行政院國家科學委員會專題研究計畫成果報告

使用 J 發散統計量衡量列聯表的對稱性

計畫編號：NSC 89-2118-M-034-002

執行期間：89 年 8 月 1 日至 90 年 7 月 31 日

計畫主持人：洪文良

處理方式：☑可立即對外提供參考
（請打√） ☐一年後可對外提供參考
☐兩年後可對外提供參考
（必要時，本會得展延發表時限）

執行單位：中國文化大學應用數學系

中華民國 九十 年 七 月 三十一 日

中文摘要

關鍵詞：J發散統計量，列聯表，對稱性。

Taneichi (1993)考慮多項分配的適合度檢定問題，提出J發散統計量並且探討其大樣本性質。因此，本計畫的目的是：根據Taneichi (1993)的想法，提出J發散型態度量(J-divergence-type measure)來衡量列聯表的對稱性。

英文摘要

Keywords:contingency table, J-divergence-type discrepancy, symmetry.

Taneichi (1993) considered the J-divergence statistics for tseting the goodness-of-fit of the multinomial distribution, which includes the likelihood ratio is special case. Therefore, we are interested in a measure of departure from symmetry about the main diagonal of square contingency tables having the same nomial row and column classifications, based on the J-divergence type of discrepancy. The purpose of this project is to propose a J-divergence-type measure which represents the degree of departure from symmetry for square contingency tables. The measure would be useful for comparing the degree of departure from symmetry in several tables.

1. Introduction

Consider an $R \times R$ contingency table that has the same nominal row and column classifications. For the analysis of such tables, interest would be in the symmetry about the main diagonal of the table rather than the independence between the row and column variables. Let p_{ij} denote the probability that an observation will fall in the ith row and jth column of the table ($i = 1, 2, \ldots, R; j = 1, 2, \ldots, R$). The usual symmetry model is defined as

$$p_{ij} = p_{ji} \text{ for } i = 1, 2, \ldots, R; j = 1, 2, \ldots, R; i \neq j.$$

See Bowker (1948), Bishop *et al.* (1975, p.282) and Agresti (1984, p.202; 1990, p.353).

Assuming that $\{p_{ij} + p_{ji}\}$ for $i \neq j$ are all positive, Tomizawa (1994) proposed two kinds of measures to represent the degree of departure from symmetry. The two kinds of measures are defined in the population form by

$$\phi_s = \frac{1}{\delta \log 2} \sum_{i \neq j} p_{ij} \log \frac{2p_{ij}}{p_{ij} + p_{ji}},$$

$$\psi_s = \frac{1}{\delta} \sum_{i<j} \frac{(p_{ij} - p_{ji})^2}{p_{ij} + p_{ji}},$$

where $\delta = \sum_{i \neq j} p_{ij}$ and $0 \log 0 = 0$. Let $p^*_{ij} = p_{ij}/\delta$ and $p^s_{ij} = (p^*_{ij} + p^*_{ji})/2$ for $i = 1, 2, \ldots, R; j = 1, 2, \ldots, R; i \neq j$. Then ϕ_s and ψ_s may be also expressed as

$$\phi_s = \frac{I(p^*; p^s)}{\log 2}, \ \psi_s = D(p^*; p^s),$$

where

$$I(p^*; p^s) = \sum_{i \neq j} p^*_{ij} \log \frac{p^*_{ij}}{p^s_{ij}}, \ D(p^*; p^s) = \sum_{i \neq j} \frac{(p^*_{ij} - p^s_{ij})^2}{p^*_{ij}}.$$

Note that $I(p^*; p^s)$ and $D(p^*; p^s)$ are the Kullback-Leibler information and the Pearson's chi-square type discrepancy, respectively, between $\{p^*_{ij}\}_{i\neq j}$ and $\{p^s_{ij}\}_{i\neq j}$. The sample version of these measures are also expressed by modifying the likelihood ratio and the Pearson chi-squared statistics for testing the goodness of fit of the symmetry model (see Tomizawa (1994) for details).

