Journal of Intelligent & Fuzzy Systems 36 (2019) 3005–3018
DOI:10.3233/JIFS-172291
IOS Press

Similarity measures for 2 × 2 tables

Matthijs J. Warrens*
Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

Abstract. 2 × 2 tables are encountered in various scientific disciplines, including biomedical, social and behavioral sciences, economics and ecology. In the literature many different similarity measures have been proposed that can be used to further summarize the information in a 2 × 2 table. Many of these measures are just functions of the four cells. In this paper the important ones are reviewed. Furthermore, it is shown how various similarity measures are related to one another by considering certain general similarity measures that have various important similarity measures as special cases. The presented overview may provide insights that may be helpful to researchers from various scientific disciplines in deciding what similarity measure to use in applications or for studying theoretical properties.

Keywords: Binary variables, presence/absence data, comparing partitions, agreement indices, association coefficients

1. Introduction

In various research situations the data can be summarized in a 2 × 2 table. For example, in psychology and biometrics it may be the result of a reliability study where two observers classify a sample of objects using a dichotomous response [33, 69, 71]. In epidemiology, a 2 × 2 table may be the result of a randomized clinical trial with a binary outcome of success [62]. Furthermore, in ecology it may be the cross-classification of the presence/absence codings of two species types in a number of locations [56, 98, 99]. Finally, in cluster analysis a 2 × 2 table may be the cross-classification of two different partitions of the same object set [1, 94].

The frequent occurrences of binary data has led to the fact that there are many similarity measures that can be used to further summarize the numbers of a 2 × 2 table [1, 5, 6, 16, 24, 40, 49, 56, 69, 105]. Well-known examples are the simple matching coefficient [82, 91], the phi coefficient [112] and Cohen's kappa [18]. Sometimes reporting a single similarity measure concludes the data-analytic part of a research study. In other cases, multiple measures or matrices of similarity measures are used as input in techniques in data mining and cluster analysis.

In this paper an overview of the various similarity measures for 2 × 2 tables is presented. There is no space to consider all similarity measures that have been proposed, but the most important ones are reviewed. The review is an extended version of Heiser and Warrens [45], and contains material from [98–100, 104, 108]. In choosing a particular similarity measure, a measure has to be considered in the context of the data analysis of which it is a part [40]. The overview provides insights that may be beneficial to both practitioners in deciding what measures to use, and theorists for studying properties of this type of similarity measures.

The paper is organized as follows. In section 2, definitions are presented. In section 3, examples of similarity measures for 2 × 2 tables are presented and various application domains are discussed. In section 4 to section 6 several general similarity measures are discussed. The aim of these sections is to show how the different measures can be classified and are

*Corresponding author. Matthijs J. Warrens, Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands. E-mail: m.j.warrens@rug.nl.

ISSN 1064-1246/19/$35.00

Table 1
Break-down of relative frequencies for binary variables X and Y

	$Y = 1$	$Y = 0$	Totals
$X = 1$	a	b	$a + b$
$X = 0$	c	d	$c + d$
Totals	$a + c$	$b + d$	1

related to one another. Rational functions are considered in section 4, chance-corrected measures in section 5, and power means in section 6.

2. Definitions

Similarity measures for 2×2 tables have been classified in a number of different ways [5, 63, 65, 92]. Furthermore, similarity measures may have different names depending on the field of science or the analytic context. Example are, association coefficients, agreement indices, reliability statistics or presence/absence coefficients.

In this review, three general types of measures are distinguished, called type A, type B and type C. Of course, many other classifications are possible. Before considering the three types, some preliminaries are discussed.

2.1. Preliminaries

In general, a 2×2 table is obtained if two objects are compared on the presence/absence of a set of attributes, or if a set of objects is cross-classified by two binary variables. To simplify the presentation, it is presupposed that the 2×2 table is a cross-classification of two binary variables X and Y.

Table 1 is an example of a 2×2 table. The four relative frequencies a, b, c and d characterize the joint distribution of the variables X and Y. Quantities a and d are often called, respectively, the positive and negative matches, whereas b and c are the mismatches. The row and column totals of Table 1 are the marginal totals that result from summing the joint proportions. Instead of relative frequencies, Table 1 may also be defined on counts or frequencies; relative frequencies are used here for notational convenience.

Similarity measures for 2×2 tables are functions that quantify the extent to which two binary variables are associated or the extent to which two objects resemble one another. The measures are functions that take as arguments the relative frequencies a, b, c and d and return numerical values that are higher if the variables are more associated [6]. In general, the symbol S is used to denote a similarity measure, but sometimes other symbols are used as well.

2.2. Symmetry

A measure is called symmetric if the values of b and c can be interchanged without changing the value of the similarity measure. Although the majority of measures discussed in this paper are symmetric, similarity measures are not required to be symmetric [65]. Asymmetric measures have natural interpretations if, for example, the variable X is a criterion against which variable Y is evaluated.

The simple matching coefficient [91] is given by

$$\frac{a + d}{a + b + c + d}.$$

This measure is also called the proportion of observed agreement, and the Rand index [82] in the context of comparing partitions. The measure is symmetric in b and c.

The Dice indices [25, 97] given by

$$\frac{a}{a + b} \quad \text{and} \quad \frac{a}{a + c}$$

are examples of asymmetric measures. Both measures may be interpreted as conditional probabilities. The first index measures the extent to which X is included in Y, whereas the second measure reciprocally measures the extent to which Y is included in X [65].

2.3. Positive and negative matches

Sokal and Sneath [92] make the classical distinction between measures that include the positive matches a only and functions that include both the positive and negative matches a and d [5, 40, 65, 98, 99]. A binary variable can be an ordinal or a nominal variable. If X is an ordinal variable, then $X = 1$ is more in some sense than $X = 0$.

For example, if a binary variable is a coding of the presence or absence of a list of attributes or features, then d reflects the number of negative matches. In the field of numerical taxonomy the quantity d is generally felt not to contribute to similarity, and hence should not be included in the definition of a similarity measure.

2.4. Type A similarity measures

Type A measures satisfy the two requirements

$$(A1) \quad S = 1 \quad \Leftrightarrow \quad b + c = 0;$$
$$(A2) \quad S = 0 \quad \Leftrightarrow \quad a = 0.$$

Property $(A1)$ states that $S = 1$ if and only if there are no mismatches (e.g., two species types always occur together), whereas $(A2)$ states that $S = 0$ if and only if the proportion $a = 0$ (e.g., two species types do not coexist). Type A similarity measures are typically functions that are increasing in a, decreasing in b and c, and have a range [0, 1]. Type A measures are suitable for ordinal variables and are similar to what are called type 1 similarity measures in Lesot, Rifgi and Benhadda [65] (see also [56]).

Symmetric examples of type A measures are the Jaccard [52] index

$$\frac{a}{a+b+c},$$

the Kulczyński [64] measure

$$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right),$$

and the Driver-Kroeber [28] measure or Ochiai [75] measure

$$\frac{a}{\sqrt{(a+b)(a+c)}}.$$

The Dice indices presented in section 2.2 are asymmetric examples of type A measure. The Kulczyński and Driver-Kroeber-Ochiai measures are, respectively, the arithmetic mean and geometric mean of the Dice indices.

