Part I

Similarity coefficients

CHAPTER 1

Coefficients for binary variables

Sequences of binary data are encountered in many different realms of research. For example, a rater may check whether or not a person possesses a certain psychological characteristic; it can be assessed if certain species types are encountered in a region or not; a person may fill in a test and can either fail or pass various items; it may be investigated if a certain object does possess or does not possess certain attributes or characteristics. Moreover, various types of quantitative data may be recoded and treated as binary. Noisy quantitative data may for instance be dichotomized. Quantitative data may also be dichotomized when the pertinent information for the problem at hand depends on a known threshold value.

A so-called similarity coefficient or association index reflects in one way or another the resemblance of two or more binary variables. Most coefficients have been proposed for the bivariate or two-way case, that is, the similarity of two sequences or variables of binary scores. In this first chapter a (brief) overview is presented of several of the bivariate coefficients for binary data that are available. The similarity coefficients may be considered both as population parameters as well as sample statistics. The formulations here will be the ones, utilized in the latter case. Following Sokal and Sneath (1963, p. 128) or more recently Albatineh, Niewiadomska-Bugaj and Mihalko (2006), the convention is adopted of calling a coefficient by its originator or the first we know to propose it. The exception to this rule is the Phi coefficient.

A major distinction is made between coefficients that do and those that do not include a certain quantity d. If a binary variable is a coding of the presence or absence of a list of attributes, then d reflects the number of negative matches, which is generally felt not to contribute to similarity. A second distinction covers coefficients that have zero value if the two sequences are (statistically) independent and coefficients that have not.

Next to introducing various bivariate coefficients, the chapter is used to outline a common problem for coefficients for binary data. Since many similarity coefficients are defined as fractions, the denominator may become 0 in some cases. For these critical cases the value of the coefficient is undefined. This case of indeterminacy for some values of coefficients for binary data has been given surprisingly little attention. As it turns out, the number of critical cases differ with the coefficients.

1.1 Four dependent quantities

Suppose the data consist of two sequences of binary (1/0) scores, for example

$$\begin{bmatrix}1\\1\\0\\0\\1\\1\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}0\\1\\1\\0\\1\\0\end{bmatrix}.$$

Various data analysis techniques do not require the full information in the two binary sequences. A convenient way to summarize the information in the two vectors is by defining the four dependent quantities

a = proportion of 1s that the variables share in the same positions
b = proportion of 1s in the first variable and 0s in second variable in the same positions
c = proportion of 0s in the first variable and 1s in second variable in the same positions
d = proportion of 0s that both variables share in the same positions.

Together, the four quantities a, b, c, and d can be used to construct the 2×2 contingency table

	Variable two		
Variable one	Value 1	Value 0	Total
Value 1	a	b	p_1
Value 0	c	d	q_1
Total	p_2	q_2	1

where the marginal probabilities are given by

$$\begin{aligned} p_1 &= a + b \quad \text{proportion of 1s in the first variable} \\ p_2 &= a + c \quad \text{proportion of 1s in the second variable} \\ q_1 &= c + d \quad \text{proportion of 0s in the first variable} \\ q_2 &= b + d \quad \text{proportion of 0s in the second variable.} \end{aligned}$$

The information in the 2×2 contingency table can be summarized by an index, called here a coefficient of similarity (affinity, resemblance, association, coexistence). As a general symbol for a similarity coefficient the capital letter S will be used. An example of a similarity coefficient is the Phi coefficient, which is given by

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

The measure S_{Phi} is sometimes attributed to Yule (1912), and is equivalent to the formula that is obtained when the Pearson's product-moment correlation derived for continuous data, is applied to binary data. See Zysno (1997) for a review on the literature on S_{Phi} and some of its modifications. The marginal proportions p_1, p_2, q_1, and q_2 can be used to obtain a shorter or more parsimonious formula for S_{Phi}, which is given by

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

Following Sokal and Sneath (1963) the convention is adopted of calling a coefficient by its originator or the first we know to propose it. The exception to this rule is actually coefficient S_{Phi}. Sokal and Sneath (1963) (among others) make a major distinction between coefficients that do or do not include the quantity d. 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, which is generally felt not to contribute to similarity. Sokal and Sneath (1963, p. 130) noted the following.

'Through reduction ad absurdum we can arrive at a universe of negative character matches purporting to establish the similarity between two entities.'

Sneath (1957) felt it was difficult to decide which negative features to include in a study and which to exclude.

'It is not pertinent to count "absence of feathers" when comparing two bacteria, but that this feature is applicable in comparing bacteria and birds.'

Sokal and Sneath (1963, p. 128, 130) also note that including negative matches may depend on what attributes or features are actually considered with respect to the species. They explain the difficulty as follows.

'It may be argued that basing similarity between two species on the mutual absence of a certain character is improper. The absence of wings, when observed among a group of distantly related organisms (such as a camel, louse and nematode), would surely be an absurd indication of affinity. Yet a positive character, such as the presence of wings (or flying organs defined without qualification as to kind of wing) could mislead equally when considered for a similarly heterogeneous assemblage (for example, bat, heron, and dragonfly).'

Examples (from the field of biological ecology) that do not include the quantity d are the coefficients given by

$$S_{\text{Jac}} = \frac{a}{p_1 + p_2 - a} \qquad \text{(Jaccard, 1912)}$$

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2} \qquad \text{(Gleason, 1920; Dice, 1945; Sørenson, 1948)}$$

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) \qquad \text{(Kulczyński, 1927)}$$

$$S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \qquad \text{(Driver and Kroeber, 1932; Ochiai, 1957).}$$

Coefficient S_{Jac} may be interpreted as the number of 1s shared by the variables in the same positions, divided by the total number of positions were 1s occur ($a + b + c = p_1 + p_2 - a$). Coefficient S_{Gleas} seems to be independently proposed by both Dice (1945) and Sørenson (1948) but is often contributed to the former. Bray (1956) noted that coefficient S_{Gleas} can already be found in Gleason (1920). The coefficient has also been proposed by various other authors, for example, Czekanowski (1932) and Nei and Li (1979). Coefficient S_{DK} by Driver and Kroeber (1932) is often attributed to Ochiai (1957). Coefficient S_{DK} is also proposed by Fowlkes and Mallows (1983) for the comparison of two clustering algorithms (see Section 2.2).

With respect to coefficient S_{Jac}, coefficient S_{Gleas} gives twice as much weight to a. The latter coefficient is regularly used with presence/absence data in the case that there are only a few positive matches relatively to the number of mismatches. In addition to S_{Jac} and S_{Gleas}, Sokal and Sneath (1963, p. 129) proposed a similarity measure that gives twice as much weight to the quantity $(b + c)$ compared to a, which is given by

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

Coefficients S_{Jac}, S_{Gleas}, and S_{SS1} are rational functions which are linear in both numerator and denominator.

If a binary variable is a coding of a nominal variable, that is, one or the other of two mutually exclusive attributes (for example, correct and incorrect, or male and female), then the quantity a reflects the number of matches on the first attribute and d reflects the number of matches on the second one. In this case, it is often felt that the quantities a and d should be equally weighted.

Goodman and Kruskal (1954, p. 758) contend that, in general, the only reasonable coefficients are those based on $(a+d)$. Examples of coefficients that do include the quantity d are the coefficients given by

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \qquad \text{(Sokal and Michener, 1958; Rand, 1971)}$$

$$S_{\text{SS2}} = \frac{2(a+d)}{2a+b+c+2d} \qquad \text{(Sokal and Sneath, 1963)}$$

$$S_{\text{RT}} = \frac{a+d}{a+2(b+c)+d} \qquad \text{(Rogers and Tanimoto, 1960)}$$

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{p_1}+\frac{a}{p_2}+\frac{d}{q_1}+\frac{d}{q_2}\right) \qquad \text{(Sokal and Sneath, 1963)}$$

$$S_{\text{SS4}} = \frac{ad}{\sqrt{p_1 p_2 q_1 q_2}} \qquad \text{(Sokal and Sneath, 1963).}$$

Since a, b, c, and d are proportions, the simple matching coefficient $S_{\text{SM}} = a + d$. Coefficient S_{SM} can be interpreted as the number of 1s and 0s shared by the variables in the same positions, divided by the total length of the variables. Coefficient S_{SM} is also proposed by Rand (1971) for the comparison of two clustering algorithms and Brennan and Light (1974) for measuring agreement of two psychologists that rate people on categories not defined in advance (see Chapter 2). In addition to S_{SM} and S_{RT}, Sokal and Sneath (1963, p. 129) proposed coefficient S_{SS2}, which gives twice as much weight to the quantity $(a + d)$ compared to $(b + c)$. Moreover, Sokal and Sneath (1963) proposed coefficients S_{SS3} and S_{SS4} as alternatives (that include the quantity d) to coefficients S_{Kul} and S_{DK}. The coefficient by Rusel and Rao (1940), given by $S_{\text{RR}} = a/(a + b + c + d) = a$, is called hybrid by Sokal and Sneath (1963), since it includes the quantity d in the denominator but not in the numerator.

1.2 Axioms for (dis)similarities

Complementary to similarity or association is the concept of dissimilarity. As an alternative to a similarity measure, the fourfold table may also be summarized by some form of dissimilarity measure. A higher value of a similarity coefficient indicates there is more association between two binary variables, whereas a low value indicates that the two sequences are dissimilar. For a dissimilarity coefficient the interpretation is the other way around. A high value indicates great dissimilarity, whereas a low value indicates great resemblance. The capital letter D will be used as a general symbol for a dissimilarity coefficient in Parts I and IV. In Part III the symbol d is used.

Various authors presented more rigorous discussions on the concepts similarity and dissimilarity. A function can only be considered a similarity or dissimilarity if it satisfies certain requirements or axioms. Some interesting exposés and discussions on axioms for (dis)similarities can be found in Baroni-Urbani and Buser (1976), Baulieu (1989, 1997), Janson and Vegelius (1981) and Batagelj and Bren (1995), in the case of bivariate or two-way coefficients, and Heiser and Bennani (1997) and Joly and Le Calvé (1995), in the case of three-way or triadic coefficients. With respect to the latter, that is, three-way dissimilarities, see Chapter 11. In addition, Zegers (1986) presented an interesting overview of requirements for similarity coefficients for more general types of data.

An essential property of a similarity coefficient $S(x_1, x_2)$ that reflects the similarity between two variables x_1 and x_2, is the property that $S(x_1, x_1) \geq S(x_1, x_2)$ and $S(x_2, x_2) \geq S(x_1, x_2)$. Furthermore, it may be required that a coefficient is symmetric, that is, $S(x_1, x_2) = S(x_2, x_1)$. Examples of coefficients that are symmetric are

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}} \quad \text{and} \quad S_{\text{Jac}} = \frac{a}{a + b + c} = \frac{a}{p_1 + p_2 - a}.$$

Two-way similarity coefficients that do not satisfy the symmetry requirement are the functions that can be found in, among others, Dice (1945, p. 298), Wallace (1983), and Post and Snijders (1993), given by

$$S_{\text{Dice1}} = \frac{a}{a + b} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{a + c} = \frac{a}{p_2}.$$

Coefficient S_{Dice1} is the number of 1s that both sequences share in the same positions, relative to the total number of 1s in the first sequence. Both S_{Dice1} and S_{Dice2} can be interpreted as conditional probabilities.

If a variable is compared with itself, it may be required that the similarity equals the value 1, that is, $S(x_1, x_1) = 1$. Coefficients S_{Phi}, S_{Jac}, S_{Dice1}, and S_{Dice2} all satisfy this axiom. A coefficient that in general violates this requirement, is an interesting measure by Russel and Rao (1940), given by

$$S_{\text{RR}} = \frac{a}{a + b + c + d} \quad \text{or simply} \quad S_{\text{RR}} = a.$$

In addition to the previous two axioms, it is sometimes required that a function has a certain range before it may be called a similarity. For similarities, it is sometimes required that the absolute value of a function is restricted from above by the value 1, that is, $|S(x_1, x_2)| \leq 1$. All coefficients that are investigated in this thesis satisfy this requirement. Coefficients that do not satisfy this axiom have quantities in the numerator that are not represented in the denominator. A coefficient that can be found in Kulczyński (1927), given by $a/(b + c)$, is an example of a coefficient that does not satisfy this requirement. Most similarity coefficients considered in this thesis satisfy the three above requirements.

Analogously to the requirements for similarities, there are axioms for the concept of dissimilarity. It is usual to require that a function $D(x_1, x_2)$ is referred to as a dissimilarity if it satisfies

$$\begin{aligned} & D(x_1, x_2) \geq 0 && \text{(nonnegativity)} \\ & D(x_1, x_2) = D(x_2, x_1) && \text{(symmetry)} \\ \text{and} \quad & D(x_1, x_1) = 0 && \text{(minimality).} \end{aligned}$$

A straightforward way to transform a similarity coefficient S into a dissimilarity coefficient D is taking the complement $D = 1 - S$. This transformation requires that $S(x_1, x_1) = 1$ in order to obtain $D = 0$. Another possible transformation, closely related to the Euclidean distance, is $D = \sqrt{1 - S}$ (Gower and Legendre, 1986): D is the square root of the complement of S. For several coefficients, transformation $D = 1 - S$ gives simple formulas. For example,

$$D_{\text{Jac}} = 1 - \frac{a}{a + b + c} = \frac{b + c}{a + b + c}.$$

In order for coefficient D_{RR} to satisfy minimality, S_{RR} must be redefined as

$$S_{\text{RR}} = \begin{cases} 1 & \text{if } x_1 = x_2 \\ a & \text{otherwise.} \end{cases}$$

Dissimilarity coefficient D_{RR} is then given by

$$D_{\text{RR}} = \begin{cases} 0 & \text{if } x_1 = x_2 \\ 1 - a & \text{otherwise.} \end{cases}$$

With respect to a dissimilarity D various other requirements can be studied, which are usually not defined for a similarity coefficient S. For D to be a distance or metric, it must satisfy the metric axioms of symmetry and

$$D(x_1, x_2) = 0 \quad \text{if and only if } x_1 = x_2 \qquad \text{(definiteness)}$$

and foremost, the triangle inequality, which is given by

$$D(x_1, x_2) \leq D(x_1, x_3) + D(x_2, x_3).$$

Metric properties of various functions are studied (reviewed) in Chapter 10. In Chapter 12 various possible multi-way generalizations of the triangle inequality are studied. Another well-known inequality is the ultrametric inequality given by

$$D(x_1, x_2) \leq \max\left(D(x_1, x_3), D(x_2, x_3)\right).$$

If a dissimilarity $D(x_1, x_2)$ satisfies the ultrametric inequality, then it also satisfies the triangle inequality. Various multi-way generalizations of the ultrametric inequality are studied in Chapter 13. Axioms for multi-way or multivariate (dis)similarities are discussed in Chapter 11.

1.3 Uncorrelatedness and statistical independence

In probability theory two binary variables are called uncorrelated if they share zero covariance, that is, $ad - bc = 0$. The covariance between two binary variables is defined as the determinant of the 2×2 contingency table. In addition to being uncorrelated, two variables may be statistically independent, which is in general a stronger requirement compared to uncorrelatedness. The two concepts are equivalent if both variables are normally distributed. Probability theory tells us that two binary variables satisfy statistical independence if the odds ratio equals unity, that is

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

The odds ratio 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 condition or event is more likely in the first group.

