

Index for Proportional Reduction in Error in Two-way Contingency Tables with Ordinal Categories

Kouji Yamamoto¹, Eri Yoshida² & Sadao Tomizawa²

¹ Center for Clinical Investigation and Research, Osaka University Hospital, 2-15, Yamadaoka, Suita, Osaka 565-0871, Japan

² Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Noda City, Chiba 278-8510, Japan

Correspondence: Kouji Yamamoto, Center for Clinical Investigation and Research, Osaka University Hospital, 2-15, Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: yamamoto-k@hp-crc.med.osaka-u.ac.jp

Received: May 1, 2012 Accepted: May 13, 2012 Online Published: July 10, 2012

doi:10.5539/jmr.v4n4p40 URL: <http://dx.doi.org/10.5539/jmr.v4n4p40>

Abstract

For two-way contingency tables with nominal categories in which the explanatory and response variables are not defined clearly, Goodman and Kruskal (1954) considered a proportional reduction in error (PRE) measure, which describes the relative decrease in the probability of making an error in predicting the value of one variable when the value of the other is known, as opposed to when it is not known. The present paper proposes a new PRE measure for two-way contingency tables with ordered categories in which the explanatory and response variables are not defined clearly. The proposed measure lies between 0 and 1. The proposed measure is useful for comparing the degree of PRE in several tables with ordered categories. Examples are given.

Keywords: independence, measure, prediction

1. Introduction

Consider an $r \times c$ contingency table in which one is an explanatory variable and the other is a response variable. Then, the measures which describe the relative decrease in the probability of making an error in predicting the value of either variable when the value of the other variable was known, as opposed to when it was not known, have been proposed by e.g., Goodman and Kruskal (1954), Everitt (1992) and Yamamoto, Nozaki and Tomizawa (2011). The measures are called proportional reduction in error (PRE) measures. In this situation, we assume that we can define which of the variables are the explanatory and response variables.

In some situations, the explanatory and response variables are not defined clearly. For this case with especially both nominal variables, Goodman and Kruskal (1954) and Yamamoto and Tomizawa (2010) considered some PRE measures. Let p_{ij} denote the probability that an observation will fall in the i th category of X and the j th category of Y ($i = 1, \dots, r; j = 1, \dots, c$). Goodman and Kruskal (1954) measure is given by

$$\lambda = \frac{\left\{ (1 - p_{\bullet m_0}) - \sum_{i=1}^r p_{i\bullet} \left(1 - \frac{p_{im_i}}{p_{i\bullet}} \right) \right\} + \left\{ (1 - p_{M_0\bullet}) - \sum_{j=1}^c p_{\bullet j} \left(1 - \frac{p_{M_{jj}}}{p_{\bullet j}} \right) \right\}}{2 - p_{\bullet m_0} - p_{M_0\bullet}}$$

$$= \frac{\left(\sum_{i=1}^r p_{im_i} - p_{\bullet m_0} \right) + \left(\sum_{j=1}^c p_{M_{jj}} - p_{M_0\bullet} \right)}{2 - p_{\bullet m_0} - p_{M_0\bullet}},$$

where

$$p_{im_i} = \max_j(p_{ij}), \quad p_{\bullet m_0} = \max_j(p_{\bullet j}), \quad p_{M_{jj}} = \max_i(p_{ij}), \quad p_{M_0\bullet} = \max_i(p_{i\bullet}), \quad p_{i\bullet} = \sum_{t=1}^c p_{it}, \quad p_{\bullet j} = \sum_{s=1}^r p_{sj}.$$

Also see Yamamoto and Tomizawa (2010).

However, these measures cannot be applied to two-way contingency tables with both ordinal variables when one wants to use the information about the category ordering of the variables. So we are interested in a PRE measure for contingency tables with ordinal categories in which the explanatory and response variables are not defined clearly.

This paper proposes a PRE measure for such a situation (Section 2). Section 3 gives an approximate variance for the estimated measure and Section 4 analyzes unaided distance vision data.

2. A New PRE Measure

Consider an $r \times c$ contingency table with ordinal categories in which the explanatory and response variable are not defined clearly.