Type A measures can be functions that are increasing in d. An example is the measure

$$\frac{a+\sqrt{ad}}{a+\sqrt{ad}+b+c}$$

presented in Baroni-Urbani and Buser [3]. A measure by Russell and Rao [85]

$$\frac{a}{a+b+c+d}$$

is a hybrid type A measure [92], since it does not satisfy $(A1)$ but does satisfy $(A2)$. Furthermore, the Russell-Rao measure satisfies the requirement $S = 1 \Leftrightarrow a = 1$. Finally, the Simpson [89] measure

$$\frac{a}{a+\min(b,c)}$$

satisfies the condition

$$(A3) \quad S = 1 \quad \Leftrightarrow \quad b = 0 \vee c = 0.$$

The Simpson measure equals unity if one species type only occurs in locations where a second species type exists.

2.5. Type B similarity measures

Type B measures satisfy the two requirements

$$(A1) \quad S = 1 \quad \Leftrightarrow \quad b + c = 0;$$
$$(A4) \quad S = 0 \quad \Leftrightarrow \quad a + d = 0.$$

Condition $(A4)$ states that $S = 0$ if and only if there are no (positive and negative) matches. Type B similarity measures are typically functions that are increasing in a and d, decreasing in b and c, and have a range [0, 1]. These similarity measures are suitable for nominal variables and are similar to what are called type 2 similarity measures in Lesot, Rifgi and Benhadda [65].

The similarity measures

$$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$$

and

$$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

presented in Sokal and Sneath [92] are examples of symmetric type B similarity measures. The two measures are, respectively, the arithmetic mean and the square of the geometric mean of the Dice indices in section 2.2 and the additional conditional probabilities

$$\frac{d}{b+d} \quad \text{and} \quad \frac{d}{c+d}.$$

Sokal and Sneath [92] proposed the two measures as alternatives to the type A measures by Kulczyński [64], and Driver and Kroeber [28] and Ochiai [75]. The second measure actually satisfies the stronger requirement $S = 0 \Leftrightarrow a = 0 \vee d = 0$.

2.6. Type C similarity measures

Type C measures satisfy the three conditions

$$(A1) \quad S = 1 \quad \Leftrightarrow \quad b + c = 0;$$
$$(A5) \quad S = 0 \quad \Leftrightarrow \quad ad = bc;$$
$$(A6) \quad S = -1 \quad \Leftrightarrow \quad a + d = 0.$$

Requirement $(A6)$ states that $S = -1$ if there are no matches. It expresses perfect negative association, in which the categories of one variable must be reversed to match the categories of the other variable. Furthermore, $(A5)$ specifies that $S = 0$ if the variables X and Y are statistically independent (see section 3.1 below). Type C similarity measures are functions that are increasing in a and d, decreasing in b and c, and have a range $[-1, 1]$.

Examples of type C similarity measures are Cohen's kappa [9, 18, 41, 61]

$$\frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)},$$

and the phi coefficient [112, 114]

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

Pearson's product-moment correlation coefficient coincides with the phi coefficient when it is applied to binary variables. The measure

$$\frac{(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)}$$

is called the mean square contingency [37]. It is the square of the phi coefficient.

Some type C similarity measures satisfy the conditions

$$(A3) \quad S = 1 \quad \Leftrightarrow \quad b = 0 \vee c = 0,$$

$$(A7) \quad S = -1 \quad \Leftrightarrow \quad a = 0 \vee d = 0.$$

An example is Yule's Q [111], given by

$$\frac{ad - bc}{ad + bc}.$$

3. Application domains

In this section, several important similarity measures for 2×2 tables and their application domains are reviewed.

3.1. Tetrachoric correlation and odds ratio

The tetrachoric correlation is a traditional measure for assessing association in a 2×2 table [27, 76]. It is an important statistic because the tetrachoric correlation is an estimate of the Pearson product-moment correlation coefficient between hypothetical row and column variables with normal distributions, that would reproduce the observed contingency table if they were divided into two categories in the appropriate proportions.

Because an approximate estimate of the Pearson correlation may well be as adequate in many applications, particularly in small samples, various authors have introduced approximations to the tetrachoric correlation [26, 76]. The tetrachoric correlation is an example of a measure for 2×2 tables that cannot be expressed in terms of the relative frequencies a, b, c and d.

Another classic measure is the odds ratio or cross-product

$$\frac{ad}{bc}.$$

The odds ratio is probably the most widely used statistic in epidemiology [62]. For Table 1 it is defined as the ratio of the odds of an event occurring in one group (a/b) to the odds of it occurring in another group (c/d). These groups might be any other dichotomous classification.

An odds ratio of 1 indicates that the condition or event under study is equally likely in both groups. An odds ratio greater than 1 indicates that the event is more likely in the first group. Probability theory tells us that two binary variables are statistically independent if the odds ratio is equal to unity, i.e.

$$\frac{ad}{bc} = 1.$$

Edwards [30] considers several measures that transform the odds ratio to a correlation-like range $[-1, 1]$. One example is Yule's Q [111]. Other examples are Digby's H [26] and Yule's Y [112], given by

$$\frac{(ad)^{3/4} - (bc)^{3/4}}{(ad)^{3/4} + (bc)^{3/4}} \quad \text{and} \quad \frac{(ad)^{1/2} - (bc)^{1/2}}{(ad)^{1/2} + (bc)^{1/2}}$$

respectively. All three examples are nonlinear transformations of the odds ratio, have been proposed as alternatives (approximations) to the tetrachoric correlation, and are type C similarity measures (section 2.6) that satisfy $(A3)$ and $(A7)$. Digby [26] shows that his measure performs better than the other two as an approximation to the tetrachoric correlation.

The three similarity measures are special cases of the general measure

$$OR(q) = \frac{(ad)^q - (bc)^q}{(ad)^q + (bc)^q},$$

where $q \in (0, 1]$. Since $|OR(q)|$ is increasing in q, the double inequality $|\text{Yule's } Y| \leq |\text{Digby's } H| \leq |\text{Yule's } Q|$ holds [99].

3.2. Epidemiological studies

Although the odds ratio is probably the most widely used measure in epidemiology, a variety of other similarity measures are used as well. In general, two cases can be distinguished in epidemiology.

In the first case the variable X is a criterion against which variable Y is evaluated [62]. Examples are the evaluation of a new medical test against a gold standard diagnosis, or a risk factor against a disorder, or assessing the validity of a binary measure against a binary criterion. In these cases a and d are the proportions of true positives and true negatives, whereas b and c are the proportions of false positives and false negatives. In this case researchers are interested in measures like

$$\frac{a}{a+b} \quad \text{(sensitivity)},$$

$$\frac{d}{c+d} \quad \text{(specificity)},$$

$$\frac{a}{a+c} \quad \text{(positive predictive value)},$$

$$\frac{d}{b+d} \quad \text{(negative predictive value)}.$$

Another important measure is the weighted kappa coefficient [9, 62]

$$\kappa(\varepsilon) = \frac{ad - bc}{\varepsilon(a+b)(b+d) + (1-\varepsilon)(a+c)(c+d)},$$

where $\varepsilon \in [0, 1]$. Weighted kappa is the unique measure that is based on an acknowledgment that the clinical consequences of a false negative may be quite different from the clinical consequences of a false positive [9, 62, 106, 108]. The real number $\varepsilon \in [0, 1]$ must be specified a priori indicating the relative importance of false negatives to false positives.