The value of the odds ratio lies between zero and infinity. Yule proposed two measures

$$S_{\text{Yule1}} = \frac{\frac{ad}{bc} - 1}{\frac{ad}{bc} + 1} = \frac{ad - bc}{ad + bc} \qquad \text{(Yule, 1900)}$$

and

$$S_{\text{Yule2}} = \frac{\frac{\sqrt{ad}}{\sqrt{bc}} - 1}{\frac{\sqrt{ad}}{\sqrt{bc}} + 1} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \qquad \text{(Yule, 1912)}$$

as alternatives to the odds ratio. Both coefficients S_{Yule1} and S_{Yule2} transform the odds ratio into a correlation-like scale with a range -1 to 1.

The odds ratio equals unity if $ad = bc$ which equals the case that $ad - bc = 0$. In this respect uncorrelatedness and independence are equivalent for two binary variables. For testing statistical independence, one may calculate the χ^2-statistic (Pearson and Heron, 1913; Pearson, 1947) for the 2×2 contingency table. Different opinions have been stated on what the appropriate expectations are for the fourfold table (see Chapter 4). In the majority of applications it is assumed that the data are a product of chance concerning two different frequency distribution functions underlying the two binary variables, each with its own parameter. The case of statistical independence for this possibility, conditionally on fixed marginal probabilities p_1, p_2, q_1, and q_2, is given by

	Variable two		
Variable one	Value 1	Value 0	Total
Value 1	p_1p_2	p_1q_2	p_1
Value 0	q_1p_2	q_1q_2	q_1
Total	p_2	q_2	1

The case of statistical independence visualized in this table is considered in Yule (1912), Pearson (1947), Goodman and Kruskal (1954) and Cohen (1960).

Let $E(a)$ denote the expectation of quantity a; the latter is the observed proportion of common 1s, whereas $E(a)$ is the expected proportion of common 1s. Under the assumption of two different frequency distribution functions, we have

$$\begin{aligned} a - E(a) &= a - p_1 p_2 = a(1 - a - b - c) - bc = ad - bc; \\ b - E(b) &= b - p_1 q_2 = bc - ad; \\ c - E(c) &= c - p_2 q_1 = bc - ad; \\ d - E(d) &= d - q_1 q_2 = ad - bc. \end{aligned}$$

The χ^2-statistic for the 2×2 contingency table is then given by

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

where n is the length of, or number of elements in, the binary variables. The quantity n is used to compensate for the fact that the entries in the fourfold table are proportions, not counts. The χ^2-statistic has one degree of freedom (Pearson, 1947; Fisher, 1922). The χ^2-statistic is related to the Phi coefficient by

$$S_{\text{Phi}} = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}}.$$

Both χ^2 and S_{Phi} equal zero if $ad = bc$, that is, when the two binary variables have zero covariance or are statistically independent. Apart from coefficient S_{Phi} various other similarity coefficients are defined with the covariance $ad - bc$ in the numerator. An example is Cohen's kappa (Cohen, 1960), which in the case of two categories is given by

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

Coefficient S_{Cohen} is a measure that is corrected for similarity due to chance (see Section 2.1 and Chapter 4).

Various authors have studied the expected value and possible standard deviation of similarity coefficients (see, for example, Sokal and Sneath, 1963; Janson and Vegelius, 1981). An interesting overview of possible distributions and some new derivations for coefficients S_{SM}, S_{Jac}, and S_{Gleas}, is presented in Snijders, Dormaar, Van Schuur, Dijkman-Caes and Driessen (1990). Knowing a value of central tendency and a measure of the amount of likely dispersion for a coefficient, may be used for statistical inference. Next, it is possible to test the hypothesis whether a similarity coefficient is statistically different from the expected value or not.

1.4 Indeterminacy

In this section we work with a slightly adjusted definition of a similarity coefficient for two binary variables. Firstly, instead of proportions or probabilities, let a, b, c, and d be counts, and let $n = a + b + c + d$ denote the total number of attributes of the binary variables. Secondly, we define a presence/absence coefficient $S(a, b, c, d)$ or S to be a map $S : (\mathbb{Z}^+)^4 \to \mathbb{R}$ from the set, U, of all ordered quadruples of nonnegative integers into the reals (Baulieu, 1989).

Many similarity coefficients are defined as fractions. The denominator of these fractions may therefore become 0 for certain values of a, b, c and d. For example, it is well-known that if $d = n$, then the value of S_{Jac} given by

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

is not defined or indeterminate. As noted by Batagelj and Bren (1995, Section 4.2) this case of indeterminacy for some values of coefficients for binary data has been given surprisingly little attention. The critical case of S_{Jac} implies a situation in which two binary variables consist entirely of 0s. One may argue that it is highly unlikely that this occurs in practice. For example, in ecology it is unlikely to have an ordinal data table that has objects without species. Furthermore, the problem can be resolved by excluding zero vectors from the data. Although these may be valid arguments for S_{Jac}, it turns out that the number of cases in which the value of a coefficient is indeterminate, differs with the coefficients.

To compare the number of critical cases of two different coefficients, a domain of possible cases must be defined. Consider the set U of all ordered four-tuples (a, b, c, d) of nonnegative integers. Since $a + b + c + d = n$, the number of different quadruples for given n ($n \geq 1$) is given by the binomial coefficient

$$\binom{n+3}{3} = \frac{(n+3)!}{n!\,3!} = \frac{(n+3)(n+2)(n+1)}{6}$$

which is the number of different four-tuples one may obtain out of n objects. Thus, for $n = 1, 2, 3, 4, 5, \ldots$, the set U consists of $4, 10, 20, 35, 56, \ldots$ different four-tuples. For example, for $n = 2$ we have the ten unique four-tuples

$$\begin{array}{lll} (2,0,0,0) & (1,1,0,0) & (0,1,1,0) \\ (0,2,0,0) & (1,0,1,0) & (0,1,0,1) \\ (0,0,2,0) & (1,0,0,1) & (0,0,1,1) \\ (0,0,0,2). & & \end{array}$$

For each coefficient we may study for how many four-tuples or quadruples for fixed n the value of the coefficient is indeterminate. For twenty eight similarity coefficients for both nominal and ordinal data, the number of different quadruples

[0]Parts of this section are to appear in Warrens, M.J. (in press), On the indeterminacy of similarity coefficients for binary (presence/absence) data, *Journal of Classification*.

in U for which the denominator of the corresponding coefficient equals zero are presented in the following table

Ordinal data	Nominal data	4-tuples
S_{RR}	$S_{\text{SM}}, S_{\text{SS3}}, S_{\text{Mich}}, S_{\text{RT}}, S_{\text{Ham}}$	0
$S_{\text{Jac}}, S_{\text{Gleas}}, S_{\text{BUB}}, S_{\text{BB}}, S_{\text{SS1}}$		1
	$S_{\text{GK}}, S_{\text{Scott}}, S_{\text{Cohen}}, S_{\text{HD}}$	2
	S_{MP}	4
$S_{\text{Kul}}, S_{\text{DK}}, S_{\text{Sim}}, S_{\text{Sorg}}, S_{\text{McC}}$		$2n+1$
	$S_{\text{Phi}}, S_{\text{Yule1}}, S_{\text{Yule2}}, S_{\text{SS2}},$ $S_{\text{SS4}}, S_{\text{Fleiss}}, S_{\text{Loe}}$	$4n$

The formulas of all coefficients can be found in the appendix entitled "List of similarity coefficients". The above table may be read as follows. If $n = 5$, U has 56 elements and for 20 of these quadruples the value of the Phi coefficient S_{Phi} is indeterminate. Note that the coefficients are placed in groups with the same number of critical cases. For coefficients with the most critical cases ($4n$), the number of quadruples for which the value of the coefficient is indeterminate increases in a linear fashion as n becomes larger. Increases of the number of quadruples with the indeterminacy problem are not proportional to increases of n. Hence, the ratio

$$\frac{\text{number of critical cases in } U}{\text{total number of quadruples in } U} \quad \text{decreases as } n \text{ becomes larger.}$$

Furthermore, for most coefficients indeterminacy only occurs in the case that at least two elements of four-tuple (a, b, c, d) are zero.

As an alternative to excluding the vectors that result in zero denominators values, Batagelj and Bren (1995) proposed to eliminate the indeterminacies by appropriately defining values in critical cases. Some of the definitions presented in this section give the same results as definitions proposed in Batagelj and Bren (1995). The definitions presented here simplify the reading.

Let

$$K_y = \frac{a}{a+y} \quad \text{with} \quad y = b, c.$$

Coefficients S_{Gleas}, S_{DK}, S_{Kul} and

$$S_{\text{Sorg}} = \frac{a^2}{p_1 p_2}, \qquad S_{\text{BB}} = \frac{a}{\max(p_1, p_2)} \quad \text{and} \quad S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)}$$

are, respectively, the harmonic mean, geometric mean, arithmetic mean, product, minimum function, and maximum function of K_b and K_c.

Consider the arithmetic mean of K_b and K_c

$$S_{\text{Kul}} = \frac{K_b + K_c}{2} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right).$$

Suppose $a + c = 0$. Note that the value of S_{Kul} is indeterminate. If we set $K_c = 0$, then S_{Kul} becomes

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{a+b} + 0\right) = 0 \quad \text{since} \quad a = 0.$$

Alternatively, we may remove the part from the definition of S_{Kul} that causes the indeterminacy. Coefficient S_{Kul} becomes

$$S_{\text{Kul}} = \frac{a}{a+b} = 0 \quad \text{since} \quad a = 0.$$

Thus, either setting $K_c = 0$ or removing the indeterminate part from the definition of the coefficient, leads to the same conclusion: $S_{\text{Kul}} = 0$. We therefore define

$$S_{\text{Kul}} = \begin{cases} 0 & \text{if} \quad a+b=0 \quad \text{or} \quad a+c=0 \\ \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{DK}, S_{Sim}, and S_{Sorg}.

Coefficient

$$S_{\text{McC}} = \frac{a^2 - bc}{(a+b)(a+c)} = 2S_{\text{Kul}} - 1.$$

Suppose $a + c = 0$. The value of coefficient S_{McC} is indeterminate. Also the numerator $(a^2 - bc) = 0$. We define

$$S_{\text{McC}} = \begin{cases} 0 & \text{if} \quad a+b=0 \quad \text{or} \quad a+c=0 \\ \frac{a^2-bc}{(a+b)(a+c)} & \text{otherwise.} \end{cases}$$

Consider the harmonic mean of K_b and K_c

$$S_{\text{Gleas}} = \frac{2}{K_b^{-1} + K_c^{-1}} = \frac{2a}{2a+b+c}.$$

Suppose $a + c = 0$. The value of K_c and K_c^{-1} is indeterminate. However, $2a/(2a + b + c) = 0$. Similar to S_{Kul} we define

$$S_{\text{Gleas}} = \begin{cases} 0 & \text{if} \quad d = n \\ 2a/(2a+b+c) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Jac}, S_{SS2}, S_{BB}, and S_{BUB}.

Note that the definitions of S_{Kul} and S_{Gleas} presented here do not ensure that $S_{\text{Kul}} = 1$ or $S_{\text{Gleas}} = 1$ if variable x_1 is compared with itself. If $x_1 = x_2 = \overbrace{(0, 0, ..., 0)}^{n}$, that is, the two variables have nothing in common, $S_{\text{Kul}} = S_{\text{Gleas}} = 0$. Furthermore, if variable $x_1 = \overbrace{(0, 0, ..., 0)}^{n}$ is compared with itself, $S_{\text{Kul}} = S_{\text{Gleas}} = 0$. Since these coefficients are appropriate for ordinal data, it is a moot point what the value of the coefficient should be if variables x_1 and x_2, or just variable x_1 if x_2 is compared with itself, are zero vectors. From a philosophical point of view it might be better to leave the coefficients for ordinal data undefined for the critical case $d = n$.

Consider coefficient

$$S_{\text{HD}} = \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right) \qquad \text{(Hawkins and Dotson, 1968).}$$

The value of S_{HD} is indeterminate if either $a = n$ or $d = n$. If $a = n$ then variables x_1 and x_2 are unit vectors; if $d = n$ then variables x_1 and x_2 are zero vectors. If both variables are zero vectors or unit vectors, we may speak of perfect agreement if x_1 and x_2 are nominal variables. We therefore define

$$S_{\text{HD}} = \begin{cases} 1 & \text{if} \quad a = n \quad \text{or} \quad d = n \\ \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Cohen}, S_{GK} and S_{Scott}. We also define

$$S_{\text{MP}} = \begin{cases} 1 & \text{if} \quad a = n \quad \text{or} \quad d = n \\ 0 & \text{if} \quad b = n \quad \text{or} \quad c = n \\ \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)} & \text{otherwise.} \end{cases}$$

Consider the Phi coefficient

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

The value of S_{Phi} is indeterminate if $a + b = 0$, $a + c = 0$, $b + d = 0$, or $c + d = 0$. For these critical cases the covariance $(ad - bc) = 0$. We define

$$S_{\text{Phi}} = \begin{cases} 1 & \text{if} \quad a = n \quad \text{or} \quad d = n \\ 0 & \text{if} \quad a + b = 0, \quad a + c = 0, \quad b + d = 0 \quad \text{or} \quad c + d = 0 \\ \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{SS4}, S_{Yule1}, S_{Yule2}, S_{Fleiss}, and S_{Loe}.

Let

$$K_y = \frac{a}{a+y} \quad \text{and} \quad K_y^* = \frac{d}{y+d} \quad \text{with} \quad y = b, c.$$

Consider the arithmetic mean of K_b, K_c, K_b^* and K_c^*

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right).$$

Suppose $c + d = 0$. Note that the value of K_c^* is indeterminate. To eliminate the critical case, we may set $K_c^* = 0$, and S_{SS3} becomes

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{a+b} + 1 + 0 + 0\right) = \frac{2a+b}{4(a+b)}. \tag{1.1}$$

Note that coefficient S_{SS3} in (1.1) has a range $[\frac{1}{4}, \frac{1}{2}]$. We may define

$$S_{\text{SS3}} = \begin{cases} \frac{2a+b}{4(a+b)} & \text{if} \quad c+d=0 \\ \frac{2a+c}{4(a+c)} & \text{if} \quad b+d=0 \\ \frac{b+2d}{4(b+d)} & \text{if} \quad a+c=0 \\ \frac{c+2d}{4(c+d)} & \text{if} \quad a+b=0 \\ \frac{1}{2} & \text{if} \quad a=n \quad \text{or} \quad d=n \\ 0 & \text{if} \quad b=n \quad \text{or} \quad c=n \\ \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) & \text{otherwise.} \end{cases}$$

As an alternative to the above robust definition of S_{SS3}, we propose to eliminate the critical case by removing the part from the definition of S_{SS3} that causes the indeterminacy. Suppose $c + d = 0$. The arithmetic mean of K_b, K_c and K_b^* is given by

$$S_{\text{SS3}}^* = \frac{1}{3}\left(\frac{a}{a+b} + 0 + 1\right) = \frac{2a+b}{3(a+b)}. \tag{1.2}$$

Note that coefficient S_{SS3}^* in (1.2) has a range $[\frac{1}{3}, \frac{2}{3}]$. We define

$$S_{\text{SS3}}^* = \begin{cases} \frac{2a+b}{3(a+b)} & \text{if} \quad c+d=0 \\ \frac{2a+c}{3(a+c)} & \text{if} \quad b+d=0 \\ \frac{b+2d}{3(b+d)} & \text{if} \quad a+c=0 \\ \frac{c+2d}{3(c+d)} & \text{if} \quad a+b=0 \\ 1 & \text{if} \quad a=n \quad \text{or} \quad d=n \\ 0 & \text{if} \quad b=n \quad \text{or} \quad c=n \\ \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) & \text{otherwise.} \end{cases}$$

1.5 Epilogue

In this first chapter basic notation and several concepts of similarity coefficients for binary data were introduced. A coefficient summarizes the two-way information in two sequences of binary (0/1) scores. A coefficient may be used to compare two variables over several cases or persons, two cases over variables, two objects over attributes, or two attributes over objects. Although the data analysis literature distinguishes between, for example, bivariate information between variables or dyadic information between cases, the terms bivariate and two-way are used for any two sequences of binary scores (the terms are considered interchangeable) in this dissertation.

