A comment on Zysno's "The Modification of the Phi-coefficient reducing its Dependence on the Marginal Distributions"

Matthijs J. Warrens

Leiden University Institute
for Psychological Research

October 2, 2008

Association coefficients are quantities that play an important role in many domains of data analysis. An example is the Phi-coefficient (Yule 1912) for two dichotomous variables. Suppose the N elements of the dichotomous variables fall into the classes "+" and "−". Many association coefficients for dichotomous variables can be defined using the four counts a, b, c, and d presented in Table 1. The quantities a, b, c, and d characterize the joint distribution of the two variables. The row and column totals of Table 1 are the marginal distributions that result from summing the joint counts. We denote the marginals by n_1 and u_1 for the first variable and by n_2 and u_2 for the second variable.

Table 1: Bivariate counts table for binary variables with classes + and −.

	Variable two		
Variable one	+	−	Total
+	a	b	n_1
−	c	d	u_1
Total	n_2	u_2	N

Using the parameters from Table 1, the Phi-coefficient can be written as

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

The maximum and minimum value of coefficient ϕ, regardless of the marginal distributions, are 1 and -1 respectively. However, the values 1 and -1 only occur in the extreme cases that $n_1 = n_2$ or $n_1 = u_2$.

Zysno (1997) reviews modifications of coefficient ϕ from the literature that reduce its dependence on the marginal distributions. The various proposals are considered unsatisfactory by Zysno, and therefore he proposes the modification

$$\phi^* = \frac{ad - bc}{|ad - bc| + Ne} \qquad \text{(Zysno 1997, p. 49)}$$

where

$$e = \begin{cases} \min(b, c) & \text{if} \quad ad \geq bc \\ \min(a, d) & \text{if} \quad ad \leq bc \end{cases}$$

(Zysno 1997, p. 47). The modified coefficient ϕ^* has indeed all of the desirable properties discussed by Zysno (1997): zero value if the variables are statistically independent, maximum value 1 and minimum value -1 for all marginal distributions. However, coefficient ϕ^* is not a new coefficient, since it has been proposed elsewhere.

The first modification discussed in Zysno (1997, p. 45) is a measure introduced by Cole (1949). Cole proposed a coefficient of ecological association that measures the degree to which the observed proportion of joint occurrences of two species types exceeds or falls short of the proportion of joint occurrences expected on the basis of chance alone. The measure, denoted by C_7, can be written as

$$\text{(i)} \qquad C_7 = \frac{ad - bc}{(a + b)(b + d)} \qquad \text{if} \quad ad \geq bc, \quad b \leq c$$

$$\text{(iii)} \qquad C_7 = \frac{ad - bc}{(a + b)(a + c)} \qquad \text{if} \quad ad < bc, \quad a \leq d$$

$$\text{(iv)} \qquad C_7 = \frac{ad - bc}{(b + d)(c + d)} \qquad \text{if} \quad ad < bc, \quad a > d.$$

Cole (1949) discussed various coefficients and their properties, and, for convenience, he assumed that $n_1 \leq n_2$ (p. 417, 420). As noted by Ratliff (1982), Cole's coefficient is not formally defined for the case $ad \geq bc$ and $b > c$. Given the general tenor of Cole's (1949) paper, it is likely that he would have used

$$\text{(ii)} \qquad C_7 = \frac{ad - bc}{(a + c)(c + d)} \qquad \text{if} \quad ad \geq bc, \quad b > c.$$

The formulas (i), (ii), (iii), and (iv) may already be found in Ratliff (1982, p. 1606). The formulas may be summarized in the formula

$$C_7 = \begin{cases} (ad - bc)/\min(n_1 u_2, n_2 u_1) & \text{if} \quad ad \geq bc \\ (ad - bc)/\min(n_1 n_2, u_1 u_2) & \text{if} \quad ad \leq bc. \end{cases}$$

Coefficient C_7 is equivalent to Loevinger's (1947, 1948) H if $ad \geq bc$, that is, if the two binary variables are positively dependent. It can be verified that the formulas ϕ^* and C_7 are equivalent.

Correspondence:
Matthijs J. Warrens
Psychometrics and Research Methodology Group
Leiden University Institute for Psychological Research
Leiden University
Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden
The Netherlands
e-mail: warrens@fsw.leidenuniv.nl

References

Cole, L. C. (1949). The measurement of interspecific association. *Ecology*, *30*, 411-424.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychometrika*, *Monograph No. 4*.

Loevinger, J. A. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, *45*, 507-530.

Ratliff, R. D. (1982). A correction of Cole's C_7 and Hurlbert's C_8 coefficients of interspecific association. *Ecology*, *50*, 1-9.

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, *75*, 579-652.

Zysno, P. V. (1997). The modification of the Phi-coefficient reducing its dependence on the marginal distributions. *Methods of Psychological Research Online*, *2*, 41-52.