Measure $\kappa(\frac{1}{2})$ is identical to

$$\frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)},$$

which is Cohen's kappa [9, 18, 31, 61]. The similarity measures $\kappa(0)$ and $\kappa(1)$ are also considered in Peirce [77] and Light [66]. All special cases of $\kappa(\varepsilon)$ are type C similarity measures (section 2.6).

In the second case the variables X and Y are equally important, for example, in studies of inter-rater reliability or test-retest reliability. Suppose the variables are observers and that Table 1 is the cross classification of the judgments by the two raters on the presence or absence of a trait. An obvious measure of agreement that has been proposed independently for this situation by various authors is the proportion of all objects on whom the two raters agree [33, 36]. This proportion of observed agreement is given by

$$\frac{a+d}{a+b+c+d}.$$

This measure is also know as the simple matching coefficient [91]. The measure can be interpreted as the number of 1s and 0s shared by the binary variables in the same positions, divided by the total number of positions.

In reliability studies it is considered a necessity that a similarity measure assesses agreement over and above chance agreement [98, 99]. Measures that control for chance agreement are Cohen's kappa and the phi coefficient [112, 114]. Although Cohen's kappa and the phi coefficient have a correlation-like range $[-1, 1]$ (type C similarity measures), the measures are commonly used to distinguish between positive agreement and no agreement. For recommendations and guidelines on what statistics to use under what circumstances in epidemiological studies, we refer to Kraemer [62].

3.3. Ecological association

In ecological biology, one may distinguish several contexts where similarity measures for 2×2 tables can be used [56, 92]. One such case deals with measuring the degree of coexistence between two species types over different locations. A second situation is measuring association between two locations over different species types. In the first situation a binary variable is a coding of the presence or absence of a species type in a number of locations. The joint proportion a then equals the proportion of locations where both species types are found.

Dice [25] discusses the two asymmetric measures (see also [81, 97])

$$\frac{a}{a+b} \quad \text{and} \quad \frac{a}{a+c}.$$

The first measure is equal to the number of locations where both species types are found divided by the number of locations where only the first species type is found. The second measure is equal to the number of locations where both species types exist divided by the number of locations of the second species type.

Popular similarity measures for ecological association are the Jaccard [52] index

$$\frac{a}{a+b+c},$$

and the Dice-Sørenson measure [25, 73, 93]

$$\frac{2a}{2a+b+c}.$$

Other options are the Kulczyński [64] measure and the Driver-Kroeber-Ochiai measure [28, 75] in section 2.4.

The Jaccard index can be interpreted as the number of 1s shared by X and Y in the same positions, divided by the total number of positions were 1s occur. The Dice-Sørenson measure is a special case of measures considered in Czekanowski [21, 22] and Gleason [35]. With respect to the Jaccard measure, the Dice-Sørenson index gives twice as much weight to relative frequentie a. The Dice-Sørenson measure is regularly used with presence/absence data in the case that there are only a few positive matches relatively to the number of mismatches.

The Jaccard index, Dice-Sørenson measure and Driver-Kroeber-Ochiai index are popular measures of ecological association, and they have been empirically compared to other measures for 2 × 2 tables in numerous studies. For example, Duarte, Santos and Melo [29] evaluated association measures in clustering and ordination of common bean cultivars analyzed by RAPD type molecular markers. The genetic distance measures obtained by taking the complement of the Dice-Sørenson index were considered the most adequate.

Boyce and Ellison [10] studied similarity measures for 2 × 2 tables in the context of fuzzy set ordination, and concluded that the Jaccard index, Dice-Sørenson measure and Driver-Kroeber-Ochiai index, are the preferred similarity measures.

3.4. Comparing two partitions

Different clustering methods perform well in different situations, and no clustering method has been shown to dominate other methods across all application domains [46, 53]. To be able to choose a clustering method that is suitable for the task at hand, it is required that the characteristics of the method are well understood. An important and fundamental topic in cluster analysis research is therefore the validation of the cluster results [46].

To evaluate the performance of clustering methods researchers typically assess the agreement between a reference standard partition that purports to represent the true cluster structure of the objects, and a trial partition produced by the method that is being evaluated [46, 97]. So-called external validity indices can be used to assess the agreement between two partitions [1, 34, 50, 82, 94]. High agreement between the two partitions then indicates good recovery of the true cluster structure [2, 97].

A related problem in social and behavioral sciences is that of measuring agreement among judges in classifying answers to open-ended questions, or psychologists rating people on categories not defined in advance [12, 57, 79, 80]. The classifications can be seen as partitions and agreement between judges can be assessed by quantifying the similarity between two partitions.

Many different similarity measures have been proposed for quantifying the agreement between two partitions. These so-called external validity indices can be divided into three different approaches, namely 1) counting pairs of objects, 2) information theory based, and 3) matching sets based [78]. Validity indices that are based on counting pairs of objects can be defined using a 2 × 2 table with quantities a, b, c, and d, by counting the number of object pairs that were placed in the same cluster in both partitions (a), in the same cluster in one partition but in different clusters in the other partition (b and c), and in different clusters in both (d).

Two popular measures for comparing partitions that are based on the pair counting approach are the Rand index [82]

$$\frac{a+d}{a+b+c+d},$$

and the Hubert-Arabie adjusted Rand index [50, 94, 95]

$$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}.$$

The Rand index is equivalent to the simple matching coefficient (section 2.2) and a measure proposed in Brennan and Light [12] for measuring agreement among psychologists rating people on categories not defined in advance. The adjusted Rand index is equivalent to Cohen's kappa for two categories (section 3.2), where the categories are "same cluster" and "different cluster" [102]. Other similarity measures for comparing partitions are the Dice indices (section 2.2), which are discussed in Wallace [97], and the

Driver-Kroeber-Ochiai measure (section 2.4), which for the context of comparing partitions was proposed by Fowlkes and Mallows [34].

3.5. Test homogeneity

The similarity measure [7, 60, 67, 68]

$$\frac{ad - bc}{\min((a+b)(b+d), (a+c)(c+d))}$$

is a central statistic in Mokken scale analysis [72, 88], a methodology that may be used to select a subset of binary test items that are sensitive to the same underlying dimension. The measure is usually attributed to Loevinger [67, 68, 72, 88]. Goodman and Kruskal [37, 38] and Krippendorff [63] note that the measure was first proposed by Benini [7]. Goodman and Kruskal [37, 38] report that the similarity measure is also considered in Jordan [59].