Two distinctions between the large number of coefficients were made in this chapter. Coefficients may be divided in groups that do or do not include the quantity d. If a binary variable is a coding of the presence or absence of a list of attributes, then d reflects the number of negative matches. A second distinction was made between coefficients that have zero value if the two sequences are statistically independent and coefficients that have not. A full account of the possibilities of statistical testing with respect to the 2×2 contingency table can be found in Pearson (1947).

No attempt was made to present a complete overview of all proposed or all possible coefficients for binary data. An overview of bivariate coefficients for binary data from the literature can be found in the appendix entitled "List of similarity coefficients". To obtain some ideas of other possible coefficients, the reader is referred to other sources: Sokal and Sneath (1963), Cheetham and Hazel (1969), Baroni-Urbani and Buser (1976), Janson and Vegelius (1982), Hubálek (1982), Gower and Legendre (1986), Krippendorff (1987), Baulieu (1989) and Albatineh et al. (2006).

CHAPTER 2

Coefficients for nominal and quantitative variables

The main title ("Similarity coefficients for binary data") suggests that the thesis is about resemblance or association measures between objects characterized by two-state (binary) attributes. Many of the bivariate or two-way coefficients, however, were not proposed for use with binary variables only. The formulas considered in this thesis are often special cases that are obtained when more general formulas from various domains of data analysis are applied to dichotomous data. The general resemblance measures may, for example, be used for frequency data or other positive counts. Some coefficients based on proportions a, b, c, and d are special cases of not just one, but multiple coefficients. For example, coefficient

$$S_{\text{Gleas}} = \frac{2a}{2a+b+c} \quad \text{or its complement} \quad 1 - S_{\text{Gleas}} = \frac{b+c}{2a+b+c}$$

have been proposed for binary variables by Gleason (1920), Dice (1945), Sørenson (1948), Nei and Li (1979), and seem to have been popularized by Bray (1956) and Bray and Curtis (1957). Coefficient S_{Gleas} is a special case of, for example, a coefficient by Czekanowski (1932), a measure by Odum (1950), and a coefficient by Williams, Lambert and Lance (1966). The simple matching coefficient

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \quad \text{or its complement} \quad 1 - S_{\text{SM}} = \frac{b+c}{a+b+c+d}$$

can be obtained, for example, as a special case of a general coefficient by Gower (1971) or Cox and Cox (2000), the observed proportion of agreement of a bivariate table of two nominal variables, the city-block distance, or as a special case of a measure by Cain and Harrison (1958).

This chapter is used to present various interesting formulas for nominal and quantitative variables, accompanied by some measures used in set theory, of which some of the coefficients that will be frequently encountered in this thesis, like S_{Gleas} and S_{SM}, are special cases. This puts the coefficients for binary data in a more general context. In addition, from this chapter ideas or possibilities may be obtained for generalizing some of the results presented in this dissertation.

2.1 Nominal variables

When dealing with bivariate or two-way similarity coefficients for nominal variables two situations can be distinguished. The two nominal variables have either identical categories or they have different categories (Popping, 1983a; Zegers, 1986). The latter possibility is discussed in Section 2.3. Suppose that two psychologists each distribute m people among a set of k mutually exclusive categories. In addition suppose that the categories are defined in advance. To measure the agreement among the two psychologists, a first step is to obtain a contingency table or matching table $\mathbf{N}$ with elements n_{ij}, where n_{ij} indicates the number of persons placed in category i $(i = 1, 2, ..., I)$ by the first psychologist and in category j $(j = 1, 2, ..., J)$ by the second psychologist. Furthermore, let

$$n_{i+} = \sum_{j=1}^{J} n_{ij} \quad \text{and} \quad n_{+j} = \sum_{i=1}^{I} n_{ij}$$

denote the marginal counts (row and column totals) of the contingency table $\mathbf{N}$. Suppose that the categories of both nominal variables are in the same order, so that the diagonal elements of the square matrix $\mathbf{N}$ (n_{ii}) reflect the number of people put in the same category by both psychologists. If there are just two categories, then $m^{-1}\mathbf{N}$ equals the usual fourfold table. A straightforward measure of bivariate association is the observed proportion of agreement P_{o}, given by

$$P_{\text{o}} = \frac{1}{m} \sum_{i=1}^{k} n_{ii} = \frac{\text{tr}(\mathbf{N})}{m}.$$

If there are just two categories, for example, presence or absence of a psychological characteristic, then

$$P_{\text{o}} = \frac{a+d}{a+b+c+d} = S_{\text{SM}}.$$

Both Scott (1955) and Cohen (1960) proposed measures that incorporate correction for chance agreement. Both measures are corrected versions of P_{o}.

After correction a similarity coefficient S has a form

$$CS = \frac{S - E(S)}{1 - E(S)} \tag{2.1}$$

where $E(S)$ is conditional on the marginals of the contingency table of which S is the summary statistic. Furthermore, the constant 1 in the denominator of (2.1) may be replaced by the maximum value of a coefficient S (all coefficients that are studied in this thesis have a maximum value of unity). Expectation $E(S)$ depends on the marginal proportions, but the maximum value does not.

We note two expectations of P_o, which will be referred to as the expected proportion of agreement $E(P_o)$. Scott (1955) works with the assumption that the data are a product of chance of a single frequency distribution. To estimate the common parameters from the marginal counts, Scott (1955) uses

$$E(P_o)_{\text{Scott}} = \frac{1}{4} \sum_{i=1}^{k} \left(\frac{n_{i+}}{m} + \frac{n_{+i}}{m} \right)^2 . \tag{2.2}$$

Alternatively, Cohen (1960) works with the assumption that the data are a product of chance of two different frequency distributions, one for each nominal variable. The expected proportion of agreement under statistical independence is given by

$$E(P_o)_{\text{Cohen}} = \frac{1}{m^2} \sum_{i=1}^{k} n_{i+} n_{+i}. \tag{2.3}$$

Expectation (2.3) may be obtained by considering all permutations of the observations of one of the two variables, while preserving the order of the observations of the other variable. For each permutation the value of P_o can be determined. The arithmetic mean of these values is (2.3).

Using P_o and either (2.2) or (2.3) in (2.1), we obtain Scott's pi and Cohen's kappa, which are given by

$$S_{\text{Scott}} = \frac{P_o - E(P_o)_{\text{Scott}}}{1 - E(P_o)_{\text{Scott}}} \qquad \text{and} \qquad S_{\text{Cohen}} = \frac{P_o - E(P_o)_{\text{Cohen}}}{1 - E(P_o)_{\text{Cohen}}}$$

and become respectively

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

with binary variables. Other suitable measures for nominal variables with identical categories are discussed in Janson and Vegelius (1979).

2.2 Comparing two partitions

In cluster analysis one may be interested in comparing two clustering methods (Rand, 1971; Fowlkes and Mallows, 1983; Hubert and Arabie, 1985; Lerman, 1988; Steinley, 2004; Albatineh et al., 2006). Suppose we have two partitions of m data points. To compare these two clusterings, a first step is to obtain a so-called matching table $\mathbf{N}$ with elements n_{ij}, where n_{ij} indicates the number of data points placed in cluster i $(i = 1, 2, ..., I)$ according to the first clustering method and in cluster j $(j = 1, 2, ..., J)$ according to the second method.

The total number of points being clustered is given by $m = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. The cluster sizes in respective clusterings are the row and column totals of the matching table n_{i+} and n_{+j}. Furthermore, we define the quantity

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \binom{n_{ij}}{2} = \frac{1}{2} \left[\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}^2 - m \right]$$

which equals the number of object pairs that were placed in the same cluster according to both clustering methods, and the three quantities

$$P = \sum_{i=1}^{I} \binom{n_{i+}}{2}, \quad Q = \sum_{j=1}^{J} \binom{n_{+j}}{2} \quad \text{and} \quad N = \binom{m}{2}.$$

The quantity N equals the total number of pairs of objects given m points.

As a second step, one may calculate some sort of resemblance measure that summarizes the information in the matching table. A well-known measure for the similarity of two partitions is the Rand index (Rand, 1971), given by

$$S_{\text{Rand}} = \frac{N + 2T - P - Q}{N}.$$

Another measure of resemblance for comparing two partitions is the coefficient by Fowlkes and Mallows (1983), given by

$$S_{\text{FM}} = \frac{T}{\sqrt{PQ}}.$$

Similar to the proportion of observed agreement P_{o} from Section 2.1, coefficient S_{Rand} may be adjusted for agreement due to chance (Morey and Agresti, 1984; Hubert and Arabie, 1985; Albatineh et al., 2006). Fowlkes and Mallows (1983) and Hubert and Arabie (1985, p. 197) noted that, if the generalized hypergeometric distribution function is assumed appropriate for the matching table $\mathbf{N}$, then the expectation $E(T)$ under statistical independence is given by

$$E\left(T\right) = \frac{PQ}{N}. \tag{2.4}$$

[0]Parts of this section are to appear in Warrens, M.J. (in press), On the equivalence of Cohen's kappa and the Hubert-Ararbie adjusted Rand index, *Journal of Classification*.

Using (2.4), the expectation of S_{Rand} can be written as

$$E(S_{\text{Rand}}) = 1 + \frac{2PQ}{N^2} - \frac{P+Q}{N} \tag{2.5}$$

(Hubert and Arabie, 1985, p. 198). Using S_{Rand} and (2.5) in (2.1), we obtain the Hubert-Arabie adjusted Rand index, given by

$$CS_{\text{Rand}} = S_{\text{HA}} = \frac{T - PQ/N}{\frac{1}{2}(P+Q) - PQ/N} = \frac{2(NT - PQ)}{N(P+Q) - 2PQ}$$

(Hubert and Arabie, 1985, p. 198).

As noted in, for example, Steinley (2004) or Albatineh et al. (2006), the information in a matching table $\mathbf{N}$ of two clustering partitions on the same data points, can be summarized by a fourfold contingency table with quantities a, b, c, and d, where a is the number of object pairs that were placed in the same cluster according to both clustering methods, b (c) is the number of pairs that were placed in the same cluster according to one method but not according to the other, and d is the number of pairs that were not in the same cluster according to either of the methods. It then holds that $a + b + c + d = N$, where $a = T$, $b = P - T$, $c = Q - T$ and $d = N + T - P - Q$, and $p_1 = a + b = P$ and $q_1 = c + d = N - P$. The four different types of object pairs are also distinguished in Brennan and Light (1974), Hubert (1977), and Hubert and Arabie (1985, p. 194). However, the latter authors expressed their formulas in terms of the binomial coefficients in quantities T, P, Q, and N, instead of the quantities a, b, c, and d.

Expressing S_{Rand} in terms of the quantities a, b, c, and d we obtain S_{SM} (see, for example, Lerman, 1988; Steinley, 2004; Albatineh et al., 2006). Expressing S_{FM} in terms of the quantities a, b, c, and d we obtain S_{DK} (see, for example, Lerman, 1988; Albatineh et al., 2006). Expressing S_{HA} in these quantities, we obtain, following Steinley (2004, p. 388), the formula

$$S_{\text{HA}} = \frac{N(a+d) - [(a+b)(a+c) + (b+d)(c+d)]}{N^2 - [(a+b)(a+c) + (b+d)(c+d)]}. \tag{2.6}$$

The numerator of (2.6) can be written as

$$\begin{aligned} &N(a+d) - [(a+b)(a+c) + (b+d)(c+d)] \\ &= Na - p_1p_2 + Nd - q_1q_2 \\ &= 2(ad - bc) \end{aligned}$$

whereas the denominator of (2.6) equals

$$\begin{aligned} &N^2 - [(a+b)(a+c) + (b+d)(c+d)] \\ &= N^2 - p_1p_2 - q_1q_2 \\ &= (p_1+q_1)(p_2+q_2) - p_1p_2 - q_1q_2 \\ &= p_1q_2 + p_2q_1. \end{aligned}$$

Hence, expressing S_{HA} in terms of the quantities a, b, c, and d, the coefficient is equivalent to S_{Cohen}. Moreover, expectation $E(T)$ in (2.4) can be written as

$$E(T) = \frac{PQ}{N} = \frac{(a+b)(a+c)}{N} = \frac{p_1 p_2}{N}.$$

Hence, statistical independence under the generalized hypergeometric distribution function used in Hubert and Arabie (1985) for the matching table of two clusterings, is equivalent to the case of statistical independence under the binomial distribution function for the fourfold contingency table.

A practical conclusion is that we can calculate the Hubert-Arabie adjusted Rand index (S_{HA}) by first forming the fourfold contingency table counting the number of pairs of objects that were placed in the same cluster in both clusterings, in the same cluster in one clustering but in different clusters in the other clustering, and in different clusters in both, and then computing Cohen's kappa (S_{Cohen}) on this fourfold table.

2.3 Comparing two judges

A problem equivalent to that of comparing two partitions of two cluster algorithms may be encountered in psychology. In contrast to the case in Section 2.1, the categories are not defined in advance and the number of categories used by each psychologist may be different. Measures of agreement among judges in classifying answers to open-ended questions, or psychologists rating people, have been described by Brennan and Light (1974), Montgomery and Crittenden (1977), Hubert (1977), Janson and Vegelius (1982), and Popping (1983a). All these authors consider pairs of people and established for all N pairs formed from the m answers for both judges whether or not they were assigned to the same category. A comparison of the various measures is presented in Popping (1984).

We adopt the notation from Section 2.2, where quantities a, b, c, and d denote the four different types of pairs. Brennan and Light (1974) proposed the measure

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

which equals the Rand index S_{Rand} and the simple matching coefficient S_{SM}. Montgomery and Crittenden (1977) proposed the measure

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

which equals coefficient S_{Yule1} by Yule (1900). Hubert (1977) proposed a measure referred to as gamma, which is given by

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

Coefficient S_{Hub} is equal to a coefficient proposed by Hamann (1961) S_{Ham} and the $G-$index by Holley and Guilford (1964).

A discussion of properties of S_{Hub} and some adjustments to coefficient S_{Hub} can be found in Janson and Vegelius (1982). As an alternative to S_{Hub} these authors present a measure called the $J-$index. Popping (1983a, 1983b) proposed a measure based on the dot-product referred to as $D2$.

2.4 Quantitative variables

Let $\mathbf{x}_j$ and $\mathbf{x}_k$ be two column vectors of length n with positive entries, for example, counts or frequencies. In this section some examples of similarity coefficients formulated in terms of the elements of $\mathbf{x}_j$ and $\mathbf{x}_k$ are considered. Let x_{ij} denote the ith element of $\mathbf{x}_j$, and let x_{ik} denote the ith element of $\mathbf{x}_k$. In the terminology of Zegers (1986, p. 58) the measures considered in this section are coefficients for quantitative variables that consist of raw scores. These measures are either similarity functions or functions of the dissimilarity/distance type. Alternatively, one may formulate resemblance measures for normed raw scores, deviation scores, rank order scores, or combination of the previous scores. The reader is referred to Zegers (1986) and Gower and Legendre (1986) for more rigorous exposés on association coefficients for quantitative data.