Furthermore, Tomizawa (1998) *et al.* proposed a power-divergence-type measure which represents the degree of departure from symmetry for square contingency tables. The power-divergence-type measure is defined by

$$\Phi^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^\lambda - 1} I^{(\lambda)}(\{p^*_{ij}\}; \{p^s_{ij}\}), \text{ for } \lambda > -1, \tag{1}$$

where

$$I^{(\lambda)}(\cdot;\cdot) = \frac{1}{\lambda(\lambda+1)} \sum_{i\neq j} p^*_{ij}[(\frac{p^*_{ij}}{p^s_{ij}})^\lambda - 1]$$

and the value at $\lambda = 0$ is taken to be the limit as $\lambda \to 0$. Note that $\Phi^{(0)}$ and $\Phi^{(1)}$ in equation (1) are the same ϕ_s and ψ_s respectively. (For more details of the power divergence $I^{(\lambda)}(\cdot;\cdot)$, see Cressie and Read (1984) and Read and Cressie (1988).)

Taneichi (1993) considered the J-divergence statistics for testing goodness-of-fit of the multinomial distribution, which includes the likelihood ratio statistics in special case. Therefore, we are interested in a measure of departure from symmetry, based on the J-divergence type of discrepancy.

The purpose of this project is to propose a J-divergence-type measure which represents the degree of departure from symmetry for square contingency tables. The measure would be useful for comparing the degree of departure from symmetry in several tables.

2. Measures of departure from symmetry

Assume that $\{p_{ij} + p_{ji}\}$ for $i \neq j$ are all positive. Let

$$\delta = \sum_{i \neq j} p_{ij}$$

$p^*_{ij} = p_{ij}/\delta$ and $p^s_{ij} = (p^*_{ij} + p^*_{ji})/2$ for $i = 1, 2, \ldots, R$; $j = 1, 2, \ldots, R$; $i \neq j$. Note that p^*_{ij} indicates the probability that an observation falls in cell (i, j), on the condition that the observation will fall in one of the off-diagonal cells of a square table; p^s_{ij} indicates half the probability that an observation will fall in one of the off-diagonal cells of a square table; and $p^*_{ij} = p^s_{ij}$ for all i and j if and only if the symmetry model hold (i.e. p^s_{ij} indicates the probability that an observation falls in cell (i, j), on the condition that the observation will fall in one of the off-diagonal cells of a square table when the symmetry model holds).

The J-divergence type measure is defined by

$$\Psi^{(\alpha)} = C(\alpha) J^{(\alpha)}(\{p^*_{ij}\}; \{p^s_{ij}\}), \text{ for } \alpha \geq 1, \tag{2}$$

where

$$C(\alpha) = \frac{-(\alpha - 1)}{(3/4)^\alpha + (1/4)^\alpha - (1/2)^\alpha - 1/2},$$

$$J^{(\alpha)}(\{a_{ij}\}; \{b_{ij}\}) = \frac{-1}{\alpha - 1} \sum_{i<j} \{(\frac{a_{ij} + b_{ij}}{2})^\alpha - \frac{1}{2}[(a_{ij})^\alpha + (b_{ij})^\alpha]\},$$

and the value at $\alpha = 1$ is taken to be the limit as $\alpha \to 1$. Thus,

$$\Psi^{(1)} = (\frac{-3}{4} \log \frac{3}{4})^{-1} J^{(1)}(\{p^*_{ij}\}; \{p^s_{ij}\})$$

where

$$J^{(1)}(\{a_{ij}\}; \{b_{ij}\}) = \frac{-1}{2} \sum_{i<j} [(a_{ij} + b_{ij}) \log(\frac{a_{ij} + b_{ij}}{2}) - a_{ij} \log a_{ij} - b_{ij} \log b_{ij}].$$

Note that $J^{(\alpha)}(\{a_{ij}\};\{b_{ij}\})$ is the J-divergence between two distributions $\{a_{ij}\}$ and $\{b_{ij}\}$.

