On Resemblance Measures For Binary Data And Correction For Maximum Value

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

Abstract
Correction for maximum value is studied for association coefficients for two binary variables. We consider several families of coefficients of which the members become equivalent after correction. Central resemblance measures in this study are coefficients by Simpson (1943) and Loevinger (1947, 1948).

1. Introduction

Important entities in psychometrics are resemblance measures for two variables. A well-known coefficient for two continuous variables is Pearson's product-moment correlation. Resemblance measures for binary variables are studied in, e.g. Gower and Legendre (1986) and Baulieu (1989). So-called presence/absence coefficients can be defined using the four dependent proportions a, b, c, and d in Table 1. Quantities a, b, c, and d may be observed proportions resulting from, e.g., classifying n persons using a dichotomous response (Fleiss, 1975). Following Sokal and Sneath (1963), the convention is adopted of calling a resemblance measure S_{Name} by its originator or the first we know to propose it. Examples are

$$S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)} \qquad \text{(Simpson, 1943)},$$

the phi coefficient

$$S_{\text{Yule}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}} \qquad \text{(Yule, 1912)}$$

which is obtained when Pearson's product-moment correlation is applied to binary data, and

$$S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)} \qquad \text{(Loevinger, 1947, 1948)}.$$

Proportions a, b, c, and d in the fourfold table are constrained by the marginal proportions p_1, p_2, q_1, and q_2. The coefficients based on these quantities are therefore also constrained by the marginals, so that maximum values are sometimes untenable. There exist considerable literature on the maximum value of coefficient S_{Yule} given the marginal probabilities, denoted by $S^{\max}_{\text{Yule}}$. A review of this literature is presented in Zysno (1997). The suggestion was made to divide S_{Yule} by $S^{\max}_{\text{Yule}}$, because this procedure allows for the maximum value of unity. As shown by Loevinger (1947, 1948) coefficients S_{Yule} and S_{Loe} are related by

$$S_{\text{Loe}} = \frac{S_{\text{Yule}}}{S^{\max}_{\text{Yule}}}.$$

Table 1. Bivariate proportions matrix of binary variables.

	Variable two		
Variable one	Value 1	Value 2	Total
Value 1	a	b	p_1
Value 2	c	d	q_1
Total	p_2	q_2	1

In this paper we study correction

$$\frac{S_{\text{Name}}}{S_{\text{Name}}^{\max}} \tag{1.1}$$

for various resemblance measures for binary variables. In both Sections 2 and 3 we consider coefficient families of which the members coincide after correction (1.1). The measures in the next section become coefficient S_{Sim} after correction (1.1); the resemblance measures in Section 3 become coefficient S_{Loe} after correction for maximum value.

2. The Simpson (1943) coefficient

In this section we study correction (1.1) for coefficients

$$\begin{aligned}
S_{\text{BB}} &= \frac{a}{\max(p_1, p_2)} && \text{(Braun-Blanquet, 1932)} \\
S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} && \text{(Gleason, 1920; Dice, 1945)} \\
S_{\text{Och}} &= \frac{a}{\sqrt{p_1 p_2}} && \text{(Ochiai, 1957)} \\
S_{\text{Kul}} &= \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) && \text{(Kulczyński, 1927)} \\
\text{and} \quad S_{\text{Sim}} &= \frac{a}{\min(p_1, p_2)} && \text{(Simpson, 1943).}
\end{aligned}$$

Resemblance measures S_{Gleas}, S_{Och}, and S_{Kul} are, respectively, the harmonic, geometric and arithmetic means (Pythagorean means) of conditional probabilities

$$S_{\text{D1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{D2}} = \frac{a}{p_2} \qquad \text{(Dice, 1945).}$$