Although the Benini index has a correlation-like range $[-1, 1]$ (type C similarity measure), it is usual to assume that two items are at least positively dependent. The Benini measure satisfies requirement $(A3)$. It is equal to unity if the binary variables form a so-called Guttman pair. In this case, all subjects that pass the first item also pass the second item, or vice versa. The Benini index can become unity with different marginal distributions, that is, the item popularities or difficulties $a + b$ and $a + c$ may be different.

Cole [19] introduced a similarity measure which is equivalent to the Benini measure if there is positive covariance between the binary variables ($ad > bc$). In the case of negative covariance ($ad < bc$), Cole's similarity measure is given by

$$\frac{ad - bc}{\min((a+b)(a+c), (b+d)(c+d))}.$$

The formula can also be found in Ratliff [83] and Warrens [103]. The Cole measure satisfies both (A3) and (A7).

The Cole measure is one of several similarity measures that are type C similarity measures that were introduced in the context of ecological association. Several authors proposed coefficients of ecological association that measure 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 [19]. In contrast, the measures discussed in section 3.3 are typically type A similarity measures.

The Cole index has been used in various applications by animal and plant ecologists [51, 83]. A variant of the Cole measure proposed in Hurlbert [51] is less influenced by the species' frequencies. Hurlbert [51] examined both the Cole measure and the variant as approximations to the tetrachoric correlation (section 3.1).

4. Rational functions

In the following sections various general similarity measures from the literature are considered. Many similarity measures for 2 × 2 tables are special cases of a certain general similarity measure. The formulation of general similarity measures reveals and specifies how the various similarity measures may be related to one another and provide ways for interpreting them (see, e.g. the end of subsection 3.1). In this section rational functions are discussed.

Gower and Legendre [40] consider the general similarity measure

$$S(\theta) = \frac{a}{a + \theta(b + c)},$$

where θ is a positive real number to avoid negative values. All special cases of $S(\theta)$ are rational functions, linear in both numerator and denominator.

Measure $S(\theta)$ is a type A similarity measure and is also studied in Fichet [32], Gower [39] and Heiser and Bennani [44]. Similarity measure $S(1)$ is the Jaccard index and $S(\frac{1}{2})$ is the Dice-Sørenson measure. Lesot et al. [65] consider the measures $S(\frac{1}{4})$ and $S(\frac{1}{8})$.

Janson and Vegelius [56] present an interesting relationship between the special cases of $S(\theta)$ that can sometimes be useful when comparing two of them (see also [2, 65, 90]). The Jaccard index and the Dice-Sørenson measure are related by $J = D/(2 - D)$. In general it holds that

$$S(2\theta) = \frac{S(\theta)}{2 - S(\theta)}.$$

Similarity measure $S(\theta)$ is a special case of the ratio model (Tversky [96])

$$S(\theta, \delta) = \frac{a}{a + \theta b + \delta c},$$

where θ and δ are positive real numbers. In contrast to $S(\theta)$ the similarity measure $S(\theta, \delta)$ does not impose the symmetry property. Measure $S(\theta, \theta)$ is identical to the $S(\theta)$. The similarity measure $S(1, 1)$ is the Jaccard index, $S(\frac{1}{2}, \frac{1}{2})$ is the Dice-Sørenson measure, and $S(0, 1)$ and $S(1, 0)$ are the Dice measures.

A second general similarity measure considered in Gower and Legendre [40] is

$$T(\theta) = \frac{a+d}{a+\theta(b+c)+d},$$

where θ is a positive real number. Compared to measure $S(\theta)$, the similarity measure $T(\theta)$ includes the negative matches d in the numerator and denominator. Measure $T(\theta)$ is a type B similarity measure (section 2.5). Measure $T(1)$ is the simple matching coefficient [91] or Rand index [82]. Measure $T(2)$ is considered in Rogers and Tanimoto [84], and measure $T(\frac{1}{2})$ is presented in Sokal and Sneath [92]. Furthermore, the identity

$$T(2\theta) = \frac{T(\theta)}{2 - T(\theta)}$$

holds.

Similarity measure $T(\theta)$ is a special case of the complement of the dissimilarity measure

$$T(\theta, \delta) = \frac{b+c}{\delta a + b + c + \varepsilon d},$$

that is considered in Baulieu [5]. The dissimilarity measure $T(\theta, \delta)$ satisfies a variety of desiderata in a formal framework considered in [5].

Warrens [108] considers another type of family of rational functions, namely

$$R(\omega, \varepsilon) = \frac{\omega a + (1-\omega)d}{\omega a + \varepsilon b + (1-\varepsilon)c + (1-\omega)d},$$

where $\omega, \varepsilon \in [0, 1]$. The similarity measure $R(0, \frac{1}{2})$ is considered in Cicchetti and Feinstein [17], the measure $R(\frac{1}{2}, \frac{1}{2})$ is the simple matching coefficient [91], and the measure $R(1, \frac{1}{2})$ is the Dice-Sørenson measure.

The formulation of $S(\theta)$ and $T(\theta)$ is closely related to the concept of order equivalence [4, 5, 39, 65, 87]. If two similarity measures are order equivalent, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations. The relevant information for these analysis methods is in the ranking induced by the similarity measures, not in the values themselves. Application examples are in image retrieval [65] and monotone equivariant cluster analysis [58].

Any two special cases of $S(\theta)$ are order equivalent, and any two special cases of $T(\theta)$ are order equivalent. Omhover, Rifqi and Detyniecki [74] showed that two special cases of $S(\theta, \delta)$ with parameters (θ, δ) and (θ', δ') are order equivalent if $\theta\delta' = \theta'\delta$. Similarly, Baulieu [5] showed that two special cases of $T(\theta, \delta)$ with parameters (θ, δ) and (θ', δ') are order equivalent if $\theta\delta' = \theta'\delta$.

Warrens [99, 110] presented various inequalities between similarity measures. Several insights can be obtained from studying inequalities between similarity measures. For example, if several similarity measures defined on the same quantities have unconditional inequalities between them it is likely that these similarity measures reflect the association or agreement between the binary variables X and Y in a similar way, but to a different extent (some have lower/higher values than others).

The similarity measures $S(\theta)$ and $T(\theta)$ are strictly decreasing in θ. Hence, the inequalities $S(\theta) > S(\theta')$ if $\theta < \theta'$, and $T(\theta) > T(\theta')$ if $\theta < \theta'$ hold. For example, if agreement is not perfect, the Jaccard index $S(1)$ always produces a lower value that the Dice-Sørenson measure $S(\frac{1}{2})$. Furthermore, the simple matching coefficient $T(1)$ always produces a higher value that the Rogers-Tanimoto measure $S(2)$.

5. Correction for chance agreement

In section 2.6 several similarity measures were presented that satisfy requirement $(A5)$, i.e., that have zero value if binary variables X and Y are statistically independent. In several domains of data analysis this requirement is a natural desideratum. For example, in reliability studies and when comparing partitions in cluster analysis, property $(A5)$ is considered a necessity. However, requirement $(A5)$ is less important for measures of ecological association (section 3.3), although some authors have argued to look at agreement beyond chance (see the Cole measure in section 3.5).