Then, we shall consider the following measure which represents the PRE in predicting the category of either variable as between knowing and not knowing the category of the other variable, defined by

$$\begin{aligned} \Lambda &= \frac{\sum_{j=1}^{c-1} \left\{ \left(1 - F_{\bullet m_0}^{(j)} \right) - \sum_{i=1}^r p_{i\bullet} \left(1 - \frac{F_{im_i}^{(j)}}{p_{i\bullet}} \right) \right\} + \sum_{i=1}^{r-1} \left\{ \left(1 - F_{M_0\bullet}^{(i)} \right) - \sum_{j=1}^c p_{\bullet j} \left(1 - \frac{F_{M_{jj}}^{(i)}}{p_{\bullet j}} \right) \right\}}{\sum_{j=1}^{c-1} \left(1 - F_{\bullet m_0}^{(j)} \right) + \sum_{i=1}^{r-1} \left(1 - F_{M_0\bullet}^{(i)} \right)} \\ &= \frac{\sum_{j=1}^{c-1} \left(\sum_{i=1}^r F_{im_i}^{(j)} - F_{\bullet m_0}^{(j)} \right) + \sum_{i=1}^{r-1} \left(\sum_{j=1}^c F_{M_{jj}}^{(i)} - F_{M_0\bullet}^{(i)} \right)}{\sum_{j=1}^{c-1} \left(1 - F_{\bullet m_0}^{(j)} \right) + \sum_{i=1}^{r-1} \left(1 - F_{M_0\bullet}^{(i)} \right)}, \end{aligned}$$

where

$$\begin{aligned} p_{i\bullet} &= \sum_{t=1}^c p_{it}, \quad p_{\bullet j} = \sum_{s=1}^r p_{sj}, \\ F_{im_i}^{(j)} &= \max(F_{i1}^{(j)}, F_{i2}^{(j)}), \quad F_{\bullet m_0}^{(j)} = \max(F_{\bullet 1}^{(j)}, F_{\bullet 2}^{(j)}), \quad F_{M_{jj}}^{(i)} = \max(F_{1j}^{(i)}, F_{2j}^{(i)}), \quad F_{M_0\bullet}^{(i)} = \max(F_{1\bullet}^{(i)}, F_{2\bullet}^{(i)}), \end{aligned}$$

with

$$\begin{aligned} F_{i1}^{(j)} &= \sum_{t=1}^j p_{it}, \quad F_{i2}^{(j)} = \sum_{t=j+1}^c p_{it}, \quad F_{\bullet k}^{(j)} = \sum_{i=1}^r F_{ik}^{(j)} \quad (k = 1, 2), \\ F_{1j}^{(i)} &= \sum_{s=1}^i p_{sj}, \quad F_{2j}^{(i)} = \sum_{s=i+1}^r p_{sj}, \quad F_{k\bullet}^{(i)} = \sum_{j=1}^c F_{kj}^{(i)} \quad (k = 1, 2). \end{aligned}$$

The measure Λ has the properties that (i) Λ lies between 0 and 1, (ii) $\Lambda = 0$ if and only if the information about one variable does not reduce the probability of making an error in predicting the categories of the other variable, and (iii) $\Lambda = 1$ if and only if no error is made, given knowledge of one variable; namely there is complete predictive association. In addition, we note that if the variables X and Y are independent, then the measure Λ takes 0, but the converse does not necessarily hold (also see Section 6).

3. Approximate Confidence Interval for the Measure

Let n_{ij} denote the observed frequency in the i th row and j th column of the table ($i = 1, \dots, r; j = 1, \dots, c$). Assuming that a multinomial distribution applies to the $r \times c$ table, we consider an approximate standard error and large-sample confidence interval for Λ , using the delta method, descriptions of which are given by, e.g., Bishop et al. (1975, Sec. 14.6). The sample version of Λ , i.e., $\hat{\Lambda}$, is given by Λ 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, $\sqrt{n}(\hat{\Lambda} - \Lambda)$ has asymptotically (as $n \rightarrow \infty$) a normal distribution with

mean zero and variance $\sigma^2[\Lambda]$, where

$$\sigma^2[\Lambda] = \frac{1}{\left\{ \sum_{j=1}^{c-1} (1 - F_{\bullet m_0}^{(j)}) + \sum_{i=1}^{r-1} (1 - F_{M_0 \bullet}^{(i)}) \right\}^4} \left[\sum_{k=1}^r \sum_{l=1}^c (\omega_{kl})^2 p_{kl} - (r+c-2)^2 \left\{ \sum_{j=1}^{c-1} \left(\sum_{i=1}^r F_{im_i}^{(j)} - F_{\bullet m_0}^{(j)} \right) + \sum_{i=1}^{r-1} \left(\sum_{j=1}^c F_{M_{jj}}^{(i)} - F_{M_0 \bullet}^{(i)} \right) \right\}^2 \right],$$

where

$$\omega_{kl} = \left(\sum_{j=1}^{c-1} A_{kl}^{(j)} + \sum_{i=1}^{r-1} C_{kl}^{(i)} \right) \left\{ \sum_{j=1}^{c-1} (1 - F_{\bullet m_0}^{(j)}) + \sum_{i=1}^{r-1} (1 - F_{M_0 \bullet}^{(i)}) \right\} - \left(\sum_{j=1}^{c-1} B_l^{(j)} + \sum_{i=1}^{r-1} D_k^{(i)} \right) \left\{ \sum_{j=1}^{c-1} \left(1 - \sum_{i=1}^r F_{im_i}^{(j)} \right) + \sum_{i=1}^{r-1} \left(1 - \sum_{j=1}^c F_{M_{jj}}^{(i)} \right) \right\},$$