The complement of the simple matching coefficient $1 - S_{\text{SM}}$ is a special case of the city-block or Manhattan distance

$$\frac{1}{n}\sum_{i=1}^{n}|x_{ij} - x_{ik}|.$$

The Jaccard (1912) coefficient

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

is obtained if in functions

$$\frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2 - \sum_{i=1}^{n} x_{ij}x_{ik}} \quad \text{or} \quad \frac{\sum_{i=1}^{n}\min(x_{ij}, x_{ik})}{\sum_{i=1}^{n}\max(x_{ij}, x_{ik})}$$

x_{ij} and x_{ik} take on values 1 and 0 only. The complement of the Jaccard coefficient S_{Jac} is a special case of

$$\frac{\sum_{i=1}^{n}|x_{ij} - x_{ik}|}{\sum_{i=1}^{n}\max(x_{ij}, x_{ik})} \quad \text{or} \quad \frac{\sum_{i=1}^{n}(x_{ij} - x_{ik})^2}{\sum_{i=1}^{n}\max(x_{ij}, x_{ik})}.$$

A member of a more general family of coefficients considered in Zegers and Ten Berge (1985) is given by

$$\frac{2\mathbf{x}_j^T\mathbf{x}_k}{\mathbf{x}_j^T\mathbf{x}_j + \mathbf{x}_k^T\mathbf{x}_k} = \frac{2\sum_{i=1}^{n} x_{ij}x_{ik}}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2}.$$

The latter coefficient is called the coefficient of identity and becomes S_{Gleas} if x_{ij} and x_{ik} take on values 1 and 0 only.

The measure

$$\frac{\sum_{i=1}^{n}|x_{ij}-x_{ik}|}{\sum_{i=1}^{n}(x_{ij}+x_{ik})} \quad \text{becomes} \quad 1-S_{\text{Gleas}}=\frac{b+c}{2a+b+c}$$

if x_{ij} and x_{ik} take on values 1 and 0 only, which is the complement of S_{Gleas} (Gower and Legendre, 1986, p. 27). Coefficient

$$\frac{\mathbf{x}_j^T\mathbf{x}_k}{(\mathbf{x}_j^T\mathbf{x}_j)^{1/2}(\mathbf{x}_k^T\mathbf{x}_k)^{1/2}}$$

is referred to as the coefficient of proportionality in Zegers and Ten Berge (1985), commonly known as Tucker's congruence coefficient (Tucker, 1951), also proposed by Burt (1948). The congruence coefficient for binary variables is given by $S_{\text{DK}} = a/\sqrt{p_j p_k}$. Three similarity coefficients, namely

$$\begin{aligned} S_{\text{Kul}} &= \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right) \\ S_{\text{Gleas}} &= \frac{2a}{p_j+p_k} \\ \text{and} \quad S_{\text{Sim}} &= \frac{a}{\min(p_j,p_k)} \end{aligned}$$

are sometimes attributed to Kulczyński (1927), Czekanowski (1932) and Simpson (1943). These authors proposed the coefficients for quantitative variables, which are given respectively by

$$\begin{aligned} S_{\text{Kul}} &= \frac{1}{2}\left[\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ij}}+\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ik}}\right] \\ S_{\text{Cze}} &= \frac{2\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}(x_{ij}+x_{ik})} \\ \text{and} \quad S_{\text{Sim}} &= \max\left[\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ij}},\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ik}}\right]. \end{aligned}$$

Sepkoski (1974) argues that, although similarity coefficients have been widely employed in cluster analysis, their use has been, for the most part, restricted to binary data. This author proposed quantified coefficients using basic rules like

$$\begin{aligned} a &= \frac{1}{n}\sum_{i=1}^{n}\min(x_{ij},x_{ik}) \\ b+c &= \frac{1}{n}\sum_{i=1}^{n}[\max(x_{ij},x_{ik})-\min(x_{ij},x_{ik})] \\ p_j &= \frac{1}{n}\sum_{i=1}^{n}x_{ij} \quad \text{and} \quad p_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}. \end{aligned}$$

The similarity coefficient used by Robinson (1951) can be written as

$$S_{\text{Rob}} = 1 - \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} - \frac{x_{ik}}{\sum_{i=1}^{n} x_{ik}}\right|.$$

When the data are binary, S_{Rob} becomes

$$S_{\text{BB}} = \frac{a_{jk}}{\max(p_j, p_k)} \qquad \text{(Braun-Blanquet, 1932).}$$

Proposition 2.1. *If S_{Rob} is applied to binary (1/0) data, then $S_{\text{Rob}} = S_{\text{BB}}$.*
Proof: For $p_j \geq p_k$, S_{Rob} can be written as

$$S_{\text{Rob}} = 1 - \frac{1}{2}\left(\frac{a_{jk}}{p_k} - \frac{a_{jk}}{p_j} + \frac{p_j - a_{jk}}{p_j} + \frac{p_k - a_{jk}}{p_k}\right) = \frac{1}{2} - \frac{p_j - 2a_{jk}}{2p_j} = \frac{a_{jk}}{p_j}.$$

Furthermore, for $p_j \leq p_k$, S_{Rob} can be written as

$$S_{\text{Rob}} = 1 - \frac{1}{2}\left(\frac{a_{jk}}{p_j} - \frac{a_{jk}}{p_k} + \frac{p_j - a_{jk}}{p_j} + \frac{p_k - a_{jk}}{p_k}\right) = \frac{a_{jk}}{p_k}.$$

This completes the proof. □

2.5 Measures from set theory

Similarity and distance functions can also be defined on sets of arbitrary elements. The following notation is used. Let a set be denoted by A and let $\overline{A}$ denote its complement. Symbol $\cup$ denotes union or set sum, and $A \cup B$ is the set containing everything in either A or B or both. Also, $\cap$ denotes intersection or set product, and $A \cap B$ is the set containing just those elements common to both A and B. Furthermore, let $|A|$ denote the cardinality of set A, which is a measure of the number of elements of the set. Some examples of similarity coefficients for two sets A and B that are frequently used, are

$$\frac{2|A \cap B|}{|A| + |B|} \quad \text{Dice coefficient}$$

$$\frac{|A \cap B|}{|A \cup B|} \quad \text{Jaccard coefficient}$$

$$\frac{|A \cap B|}{|A|^{1/2}|B|^{1/2}} \quad \text{Cosine coefficient}$$

$$\text{and} \quad \frac{|A \cap B|}{\min(|A|, |B|)} \quad \text{Overlap coefficient.}$$

Special cases of these measures are the respective similarity coefficients

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2},\ S_{\text{Jac}} = \frac{a}{a + b + c},\ S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \quad \text{and} \quad S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)}.$$

Restle (1959) studied the symmetric set difference

$$|(A \cup B) \cap (\overline{A \cap B})|$$

which is a more general form of the complement of the simple matching coefficient, $1 - S_{\text{SM}}$. Boorman and Arabie (1972) discuss several set-theoretical measures, including the minimum lattice-moves distance

$$|A| + |B| - 2|A \cap B|$$

which is equivalent to the above measure studied by Restle (1959), and the minimum set-moves distance which may be approximated by

$$|A \cap B| - \min(|A|, |B|).$$

2.6 Epilogue

In this second chapter, various general formulas from different domains of data analysis were considered. Some of the similarity coefficients for binary data considered throughout this thesis are special cases of these formulas. The chapter puts the coefficients for binary variables in a broader perspective. Furthermore, the more general formulas provide some ideas for possible generalizations of various results in this thesis. The thesis by Zegers (1986) is a good source for the vast amount of different contexts in which similarity coefficients may be considered.

It was shown that several similarity measures used in cluster analysis for the matching table of two clustering algorithms are in fact equivalent to similarity coefficients defined on the four dependent quantities from the 2×2 contingency table, after a simple recoding. Two well-known measures are the Rand index and the Hubert-Arabie adjusted Rand index, given respectively by

$$S_{\text{Rand}} = 1 - \frac{P + Q - 2T}{N} \quad \text{and} \quad S_{\text{HA}} = \frac{2(NT - PQ)}{N(P + Q) - 2PQ}.$$

Both measures are calculated using the information in the matching of two clusterings on the same data points. Coefficient S_{Rand} was also proposed by Brennan and Light (1974) for comparing ratings by two psychologists. If the Rand index S_{Rand} is formulated in terms of the quantities a, b, c, and d, it is equivalent to the simple matching coefficient S_{SM}. Furthermore, if the Hubert-Arabie adjusted Rand index S_{HA} is formulated in terms of the quantities a, b, c, and d, it is equivalent to Cohen's kappa for two categories (S_{Cohen}).

Interestingly, both Cohen (1960) and Hubert and Arabie (1985) proposed a similarity measure that has been, or still is, the preferred coefficient, or at least the best-known coefficient, in their particular domain of data analysis (respectively interrater reliability and cluster analysis). Moreover, both measures were proposed in response to, or as alternative to, earlier coefficients (Scott, 1955, in the case of Cohen, 1960; Morey and Agresti, 1984, in the case of Hubert and Arabie, 1985).

CHAPTER 3

Coefficient families

In this chapter it is shown how various similarity coefficients may be related. Similarity measures may be members of some sort of parameter family or can be related in the sense that several coefficients have a similar form. Various well-known coefficients belong to parameter families of which all members are fractions, linear in both numerator and denominator. A distinction is made between coefficients that do include the quantity d (representing negative matches), like

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \quad \text{and} \quad S_{\text{Ham}} = \frac{a-b-c+d}{a+b+c+d} \quad \text{(Hamann, 1961)}$$

and those that do not include the quantity d, like

$$S_{\text{Jac}} = \frac{a}{p_1+p_2-a} \quad \text{and} \quad S_{\text{Gleas}} = \frac{2a}{p_1+p_2}.$$

A variety of similarity coefficients can be defined as some sort of mean value of two different quantities. For example, resemblance measures S_{Gleas} and

$$S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \quad \text{and} \quad S_{\text{Kul}} = \frac{a(p_1+p_2)}{2p_1p_2}$$

are respectively the harmonic, geometric and arithmetic mean of the conditional probabilities $p_1^{-1}a$ and $p_2^{-1}a$.

Different types of coefficients may be obtained by considering abstractions of these Pythagorean means. One type of generalized mean that is considered in this chapter is the so-called power mean.

A very general family of coefficients is the class of all functions of the form $\lambda+\mu a$, where a is the proportion of 1s that two variables share in the same positions, and λ and μ are functions of p_1 and p_2 only. This family includes coefficients S_{Gleas}, S_{DK}, and S_{Kul} and various other measures. Properties of this family with respect to correction for similarity due to chance, are considered in Chapter 4.

There are some advantages to studying families of coefficients instead of individual coefficients. First of all, from the family formulation it is often apparent how different members are related. Coefficient properties like bounds are easily investigated using parameter families. Another advantage of studying parameter families instead of individual coefficients, is that often more general results can be obtained. As an example, results on linearity given in Hubálek (1982) for individual coefficients are here studied for families of coefficients.

3.1 Parameter families

Gower and Legendre (1986, p. 13) define two parameter families of which all members are linear in both numerator and denominator. They make a distinction between coefficients that do and do not include the quantity d. The first family for presence/absence data is given by

$$S_{\text{GL1}}(\theta) = \frac{a}{a+\theta(b+c)} = \frac{a}{\theta(p_1+p_2)+(1-2\theta)a}.$$

where $\theta > 0$ to avoid negative values. Members of $S_{\text{GL1}}(\theta)$ are

$$\begin{aligned} S_{\text{GL1}}(\theta=1) &= S_{\text{Jac}} = \frac{a}{p_1+p_2-a} \\ S_{\text{GL1}}(\theta=1/2) &= S_{\text{Gleas}} = \frac{2a}{p_1+p_2} \\ S_{\text{GL1}}(\theta=2) &= S_{\text{SS1}} = \frac{a}{a+2(b+c)} \quad \text{(Sokal and Sneath, 1963).} \end{aligned}$$

Members with $0 < \theta < 1$ give more weight to a. With presence/absence data this is regularly done in the case that there are only a few positive matches relatively to the number of mismatches, that is, a is much smaller than $(b+c)$. Similar arguments can be used for the opposite case and $\theta > 1$.

All members of $S_{\text{GL1}}(\theta)$ are bounded by 0 and 1, that is, $0 \le S_{\text{GL1}}(\theta) \le 1$. In addition, members are bounds of each other:

$$0 \le S_{\text{SS1}} \le S_{\text{Jac}} \le S_{\text{Gleas}} \le 1$$

or more generally

$$S_{\text{GL1}}(\theta_1) \le S_{\text{GL1}}(\theta_2) \quad \text{for} \quad \theta_1 > \theta_2 > 0.$$

The formulation of $S_{\mathrm{GL1}}(\theta)$ (and that of $S_{\mathrm{GL2}}(\theta)$ below) is closely related to the concept of global order equivalence (Sibson, 1972; Batagelj and Bren, 1995). Let $S(a, b, c, d)$ denote a function of the quantities a, b, c, and d. Two coefficients S and S^* are said to be globally order equivalent if

$$\begin{array}{rrcl} & S(a_1, b_1, c_1, d_1) & > & S(a_2, b_2, c_2, d_2) \\ \text{if and only if} & S^*(a_1, b_1, c_1, d_1) & > & S^*(a_2, b_2, c_2, d_2). \end{array}$$

If two coefficients are globally order equivalent, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations (see, for example, Gower, 1986; Batagelj and Bren, 1995).

Theorem 3.1. *Two members of $S_{\mathrm{GL1}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL1}}(\theta)$, we have

$$\begin{aligned} \frac{a_1}{a_1 + \theta(b_1 + c_1)} &> \frac{a_2}{a_2 + \theta(b_2 + c_2)} \\ a_1 a_2 + a_1\theta(b_2 + c_2) &> a_1 a_2 + a_2\theta(b_1 + c_1) \\ \frac{a_1}{b_1 + c_1} &> \frac{a_2}{b_2 + c_2}. \end{aligned}$$

Since an ordinal comparison with respect to $S_{\mathrm{GL1}}(\theta)$ does not depend on the value of θ, any two members of $S_{\mathrm{GL1}}(\theta)$ are globally order equivalent. □

Janson and Vegelius (1981) pointed out an interesting relationship between various members of $S_{\mathrm{GL1}}(\theta)$. With respect to S_{Gleas}, S_{Jac}, and S_{SS1}, we have

$$S_{\mathrm{Jac}} = \frac{S_{\mathrm{Gleas}}}{2 - S_{\mathrm{Gleas}}} \qquad \text{and} \qquad S_{\mathrm{SS1}} = \frac{S_{\mathrm{Jac}}}{2 - S_{\mathrm{Jac}}}.$$

In general we have the following result.