Let $p_{ij}^c = p_{ij}/(p_{ij}+p_{ji})$ for $i=1,2,\ldots,R;\ j=1,2,\ldots,R;\ i\neq j$. Note that p_{ij}^c indicates the probability that an observation falls in cell (i,j), on the condition that the observation will fall in cell (i,j) or (j,i) of the $R\times R$ table; and $p_{ij}^c = 1/2$ for all i and j if and only if the symmetry model holds. Then $\Psi^{(\alpha)}$ may be expressed as

$$\Psi^{(\alpha)} = C(\alpha)\sum_{i<j}(p_{ij}^* + p_{ji}^*)^\alpha J_{ij}^{(\alpha)}(\{p_{ij}^c, p_{ji}^c\};\{\frac{1}{2},\frac{1}{2}\}) \text{ for } \alpha \geq 1,$$

where

$$J_{ij}^{(\alpha)}(\cdot;\cdot) = \frac{-1}{\alpha-1}\{(\frac{p_{ij}^c+1/2}{2})^\alpha - \frac{1}{2}[(p_{ij}^c)^\alpha + (\frac{1}{2})^\alpha] + (\frac{p_{ji}^c+1/2}{2})^\alpha - \frac{1}{2}[(p_{ji}^c)^\alpha + (\frac{1}{2})^\alpha]$$

and the value at $\alpha=1$ is taken to be the limit as $\alpha\to 1$. Thus,

$$\Psi^{(1)} = (\frac{-3}{4}\log\frac{3}{4})^{-1}\sum_{i<j}(p_{ij}^* + p_{ji}^*)^\alpha J_{ij}^{(1)}(\{p_{ij}^c, p_{ji}^c\};\{\frac{1}{2},\frac{1}{2}\}),$$

where

$$\begin{aligned} J_{ij}^{(1)}(\cdot;\cdot) \;=\; & \frac{-1}{2}[(p_{ij}^c+\frac{1}{2})\log(\frac{p_{ij}^c+1/2}{2}) - p_{ij}^c\log p_{ij}^c - \frac{1}{2}\log\frac{1}{2} + \\ & (p_{ji}^c+\frac{1}{2})\log(\frac{p_{ji}^c+1/2}{2}) - p_{ji}^c\log p_{ji}^c - \frac{1}{2}\log\frac{1}{2}]. \end{aligned}$$

We see that the measure $\Psi^{(\alpha)}$ must lie between 0 and 1. Also, for each α (≥ 1), (i) there is a structure of symmetry in the $R\times R$ table if and only if $\Psi^{(\alpha)}=0$, and (ii) the degree of departure from symmetry is the largest in the sense that $p_{ij}^c = 1$ (then $p_{ji}^c = 0$) or $p_{ji}^c = 1$ (then $p_{ij}^c = 0$) for $i=1,2,\cdots,R;\ j=1,2,\cdots,R;\ i\neq j$; if and only if $\Psi^{(\alpha)}=1$. According to the weight sum of J-divergence, $\Psi^{(\alpha)}$ represents the degree of departure from symmetry, and the degree increases as the value of $\Psi^{(\alpha)}$ increases.

3. Measures for each pair of symmetric cells

When the symmetry model does not hold, we are interested in finding which pair of cells (i,j) and (j,i) (for $1 \leq i < j \leq R$) contribute most to the degree of departure from symmetry. The larger the dimension of the table becomes, the more important this may be. For the pair of cells (i,j) and (j,i) (for $1 \leq i < j \leq R$), assuming that $p_{ij} + p_{ji} > 0$, consider the measure defined by

$$\Phi_{ij}^{(\alpha)} = C(\alpha) J_{ij}^{(\alpha)}(\{p_{ij}^c, p_{ji}^c\}; \{\frac{1}{2}, \frac{1}{2}\}), \text{ for } \alpha \geq 1$$

where the value at $\lambda = 1$ is taken to be the limit as $\lambda \to 1$.

4. Approximate confidence intervals for measures

Let n_{ij} denote the observed frequency in the ith row and jth column of the table ($i = 1, 2, \ldots, R; j = 1, 2, \ldots, R$). Assuming that a multinomial distribution applies to the $R \times R$ table, we shall consider an approximate standard error and large-sample confidence interval for each measure, say Ψ, using the delta method, descriptions of which are given by Bishop *et al.* (1975, Section 14.6) and Agresti (1984, p.185, Appendix C). The sample version of Ψ, i.e., $\hat{\Psi}$, is given by Ψ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$, where $\hat{p}_{ij} = n_{ij}/n$ and $n = \sum n_{ij}$. Using the delta method, $\sqrt{n}(\hat{\Psi} - \Psi)$ has asymptotically (as $n \to \infty$) a normal distribution with mean zero and variance $\sigma^2(\Psi)$. (See Appendix for the values of variance for the measures $\Psi^{(\alpha)}$ and $\Psi_{ij}^{(\alpha)}$.) Let $\hat{\sigma}^2(\Psi)$ denote $\sigma^2(\Psi)$ with $\{p_{ij}\}$ replace by $\{\hat{p}_{ij}\}$. Then, $\hat{\sigma}(\Psi)/n^{1/2}$ is an estimated standard error for $\hat{\Psi}$, and $\hat{\Psi} \pm z_{p/2}\hat{\sigma}(\Psi)/n^{1/2}$ is an approximate $100(1-p)\%$ confidence interval for Ψ, where $z_{p/2}$ is the percentage point from the standard normal distribution that corresponds to a two-tail probability equal to p.