with

$$\begin{aligned} A_{kl}^{(j)} &= I[\text{for } j \text{ fixed } (m_k = 1 \text{ and } 1 \leq l \leq j) \text{ or } (m_k = 2 \text{ and } j+1 \leq l \leq c)], \\ B_l^{(j)} &= I[\text{for } j \text{ fixed } (m_0 = 1 \text{ and } 1 \leq l \leq j) \text{ or } (m_0 = 2 \text{ and } j+1 \leq l \leq c)], \\ C_{kl}^{(i)} &= I[\text{for } i \text{ fixed } (M_l = 1 \text{ and } 1 \leq k \leq i) \text{ or } (M_l = 2 \text{ and } i+1 \leq k \leq r)], \\ D_k^{(i)} &= I[\text{for } i \text{ fixed } (M_0 = 1 \text{ and } 1 \leq k \leq i) \text{ or } (M_0 = 2 \text{ and } i+1 \leq k \leq r)], \end{aligned}$$

and $I(\cdot)$ is the indicator function.

Let $\hat{\sigma}^2[\Lambda]$ denote $\sigma^2[\Lambda]$ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$. Then, $\hat{\sigma}[\Lambda]/\sqrt{n}$ is an estimated standard error for $\hat{\Lambda}$, and $\hat{\Lambda} \pm z_{p/2} \hat{\sigma}[\Lambda]/\sqrt{n}$ is an approximate 100(1-p)% confidence interval for Λ , where $z_{p/2}$ is the (1-p/2) percentile of the standard normal distribution.

4. An Example

Consider the data in Table 1 on unaided distance vision. Table 1a is the data, taken from Tomizawa (1984), on unaided distance vision of 4746 students aged 18 to about 25 including about 10% women in Faculty of Science and Technology, Science University of Tokyo in Japan examined in April 1982. Table 1b is the data, taken from Tomizawa (1985), on unaided distance vision of 3168 pupils comprising nearly equal number of boys and girls aged 6-12 at elementary schools in Tokyo, Japan, examined in June 1984.

For the data in Tables 1a and 1b, two variables, right and left eye grades, in each of tables have ordinal categories and we cannot define clearly which of the right and left eye grades is the explanatory variable and the response variable. Thus for these data, we are interested in applying the measure Λ . The value of $\hat{\Lambda}$ is 0.735 for Table 1a and 0.496 for Table 1b (see Table 2). This shows that the information about either eye grades reduces the probability of making an error in predicting the other by 73.5% for Table 1a, and by 49.6% for Table 1b, as opposed to when it is not known.

When the degrees of the relative decrease for Tables 1a and 1b are compared by using the 95% confidence interval for Λ , the value of $\hat{\Lambda}$ is greater for Table 1a than for Table 1b. Namely, the information about either eye grades reduces the probability of making an error in prediction more for college students than for pupils.

Table 1. Unaided distance vision data of (a) 4746 students (Tomizawa, 1984) and (b) 3168 pupils (Tomizawa, 1985)

(a) Students					
Right eye grade	Left eye grade				Total
	Best (1)	Second (2)	Third (3)	Worst (4)	
Best (1)	1291	130	40	22	1483
Second (2)	149	221	114	23	507
Third (3)	64	124	660	185	1033
Worst (4)	20	25	249	1429	1723
Total	1524	500	1063	1659	4746

(b) Pupils					
Right eye grade	Left eye grade				Total
	Best (1)	Second (2)	Third (3)	Worst (4)	
Best (1)	2470	126	21	10	2627
Second (2)	96	138	33	5	272
Third (3)	10	42	75	15	142
Worst (4)	12	7	16	92	127
Total	2588	313	145	122	3168

Table 2. Values of $\hat{\Lambda}$, approximate standard errors for them and approximate 95% confidence intervals for Λ , applied to Tables 1a and 1b

Table	$\hat{\Lambda}$	Standard error	Confidence interval
Table 1a (Students)	0.735	0.008	[0.718, 0.751]
Table 1b (Pupils)	0.496	0.027	[0.444, 0.549]

Note: Though the measure λ should be used for nominal case, we apply λ to the data in Tables 1a and 1b for comparison of Λ and λ . The values of estimated λ are 0.625 for Table 1a and 0.299 for Table 1b.