Proposition 3.1. *It holds that*

$$S_{\mathrm{GL1}}(2\theta) = \frac{S_{\mathrm{GL1}}(\theta)}{2 - S_{\mathrm{GL1}}(\theta)}.$$

Proof: Define $x = a + \theta(b + c)$. Then

$$\frac{S_{\mathrm{GL1}}(\theta)}{2 - S_{\mathrm{GL1}}(\theta)} = \frac{x^{-1}a}{x^{-1}(2x - a)} = S_{\mathrm{GL1}}(2\theta). \quad \square$$

A parameter family closely related to $S_{\mathrm{GL1}}(\theta)$ may be obtained using the transformation $2S - 1$, that is,

$$S_{\mathrm{GL3}}(\theta) = \frac{2a}{a + \theta(b + c)} - 1 = \frac{a - \theta(b + c)}{a + \theta(b + c)}$$

with $\theta > 0$. A member of $S_{\mathrm{GL3}}(\theta)$ is

$$S_{\mathrm{GL3}}(\theta = 1/2) = S_{\mathrm{NS1}} = \frac{2a - b - c}{2a + b + c} \qquad \text{(No source)}.$$

Members with $0 < \theta < 1$ give more weight to a. All members of $S_{\mathrm{GL3}}(\theta)$ are bounded by -1 and 1, that is, $-1 \leq S_{\mathrm{GL3}}(\theta) \leq 1$. Parameter family $S_{\mathrm{GL3}}(\theta)$ is a transformation that preserves the scale of $S_{\mathrm{GL1}}(\theta)$ but uses a different range. The value zero for $S_{\mathrm{GL3}}(\theta)$ is equal to the value 0.5 for $S_{\mathrm{GL1}}(\theta)$ for fixed θ. For example, we have

$$S_{\mathrm{Gleas}} = \frac{2a}{2a + b + c} = 0.5 \quad \text{if and only if} \quad 2a = b + c$$

and

$$S_{\mathrm{NS1}} = \frac{2a - b - c}{2a + b + c} = 0 \quad \text{if and only if} \quad 2a = b + c.$$

The zero value case of coefficient S_{NS1} is not the same as the zero value case for coefficients with the covariance $ad - bc$ in the numerator. Two variables are not necessarily statistically independent if $S_{\mathrm{NS1}} = 0$ (Section 1.3). The formulation of $S_{\mathrm{GL3}}(\theta)$ is not completely arbitrary, because it is related to $S_{\mathrm{GL1}}(\theta)$ by the concept of global order equivalence.

Proposition 3.2. *Two members of $S_{\mathrm{GL3}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL3}}(\theta)$, we have

$$\frac{a_1 - \theta(b_1 + c_1)}{a_1 + \theta(b_1 + c_1)} > \frac{a_2 - \theta(b_2 + c_2)}{a_2 + \theta(b_2 + c_2)} \quad \text{if and only if} \quad \frac{a_1}{b_1 + c_1} > \frac{a_2}{b_2 + c_2}.$$

Since an ordinal comparison with respect to $S_{\mathrm{GL3}}(\theta)$ does not depend on the value of θ, any two members of $S_{\mathrm{GL3}}(\theta)$ are globally order equivalent. □

Corollary 3.1 *Members of $S_{\mathrm{GL1}}(\theta)$ and $S_{\mathrm{GL3}}(\theta)$ are globally order equivalent.*

The second family in Gower and Legendre (1986, p. 13), the counterpart of $S_{\mathrm{GL1}}(\theta)$ for nominal data, is given by

$$S_{\mathrm{GL2}}(\theta) = \frac{a + d}{a + \theta(b + c) + d} = \frac{1 + 2a - p_1 - p_2}{1 + (\theta - 1)(p_1 + p_2) + 2a(1 - \theta)}$$

where $\theta > 0$ to avoid negative values.

Members of $S_{\mathrm{GL2}}(\theta)$ are

$$S_{\mathrm{GL2}}(\theta = 1) = S_{\mathrm{SM}} = \frac{a+d}{a+b+c+d} = a+d$$

$$S_{\mathrm{GL2}}(\theta = 1/2) = S_{\mathrm{SS2}} = \frac{2(a+d)}{2a+b+c+2d} = \frac{2(a+d)}{1+a+d}$$

(Sokal and Sneath, 1963)

$$S_{\mathrm{GL2}}(\theta = 2) = S_{\mathrm{RT}} = \frac{a+d}{a+2(b+c)+d} = \frac{a+d}{1+b+c}$$

(Rogers and Tanimoto, 1960).

Similar to $S_{\mathrm{GL1}}(\theta)$, the members of $S_{\mathrm{GL2}}(\theta)$ are bounded by 0 and 1, that is, $0 \leq S_{\mathrm{GL2}}(\theta) \leq 1$. Also, members with $0 < \theta < 1$ give more weight to $(a+d)$.

Theorem 3.2. *Two members of $S_{\mathrm{GL2}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL2}}(\theta)$, we have

$$\frac{a_1+d_1}{a_1+\theta(b_1+c_1)+d_1} > \frac{a_2+d_2}{a_2+\theta(b_2+c_2)+d_2}$$
$$\frac{a_1+d_1}{b_1+c_1} > \frac{a_2+d_2}{b_2+c_2}.$$

Since an ordinal comparison with respect to $S_{\mathrm{GL2}}(\theta)$ does not depend on the value of θ, any two members of $S_{\mathrm{GL2}}(\theta)$ are globally order equivalent. □

Families $S_{\mathrm{GL1}}(\theta)$ and $S_{\mathrm{GL2}}(\theta)$ are related in the following way.

Proposition 3.3. *It holds that $S_{\mathrm{GL2}}(\theta) \geq S_{\mathrm{GL1}}(\theta)$.*
Proof: $S_{\mathrm{GL2}}(\theta) \geq S_{\mathrm{GL1}}(\theta)$ if and only if $\theta d(b+c) \geq 0$. □

Similar to S_{Gleas}, S_{Jac}, and S_{SS1}, we have with respect to S_{SS2}, S_{SM}, and S_{RT}

$$S_{\mathrm{SM}} = \frac{S_{\mathrm{SS2}}}{2 - S_{\mathrm{SS2}}} \qquad \text{and} \qquad S_{\mathrm{RT}} = \frac{S_{\mathrm{SM}}}{2 - S_{\mathrm{SM}}}.$$

In general we have the following result.

Proposition 3.4. *It holds that*

$$S_{\mathrm{GL2}}(2\theta) = \frac{S_{\mathrm{GL2}}(\theta)}{2 - S_{\mathrm{GL2}}(\theta)}.$$

Proof: Define $x = a + \theta(b+c) + d$. Then

$$\frac{S_{\mathrm{GL2}}(\theta)}{2 - S_{\mathrm{GL2}}(\theta)} = \frac{x^{-1}(a+d)}{x^{-1}(2x-a-d)} = S_{\mathrm{GL1}}(2\theta). \quad \square$$

A parameter family closely related to $S_{\mathrm{GL2}}(\theta)$ may be obtained using the transformation $2S - 1$,

$$S_{\mathrm{GL4}}(\theta) = \frac{2(a+d)}{a+\theta(b+c)+d} - 1 = \frac{a-\theta(b+c)+d}{a+\theta(b+c)+d}$$

with $\theta > 0$. A member of $S_{\mathrm{GL4}}(\theta)$ is

$$S_{\mathrm{GL4}}(\theta = 1) = S_{\mathrm{Ham}} = \frac{a-b-c+d}{a+b+c+d} = a-b-c+d \quad \text{(Hamann, 1961).}$$

Members with $0 < \theta < 1$ give more weight to $(a+d)$. We have

$$S_{\mathrm{SM}} = a + d = 0.5 \quad \text{if and only if} \quad a + d = b + c$$

and

$$S_{\mathrm{Ham}} = a - b - c + d = 0 \quad \text{if and only if} \quad a + d = b + c.$$

The zero value case of coefficient S_{Ham} is not the same as the zero value case for coefficients with the covariance $ad - bc$ in the numerator (Section 1.3), nor the zero value case of S_{NS1}. Two variables are not necessarily independent if $S_{\mathrm{Ham}} = 0$. The formulation of $S_{\mathrm{GL4}}(\theta)$ is not completely arbitrary, since it is related to $S_{\mathrm{GL2}}(\theta)$ by the concept of global order equivalence.

Proposition 3.5. *Two members of $S_{\mathrm{GL4}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL4}}(\theta)$, we have

$$\frac{a_1 - \theta(b_1 + c_1) + d_1}{a_1 + \theta(b_1 + c_1) + d_1} > \frac{a_2 - \theta(b_2 + c_2) + d_2}{a_2 + \theta(b_2 + c_2) + d_2}$$
$$\frac{a_1 + d_1}{b_1 + c_1} > \frac{a_2 + d_2}{b_2 + c_2}.$$

Since an ordinal comparison with respect to $S_{\mathrm{GL4}}(\theta)$ does not depend on the value of θ, any two members of $S_{\mathrm{GL4}}(\theta)$ are globally order equivalent. □

Corollary 3.2 *Members of $S_{\mathrm{GL2}}(\theta)$ and $S_{\mathrm{GL4}}(\theta)$ are globally order equivalent.*

3.2 Power means

There are several functions that may reflect the mean value of two real positive values x and y. The harmonic, geometric and arithmetic means, also known as the Pythagorean means, are given by respectively

$$\frac{2}{x^{-1}+y^{-1}}, \quad \sqrt{xy} \quad \text{and} \quad \frac{x+y}{2}.$$

Several coefficients can be expressed in terms of these Pythagorean means. For example, consider the quantities

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

(Dice, 1945; Post and Snijders, 1993). The harmonic, geometric and arithmetic means of the quantities S_{Dice1} and S_{Dice2} are respectively

$$S_{\text{Gleas}} = \frac{2a}{p_1+p_2}, \; S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \quad \text{and} \quad S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1}+\frac{a}{p_2}\right).$$

Different types of coefficients may be obtained by considering abstractions of the Pythagorean means. One type of so-called generalized means is the power mean, sometimes referred to as the Hölder mean (see, for example, Bullen, 2003, Chapter 3). Let θ be a real value. The power mean $M_\theta(x,y)$ of x and y is then given by

$$M_\theta(x,y) = \left(\frac{x^\theta + y^\theta}{2}\right)^{1/\theta}.$$

Special cases of $M_\theta(x,y)$ are

$$\begin{aligned}
\lim_{\theta\to-\infty} M_\theta(x,y) &= \min(x,y) && \text{(minimum)}\\
M_{-1}(x,y) &= \frac{2}{x^{-1}+y^{-1}} && \text{(harmonic mean)}\\
\lim_{\theta\to 0} M_\theta(x,y) &= \sqrt{xy} && \text{(geometric mean)}\\
M_1(x,y) &= \frac{x+y}{2} && \text{(arithmetic mean)}\\
\lim_{\theta\to\infty} M_\theta(x,y) &= \max(x,y) && \text{(maximum).}
\end{aligned}$$

[0]Parts of this section are to appear in Warrens, M.J. (in press), Bounds of resemblance measures for binary (presence/absence) variables, *Journal of Classification*.

A variety of coefficients turn out to be special cases of a power mean. In terms of S_{Dice1} and S_{Dice2} we characterize the following coefficients from the literature.

$$\begin{aligned}
S_{\text{BB}} &= \frac{a}{\max(p_1, p_2)} && \text{(minimum; Braun-Blanquet, 1932)} \\
S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} && \text{(harmonic mean)} \\
S_{\text{DK}} &= \frac{a}{\sqrt{p_1 p_2}} && \text{(geometric mean; Driver and Kroeber, 1932)} \\
S_{\text{Kul}} &= \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) && \text{(arithmetic mean; Kulczyński, 1927)} \\
S_{\text{Sim}} &= \frac{a}{\min(p_1, p_2)} && \text{(maximum; Simpson, 1943).}
\end{aligned}$$

The product of the two quantities (or the square of the geometric mean S_{DK}) is not a special case of a power mean. It is given by

$$S_{\text{Sorg}} = \frac{a^2}{p_1 p_2} \quad \text{(Sorgenfrei, 1958; Cheetham and Hazel, p. 1131).}$$

Coefficient S_{Sorg} is sometimes referred to as the correlation ratio. The various coefficients for presence/absence data (without the quantity d) are related in the following way.

Proposition 3.6. *It holds that*

$$0 \leq S_{\text{Sorg}} \overset{(i)}{\leq} S_{\text{Jac}} \overset{(ii)}{\leq} S_{\text{BB}} \leq S_{\text{Gleas}} \leq S_{\text{DK}} \leq S_{\text{Kul}} \leq S_{\text{Sim}} \leq 1.$$

Proof: Inequality (i) holds if and only if $p_1 p_2 \geq a(a+b+c)$ if and only if $bc \geq 0$. Inequality (ii) holds if and only if $b + c \geq \max(b, c)$. The remaining inequalities follow from a property of a power mean:

$$M_{\theta_1}\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \leq M_{\theta_2}\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{for} \quad \theta_1 < \theta_2. \qquad \square$$

As a second example of a power mean, consider the quantities

$$S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1} \quad \text{(Cole, 1949).}$$

The quantity $(ad - bc)$ is known as the covariance between two binary vectors. If $p_1 \leq p_2$ then $p_1 q_2$ is the maximum value of the covariance $(ad - bc)$ given the marginal proportions. Note that the covariance may become negative and strictly speaking we have defined the power mean for two real positive values only. However, as it turns out, the power mean of two real negative values has very similar properties as the power mean of two positive values. As long as the two values have the same sign, the distinction between positive and negative values appears not to be important.

With respect to S_{Cole1} and S_{Cole2} we have the special cases

$$\begin{aligned} S_{\text{Cohen}} &= \frac{2(ad-bc)}{p_1q_2+p_2q_1} && \text{(harmonic mean)} \\ S_{\text{Phi}} &= \frac{ad-bc}{\sqrt{p_1p_2q_1q_2}} && \text{(geometric mean)} \\ S_{\text{Loe}} &= \frac{ad-bc}{\min(p_1q_2, p_2q_1)} && \text{(maximum; Loevinger, 1947, 1948).} \end{aligned}$$

Coefficient S_{Loe} is attributed to Loevinger (1947, 1948) by Mokken (1971) and Sijtsma and Molenaar (2002). However, Krippendorff (1987) reports Benini (1901) as probably the first to put forward this coefficient. Some new properties of this coefficient are considered in Chapter 5. Similar to Proposition 3.6, the next result follows from a property of power means, more specifically the harmonic-geometric mean inequality.

Proposition 3.7. *It holds that*

$$0 \leq |S_{\text{Cohen}}| \leq |S_{\text{Phi}}| \leq |S_{\text{Loe}}| \leq 1.$$

3.3 A general family

Albatineh et al. (2006) define yet another way on how various coefficients can be related. These authors study correction for chance with respect to a family $\mathcal{L}$ of the form $\lambda + \mu x$. Coefficients in the $\mathcal{L}$ family are linear functions of the quantity x, and the expectation of $S = \lambda + \mu x$ depends on the quantity x only, that is, $E(S) = \lambda + \mu E(x)$. Properties of the $\mathcal{L}$ family with respect to correction for chance are considered in the next chapter. For the moment it will be shown that $\mathcal{L}$ defines a very general family.

