Similarity coefficients for binary data : properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients
Warrens, M.J.

List of similarity coefficients

In this appendix we present a list of the two-way coefficients for binary data that one may find in the literature. The coefficients are ordered on year of appearance.

Peirce (1884):

$$S_{\text{Peir1}} = \frac{ad - bc}{p_1 q_1} \quad \text{and} \quad S_{\text{Peir2}} = \frac{ad - bc}{p_2 q_2}$$

Doolittle (1885), Pearson (1926):

$$S_{\text{Doo}} = \frac{(ad - bc)^2}{p_1 p_2 q_1 q_2}$$

Yule (1900), Montgomery and Crittenden (1977):

$$S_{\text{Yule1}} = \frac{ad - bc}{ad + bc}$$

Pearson (1905) (quoted by Yule and Kendall, 1950):

$$\text{Chi-square} \qquad \chi^2 = \frac{n(ad - bc)^2}{p_1 p_2 q_1 q_2}$$

Forbes (1907):

$$S_{\text{Forbes}} = \frac{na}{p_1 p_2}$$

Jaccard (1912):

$$S_{\text{Jac}} = \frac{a}{a + b + c}$$

Yule (1912), Pearson and Heron (1913):

$$\text{phi coefficient} \qquad S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}}$$

Yule (1912):

$$S_{\text{Yule2}} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Gleason (1920), Dice (1945), Sørenson (1948), Nei and Li (1979):

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2}$$

Michael (1920):

$$S_{\text{Mich}} = \frac{4(ad - bc)}{(a+d)^2 + (b+c)^2}$$

Kulczyński (1927), Driver and Kroeber (1932):

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) \quad \text{and} \quad S_{\text{Kul2}} = \frac{a}{b+c}$$

Braun-Blanquet (1932):

$$S_{\text{BB}} = \frac{a}{\max(p_1, p_2)}$$

Driver and Kroeber (1932), Ochiai (1957), Fowlkes and Mallows (1983):

$$S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}}$$

Kuder and Richardson (1937), Cronbach (1951) for two binary variables:

$$S_{\text{KR}} = \frac{4(ad - bc)}{p_1 q_1 + p_2 q_2 + 2(ad - bc)}$$

Russel and Rao (1940):

$$S_{\text{RR}} = \frac{a}{a + b + c + d}$$

Simpson (1943):

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

Dice (1945), Wallace (1983), Post and Snijders (1993):

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

Loevinger (1947, 1948), Mokken (1971), Sijtsma and Molenaar (2002):

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

Cole (1949):

$$S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1}$$

Goodman and Kruskal (1954):

$$S_{\text{GK}} = \frac{2\min(a, d) - b - c}{2\min(a, d) + b + c}$$

Scott (1955):

$$S_{\text{Scott}} = \frac{4ad - (b+c)^2}{(p_1 + p_2)(q_1 + q_2)}$$

Sokal and Michener (1958), Rand (1971), Brennan and Light (1974):

$$\text{Simple matching coefficient} \qquad S_{\text{SM}} = \frac{a+d}{a+b+c+d}$$

Sorgenfrei (1958), Cheetham and Hazel (1969):

$$\text{Correlation ratio} \quad S_{\text{Sorg}} = \frac{a^2}{p_1 p_1}$$

Cohen (1960):

$$S_{\text{Cohen}} = \frac{2(ad-bc)}{p_1 q_2 + p_2 q_1}$$

Rogers and Tanimoto (1960), Farkas (1978):

$$S_{\text{RT}} = \frac{a+d}{a+2(b+c)+d}$$

Stiles (1961):

$$S_{\text{Sti}} = \log_{10} \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{p_1 p_2 q_1 q_2}$$

Hamann (1961), Holley and Guilford (1964), Hubert (1977):

$$S_{\text{Ham}} = \frac{a-b-c+d}{a+b+c+d}$$

Mountford (1962):

$$S_{\text{Mount}} = \frac{2a}{a(b+c)+2bc}$$

Fager and McGowan (1963):

$$S_{\text{FM}} = \frac{a}{\sqrt{p_1 p_2}} - \frac{1}{2\sqrt{\max(p_1, p_2)}}$$

Sokal and Sneath (1963):

$$S_{\text{SS1}} = \frac{a}{a+2(b+c)} \qquad S_{\text{SS2}} = \frac{2(a+d)}{2a+b+c+2d}$$