If a similarity measure does not satisfy desideratum $(A5)$, it may be corrected for agreement due to chance [1, 33, 63, 100, 101, 107, 108, 113]. After correction for chance agreement a similarity coefficient S has a form

$$\frac{S - E(S)}{1 - E(S)},$$

where the number 1 is the maximum value of the similarity measure S, and expectation $E(S)$ is conditional upon fixed marginal totals of the 2×2 table. The expectation $E(S)$ can have different forms, depending on what assumptions are deemed appropriate, i.e. two, one, or no relevant underlying continua [100, 107].

Table 2
Expectations for Table 1 under statistical independence.

	$Y=1$	$Y=0$	Totals
$X=1$	$(a+b)(a+c)$	$(a+b)(b+d)$	$a+b$
$X=0$	$(a+c)(c+d)$	$(b+d)(c+d)$	$c+d$
Totals	$a+c$	$b+d$	1

Statistical independence is based on the particular assumption that the data are a product of chance concerning two different frequency distributions, one for each variable [18, 63]. The expectations of a, b, c and d under statistical independence are presented in Table 2. The values may be obtained by considering all permutations of the observations of one of the binary variables, while preserving the order of the observations of the other variable. For each permutation the value of a, b, c or d can be determined. The arithmetic mean of these values are the quantities in Table 2.

A second possibility is that the frequency distribution underlying the two binary variables is the same for both variables [63, 86]. The expectation of a, b, c and d must be estimated from the marginal totals, and different functions may be used. Scott [86] proposed the arithmetic mean. The corresponding expectations are presented in Table 3. See Krippendorff [63] for a more complete discussion of the topic.

A third possibility is that there are no relevant underlying continua. If, for example, two raters randomly allocate objects to categories, then, for each rater, the expected marginal probability for each category is $1/2$. The probability that two raters assign, by chance, any object to the same category is $(1/2)(1/2) = 1/4$. The corresponding expectations are presented in Table 4.

Consider the general similarity measure of the form

$$CS = \frac{a+d-E(a+d)}{1-E(a+d)}.$$

The measure CS is obtained if the simple matching coefficient [82, 91] is corrected for chance agreement. The quantity $a+d$ is a simplification of the simple matching coefficient or the proportion of observed agreement, and the number 1 is the maximum value of $a+d$. The specific form of the $E(a+d)$ depends on the assumption that are deemed fit or appropriate. Using different forms for $E(a+d)$ yields different special cases of CS.

Table 4
Possible expectations for Table 1 when there is no underlying frequency distribution.

	$Y=1$	$Y=0$	Totals
$X=1$	1/4	1/4	1/2
$X=0$	1/4	1/4	1/2
Totals	1/2	1/2	1

Interestingly, similarity measure CS is also obtained if the similarity measure

$$R\left(\omega, \frac{1}{2}\right) = \frac{2\omega a + 2(1-\omega)d}{2\omega a + b + c + 2(1-\omega)d}$$

is corrected for chance agreement [108]. Thus, all special cases of $R(\omega, \frac{1}{2})$, which include the Dice-Sørenson measure [25, 93], the simple matching coefficient [91], and a measure by Cicchetti and Feinstein [17], coincide after correction for chance agreement, irrespective of the value of ω [108].

Using the expectations from Table 2 in CS yields Cohen's kappa [18] or the adjusted Rand index [50]

$$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}.$$

Furthermore, using

$$\left(\frac{2a+b+c}{2}\right)^2 + \left(\frac{b+c+2d}{2}\right)^2$$

for $E(a+d)$ in CS yields Scott's pi [86] or the intraclass kappa [9]

$$\frac{4ad-(b+c)^2}{(2a+b+c)(b+c+2d)}.$$

Table 3
Expectations for Table 1 with one underlying frequency distribution.

	$Y=1$	$Y=0$	Totals
$X=1$	$\left(\frac{2a+b+c}{2}\right)^2$	$\left(\frac{2a+b+c}{2}\right)\left(\frac{b+c+2d}{2}\right)$	$\frac{2a+b+c}{2}$
$X=0$	$\left(\frac{2a+b+c}{2}\right)\left(\frac{b+c+2d}{2}\right)$	$\left(\frac{b+c+2d}{2}\right)^2$	$\frac{b+c+2d}{2}$
Totals	$\frac{2a+b+c}{2}$	$\frac{b+c+2d}{2}$	1

Moreover, using $E(a+d)=\frac{1}{2}$ in CS yields Bennett, Alpert and Goldstein's S [8, 109]

$$\frac{a+d-(b+c)}{a+b+c+d}.$$

This similarity measure is actually a special case of the measure proposed in Bennett et al. [8]. The original measure can be applied to square agreement tables with two or more categories and is equivalent to the measure C proposed in Janson and Vegelius [55], the measure κ_n discussed in Brennan and Prediger [13] and RE proposed in Janes [54]. The 2×2 version of the similarity measure is also dicussed or derived in Hamann [43], Holley and Guilford [47], Maxwell [70] and Byrt, Bishop and Carlin [15].

Finally, if there are no relevant underlying continua one may also use the expectation

$$E(a+d)=\frac{2\max(a,d)+b+c}{2}.$$

This formula is appropriate if one focuses on the most abundant category. Using this expectation in CS yields Goodman and Kruskal's lambda [36]

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

Cohen's kappa, Scott's pi, Bennett, Alpert and Goldstein's S and Goodman and Kruskal's lambda are based on different assumptions and may therefore not be appropriate in all contexts. The assumptions are hidden in the different definitions of $E(a+d)$. Reviews of the rationales behind these similarity measures can be found in Zwick [113] and Hsu and Field [48].

Similarity measure $R(\omega,\frac{1}{2})$ is a particular case of

$$R(\omega,\varepsilon)=\frac{\omega a+(1-\omega)d}{\omega a+\varepsilon b+(1-\varepsilon)c+(1-\omega)d}$$

from section 4. It turns out that all special cases of $R(\omega,\varepsilon)$ with respect to ω coincided after correction for chance agreement [108].

Using the expectations from Table 2 $R(\omega,\varepsilon)$ becomes equal to the weighted kappa

$$\kappa(\varepsilon)=\frac{ad-bc}{\varepsilon(a+b)(b+d)+(1-\varepsilon)(a+c)(c+d)}.$$

from section 3.2 [108]. For example, after correction for chance agreement both the sensitivity and the negative predictive value becomes $\kappa(1)$. Furthermore, both the specificity and the positive predictive value become $\kappa(0)$ after correction for chance agreement.

Warrens [99, 100, 107] presented various inequalities between similarity measures. The similarity measure CS is strictly decreasing in $E(a+d)$ [99, 100]. This property can be used to derive various inequalities between the similarity measures considered in this section.

Since the $E(a+d)$ associated with Cohen's kappa is smaller than the $E(a+d)$ associated with Scott's pi, which in turn is smaller than the $E(a+d)$ with Goodman and Kruskal's lambda, the double inequality Cohen's kappa $\geq$ Scott's pi $\geq$ Goodman and Kruskal's lambda holds. Furthermore, since the $E(a+d)$ associated with Scott's pi is never smaller than $\frac{1}{2}$, the double inequality Bennett's $S\geq$ Scott's pi $\geq$ Goodman and Kruskal's lambda also holds.