For example, coefficients in Section 2.1 belong to $\mathcal{L}$ family. Using $x = P_{\text{o}}$ we have

$$S_{\text{SM}} = P_{\text{o}} \;\rightarrow\; \lambda = 0 \quad \text{and} \quad \mu = 1$$

$$S_{\text{Scott}} \;\rightarrow\; \lambda = \frac{-E(P_{\text{o}})_{\text{Scott}}}{1 - E(P_{\text{o}})_{\text{Scott}}} \quad \text{and} \quad \mu = \frac{1}{1 - E(P_{\text{o}})_{\text{Scott}}}$$

$$\text{and} \quad S_{\text{Cohen}} \;\rightarrow\; \lambda = \frac{-E(P_{\text{o}})_{\text{Cohen}}}{1 -_E (P_{\text{o}})_{\text{Cohen}}} \quad \text{and} \quad \mu = \frac{1}{1 - E(P_{\text{o}})_{\text{Cohen}}}.$$

As a second example, take $x = a$, the proportion of 1s that two binary variables share in the same positions, and λ and μ are functions of p_1 and p_2 only. Then we

have

$$\begin{aligned} S_{\mathrm{SM}} &= a + d \\ &= 1 + 2a - p_1 - p_2 && \to \lambda = 1 - p_1 - p_2,\ \mu = 2 \\ S_{\mathrm{Ham}} &= a - b - c + d \\ &= 2a + 1 - 2p_1 - 2p_2 && \to \lambda = 1 - 2p_1 - 2p_2,\ \mu = 2 \\ \text{and} \quad S_{\mathrm{Gleas}} &= \frac{2a}{p_1 + p_2} && \to \lambda = 0,\ \mu = \frac{2}{p_1 + p_2}. \end{aligned}$$

In Proposition 3.8 it is shown that the power mean of the quantities S_{Dice1} and S_{Dice2}, and the power mean of S_{Cole1} and S_{Cole2} are in the $\mathcal{L}$ family.

Proposition 3.8. *Power means*

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{and} \quad M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_2q_1}\right)$$

are members of the $\mathcal{L}$ family.
Proof:

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) = \left[\frac{a^\theta(p_1^\theta + p_2^\theta)}{2p_1^\theta p_2^\theta}\right]^{1/\theta} = \frac{a}{p_1p_2}\left[\frac{p_1^\theta + p_2^\theta}{2}\right]^{1/\theta}.$$

Thus, for

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{we have} \quad \mu = \frac{1}{p_1p_2}\left(\frac{p_1^\theta + p_2^\theta}{2}\right)^{1/\theta}.$$

Similarly, for

$$M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_2q_1}\right)$$

we have

$$\mu = \frac{1}{p_1p_2q_1q_2}\left[\frac{(p_1q_2)^\theta + (p_2q_1)^\theta}{2}\right]^{1/\theta} \quad \text{and} \quad \lambda = -\frac{1}{q_1q_2}\left[\frac{(p_1q_2)^\theta + (p_2q_1)^\theta}{2}\right]^{1/\theta}$$

because $ad - bc = a - p_1p_2$. □

Let $f(p_1, p_2)$ be a function of the marginals p_1 and p_2. Then, all coefficients of the form

$$\frac{a}{f(p_1,p_2)} \quad \text{or} \quad \frac{ad-bc}{f(p_1,p_2)} = \frac{a - p_1p_2}{f(p_1,p_2)}$$

belong to the $\mathcal{L}$ family. Examples are

$$\begin{aligned} S_{\mathrm{RR}} &= \frac{a}{a+b+c+d} \\ S_{\mathrm{MP}} &= \frac{2(ad-bc)}{p_1q_1 + p_2q_2} && \text{(Maxwell and Pilliner, 1968)} \\ \text{and} \quad S_{\mathrm{Fleiss}} &= \frac{(ad-bc)(p_1q_1 + p_2q_2)}{2p_1q_2p_2q_1} && \text{(Fleiss, 1975).} \end{aligned}$$

Moreover, if two coefficients $S_1 = \lambda_1 + \mu_1 a$ and $S_2 = \lambda_2 + \mu_2 a$ are in $\mathcal{L}$, then the arithmetic mean

$$\frac{S_1 + S_2}{2} = \frac{\lambda_1 + \mu_1 a + \lambda_2 + \mu_2 a}{2} = \frac{\lambda_1 + \lambda_2}{2} + \frac{a(\mu_1 + \mu_2)}{2}$$

is also in $\mathcal{L}$. Finally, if $S_1 = \lambda + \mu a$ is in the $\mathcal{L}$ family, then

$$S_2 = 2S_1 - 1 = 2\lambda - 1 + 2\mu a$$

also belongs to $\mathcal{L}$.

3.4 Linearity

Instead of proportions, let a, b, c, and d be the number of 1s and 0s that two binary variables may share or not share in the same positions. Furthermore, let $S(a)$ be short for $S(a, b, c, d)$ (S is a function of quantities a, b, c and d) and let $S(a+1)$ be short for $S(a+1, b-1, c-1, d+1)$. Hubálek (1982) gives the following definition of linearity. A function $S(a)$ is called linear if

$$S(a+1) - S(a) = S(a+2) - S(a+1),$$

or equivalently, if

$$2 \times S(a+1) = S(a+2) + S(a).$$

Using this definition of linearity, non-linearity can be defined in two ways. A function $S(a)$ is called convex if $2 \times S(a+1) < S(a+2) + S(a)$; $S(a, b, c, d)$ is called concave if $2 \times S(a+1) > S(a+2) + S(a)$.

Using numerical examples, Hubálek (1982) determined for various coefficients which ones are linear and which are non-linear. In this section the above definition of linearity is studied for several parameter families, instead of individual coefficients. The result below concerns coefficients that are rational functions, linear in both numerator and denominator.

Let $x = f(a, d)$ denote a linear function of a and d, and let $y = g(b, c)$ denote a linear function of b and c. Furthermore, let

$$u = \begin{cases} 1 & \text{if } x \text{ is a function of } a \text{ or } d \text{ only} \\ 2 & \text{if } x \text{ is a function of both } a \text{ and } d \end{cases}$$

and let

$$v = \begin{cases} 1 & \text{if } y \text{ is a function of } b \text{ or } c \text{ only} \\ 2 & \text{if } y \text{ is a function of both } b \text{ and } c. \end{cases}$$

Proposition 3.9. *Parameter families of the form*

$$(i) \qquad S(x,y) = \frac{x}{x+y} \qquad \left(\textit{with} \quad S(x+u,y-v) = \frac{x+u}{x+u+y-v} \right)$$

and

$$(ii) \qquad S(x,y) = \frac{x-y}{x+y} \qquad \left(\textit{with} \quad S(x+u,y-v) = \frac{x+u-y+v}{x+u+y-v} \right)$$

are convex for $u < v$, linear for $u = v$, and concave for $u > v$.
Proof: We consider (i) first. Using (i) in $2 \times S(a+1) \leq S(a+2) + S(a)$ we obtain

$$\frac{2(x+u)}{x+u+y-v} \leq \frac{x+2u}{x+2u+y-2v} + \frac{x}{x+y}. \tag{3.1}$$

Bringing all fractions under the same denominator, (3.1) becomes

$$\begin{aligned} (x+y)(2x+2u)(x+2u+y-2v) \;&\leq\; (x+y)(x+2u)(x+u+y-v) \\ &+\; x(x+u+y-v)(x+2u+y-2v) \end{aligned}$$

which, after some algebra, equals

$$(x+y)(x^2+3ux+xy-3vx+2u^2-2uv) \leq x(x+u+y-v)(x+2u+y-2v)$$

which, after some more algebra, can be written as $u^2y + uvx \leq uvy + v^2x$ if and only if $u \leq v$.

Next, we consider (ii). Parameter families (i) and (ii) are related by

$$\frac{x-y}{x+y} \;=\; \frac{2x}{x+y} - 1. \tag{3.2}$$

Using (3.2) in $2 \times S(a+1) \leq S(a+2) + S(a)$ we obtain

$$\frac{4(x+u)}{x+u+y-v} - 2 \;\leq\; \frac{2(x+2u)}{x+2u+y-2v} + \frac{2x}{x+y} - 2$$

which equals (3.1). □

Corollary 3.3. *Parameter families*

$$S_{\mathrm{GL1}}(\theta) = \frac{a}{a+\theta(b+c)} \quad \textit{and} \quad S_{\mathrm{GL3}}(\theta) = \frac{a-\theta(b+c)}{a+\theta(b+c)}$$

are convex for $\theta > \frac{1}{2}$, linear for $\theta = \frac{1}{2}$, and concave for $0 < \theta < \frac{1}{2}$.
Proof: With respect to these families we have $x = a$ and $y = \theta(b+c)$, and hence $u = 1$ and $v = 2\theta$. The family is then convex if $1 < 2\theta$. □

Corollary 3.4. *Parameter families*

$$S_{\mathrm{GL2}}(\theta) = \frac{a+d}{a+\theta(b+c)+d} \quad \textit{and} \quad S_{\mathrm{GL4}}(\theta) = \frac{a-\theta(b+c)+d}{a+\theta(b+c)+d}$$

are convex for $\theta > 1$, linear for $\theta = 1$, and concave for $0 < \theta < 1$.
Proof: For these families $u = 2$ and $v = 2\theta$. The families are then convex if $2 < 2\theta$. □

3.5 Epilogue

In this chapter it was shown how various similarity coefficients may be related. Similarity measures may be members of some sort of parameter family or can be related in the sense that several coefficients have a similar form. Various well-known coefficients belong to parameter families of which all members are rational functions, linear in both numerator and denominator. Some coefficients are members of more than one family. As an example, consider

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

Coefficient S_{Gleas} is the harmonic mean of

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

and is therefore a special case of a power mean. In addition, S_{Gleas} is a member ($\theta = 1/2$) of the family given by

$$S_{\text{GL1}}(\theta) = \frac{a}{a + \theta(b + c)}.$$

Due to this double membership, S_{Gleas} is a key coefficient in Chapter 16, where various multivariate formulations of coefficients are presented. In terms of linearity as defined by Hubálek (1982), S_{Gleas} is the linear coefficient in family $S_{\text{GL1}}(\theta)$. For other values than $\theta = 1/2$ we obtain either convex or concave coefficients. With respect to the linearity,

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

is the linear coefficient in the second family of rational functions, $S_{\text{GL2}}(\theta)$. Similar to S_{Gleas}, S_{SM} can be introduced as a special case of a power mean. For example, S_{SM} is equal to the harmonic mean of the quantities

$$\frac{a + d}{p_1 + q_2} \quad \text{and} \quad \frac{a + d}{p_2 + q_1}.$$

Both S_{Gleas} and S_{SM} can be written as linear functions of the quantity a and are therefore members in the $\mathcal{L}$ family. Some of the consequences of this property are studied in the next chapter: S_{Gleas} and S_{SM} become equivalent after correction for chance. Moreover given a certain expectation of the quantity a, S_{Gleas} and S_{SM} become

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

after correction for similarity due to chance.

There are some properties in which S_{Gleas} and S_{SM} do differ. With respect to indeterminacy, S_{Gleas} has more critical cases compared to S_{SM}. Moreover, in Chapter 10 it is shown that $1 - S_{\text{SM}}$ is metric, that is, $1 - S_{\text{SM}}$ is a function that satisfies the triangle inequality, whereas the function $1 - S_{\text{Gleas}}$ does not.

Instead of using the power mean, new coefficients may be created by considering other type of means (Bullen, 2003). For example, the Heronian mean of

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

is given by

$$\frac{1}{3}\left(\frac{a}{p_1} + \frac{a}{\sqrt{p_1 p_2}} + \frac{a}{p_2}\right)$$

whereas the Heinz mean is given by

$$\left(\frac{a}{p_1}\right)^u \left(\frac{a}{p_2}\right)^{1-u} + \left(\frac{a}{p_1}\right)^{1-u} \left(\frac{a}{p_2}\right)^u \quad \text{with} \quad 0 \leq u \leq \frac{1}{2}.$$

New coefficients can also be created by including the quantities

$$\frac{d}{b+d} = \frac{d}{q_2} \quad \text{and} \quad \frac{d}{c+d} = \frac{d}{q_1}.$$

For example, the function

$$\frac{4ad}{4ad + (a+d)(b+c)}$$

is the harmonic mean of conditional probabilities

$$\frac{a}{p_1}, \frac{a}{p_2}, \frac{d}{q_1} \quad \text{and} \quad \frac{d}{q_2}.$$

CHAPTER 4

Correction for chance agreement

When comparing two variables some degree of similarity or agreement may be expected due to chance alone, except for the most extreme circumstances (either $p_1 = q_2 = 0$ or $p_2 = q_1 = 0$). Different opinions have been stated on the need to incorporate chance similarity. Goodman and Kruskal (1954, p. 758) contend that similarity due to chance in the measurement of resemblance need not be of much concern, since the observed degree of similarity may usually be assumed to be in excess of chance. In contrast, Zegers (1986) and Popping (1983a) find it quite natural that in absence of association between two variables, the value of a similarity coefficient is zero. Whether or not correction for chance is desirable, depends on the domain or field of data analysis that is considered.

Consider the situation where two variables are the ratings of m people by two observers on two mutually exclusive categories, for example, the observers rate various persons on the presence or absence of a certain trait. In this field, Scott (1955), Cohen (1960), Fleiss (1975), Krippendorff (1987), and Zegers (1986), among others, have proposed measures that are corrected for chance. The best-known example is perhaps the kappa-statistic (Cohen, 1960; S_{Cohen}). Alternatively, the quantities a, b, c, and d can be the result of a comparison between two clustering methods (Section 2.2). In cluster analysis it is general consensus that the popular coefficient S_{SM}, called the Rand index, should be corrected for chance agreement (Morey and Agresti, 1984; Hubert and Arabie, 1985), although there is some debate on what expectation is appropriate (Steinley, 2004; Albatineh et al., 2006).

With respect to correction for chance, various authors have reported results on equivalence of coefficients after correction for similarity due to chance (Fleiss, 1975; Zegers, 1986). Albatineh et al. (2006) studied correction for chance for a family $\mathcal{L}$ of coefficients of the form $S = \lambda + \mu x$ (Section 3.3). These authors appear to be the first to study correction for chance irrespective of the used expectation $E(S)$. The present chapter continues and extends this general approach. Furthermore, the results in this chapter unify various findings in Fleiss (1975), Zegers (1986) and Krippendorff (1987).

Clearly, not all coefficients studied in this thesis have been proposed for, or are used in, data-analytic circumstances where it is desirable to incorporate chance similarity. This practical limitation is however ignored in this chapter. Correction for chance is studied for a general family of coefficients, while ignoring the data-analytic context in which the individual members are usually applied. Using the powerful result from Albatineh et al. (2006), some additional properties of coefficients of the form $\lambda+\mu x$ with respect to correction for chance are presented. For both uncorrected and corrected similarity coefficients properties are derived. Some specific results are obtained by considering different expectations.

4.1 Some equivalences

A corrected similarity coefficient (denoted CS) has, after elimination of the effect of similarity due to chance, a form (2.1)

$$CS = \frac{S - E(S)}{1 - E(S)} \tag{4.1}$$

where S is the similarity coefficient, $E(S)$ the similarity coefficient under chance, and 1 embodies the maximum value of S regardless of the marginal proportions. Most coefficients in this thesis have maximum value unity. Albatineh et al. (2006) showed that correction (4.1) is relatively simple for members in $\mathcal{L}$ family.

Theorem 4.1 [Albatineh et al., 2006, p. 309]. *Two members in the $\mathcal{L}$ family become identical after correction* (4.1) *if they have the same ratio*

$$\frac{1-\lambda}{\mu}. \tag{4.2}$$

Proof: $E(S) = E(\lambda + \mu x) = \lambda + \mu E(x)$ and consequently the CS becomes

$$CS = \frac{S - E(S)}{1 - E(S)} = \frac{\lambda + \mu x - \lambda - \mu E(x)}{1 - \lambda - \mu E(x)} = \frac{x - E(x)}{\mu^{-1}(1-\lambda) - E(x)}. \tag{4.3}$$

□

Thus, the value of a similarity coefficient after correction for chance depends on ratio (4.2), where λ and μ characterize the particular measure within the $\mathcal{L}$ family. Two members in $\mathcal{L}$ become identical after correction (4.1) if they have the same ratio (4.2).