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{p_1} + \frac{a}{p_2} + \frac{d}{q_1} + \frac{d}{q_2}\right) \qquad S_{\text{SS4}} = \frac{ad}{\sqrt{p_1 p_2 q_1 q_2}}$$

$$\text{and} \quad S_{\text{SS5}} = \frac{a+d}{b+c}$$

McConnaughey (1964):

$$S_{\text{McC}} = \frac{a^2 - bc}{p_1 p_2}$$

Rogot and Goldberg (1966):

$$S_{\text{RG}} = \frac{a}{p_1 + p_2} + \frac{d}{q_1 + q_2}$$

Johnson (1967):

$$S_{\text{John}} = \frac{a}{p_1} + \frac{a}{p_2}$$

Hawkins and Dotson (1968):

$$S_{\text{HD}} = \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right)$$

Maxwell and Pilliner (1968):

$$S_{\text{MP}} = \frac{2(ad-bc)}{p_1q_1 + p_2q_2}$$

Fleiss (1975):

$$S_{\text{Fleiss}} = \frac{(ad-bc)[p_1q_2 + p_2q_1]}{2p_1p_2q_1q_2}$$

Clement (1976):

$$S_{\text{Clem}} = \frac{aq_1}{p_1} + \frac{dp_1}{q_1}$$

Baroni-Urabani and Buser (1976):

$$S_{\text{BUB}} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}} \quad \text{and} \quad S_{\text{BUB2}} = \frac{a-b-c+\sqrt{ad}}{a+b+c+\sqrt{ad}}$$

Kent and Foster (1977):

$$S_{\text{KF1}} = \frac{-bc}{bp_1 + cp_2 + bc} \quad \text{and} \quad S_{\text{KF2}} = \frac{-bc}{bq_1 + cq_2 + bc}$$

Harris and Lahey (1978):

$$S_{\text{HL}} = \frac{a(q_1+q_2)}{2(a+b+c)} + \frac{d(p_1+p_2)}{2(b+c+d)}$$

Digby (1983):

$$S_{\text{Digby}} = \frac{(ad)^{3/4} - (bc)^{3/4}}{(ad)^{3/4} + (bc)^{3/4}}$$

Some coefficients for which no source was found in the literature:

$$\frac{2a-b-c}{2a+b+c}, \quad \frac{2d}{b+c+2d}, \quad \frac{2d-b-c}{b+c+2d}$$

$$\frac{4ad}{4ad+(a+d)(b+c)}$$ which is the harmonic mean of $\frac{a}{p_1}$, $\frac{a}{p_2}$, $\frac{d}{q_1}$ and $\frac{d}{q_2}$

$$\frac{ad-bc}{\min(p_1p_2, q_1q_2)}$$ for which its minimum value of -1 is tenable.

Summary of coefficient properties

For some of the vast amount of similarity coefficients in the appendix entitled "List of similarity coefficients", several mathematical properties were studied in this thesis. Seven coefficients stand out in the sense that for these coefficients multiple attractive properties were established in this thesis. A practical conclusion is that in most data-analytic applications the choice for the right coefficient for binary variables can probably be limited to the following seven coefficients.

Source	Jaccard (1912)
Formula	$S_{\text{Jac}} = a/(a+b+c)$
Properties	– Value indeterminate if $d = 1$ – Member of parameter family $S_{\text{GL1}} = a/[a + \theta(b+c)]$; members are interchangeable with respect to an ordinal comparison – Bounded below by correlation ratio $S_{\text{Sorg}} = a^2/p_1p_2$ – Bounded above by $S_{\text{BB}} = a/\max(p_1, p_2)$ – $D_{\text{Jac}} = 1 - S_{\text{Jac}}$ satisfies the triangle inequality – Coefficient matrix is a Robinson matrix if $\mathbf{X}$ is double Petrie – A multivariate generalization satisfies a strong generalization of the triangle inequality