Different types of coefficients may be obtained by considering abstractions of the Pythagorean means. One type of so-called generalized means are power means, sometimes also referred to as Hölder means (see, for example, Bullen, 2003, Chapter 3). The power mean of S_{D1} and S_{D2} is given by

$$\begin{aligned}
M_\theta(S_{\text{D1}}, S_{\text{D2}}) = M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) &= \left[\frac{1}{2}\left(\frac{a}{p_1}\right)^\theta + \frac{1}{2}\left(\frac{a}{p_2}\right)^\theta\right]^{1/\theta} \\
&= \frac{a}{p_1 p_2}\left(\frac{p_1^\theta + p_2^\theta}{2}\right)^{1/\theta}
\end{aligned} \tag{1.2}$$

where θ is a real number. We have

$$\begin{aligned} S_{\mathrm{BB}} &= \lim_{\theta\to-\infty} M_\theta\left(\frac{a}{p_1},\frac{a}{p_2}\right) \\ S_{\mathrm{Gleas}} &= M_{-1}\left(\frac{a}{p_1},\frac{a}{p_2}\right) \\ S_{\mathrm{Och}} &= \lim_{\theta\to 0} M_\theta\left(\frac{a}{p_1},\frac{a}{p_2}\right) \\ S_{\mathrm{Kul}} &= M_1\left(\frac{a}{p_1},\frac{a}{p_2}\right) \\ \text{and}\quad S_{\mathrm{Sim}} &= \lim_{\theta\to\infty} M_\theta\left(\frac{a}{p_1},\frac{a}{p_2}\right). \end{aligned}$$

Note that the maximum value of $M_\theta(S_{\mathrm{D1}}, S_{\mathrm{D2}})$ depends on proportion a. Probability a is bounded from above by marginal proportion p_1 and p_2, and its maximum value, denoted by $a^{\max}$ is obtained if either proportion b, c, or both equal zero (Zysno, 1997). Hence, $a^{\max} = \min(p_1, p_2)$. The maximum value of power mean $M_\theta(S_{\mathrm{D1}}, S_{\mathrm{D2}})$ equals

$$M_\theta\left(\frac{a}{p_1},\frac{a}{p_2}\right)^{\max} = M_\theta\left(1,\frac{\min(p_1,p_2)}{\max(p_1,p_2)}\right) \tag{1.3}$$

where

$$\frac{\min(p_1,p_2)}{\max(p_1,p_2)} = S_{\mathrm{BB}}^{\max}$$

which is the maximum value of the minimum function of S_{D1} and S_{D2}. Thus, the maximum value of a power mean of S_{D1} and S_{D2}, is equal to the power mean corresponding to the same θ of the value 1 and $S_{\mathrm{BB}}^{\max}$. Furthermore, for the maximum function we have

$$S_{\mathrm{Sim}}^{\max} = \max\left(1,\frac{\min(p_1,p_2)}{\max(p_1,p_2)}\right) = 1.$$

Replacing a by $\min(p_1, p_2)$ in (1.2), we obtain a different expression of (1.3)

$$M_\theta\left(\frac{a}{p_1},\frac{a}{p_2}\right)^{\max} = \frac{\min(p_1,p_2)}{p_1 p_2}\left(\frac{p_1^\theta + p_2^\theta}{2}\right)^{1/\theta}. \tag{1.4}$$

Using (1.2) and (1.4) in (1.1) we obtain

$$\frac{M_\theta(S_{\mathrm{D1}}, S_{\mathrm{D2}})}{M_\theta(S_{\mathrm{D1}}, S_{\mathrm{D2}})^{\max}} = \frac{a}{\min(p_1,p_2)} = S_{\mathrm{Sim}}.$$

Thus, resemblance measures S_{BB}, S_{Gleas}, S_{Ots}, S_{Kul} and S_{Sim} coincide after correction (1.1). Moreover, power mean $M_\theta(S_{\mathrm{D1}}, S_{\mathrm{D2}})$ becomes the maximum function S_{Sim} after correction (1.1).