6. Generalized means

There are several functions that may reflect the mean value of two real non-negative numbers. Examples are the harmonic, geometric and arithmetic means, also known as the Pythagorean means. Various similarity measures can be expressed as a mean function of certain basic building blocks. Examples of these building blocks are the Dice measures [25, 81, 97]

$$\frac{a}{a+b}\quad\text{and}\quad\frac{a}{a+c}.$$

The Dice-Sørenson measure [25, 93], Driver-Kroeber-Ochiai measure [28, 75] and the Kulczyński measure [64] are, respectively, the harmonic, geometric and arithmetic means of the Dice measures [56, 99]. Furthermore, the Braun-Blanquet [11] and Simpson [89] measures

$$\frac{a}{a+\max(b,c)}\quad\text{and}\quad\frac{a}{a+\min(b,c)}$$

are, respectively the minimum and maximum of the Dice measures.

Other examples of building blocks are the weighted kappas [9, 19, 77]

$$\kappa(1)=\frac{ad-bc}{(a+b)(b+d)}.$$

and

$$\kappa(0)=\frac{ad-bc}{(a+c)(c+d)}.$$

Cohen's kappa [18, 50], the phi coefficient [112, 114] and the Benini index [7, 67, 68] from section 3.5

are, respectively the harmonic mean, the geometric mean and the maximum value of the weighted kappas.

One so-called generalized mean is the power mean, sometimes referred to as the Hölder mean [14]. The minimum, maximum and the Pythagorean means are special cases of this generalized mean. Let p be a real number. The power mean of the Dice measures

$$M\left(p, \frac{a}{a+b}, \frac{a}{a+c}\right)$$

can be written as

$$\frac{a}{(a+b)(a+c)}\left[\frac{(a+b)^p+(a+c)^p}{2}\right]^{1/p}.$$

The general similarity measure $M(p)$ becomes the Braun-Blanquet measure for $p \to -\infty$, the Dice-Sørenson measure for $p = -1$, the Driver-Kroeber-Ochiai measure for $p \to 0$, the Kulczyński measure for $p = 1$, and the Simpson measure for $p \to \infty$.

The power mean of weighted kappas $\kappa(1)$ and $\kappa(0)$

$$N\left(p, \frac{ad-bc}{(a+b)(b+d)}, \frac{ad-bc}{(a+c)(c+d)}\right)$$

can be written as

$$\frac{ad-bc}{(a+b)(a+c)(b+d)(c+d)} \times \left[\frac{[(a+b)(b+d)]^p+[(a+c)(c+d)]^p}{2}\right]^{1/p}.$$

The general similarity measure $N(p)$ becomes Cohen's kappa for $p = -1$, the phi coefficient for $p \to 0$, and the Benini measure for $p \to \infty$.

Warrens [99, 100, 107] presented various inequalities between similarity measures. Since $M(p)$ is increasing in p, the quadruple inequality Braun-Blanquet measure $\leq$ Dice-Sørenson measure $\leq$ Driver-Kroeber-Ochiai measure $\leq$ Kulczyński measure $\leq$ Simpson measure holds. Furthermore, since $|N(p)|$ is increasing in p, the double inequality |Cohen's kappa| $\leq$ |phi coefficient| $\leq$ |Benini measure| holds.

For many similarity measures for 2×2 tables the maximal attainable value depends on the marginal distributions. For example, the relative frequency a in Table 1 cannot exceed its marginal probabilities $a + b$ and $a + c$. The Jaccard index and Dice-Sørenson measure (section 3.3), for example, can therefore only attain the maximum value of unity if $b = c$, that is, in the case of marginal symmetry.

The maximum value of a is given by

$$a_{\max} = a + \min(b, c).$$

The maximum value of the Dice-Sørenson measure given the marginal totals is equal to

$$\frac{2a + 2\min(b, c)}{2a + b + c}.$$

The maximum value of the covariance $ad - bc$ between two binary variables, given the marginal distributions, is equal to

$$(ad-bc)_{\max} = \min[(a+b)(b+d), (a+c)(c+d)].$$

The maximum value of the phi coefficient and similarity measures with the covariance $ad - bc$ in the numerator is thus also restricted by the marginal distributions [20, 42, 60, 114]. In the literature on this phenomenon, it was suggested to divide the phi coefficient by its maximum value given the marginal probabilities (phi/phimax) [23].

In general, this transformation can be applied to any similarity measure that has a maximum value that is restricted by the marginal totals. After correction for maximum value a similarity measure S has a form

$$\frac{S}{S_{\max}},$$

where $S_{\max}$ is the maximum value of S given the marginal totals.

Warrens [103] showed that all special cases of $M(p)$ coincide after correction for maximum value. This similarity measure happens to be the Simpson measure [89]. Furthermore, various authors have observed that phi/phimax is equal to kappa/kappamax [33]. Warrens [103] showed that all special cases of $N(p)$ become the Benini measure [7] (section 3.5) after the linear transformation $S/S_{\max}$.

References

[1] A.N. Albatineh, M. Niewiadomska-Bugaj and D. Mihalko, On similarity indices and correction for chance agreement, *Journal of Classification* **23** (2006), 301–313.

[2] A.N. Albatineh and M. Niewiadomska-Bugaj, MCS: A method for finding the number of clusters, *Journal of Classification* **28** (2011), 184–209.

[3] C. Baroni-Urbani and M.W. Buser, Similarity of binary data, *Systematic Zoology* **25** (1976), 251–259.

[4] V. Batagelj and M. Bren, Comparing resemblance measures, *Journal of Classification* **12** (1995), 73–90.

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

[6] F.B. Baulieu, Two variant axiom systems for presence/absence based dissimilarity coefficients, *Journal of Classification* **14** (1997), 159–170.

[7] R. Benini, *Principii di Demografia*, G. Barbèra, Firenzi. No. 29 of *Manuali Barbèra di Scienza Giuridiche Sociale e Politiche*, 1901.

[8] E.M. Bennett, R. Alpert and A.C. Goldstein, Communications through limited response questioning, *Public Opinion Quarterly* **18** (1954), 303–308.

[9] D.A. Bloch and H.C. Kraemer, 2×2 Kappa coefficients: Measures of agreement or association, *Biometrics* **45** (1989), 269–287.

[10] R.L. Boyce and P.C. Ellison, Choosing the best similarity index when performing fuzzy set ordination on binary data, *Journal of Vegetational Science* **12** (2001), 711–720.

[11] J. Braun-Blanquet, *Plant Sociology: The Study of Plant Communities*, authorized English translation of Pflanzensoziologie, McGraw-Hill, New York 1932.

[12] R.L. Brennan and R.J. Light, Measuring agreement when two observers classify people into categories not defined in advance, *British Journal of Mathematical and Statistical Psychology* **27** (1974), 154–163.

[13] R.L. Brennan and D.J. Prediger, Coefficient kappa: Some uses, misuses, and alternatives, *Educational and Psychological Measurement* **41** (1981), 687–699.