5. Simulation Study

Consider now random variables Z_1 and Z_2 having a joint bivariate normal distribution with means $E(Z_1) = \mu_1$ and $E(Z_2) = \mu_2$, variances $Var(Z_1) = \sigma_1^2$ and $Var(Z_2) = \sigma_2^2$, and correlation $Corr(Z_1, Z_2) = \rho$. Suppose that there is an underlying bivariate normal distribution with the conditions, for example, $\mu_2 = \mu_1 + 0.2$, $\sigma_2^2 = 1.2\sigma_1^2$, and suppose that a 4×4 table is formed using cutpoints for each variable at $\mu_1, \mu_1 \pm 0.6\sigma_1$. Then, in terms of simulation studies, each subtable of Table 3 gives a 4×4 table of sample size 10000, formed from an underlying bivariate normal distribution with a fixed ρ ($\rho = 0, \pm 0.3, \pm 0.6, \pm 0.9$). Table 4 gives the estimated values of Λ for each value of ρ . From Table 4, we see that the estimated value of Λ increases as $|\rho|$ increases. Therefore, when there is an underlying bivariate normal distribution, the proposed measure Λ may be appropriate as a PRE measure which describes the relative decrease in the probability of making an error in predicting the value of one variable when the value of the other is known, as opposed to when it is not known.

6. Concluding Remarks

For analyzing the ordinal-ordinal contingency table in which the explanatory and response variables are not defined clearly, we have proposed the measure Λ . The measure Λ is not invariant under the arbitrary permutations of row and/or column categories. Thus this measure should be applied for the ordinal-ordinal contingency table. On the other hand, the measure λ is invariant under the arbitrary permutations of row and/or column categories. Thus λ would not be appropriate for the ordinal-ordinal contingency table.

As described in Section 2, the measure $\Lambda = 0$ if and only if the information about categories of either variable does not reduce the probability of making an error in predicting the categories of the other. However, $\Lambda = 0$ is not always equivalent to the independence between two variables. We illustrate such an example in Table 5. Obviously,

X is not independent of Y , but the measure Λ takes 0. Namely, X is not always independent of Y just because the measure Λ takes 0.

Table 3. The 4×4 tables of sample size 10000, formed by using cutpoints for each variable at $\mu_1, \mu_1 \pm 0.6\sigma_1$, from an underlying bivariate normal distribution with the conditions $\mu_2 = \mu_1 + 0.2$, $\sigma_2^2 = 1.2\sigma_1^2$, and $\rho = 0, \pm 0.3, \pm 0.6, \pm 0.9$

(a) $\rho = -0.9$				(b) $\rho = -0.6$			
2	20	279	2442	127	294	555	1824
37	352	901	925	310	433	621	853
357	953	821	185	568	546	571	572
1956	622	133	15	1390	660	424	252

(c) $\rho = -0.3$				(d) $\rho = 0$			
342	417	618	1396	687	537	582	949
472	425	528	888	537	457	512	805
553	413	484	686	508	421	483	810
1044	595	553	586	638	535	575	964

(e) $\rho = 0.3$				(f) $\rho = 0.6$			
999	598	577	589	1374	703	427	276
566	472	505	755	556	541	539	575
466	449	510	828	279	490	606	916
322	425	600	1339	105	280	522	1811

(g) $\rho = 0.9$			
1994	637	135	9
335	935	781	176
43	323	925	955
3	23	275	2451

Table 4. The values of $\hat{\Lambda}$ applied to each subtable of Table 3

	Values of ρ						
	-0.9	-0.6	-0.3	0	0.3	0.6	0.9
$\hat{\Lambda}$	0.597	0.232	0.071	0.001	0.062	0.231	0.606

Table 5. An artificial data on cell probabilities $\{p_{ij}\}$

X	Y				Total
	1	2	3	4	
1	0.35	0.13	0.09	0.03	0.6
2	0.16	0.1	0.03	0.01	0.3
3	0.06	0.02	0.01	0.01	0.1
Total	0.57	0.25	0.13	0.05	1

Acknowledgments

The authors would like to express our thanks to Ms. Yuri Nozaki for many advices to improve this paper. In addition, we would like to thank the editor and two anonymous reviewers for their helpful comments and suggestions.

References

- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Massachusetts, Cambridge: The MIT Press.
- Everitt, B. S. (1992). *The Analysis of Contingency Tables* (2nd ed.). London: Chapman and Hall.

- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732-764. <http://dx.doi.org/10.2307/2281536>
- 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.
- Tomizawa, S. (1985). Analysis of data in square contingency tables with ordered categories using the conditional symmetry model and its decomposed models. *Environmental Health Perspectives*, 63, 235-239. <http://dx.doi.org/10.2307/3430051>
- Yamamoto, K., Nozaki, Y., & Tomizawa, S. (2011). On measure of proportional reduction in error for nominal-ordinal contingency tables. *Journal of Statistics and Management Systems*, 14, 767-773.
- Yamamoto, K., & Tomizawa, S. (2010). Measures of proportional reduction in error for two-way contingency tables with nominal categories. *Biostatistics, Bioinformatics and Biomathematics*, 2, 43-52.