{t=1}^{J} \left(\frac{\partial \Phi^{(\lambda)}}{\partial p_{st}} \right) p_{st} \right)^2,$$

where

$$\frac{\partial \Phi^{(\lambda)}}{\partial p_{st}} = \frac{2^{\lambda-1}}{2^{\lambda}-1} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} \left[\left(\frac{\partial c_{ij}}{\partial p_{st}} \right) (c_{ij}^*)^{\lambda} + \left(\frac{\partial d_{ij}}{\partial p_{st}} \right) (d_{ij}^*)^{\lambda} + \lambda \left((c_{ij}^*)^{\lambda} - (d_{ij}^*)^{\lambda} \right) \left(\left(\frac{\partial c_{ij}}{\partial p_{st}} \right) d_{ij}^* - \left(\frac{\partial d_{ij}}{\partial p_{st}} \right) c_{ij}^* \right) \right],$$

$$\frac{\partial c_{ij}}{\partial p_{st}} = \frac{1}{C^*} \left[\delta_{ij;st} p_{i+1,j+1} + \delta_{i+1,j+1;st} p_{ij} - c_{ij}(p_{s-1,t-1} + p_{s+1,t+1}) \right],$$

$$\frac{\partial d_{ij}}{\partial p_{st}} = \frac{1}{D^*} \left[\delta_{i,j+1;st} p_{i+1,j} + \delta_{i+1,j;st} p_{i,j+1} - d_{ij}(p_{s-1,t+1} + p_{s+1,t-1}) \right],$$

$$\delta_{ij;st} = \begin{cases} 1 & \text{when} \quad (i,j) = (s,t), \\ 0 & \text{when} \quad (i,j) \neq (s,t), \end{cases}$$

$$p_{00} = p_{0t} = p_{s0} = p_{s,J+1} = p_{I+1,t} = p_{I+1,J+1} = 0.$$

We note that the asymptotic distribution of $\sqrt{n}(\hat{\Phi}^{(\lambda)} - \Phi^{(\lambda)})$ is not applicable when $\Phi^{(\lambda)} = 0$ and $\Phi^{(\lambda)} = 1$ because then $\sigma^2 = 0$. Let $\hat{\sigma}^2$ denote σ^2 with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$. Then $\hat{\sigma}/\sqrt{n}$ is an estimated standard error for $\hat{\Phi}^{(\lambda)}$. An approximate $100(1-p)$ percent confidence interval for $\Phi^{(\lambda)}$ is given by $\hat{\Phi}^{(\lambda)} \pm z_{p/2}\hat{\sigma}/\sqrt{n}$, where $z_{p/2}$ is the percentage point from the standard normal distribution corresponding to a two-tail probability equal to p.

An example

Table 1. (a) Happiness and relative family income; taken from Agresti (2010, p.21) and (b) happiness and number of sex partners; taken from Agresti (2010, p.202)

(a) Happiness and relative family income

	Happiness			
Family Income	Very Happy	Pretty Happy	Not Too Happy	Total
Above average	272	294	49	615
Average	454	835	131	1420
Below average	185	527	208	920
Total	911	1656	388	2955

(b) Happiness and number of sex partners

	Happiness			
No. sex partners	Very Happy	Pretty Happy	Not Too Happy	Total
≥ 2	57	198	57	312
1	535	832	118	1485
0	154	329	112	595
Total	746	1359	287	2392

Table 1a shows cross-classification of family income and happiness, taken from Agresti (2010, p.21). We denote the likelihood ratio chi-square test statistic for testing goodness-of-fit of the uniform association model by G^2. The uniform association model fits these data poorly, yielding $G^2 = 15.6$ with 3 degrees of freedom. Thus we should apply the measure $\Phi^{(\lambda)}$. For instance we may set $\lambda = 1.0$.

From Table 2a, the degree of departure from uniform association is estimated to be 1.54 percent of the maximum degree of departure from uniform association.

Table 1b shows cross-classification of the number of sex partners and happiness, taken from Agresti (2010, p.202). Since the uniform association model fits these data poorly, yielding $G^2 = 89.8$ with 3 degrees of freedom, we apply the measure for these data. If we set $\lambda = 1.0$, the degree of departure from uniform association is estimated to be 9.05 percent of the maximum degree of departure from uniform association.

In addition, we can describe from comparison between the confidence intervals for $\Phi^{(\lambda)}$ that the degree of departure from uniform association is stronger in Table 1b than in Table 1a.

Table 2. Estimate of $\Phi^{(\lambda)}$, estimated approximate standard error for $\hat{\Phi}^{(\lambda)}$, and approximate 95% confidence interval for $\Phi^{(\lambda)}$, applied to Tables 1a and 1b

(a) For Table 1a

λ	Estimated measure	Standard error	Confidence interval
-0.6	0.0055	0.0029	$(-0.0002, 0.0113)$
0	0.0112	0.0059	$(-0.0003, 0.0227)$
0.6	0.0144	0.0075	$(-0.0002, 0.0290)$
1	0.0154	0.0080	$(-0.0002, 0.0311)$

(b) For Table 1b

λ	Estimated measure	Standard error	Confidence interval
-0.6	0.0337	0.0071	(0.0197, 0.0477)
0	0.0669	0.0138	(0.0399, 0.0940)
0.6	0.0848	0.0172	(0.0511, 0.1185)
1	0.0905	0.0182	(0.0548, 0.1262)

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data,* Second Edition. Wiley, New York.

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

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

Patil, G. P. and Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, **77**, 548-561.

Read, T. R. C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, New York.

Tahata, K. and Tomizawa, S. (2011). Measure for uniform association based on concordant and discordant pairs for cross-classifications. *Journal of Statistical Theory and Practice*, to appear.