Source	Gleason (1920), Dice (1945), Sørenson (1948), Bray (1956), Bray and Curtis (1957), Nei and Li (1979)
Formula	$S_{\text{Gleas}} = 2a/(p_1 + p_2)$
Properties	– Value indeterminate if $d = 1$ – Member of parameter family $S_{\text{GL1}} = a/[a + \theta(b + c)]$; members are interchangeable with respect to an ordinal comparison – Special case of a coefficient by Czekanowski (1932) – Bounded below by $S_{\text{BB}} = a/\max(p_1, p_2)$ – Bounded above by $S_{\text{DK}} = a/\sqrt{p_1p_2}$ – Becomes S_{Cohen} after correction for chance using $E(a + d) = p_1p_2 + q_1q_2$ – Coefficient matrix is a Robinson matrix if **X** is double Petrie – Three straightforward multivariate generalizations

Source	Braun-Blanquet (1932)
Formula	$S_{\text{BB}} = a/\max(p_1, p_2)$
Properties	– Value indeterminate if $d = 1$ – Special case of a coefficient by Robinson (1951) – Bounded below by $S_{\text{Jac}} = a/(a + b + c)$ – Bounded above by $S_{\text{Gleas}} = 2a/(p_1 + p_2)$ – Coefficient matrix is a Robinson matrix if **X** is double Petrie – Coefficient matrix is a Robinson matrix with a monotonic stochastic model – First eigenvector of coefficient matrix reflects a stochastic model

Source	Russel-Rao (1940)
Formula	$S_{\text{RR}} = a/(a+b+c+d)$
Properties	– No indeterminate values – $D_{\text{RR}} = 1 - S_{\text{RR}}$ satisfies the triangle inequality – Coefficient matrix is a Robinson matrix if $\mathbf{X}$ is row Petrie – Coefficient matrix is totally positive of order 2 if $\mathbf{X}$ is double Petrie – First eigenvector of coefficient matrix reflects an ordering of a stochastic model – Two multivariate generalizations satisfy a strong generalization of the triangle inequality

Source	Loevinger (1947, 1948)
Formula	$S_{\text{Loe}} = (ad - bc)/\min(p_1 q_2, p_2 q_1)$
Properties	– $S_{\text{Loe}} = [a - E(a)]/[a_{\max} - E(a)]$ with $E(a) = p_1 p_2$ and $a_{\max} = \min(p_1, p_2)$ – Coefficient $S_{\text{Sim}} = a/\min(p_1, p_2)$ becomes S_{Loe} after correction for chance using $E(a) = p_1 p_2$ – Various coefficients, including S_{Cohen} and S_{Phi}, become S_{Loe}, after correction for maximum value – Coefficients that are linear in $(a+d)$ become S_{Loe} after correction for chance using $E(a+d) = p_1 p_2 + q_1 q_2$ and correction for maximum value; the result is irrespective of what correction is applied first

Source **Formula**	Sokal and Michener (1958) $S_{\text{SM}} = (a+d)/(a+b+c+d)$ "Simple matching coefficient"
Properties	– No indeterminate values – Is a special case of proportion of agreement for two nominal variables – Is equivalent to coefficients by Rand (1971) and Brennan and Light (1974) – Member of parameter family $S_{\text{GL2}} = (a+d)/[a+\theta(b+c)+d]$; members are interchangeable with respect to an ordinal comparison – Becomes S_{Cohen} after correction for chance using $E(a+d) = p_1p_2 + q_1q_2$ – $D_{\text{SM}} = 1 - S_{\text{SM}}$ satisfies the triangle inequality – Two multivariate generalizations satisfy a strong generalization of the triangle inequality

Source **Formula**	Cohen (1960) $S_{\text{Cohen}} = 2(ad-bc)/(p_1q_2 + p_2q_1)$
Properties	– S_{Cohen} is a special case of Cohen's kappa for two nominal variables – Bounded below by $S_{\text{Scott}} = (4ad - (b+c)^2)/(p_1+p_2)(q_1+q_2)$ – A variety of coefficients that are linear in $(a+d)$, like S_{SM} and S_{Gleas}, become S_{Cohen} after correction for chance using $E(a+d) = p_1p_2 + q_1q_2$ – Is equivalent to the Adjusted Rand index by Hubert and Arabie (1985)

Coefficient index

Author index