[14] P.S. Bullen, *Handbook of Means and Their Inequalities*, Kluwer, Dordrecht, 2003.

[15] T. Byrt, J. Bishop and J.B. Carlin, Bias, prevalence and kappa, *Journal of Clinical Epidemiology* **46** (1993), 423–429.

[16] A.H. Cheetham and J.E. Hazel, Binary (presence-absence) similarity coefficients, *Journal of Paleontology* **43** (1969), 1130–1136.

[17] D.V. Cicchetti and A.R. Feinstein, High agreement but low kappa: II. Resolving the paradoxes, *Journal of Clinical Epidemiology* **43** (1990), 551–558.

[18] J.A. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20** (1960), 213–220.

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

[20] E.E. Cureton, Note on $\phi/\phi_{\max}$, *Psychometrika* **24** (1959), 89–91.

[21] J. Czekanowski, *Zarys Metod Statystycnck*, E. Wendego, Warsaw, 1913.

[22] J. Czekanowski, "Coefficient of racial likeness" und "Durchschnittliche Differenz", *Anthropologischer Anzeiger* **9** (1932), 227–249.

[23] E.C. Davenport and N.A. El-Sanhurry, Phi/phimax: Review and synthesis, *Educational and Psychological Measurement* **51** (1991), 821–828.

[24] E. Deza and M.M. Deza, *Dictionary of Distances*, Elsevier, Amsterdam, 2006.

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

[26] P.G.N. Digby, Approximating the tetrachoric correlation coefficient, *Biometrics* **39** (1983), 753–757.

[27] D.R. Divgi, Calculation of the tetrachoric correlation coefficient, *Psychometrika* **44** (1979), 169–172.

[28] H.E. Driver and A.L. Kroeber, Quantitative expression of cultural relationship, *The University of California Publications in American Archaeology and Ethnology* **31** (1932), 211–256.

[29] J.M. Duarte, J.B. Santos and L.C. Melo, Comparison of similarity coefficients based on RAPD markers in the common bean, *Genetics and Molecular Biology* **22** (1999), 427–432.

[30] A.W.F. Edwards, The measure of association in a 2×2 table, *Journal of the Royal Statistical Society, Series A* **126** (1963), 109–114.

[31] A.R. Feinstein and D.V. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes, *Journal of Clinical Epidemiology* **43** (1990), 543–549.

[32] B. Fichet, *Distances and Euclidean distances for presenceabsence characters and their application to factor analysis*, in: Multidimensional Data Analysis, J. de Leeuw, W.J. Heiser, J.J. Meulman and F. Critchley, eds., DSWO Press, Leiden, 1986, pp. 23–46.

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

[34] E.B. Fowlkes and C.L. Mallows, A method for comparing two hierarchical clusterings, *Journal of the American Statistical Association* **78** (1983), 553–569.

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

[36] L.A. Goodman and W.H. Kruskal, Measures of association for cross classifications, *Journal of the American Statistical Association* **49** (1954), 732–764.

[37] L.A. Goodman and W.H. Kruskal, Measures of association for cross classifications II: Further discussion and references, *Journal of the American Statistical Association* **54** (1959), 123–163.

[38] L.A. Goodman and W.H. Kruskal, *Measures of Association for Cross Classifications*, Springer-Verlag, New York, 1979.

[39] J.C. Gower, *Euclidean distance matrices*, in: Multidimensional Data Analysis, J. de Leeuw, W.J. Heiser, J.J. Meulman and F. Critchley, eds., DSWO Press, Leiden, 1986, pp. 11–22.

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

[41] I. Guggenmoos-Holzmann, The meaning of kappa: Probabilistic concepts of reliability and validity revisited, *Journal of Clinical Epidemiology* **49** (1996), 775–783.

[42] J.P. Guilford, The minimal phi coefficient and the maximal phi, *Educational and Psychological Measurement* **25** (1965), 3–8.

[43] U. Hamann, Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen, *Willdenowia* **2** (1961), 639–768.

[44] W.J. Heiser and M. Bennani, Triadic distance models: Axiomatization and least squares representation, *Journal of Mathematical Psychology* **41** (1997), 189–206.

[45] W.J. Heiser and M.J. Warrens, *Families of relational statistics for* 2×2 *tables*, in: Advances in Interdisciplinary Applied Discrete Mathematics, H. Kaul, H.M. Mulder, eds., World Scientific, Singapore, 2010, pp. 25–52.

[46] C. Hennig, M. Meilă, F. Murtagh and R. Rocci, *Handbook of Cluster Analysis*, Chapman and Hall/CRC, New York, 2015.

[47] J.W. Holley and J.P. Guilford, A note on the G index of agreement, *Educational and Psychological Measurement* **24** (1964), 749–753.

[48] L.M. Hsu and R. Field, Interrater agreement measures: Comments on kappa$_n$, Cohen's kappa, Scott's π and Aickin's α, *Understanding Statistics* **2** (2003), 205–219.

[49] Z. Hubálek, Coefficients of association and similarity based on binary (presence-absence) data: An evaluation, *Biological Reviews* **57** (1982), 669–689.

[50] L.J. Hubert and P. Arabie, Comparing partitions, *Journal of Classification* **2** (1985), 193–218.

[51] S.H. Hurlbert, A coefficient of interspecific association, *Ecology* **50** (1969), 1–9.

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

[53] A.K. Jain, Data clustering: 50 years beyond k-means, *Pattern Recognition Letters* **31** (2010), 651–666.

[54] C.L. Janes, An extension of the random error coefficient of agreement to N × N tables, *British Journal of Psychiatry* **134** (1979), 617–619.

[55] S. Janson and J. Vegelius, On generalizations of the G index and the Phi coefficient to nominal scales, *Multivariate Behavioral Research* **14** (1979), 255–269.

[56] S. Janson and J. Vegelius, Measures of ecological association, *Oecologia* **49** (1981), 371–376.

[57] S. Janson and J. Vegelius, The J-index as a measure of nominal scale response agreement, *Applied Psychological Measurement* **6** (1982), 111–121.

[58] M.F. Janowitz, Monotone equivariant cluster analysis, *Journal of Mathematical Psychology* **37** (1979), 148–165.

[59] K. Jordan, A Korreláció számitása I, *Magyar Statisztikai Szemle Kiadványai* **1** (1941), Szám.

[60] H.M. Johnson, Maximal selectivity, correctivity and correlation obtainable in a 2 × 2 contingency table, *American Journal of Psychology* **58** (1945), 65–68.

[61] H.C. Kraemer, Ramifications of a population model for k as a coefficient of reliability, *Psychometrika* **44** (1979), 461–472.

[62] H.C. Kraemer, Reconsidering the odds ratio as a measure of 2 × 2 association in a population, *Statistics in Medicine* **23** (2004), 257–270.

[63] K. Krippendorff, Association, agreement, and equity, *Quality and Quantity* **21** (1987), 109–123.

[64] S. Kulczyński, Die Pflanzenassociationen der Pienenen, *Bulletin International de... Polonaise des Sciences et des Letters, classe des sciences mathematiques et naturelles, Serie B, Supplement II*, **2**, 1927, pp. 57–203.