The following corollary concerns the coefficients from Section 2.1 that are linear in the observed proportion of agreement P_{o}.

Corollary 4.1. *Coefficients*

$$S_{\mathrm{SM}} = P_{\mathrm{o}}$$
$$S_{\mathrm{Scott}} = \frac{P_{\mathrm{o}} - E(P_{\mathrm{o}})_{\mathrm{Scott}}}{1 - E(P_{\mathrm{o}})_{\mathrm{Scott}}}$$
$$\textit{and} \qquad S_{\mathrm{Cohen}} = \frac{P_{\mathrm{o}} - E(P_{\mathrm{o}})_{\mathrm{Cohen}}}{1 - E(P_{\mathrm{o}})_{\mathrm{Cohen}}}$$

become equivalent after correction (4.1).
Proof: By Theorem 4.1 it suffices to look at ratio (4.2). Using the formulas of λ and μ corresponding to each coefficient (see Section 3.3), ratio (4.2)

$$\frac{1-\lambda}{\mu} = 1 \tag{4.4}$$

for all three coefficients. □

The next corollary extends Corollary 4.2 (i) in Albatineh et al. (2006) from three measures (S_{SM}, S_{Ham}, S_{Gleas}) to ten coefficients. All ten coefficients are linear in the quantity a.

Corollary 4.2. *Coefficients*

$$
\begin{aligned}
S_{\text{SM}} &= 1 + 2a - p_1 - p_2 \\
S_{\text{Ham}} &= 1 + 2a - 2p_1 - 2p_2 \\
S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} \\
S_{\text{GK}} &= \frac{2\min(a,d) - b - c}{2\min(a,d) + b + c} && \text{(Goodman and Kruskal, 1954)} \\
S_{\text{NS1}} &= \frac{2a - b - c}{2a + b + c} = \frac{4a - 2p_1 + 2p_2}{p_1 + p_2} && \text{(no source)} \\
S_{\text{NS2}} &= \frac{2d}{b + c + 2d} = \frac{2(a + q_1 + q_2 - 1)}{q_1 + q_2} && \text{(no source)} \\
S_{\text{NS3}} &= \frac{2d - b - c}{b + c + 2d} = \frac{4a + 3q_1 + 3q_2 - 4}{q_1 + q_2} && \text{(no source)} \\
S_{\text{RG}} &= \frac{a}{p_1 + p_2} + \frac{a + q_1 + q_2 - 1}{q_1 + q_2} && \text{(Rogot and Goldberg, 1966)} \\
S_{\text{Scott}} &= \frac{4a - (p_1 + p_2)^2}{4 - (p_1 + p_2)^2} \\
S_{\text{Cohen}} &= \frac{2(a - p_1 p_2)}{p_1 q_2 + p_2 q_1}
\end{aligned}
$$

become equivalent after correction (4.1).
Proof: By Theorem 4.1 it suffices to look at ratio (4.2). Using the formulas of λ and μ corresponding to each coefficient, ratio (4.2)

$$\frac{1 - \lambda}{\mu} = \frac{p_1 + p_2}{2} \tag{4.5}$$

for all ten coefficients. □

Note that ratio (4.5) is the arithmetic mean of marginal probabilities p_1 and p_2. The interpretation of (4.5) depends on how x was specified in in $\lambda + \mu x$, and ratio (4.5) is different from (4.4). Alternatively, we may formulate the ten coefficients as functions that are linear in the quantity $x = a + d$ instead of $x = a$. The result with respect to correction for chance agreement is of course the same, but ratio (4.6) now equals ratio (4.4).

Corollary 4.2b. *Coefficients*

$$
\begin{aligned}
S_{\text{SM}} &= a + d \\
S_{\text{Ham}} &= 2(a + d) - 1 \\
S_{\text{Gleas}} &= \frac{(a + d) - 1}{p_1 + p_2} + 1 \\
S_{\text{GK}} &= \frac{2(a + d) - 2}{\min(p_1 + p_2, q_1 + q_2)} + 1 \\
S_{\text{NS1}} &= \frac{2(a + d) - 2}{p_1 + p_2} + 1 \\
S_{\text{NS2}} &= \frac{(a + d) - 1}{q_1 + q_2} + 1 \\
S_{\text{NS3}} &= \frac{2(a + d) - 2}{q_1 + q_2} + 1 \\
S_{\text{RG}} &= \frac{(a + d) - 1}{2(p_1 + p_2)} + \frac{(a + d) - 1}{2(q_1 + q_2)} + 1 \\
S_{\text{Scott}} &= \frac{4(a + d) - (p_1 + p_2)^2 - (q_1 + q_2)^2}{4 - (p_1 + p_2)^2 - (q_1 + q_2)^2} \\
S_{\text{Cohen}} &= \frac{(a + d) - p_1 p_2 - q_1 q_2}{p_1 q_2 + p_2 q_1}
\end{aligned}
$$

become equivalent after correction (4.1).
Proof: By Theorem 4.1 it suffices to look at ratio (4.2). Using the formulas of λ and μ corresponding to each coefficient, ratio (4.2)

$$\frac{1 - \lambda}{\mu} = 1 \tag{4.6}$$

for all ten coefficients. □

Since $a = p_2 - q_1 + d$, probabilities a and d are also linear in $(a+d)$. Linear in $(a+d)$ is therefore equivalent to linear in a and linear in d. Furthermore, Albatineh et al. (2006) studied coefficients that are linear in $\sum\sum n_{ij}^2$, where n_{ij} is the number of data points placed in cluster i according to the first clustering method and in cluster j according to the second clustering method. Because $ma = (\sum\sum n_{ij}^2 - m)/2$, linear in $\sum\sum n_{ij}^2$ is equivalent to linear in a and equivalent to linear in $(a + d)$.

The corrected coefficient corresponding to the nine resemblance measures in Corollary 4.2 has a form

$$CS = \frac{(a + d) - E(a + d)}{1 - E(a + d)}. \tag{4.7}$$

Coefficient (4.7) may be obtained by using $(a+d)$, $E(a+d)$, and (4.6) in the extreme-right part of (4.3). Since expectation $E(a + d)$ is unspecified, coefficient (4.7) is a general corrected coefficient.

4.2 Expectations

A commonly used expectation was briefly considered in Section 1.3. Different opinions have been stated on what the appropriate expectations are for the 2×2 contingency table. Detailed discussions on the various ways of regarding data as the product of chance can be found in Krippendorff (1987), Mak (1988), Bloch and Kraemer (1989) and Pearson (1947). In cluster analysis it is general consensus that the popular coefficient S_{SM}, called the Rand index, should be corrected for agreement due to chance (Morey and Agresti, 1984; Hubert and Arabie, 1985), although there is some debate on what expectation is appropriate (Hubert and Arabie, 1985; Steinley, 2004; Albatineh et al., 2006). We consider five examples of $E(a+d)$.

Suppose it is assumed that the frequency distribution underlying the two variables in the 2×2 contingency table is the same for both variables (Scott, 1955; Krippendorff, 1987, p. 113). Coefficients used in this context are sometimes referred to as agreement indices. The common parameter p must be either known or it must be estimated from p_1 and p_2. Different functions may be used. For example, Scott (1955) and Krippendorff (1987) use the arithmetic mean

$$p = \frac{p_1 + p_2}{2}.$$

Following Scott (1955) and Krippendorff (1987, p. 113) we have

$$E(a+d)_{\text{Scott}} = \left(\frac{p_1 + p_2}{2}\right)^2 + \left(\frac{q_1 + q_2}{2}\right)^2.$$

Let n denote the number of elements of the binary variables. Mak (1988) proposed the expectation

$$E(a+d)_{\text{Mak}} = 1 - \frac{n(p_1 + p_2)(q_1 + q_2) - (b+c)}{2(n-1)}$$

(see also, Blackman and Koval, 1993).

Instead of a single distribution function, it may be assumed that the data in the fourfold table are a product of chance concerning two different frequency distributions, each with its own parameter (Cohen, 1960; Krippendorff, 1987). Coefficients used in this context are sometimes referred to as association indices. The expectation of an entry in the 2×2 contingency table under statistical independence, is defined by the product of the marginal probabilities. We have

$$E(a+d)_{\text{Cohen}} = p_1 p_2 + q_1 q_2.$$

Expectation $E(a+d)_{\text{Cohen}}$ can be obtained by considering all permutations of the observations of one of the two variables, while preserving the order of the observations of the other variable. For each permutation the value of $(a+d)$ can be determined. The arithmetic mean of these values is $p_1 p_2 + q_1 q_2$.

A third possibility is that there are no relevant underlying continua. For this case two forms of $E(a+d)$ may be found in the literature. Goodman and Kruskal (1954, p. 757) use expectation

$$E(a+d)_{\mathrm{GK}} = \frac{\max(p_1+p_2, q_1+q_2)}{2}.$$

According to Krippendorff (1987, p. 114) an equity coefficient is characterized by expectation

$$E(a+d)_{\mathrm{Kripp}} = \frac{1}{2}.$$

Let us summarize the three situations. In the case of association the observations are regarded as ordered pairs. In the case of agreement the observations are considered as pairs without regard for their order; a mismatch is a mismatch regardless of the kind. In the case of equity one only distinguishes between matching and non-matching observations (cf. Krippendorff, 1987).

Proposition 4.1 below unifies and extends findings in Fleiss (1975) and Zegers (1986) on what coefficients become Cohen's kappa after correction for chance. Depending on what expectation $E(a+d)$ is used, the coefficients in Corollary 4.2 become, after correction for chance, either Scott's (1955) pi (S_{Scott}), Cohen's (1960) kappa (S_{Cohen}), Goodman and Kruskal's (1954) lambda (S_{GK}), Hamann's (1961) eta (S_{Ham}), or Mak's (1988) rho. The latter coefficient can be written as

$$S_{\mathrm{Mak}} = \frac{4nad - n(b+c)^2 + (b+c)}{n(p_1+p_2)(q_1+q_2) - (b+c)} \qquad \text{(Mak, 1988)}$$

where n is length of the binary variables. With respect to Proposition 4.1, let $\mathcal{L}$ family consists of functions $\lambda + \mu(a+d)$.

Proposition 4.1. *Let S be a member in $\mathcal{L}$ family for which ratio* (4.6) *holds. If the appropriate expectation is*

$$\begin{array}{rlll}
(i) & E(a+d)_{\mathrm{Scott}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Scott}} \\
(ii) & E(a+d)_{\mathrm{Mak}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Mak}} \\
(iii) & E(a+d)_{\mathrm{Cohen}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Cohen}} \\
(iv) & E(a+d)_{\mathrm{GK}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{GK}} \\
(v) & E(a+d)_{\mathrm{Kripp}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Ham}}
\end{array}$$

after correction (4.1).
Proof (i): Using $E(a+d)_{\mathrm{Scott}}$ in (4.7) we obtain an index with numerator

$$a + d - \left(\frac{p_1+p_2}{2}\right)^2 - \left(\frac{q_1+q_2}{2}\right)^2 = 2ad - \frac{(b+c)^2}{2} \tag{4.8}$$

and denominator

$$\frac{(p_1+p_2+q_1+q_2)^2 - (p_1+p_2)^2 - (q_1+q_2)^2}{4} = \frac{(p_1+p_2)(q_1+q_2)}{2}. \tag{4.9}$$

Dividing the right-hand part of (4.8) by the right-hand part of (4.9) we obtain

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

Proof (*ii*): Using $E(a+d)_{\text{Mak}}$ in (4.7) and multiplying the result by $2(n-1)$ we obtain an index with numerator

$$\begin{aligned} &2(a+d-1)(n-1)+n(p_1+p_2)(q_1+q_2)-(b+c)\\ =&n(2a+b+c)(b+c+2d)-2n(b+c)+(b+c) \end{aligned} \tag{4.10}$$

and denominator

$$n(p_1+p_2)(q_1+q_2)-(b+c). \tag{4.11}$$

We have

$$\begin{aligned} &(2a+b+c)(b+c+2d)-2(b+c)\\ =&4ad+(2a+2d)(b+c)+(b+c)^2-2(b+c)\\ =&4ad+(2a+2d-2)(b+c)+(b+c)^2\\ =&4ad-2(b+c)^2+(b+c)^2\\ =&4ad-(b+c)^2. \end{aligned} \tag{4.12}$$

Using (4.12), numerator (4.10) can be written as

$$n\left[4ad-(b+c)^2\right]+(b+c). \tag{4.13}$$

Dividing (4.13) by (4.11) we obtain coefficient S_{Mak}.
Proof (*iii*): Using $E(a+d)_{\text{Cohen}}$ in (4.7) we obtain

$$\frac{a+d-p_1p_2-q_1q_2}{(p_1+q_1)(p_2+q_2)-p_1p_2-q_1q_2} = \frac{2(ad-bc)}{p_1q_2+p_2q_1} = S_{\text{Cohen}}.$$

Proof (*iv*): Using $E(a+d)_{\text{GK}}$ in (4.7) we obtain

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

Proof (*v*): Using $E(a+d)_{\text{Kripp}}$ in (4.7) we obtain

$$2(a+d)-1 = a-b-c+d = S_{\text{Ham}}. \qquad \square$$

4.3 Two transformations

In this section we consider the two functions of similarity coefficients

$$S_2 = 2S_1 - 1 \qquad \text{and} \qquad S_3 = \frac{S_1 + S_2}{2}.$$

Both transformations may be used to construct new resemblance measures from existing similarity coefficients. It holds that $S_2 = 2S_1 - 1$ is in the $\mathcal{L}$ family if and only if S_1 is in $\mathcal{L}$, and if S_1 and S_2 are in $\mathcal{L}$, then $S_3 = (S_1 + S_2)/2$ is in $\mathcal{L}$. In this section it is shown how the new coefficients are related to the old coefficients in terms of correction for similarity due to chance. With respect to Proposition 4.2, let $\mathcal{L}$ consists of functions of the form $\lambda + \mu x$.

Proposition 4.2. *Let S_1 be a member of $\mathcal{L}$. S_1 and $S_2 = 2S_1 - 1$ become identical after correction* (4.1).
Proof: $S_2 = 2\lambda + 2\mu a - 1$ and $E(S_2) = 2\lambda - 1 + 2\mu E(x)$. Consequently the CS_2 becomes

$$\begin{aligned} CS_2 &= \frac{2\lambda + 2\mu x - 1 - 2\lambda - 2\mu E(x) + 1}{1 - 2\lambda - 2\mu E(x) + 1} = \frac{\lambda + \mu x - \lambda - \mu E(x)}{1 - \lambda - \mu E(x)} \\ &= \frac{S_1 - E(S_1)}{1 - E(S_1)} = CS_1. \qquad \square \end{aligned}$$

Similarity coefficients that are related by transformation $S_2 = 2S_1 - 1$ can be found in Corollary 4.2. Examples are

$$\begin{aligned} S_{\text{Ham}} &= 2S_{\text{SM}} - 1 \\ S_{\text{NS1}} &= 2S_{\text{Gleas}} - 1 \\ \text{and} \quad S_{\text{NS3}} &= 2S_{\text{NS2}} - 1. \end{aligned}$$

Another example is $S_{\text{McC}} = 2S_{\text{Kul}} - 1$, where

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) \qquad \text{and} \qquad S_{\text{McC}} = \frac{a^2 - bc}{p_1 p_2} \quad \text{(McConnaughey, 1964).}$$

The fact that coefficient S_{Kul} and S_{McC} become equivalent after correction (4.1) irrespective of the used expectation was already proved in Corollary 4.2 (ii) in Albatineh et al. (2006).