3. The Loevinger (1947, 1948) coefficient

In this section we study correction (1.1) for coefficients

$$\begin{aligned} S_{\text{Cohen}} &= \frac{2(ad-bc)}{p_1q_2+p_2q_1} && \text{(Cohen, 1960)} \\ S_{\text{MP}} &= \frac{2(ad-bc)}{p_1q_1+p_2q_2} && \text{(Maxwell and Pilliner, 1968)} \\ S_{\text{Yule}} &= \frac{ad-bc}{\sqrt{p_1p_2q_1q_2}} && \text{(Yule, 1912)} \\ S_{\text{Fle}} &= \frac{(ad-bc)(p_1q_1+p_2q_2)}{2p_1q_2p_2q_1} && \text{(Fleiss, 1975)} \\ \text{and} \quad S_{\text{Loe}} &= \frac{ad-bc}{\min(p_1q_2,p_2q_1)} && \text{(Loevinger, 1947, 1948).} \end{aligned}$$

Resemblance measures S_{Cohen}, S_{Yule}, and S_{Loe} are, respectively, the harmonic and geometric means and the maximum function of quantities

$$S_{\text{C1}} = \frac{ad-bc}{p_1q_2} \quad \text{and} \quad S_{\text{C2}} = \frac{ad-bc}{p_2q_1} \qquad \text{(Cole, 1949).}$$

The power mean of S_{C1} and S_{C2} is given by

$$M_\theta(S_{\text{C1}}, S_{\text{C2}}) = \frac{ad-bc}{p_1p_2q_1q_2}\left[\frac{(p_1q_2)^\theta + (p_2q_1)^\theta}{2}\right]^{1/\theta}. \tag{1.5}$$

We have

$$\begin{aligned} S_{\text{Cohen}} &= M_{-1}(S_{\text{C1}}, S_{\text{C2}}) \\ S_{\text{Yule}} &= \lim_{\theta\to 0} M_\theta(S_{\text{C1}}, S_{\text{C2}}) \\ \text{and} \quad S_{\text{Loe}} &= \lim_{\theta\to\infty} M_\theta(S_{\text{C1}}, S_{\text{C2}}). \end{aligned}$$

Note that the maximum value of $M_\theta(S_{\text{C1}}, S_{\text{C2}})$ depends on the covariance $(ad-bc)$. The maximum covariance of two binary variables given the marginal probabilities is given by $(ad-bc)^{\max} = \min(p_1q_2, p_2q_1)$ (cf. Zysno, 1997). The maximum value of power mean $M_\theta(S_{\text{C1}}, S_{\text{C2}})$

$$M_\theta(S_{\text{C1}}, S_{\text{C2}})^{\max} = M_\theta\left(1, \frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}\right).$$

Thus, the maximum value of power mean $M_\theta(S_{\text{C1}}, S_{\text{C2}})$ is equal to the power mean corresponding to the same θ of the value 1 and the quantity

$$\frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}.$$

Furthermore, with respect to the maximum function S_{Loe} we have

$$S_{\text{Loe}}^{\max} = \max\left(1, \frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}\right) = 1.$$

Replacing $(ad - bc)$ by $\min(p_1q_2, p_2q_1)$ in (1.5), we obtain the maximum value of $M_\theta(S_{\text{C1}}, S_{\text{C2}})$

$$M_\theta(S_{\text{C1}}, S_{\text{C2}})^{\max} = \frac{\min(p_1q_2, p_2q_1)}{p_1p_2q_1q_2}\left[\frac{(p_1q_2)^\theta + (p_2q_1)^\theta}{2}\right]^{1/\theta}. \qquad (1.6)$$

Using (1.5) and (1.6) in (1.1) we obtain

$$\frac{M_\theta(S_{\text{C1}}, S_{\text{C2}})}{M_\theta(S_{\text{C1}}, S_{\text{C2}})^{\max}} = \frac{ad - bc}{\min(p_1q_2, p_2q_1)} = S_{\text{Loe}}.$$

Thus, power mean $M_\theta(S_{\text{C1}}, S_{\text{C2}})$ becomes the maximum function S_{Loe} after correction (1.1).