[65] M.-J. Lesot, M. Rifgi and H. Benhadda, Similarity measures for binary and numerical data: A survey, *International Journal of Knowledge Engineering and Soft Data Paradigms* **1** (2009), 63–84.

[66] R.J. Light, Measures of response agreement for qualitative data: Some generalizations and alternatives, *Psychological Bulletin* **76** (1971), 365–377.

[67] J.A. Loevinger, A systematic approach to the construction and evaluation of tests of ability, *Psychometrika Monograph* (4) (1947).

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

[69] A. Martín Andrés and P. Femia-Marzo, Chance-corrected measures of reliability and validity in 2 × 2 tables, *Communications in Statistics*, *Theory and Methods* **37** (2008), 760–772.

[70] A.E. Maxwell, Coefficients of agreement between observers and their interpretation, *British Journal of Psychiatry* **130** (1977), 79–83.

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

[72] R.J. Mokken, *A theory and procedure of scale analysis*, Mouton, Hague, 1971.

[73] M. Nei and W.-H. Li, Mathematical model for studying genetic variation in terms of restriction endonucleases, *Proceedings of the National Academy of Sciences* **76** (1979) 5269–5273.

[74] J.F. Omhover, M. Rifqi and M. Detyniecki, *Ranking invariance based on similarity measures in document retrieval*, in: M. Detyniecki, J.M. Jose, A. Nürnberger and C.J.K. Rijsbergen, eds., Adaptive Multimedia Retrieval: User, Context and Feedback. Third InternationalWorkshop, AMR 2005, Revised Selected Papers, Springer, LNCS, 2006, pp. 55–64.

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

[76] K. Pearson, Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable, *Philosophical Transactions of the Royal Society of London, Series A* **195** (1900), 1–47.

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

[78] D. Pfitzner, R. Leibbrandt and D. Powers, Characterization and evaluation of similarity measures for pairs of clusterings, *Knowledge and Information Systems* **19** (2009), 361.

[79] R. Popping, *Overeenstemmingsmaten voor nominale data*, Ph. D. Dissertation, University of Groningen, 1983.

[80] R. Popping, Traces of agreement. On some agreement indices for open-ended questions, *Quality and Quantity* **18** (1984), 147–158.

[81] W.J. Post and T.A.B. Snijders, Nonparametric unfolding models for dichotomous data, *Methodika* **7** (1993), 130–156.

[82] W.M. Rand, Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* **66** (1971), 846–850.

[83] R.D. Ratliff, A correction of Cole's C7 and Hurlbert's C8 coefficients of interspecific association, *Ecology* **63** (1982) 1605–1606Öź.

[84] D.J. Rogers and T.T. Tanimoto, A computer program for classifying plants, *Science* **132** (1960), 1115–1118.

[85] P.F. Russell and T.R., On habitat and association of species of Anopheline larvae in South-Eastern Madras, *Journal of Malaria Institute India* **3** (1940), 153–178.

[86] W.A. Scott, Reliability of content analysis: The case of nominal scale coding, *Public Opinion Quarterly* **19** (1955), 321–325.

[87] R. Sibson, Order invariant methods for data analysis, *Journal of the Royal Statistical Society, Series B* **34** (1972), 311–349.

[88] K. Sijtsma and I.W. Molenaar, *Introduction to nonparametric item response theory*, Sage, Thousand Oaks, 2002.

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

[90] T.A.B. Snijders, M. Dormaar, W.H. van Schuur, C. Dijkman-Caes and G. Driessen, Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes, *Journal of Classification* **7** (1990), 5–31.

[91] R.R. Sokal and C.D. Michener, A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin* **38** (1958), 1409–1438.
[92] R.R. Sokal and P.H. Sneath, *Principles of numerical taxonomy*, Freeman San Francisco, 1963.
[93] T. Sørenson, A method of stabilizing groups of equivalent amplitude in plant sociology based on the similarity of species content and its application to analyses of the vegetation on Danish commons, *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter* **5** (1948), 1–34.
[94] D. Steinley, Properties of the Hubert-Arabie adjusted Rand index, *Psychological Methods* **9** (2004), 386–396.
[95] D. Steinley, M.J. Brusco and L. Hubert, The variance of the adjusted Rand index, *Psychological Methods* **21** (2016), 261–272.
[96] A. Tversky, Features of similarity, *Psychological Review* **84** (1977), 327–352.
[97] D.L. Wallace, A method for comparing two hierarchical clusterings: Comment, *Journal of the American Statistical Association* **78** (1983), 569–576.
[98] M.J. Warrens, On the indeterminacy of resemblance measures for (presence/absence) data, *Journal of Classification* **25** (2008), 125–136.
[99] M.J. Warrens, Bounds of resemblance measures for binary (presence/absence) variables, *Journal of Classification* **25** (2008), 195–208.
[100] M.J. Warrens, On similarity coefficients for 2 × 2 tables and correction for chance, *Psychometrika* **73** (2008), 487–502.
[101] M.J. Warrens, On association coefficients for 2 × 2 tables and properties that do not depend on the marginal distributions, *Psychometrika* **73** (2008), 777–789.
[102] M.J. Warrens, On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index, *Journal of Classification* **25** (2008), 177–183.
[103] M.J. Warrens, *On resemblance measures for binary data and correction for maximum value*, in: New Trends in Psychometrics, K. Shigemasu, A. Okada, T. Imaizumi and T. Hoshino, eds., University Academic Press, Tokyo, 2008, pp. 543–548.
[104] M.J. Warrens, *Similarity coefficients for binary data*. Properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients, PhD thesis, Leiden University, 2008.
[105] M.J. Warrens, k-Adic similarity coefficients for binary (presence/absence) data, *Journal of Classification* **26** (2009), 227–245.
[106] M.J. Warrens, A Kraemer-type rescaling that transforms the odds ratio into the weighted kappa coefficient, *Psychometrika* **75** (2010), 328–330.
[107] M.J. Warrens, Inequalities between kappa and kappa-like statistics for k × k tables, *Psychometrika* **75** (2010), 176–185.
[108] M.J. Warrens, Chance-corrected measures for 2 × 2 tables that coincide with weighted kappa, *British Journal of Mathematical and Statistical Psychology* **64** (2011), 355–365.
[109] M.J. Warrens, Power weighted versions of Bennett, Alpert and Goldstein's S, *Journal of Mathematics ID* **2** (3190), 9.
[110] M.J. Warrens, Inequalities between similarities for numerical data, *Journal of Classification* **33** (2016), 141–148.
[111] G.U. Yule, On the association of attributes in statistics, *Philosophical Transactions of the Royal Society A* **75** (1900), 257–319.
[112] G.U. Yule, On the methods of measuring the association between two attributes, *Journal of the Royal Statistical Society* **75** (1912), 579–652.
[113] R. Zwick, Another look at interrater agreement, *Psychological Bulletin* **103** (1988), 374–378.
[114] P.V. Zysno, The modification of the phi-coefficient reducing its dependence on the marginal distributions, *Methods of Psychological Research Online* **2** (1997), 41–52.