Appendix

Using the delta method, $n^{1/2}(\hat{\Psi}-\Psi)$ has asymptotical variance $\sigma^2(\Psi)$ as follows:

$$\sigma^2(\Psi^{(\alpha)}) = \frac{1}{\delta^2}\{\sum_{i\neq j} p_{ij}(\Delta_{ij}^{(\alpha)})^2 - \delta(\Psi^{(\alpha)})^2\}, \text{ for } \alpha > 1$$

where

$$\begin{aligned}\Delta_{ij}^{(\alpha)} &= \frac{1}{(\alpha-1)C(\alpha)}\{[\frac{\alpha}{2^\alpha}(p_{ij}^c+\frac{1}{2})^{\alpha-1} - \frac{\alpha}{2}(p_{ij}^c)^{\alpha-1} - 1 \\ &\quad +\alpha p_{ij}^c[\frac{\alpha}{2^\alpha}(p_{ij}^c+\frac{1}{2})^{\alpha-1} - \frac{\alpha}{2^\alpha}(p_{ji}^c+\frac{1}{2})^{\alpha-1} - \frac{\alpha}{2}(p_{ij}^c)^{\alpha-1} + \frac{\alpha}{2}(p_{ji}^c)^{\alpha-1}]\}\end{aligned}$$

Also, for $1 \le i < j \le R$, we have

$$\sigma^2(\Psi_{ij}^{(\alpha)}) = \begin{cases} [(\alpha-1)C(\alpha)]^2\frac{p_{ij}^c p_{ji}^c}{p_{ij}+p_{ji}}[\frac{\alpha}{2^\alpha}(p_{ij}^c+\frac{1}{2})^{\alpha-1} - \frac{\alpha}{2^\alpha}(p_{ji}^c+\frac{1}{2})^{\alpha-1} \\ -\frac{\alpha}{2}(p_{ij}^c)^{\alpha-1} + \frac{\alpha}{2}(p_{ji}^c)^{\alpha-1}]^2 & \text{for } \alpha > 1\ ; \\ [\frac{1}{\frac{3}{4}\log\frac{3}{4}}]^2\frac{p_{ij}^c p_{ji}^c}{p_{ij}+p_{ji}}\{\frac{1}{2}[\log(p_{ij}^c+\frac{1}{2}) - \log(p_{ji}^c+\frac{1}{2}) \\ -\log p_{ij}^c + \log p_{ji}^c]\}^2 & \text{for } \alpha = 1. \end{cases}$$

Note that $\sigma^2(\Psi^{(0)})$ and $\sigma^2(\Psi_{ij}^{(0)})$ are equal to the limits of $\sigma^2(\Psi^{(\alpha)})$ and $\sigma^2(\Psi_{ij}^{(\alpha)})$, respectively, as $\alpha \to 1$.

References

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* John Wiley, New York.

[2] Agresti, A. (1990). *Categorical Data Analysis.* John Wiley, New York.

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

[4] Bowker, A.H. (1948). A test for symmetry in contingency tables. *J. Amer. Statist. Assoc.* **43**, 572-574.

[5] Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Royal. Statist. Soc. B* **46**, 440-464.

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

[7] Taneichi, N. (1993). Goodness-of-fit based on the J-divergence for the multinomial distribution. *J. Japan. Statist. Soc.* **23**, 9-18.

[8] Tomizawa, S. (1994). Two kinds of measures of departure from symmetry in square contingency tables having nominal categories. *Statistica Sinica* **4**, 325-334.

[9] Tomizawa, S., Seo T. and Yamamoto, H. (1998). Power-divergence-type measure of departure from symmetry for square contingency tables that have nominal categories. *J. Appl. Statis.* **25**, 387-398.