Resemblance measures S_{MP}, S_{Yule} and S_{Fle} are the harmonic, geometric and arithmetic means of quantities

$$S_{\text{P1}} = \frac{ad - bc}{p_1q_1} \quad \text{and} \quad S_{\text{P2}} = \frac{ad - bc}{p_2q_2} \qquad \text{(Peirce, 1884).}$$

The power mean of S_{P1} and S_{P2} is given by

$$M_\theta(S_{\text{P1}}, S_{\text{P2}}) = \frac{ad - bc}{p_1p_2q_1q_2}\left[\frac{(p_1q_1)^\theta + (p_2q_2)^\theta}{2}\right]^{1/\theta}. \qquad (1.7)$$

We have

$$\begin{aligned} S_{\text{MP}} &= M_{-1}(S_{\text{P1}}, S_{\text{P2}}) \\ S_{\text{Yule}} &= \lim_{\theta \to 0} M_\theta(S_{\text{P1}}, S_{\text{P2}}) \\ \text{and} \quad S_{\text{Fle}} &= M_1(S_{\text{P1}}, S_{\text{P2}}). \end{aligned}$$

The maximum value of $M_\theta(S_{\text{P1}}, S_{\text{P2}})$ depends on the covariance $(ad - bc)$. Replacing $(ad - bc)$ by $\min(p_1q_2, p_2q_1)$ in (1.7), we obtain the maximum value of $M_\theta(S_{\text{P1}}, S_{\text{P2}})$

$$M_\theta(S_{\text{P1}}, S_{\text{P2}})^{\max} = \frac{\min(p_1q_2, p_2q_1)}{p_1p_2q_1q_2}\left[\frac{(p_1q_1)^\theta + (p_2q_2)^\theta}{2}\right]^{1/\theta}. \qquad (1.8)$$

Using (1.7) and (1.8) in (1.1) we obtain

$$\frac{M_\theta(S_{\text{P1}}, S_{\text{P2}})}{M_\theta(S_{\text{P1}}, S_{\text{P2}})^{\max}} = \frac{ad - bc}{\min(p_1q_2, p_2q_1)} = S_{\text{Loe}}.$$

Thus, power mean $M_\theta(S_{\text{P1}}, S_{\text{P2}})$ becomes S_{Loe} after correction (1.1), although S_{Loe} is not a special case of $M_\theta(S_{\text{P1}}, S_{\text{P2}})$.

4. Summary

In this paper we showed which resemblance measures for binary variables coincide if they are corrected for their maximum value given fixed marginal probabilities. Resemblance measures S_{BB}, S_{Gleas}, S_{Och},and S_{Kul} coincide after correction (1.1) and become S_{Sim}, the coefficient by Simpson (1943). Coefficients S_{Cohen}, S_{Yule}, S_{MP} and S_{Fle} coincide after correction (1.1) and become S_{Loe}, which is the coefficient by Loevinger (1947, 1948).

References

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification, 6*, 233–246.

Braun-Blanquet, J. (1932). *Plant Sociology.* Authorized English translation of Pflanzensoziologie. New York: McGraw-Hill.

Bullen, P. S. (2003). *Handbook of means and their inequalities.* Dordrecht, Netherlands: Kluwer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

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

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297–302.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651–659.

Gleason, H. A. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club, 47*, 21–33.

Gower, J. C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3*, 5–48.

Jaccard, P. (1912). The Distribution of the flora in the Alpine zone. *The New Phytologist, 11*, 37–50.

Kulczyński, S. (1927). Die Pflanzenassociationen der Pienenen. *Bulletin International de L'Académie Polonaise des Sciences et des Letters, classe des sciences mathematiques et naturelles, Serie B, Supplément II, 2*, 57–203.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of test ability. *Psychological Monograph, 61*, 1–49.

Loevinger, J. A. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507–529.

Maxwell, A. E. & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology, 21*, 105–116.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bulletin of the Japanese Society for Fish Science, 22*, 526–530.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science, 4*, 453–454.

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science, 241*, 1–31.

Sokal, R. R. & Sneath, R. H. (1963). *Principles of Numerical Taxonomy.* San Francisco: W. H. Freeman and Company.

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.