Proposition 4.3. *Let S_i for $i = 1, 2, ..., m$ be members in $\mathcal{L}$ family that become identical after correction* (4.1). *Then S_i for $i = 1, 2, ..., m$ and the arithmetic mean $S^* = m^{-1} \sum_{i=1}^{m} S_i$ coincide after correction* (4.1).
Proof:

$$E(S^*) = E\left(\frac{\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i x}{m}\right) = \frac{\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i E(x)}{m}.$$

Using arithmetic mean S^* in (4.1), we obtain

$$CS^* = \frac{x - E(x)}{y - E(x)} \qquad \text{where} \quad y = \frac{m - \sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{m} \mu_i}.$$

Let

$$z = \frac{1 - \lambda_1}{\mu_1} = \frac{1 - \lambda_2}{\mu_2} = ... = \frac{1 - \lambda_m}{\mu_m}.$$

It must be shown that ratio y equals ratio z. We have

$$y = \frac{\sum_{i=1}^{m}(1 - \lambda_i)}{\sum_{i=1}^{m} \mu_i} = \frac{\sum_{i=1}^{m} z\mu_i}{\sum_{i=1}^{m} \mu_i} = \frac{z \sum_{i=1}^{m} \mu_i}{\sum_{i=1}^{m} \mu_i} = z.$$

This completes the proof. □

Coefficient

$$S_{\text{RG}} = \frac{a}{2a + b + c} + \frac{d}{b + c + 2d} = \frac{S_{\text{Gleas}} + S_{\text{NS2}}}{2}$$

in Corollary 4.2, is the arithmetic mean of S_{Gleas} and S_{NS2}.

4.4 Corrected coefficients

The coefficients in Corollary 4.2 and Proposition 4.1 become either S_{Scott}, S_{Mak}, S_{Cohen}, S_{GK}, or S_{Ham}, depending on what expectation $E(a + d)$ is used. Note that corrected coefficients S_{Scott}, S_{Cohen}, S_{GK}, and S_{Ham} belong to the class of resemblance measures that is considered in Corollary 4.2 and Proposition 4.1. This suggests that corrected coefficients may have some interesting properties. The corrected coefficients and their properties are the topic of this section. If $E(S)$ in (4.1) depends on the marginal probabilities of the 2×2 contingency table, then CS in (4.1) belongs to $\mathcal{L}$. With respect to Proposition 4.4, let $\mathcal{L}$ consists of functions of the form $\lambda + \mu(a + d)$.

Proposition 4.4. *Let $E(S)$ in* (4.1) *depend on the marginal probabilities. If S is in $\mathcal{L}$ family, then CS in* (4.1) *is in $\mathcal{L}$.*
Proof: Expectation $E(S) = E[\lambda_1 + \mu_1(a + d)]$ is a function of the marginal probabilities. Thus $E(a + d)$, λ, and μ in (4.3) are functions of the marginal proportions. Equation (4.3) can therefore be written in a form $\lambda_2 + \mu_2(a + d)$ where

$$\lambda_2 = \frac{-E(a + d)}{\mu_1^{-1}(1 - \lambda_1) - E(a + d)} \qquad \text{and} \qquad \mu_2 = \frac{1}{\mu_1^{-1}(1 - \lambda_1) - E(a + d)}. \qquad \square$$

Examples of corrected coefficients that are in the $\mathcal{L}$ family are S_{Scott}, S_{Cohen}, S_{GK}, and S_{Ham}. These coefficients may be considered as corrected coefficients as well as ordinary coefficients that may be corrected for agreement due to chance. For example, S_{Scott}, S_{GK}, and S_{Ham} (and S_{Cohen}) become S_{Cohen} after correction (4.1) if expectation $E(a+d)_{\text{Cohen}}$ is used. Coefficient S_{Mak} cannot be written in a form $\lambda + \mu(a+d)$, and does therefore not belong to $\mathcal{L}$.

Next we consider the following problem. Suppose a coefficient S in $\mathcal{L}$ is corrected twice, using two different expectations, $E(a+d)$ and $E(a+d)^*$. Let the corrected coefficients be given by

$$CS = \frac{a+d-E(a+d)}{\mu^{-1}(1-\lambda)-E(a+d)} \qquad \text{and} \qquad CS^* = \frac{a+d-E(a+d)^*}{\mu^{-1}(1-\lambda)-E(a+d)^*}.$$

Note that $\mu^{-1}(1-\lambda)$ corresponding to coefficient S, is the same in both CS and CS^*. The problem is then as follows: if $E(a+d) \geq E(a+d)^*$, how are CS and CS^* related? Proposition 4.5 below is limited to coefficients in the $\mathcal{L}$ family of which the maximum value is unity, that is

$$\lambda + \mu(a+d) \leq 1 \qquad \text{if and only if} \qquad \frac{1-\lambda}{\mu} \geq (a+d).$$

It can be verified that most (if not all) similarity coefficients in this thesis satisfy this condition.

Proposition 4.5. *$CS \leq CS^*$ if and only if $E(a+d) \geq E(a+d)^*$.*
Proof: $CS \leq CS^*$ if and only if

$$E(a+d)\left[\frac{1-\lambda}{\mu}-(a+d)\right] \geq E(a+d)^*\left[\frac{1-\lambda}{\mu}-(a+d)\right].$$

The requirement $\lambda + \mu(a+d) \leq 1$ completes the proof. □

In the following, let $S = \lambda + \mu(a+d)$ be in $\mathcal{L}$ family and let

$$CS_{\text{Name}} = \frac{a+d-E(a+d)_{\text{Name}}}{\mu^{-1}(1-\lambda)-E(a+d)_{\text{Name}}}$$

be a corrected coefficient using expectation $E(a+d)_{\text{Name}}$. Using specific expectations $E(a+d)$ in combination with Proposition 4.5, we obtain the following result.

Proposition 4.6. *It holds that* $CS_{\mathrm{GK}} \overset{(i)}{\leq} CS_{\mathrm{Scott}} \overset{(ii)}{\leq} CS_{\mathrm{Cohen}}$.
Proof (i): Due to Proposition 4.5, it suffices to show that $E(a+d)_{\mathrm{GK}} \geq E(a+d)_{\mathrm{Scott}}$. Suppose $(p_1 + p_2) \geq (q_1 + q_2)$. We have

$$\begin{aligned}
E(a+d)_{\mathrm{GK}} &\geq E(a+d)_{\mathrm{Scott}} \\
\frac{p_1+p_2}{2} &\geq \left(\frac{p_1+p_2}{2}\right)^2 + \left(\frac{q_1+q_2}{2}\right)^2 \\
\frac{p_1+p_2}{2}\left(1 - \frac{p_1+p_2}{2}\right) &\geq \left(\frac{q_1+q_2}{2}\right)^2 \\
\frac{p_1+p_2}{2}\left(\frac{q_1+q_2}{2}\right) &\geq \left(\frac{q_1+q_2}{2}\right)^2 \\
(p_1+p_2) &\geq (q_1+q_2).
\end{aligned}$$

Proof (ii): It must be shown that $E(a+d)_{\mathrm{Scott}} \geq E(a+d)_{\mathrm{Cohen}}$. We have

$$\left(\frac{p_1+p_2}{2}\right)^2 \geq p_1 p_2 \tag{4.14}$$

if and only if

$$\frac{p_1+p_2}{2} \geq \sqrt{p_1 p_2}. \tag{4.15}$$

Furthermore, we have

$$\left(\frac{q_1+q_2}{2}\right)^2 \geq q_1 q_2 \tag{4.16}$$

if and only if

$$\frac{q_1+q_2}{2} \geq \sqrt{q_1 q_2}. \tag{4.17}$$

Because the arithmetic mean of two numbers is equal or greater than the geometric mean, inequalities (4.15) and (4.17) are true. Adding (4.14) and (4.16) we obtain the desired inequality. □

Blackman and Koval (1993, p. 216) derived the inequality $S_{\mathrm{Scott}} \leq S_{\mathrm{Cohen}}$. Note that this inequality follows from the more general result Proposition 4.6 by using a coefficient S for which (4.6) is characteristic.

4.5 Epilogue

Under the assumption that $E(a+d) = p_1p_2 + q_1q_2$ is the appropriate expectation, Fleiss (1975) showed that

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} = a+d \quad \text{and} \quad S_{\text{Gleas}} = \frac{2a}{p_1+p_2}$$

and S_{GK} and S_{RG} become S_{Cohen} after correction (4.1). Zegers (1986) showed that S_{SM}, S_{Gleas} and S_{Ham} become S_{Cohen} after correction (4.1). Albatineh et al. (2006) showed that S_{SM}, S_{Gleas} and S_{Ham} become equivalent irrespective of the used expectation. These results were extended and unified by Corollary 4.2 and Proposition 4.1. Corollary 4.2 specifies up to ten coefficients that become equivalent after correction (4.1) irrespective of expectation $E(a+d)$. The coefficients in Corollary 4.2 become either S_{Scott}, S_{Mak}, S_{Cohen}, S_{GK}, or S_{Ham}, depending on what expectation $E(a+d)$ is used. Moreover, two transformations from Section 4.3 may be used to construct an infinite amount of coefficients that become equivalent after correction (4.1).

Whether $E(a+d)_{\text{Cohen}}$ or another $E(a+d)$ is the appropriate expectation depends on the context of the data analysis. However, since a large number of coefficients are defined with the covariance

$$\frac{a+d-E(a+d)_{\text{Cohen}}}{2} = \frac{(a-p_1p_2)+(d-q_1q_2)}{2} = ad-bc$$

in the numerator, it appears that $E(a+d)_{\text{Cohen}}$ is the preferred (or most appropriate) expectation in many cases.

The quantities

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

and

$$S_{\text{Cole1}} = \frac{ad-bc}{p_1q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad-bc}{p_2q_1} \quad \text{(Cole, 1949)}$$

where used in the previous chapter to construct power means

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{and} \quad M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_2q_1}\right).$$

As it turns out, if the expectation of a is $E(a) = p_1 p_2$, several members of the two power means corresponding to the same θ are related. We have, for example,

$$S_{\text{Dice1}} = \frac{a}{p_1} \qquad \text{becomes} \quad S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2}$$

$$S_{\text{Dice2}} = \frac{a}{p_2} \qquad \text{becomes} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1}$$

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

$$\text{and} \quad S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)} \qquad \text{becomes} \quad S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)}.$$

CHAPTER 5

Correction for maximum value

The proportions a, b, c, and d in the fourfold table

a	b	p_1
c	d	q_1
p_2	q_2	1

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 and minimum values are sometimes untenable. Guilford (1965), Cureton (1959) and Davenport and El-Sanhurry (1991) consider the maximum of S_{Phi} given marginals p_1 and p_2, denoted by $[S_{\text{Phi}}]_{\text{max}}$. Loevinger (1947, 1948) suggested using the ratio

$$\frac{S_{\text{Phi}}}{[S_{\text{Phi}}]_{\text{max}}}$$

since this procedure allows the corrected value to become unity. As noted by Loevinger (1947, 1948), Sijtsma and Molenaar (2002) and Davenport and El-Sanhurry (1991), coefficients S_{Phi}, S_{Cohen} and S_{Loe} are related by

$$S_{\text{Loe}} = \frac{S_{\text{Phi}}}{[S_{\text{Phi}}]_{\text{max}}} = \frac{S_{\text{Cohen}}}{[S_{\text{Cohen}}]_{\text{max}}}.$$

The relations between similarity coefficients for two binary variables suggested in this equality are the topic of this chapter.

The maximum and minimum of various coefficients and several equivalences are studied first. The maximum of a coefficient is determined by applying the formula to the case of two Guttman items (Section 6.3; Mokken, 1971; Guilford, 1965). Furthermore, it is shown what families of coefficients become equivalent after correction

$$\frac{S}{[S]_{\max}}. \tag{5.1}$$

5.1 Maximum value

In this section we derive the maximum value for a variety of coefficients. We focus on coefficients that are special cases of a power mean. Following Guilford (1965) and Cureton (1959), the maximum value of a coefficient is obtained if either quantity b, c, or both equal zero. Hence, with unequal marginal proportions $p_1 \neq p_2$, the 2×2 contingency table has the form

a	0	p_1
c	d	q_1
p_2	q_2	1

for example $\begin{bmatrix}0\\0\\0\\1\\1\\1\end{bmatrix}$ and $\begin{bmatrix}0\\1\\1\\1\\1\\1\end{bmatrix}$

if $b = 0$, or

a	b	p_1
0	d	q_1
p_2	q_2	1

for example $\begin{bmatrix}0\\1\\1\\1\\1\\1\end{bmatrix}$ and $\begin{bmatrix}0\\0\\0\\1\\1\\1\end{bmatrix}$

if $c = 0$. Note that the maximum is obtained if the two binary variables being compared are so-called Guttman items (Section 6.3; Mokken, 1971). The maximum value of proportion a given the marginals p_1 and p_2, denoted by $a_{\max}$, is given by

$$a_{\max} = \begin{cases} p_1 & \text{if} \quad b = 0 \\ p_2 & \text{if} \quad c = 0 \end{cases} \qquad \text{or} \quad a_{\max} = \min(p_1, p_2).$$

Thus, without correction for maximum value, quantity a can only reach its maximum value if $p_1 = p_2$. The maximum value of measures for binary variables that do not include quantity d, may be obtained by replacing probability a by $a_{\max}$. Assuming $p_1 \neq p_2$ we obtain

$$[S_{\text{GL1}}(\theta)]_{\max} = \frac{\min(p_1, p_2)}{\theta(p_1 + p_2) + (1 - 2\theta)\min(p_1, p_2)}$$

with

$$[S_{\mathrm{GL1}}(1)]_{\max} = [S_{\mathrm{Jac}}]_{\max} = \frac{\min(p_1, p_2)}{\max(p_1, p_2)} < 1$$

$$[S_{\mathrm{GL1}}(1/2)]_{\max} = [S_{\mathrm{Gleas}}]_{\max} = \frac{2\min(p_1, p_2)}{p_1 + p_2} < 1.$$

With respect to the inequalities

$$S_{\mathrm{Sorg}} = \frac{a^2}{p_1 p_2} \;\leq\; S_{\mathrm{Jac}} = \frac{a}{p_1 + p_2 - a} \;\leq\; S_{\mathrm{BB}} = \frac{a}{\max(p_1, p_2)}$$

we obtain the equality

$$[S_{\mathrm{Sorg}}]_{\max} = [S_{\mathrm{Jac}}]_{\max} = [S_{\mathrm{BB}}]_{\max} = \frac{\min(p_1, p_2)}{\max(p_1, p_2)}.$$

With respect to the power mean of the quantities

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

the equality $a_{\max} = \min(p_1, p_2)$ leads to

$$\left[M_\theta \left(\frac{a}{p_1}, \frac{a}{p_2} \right) \right]_{\max} = M_\theta \left(1, \frac{\min(p_1, p_2)}{\max(p_1, p_2)} \right).$$

where

$$\frac{\min(p_1, p_2)}{\max(p_1, p_2)} = [S_{\mathrm{BB}}]_{\max}.$$

Thus, the maximum value of a coefficient that is a special case of the power mean of S_{Dice1} and S_{Dice2}, is equal to the coefficient corresponding to the same θ of the value 1 and $[S_{\mathrm{BB}}]_{\max}$, where the latter is the maximum value of the minimum function of S_{Dice1} and S_{Dice2}. Hence, only for the maximum function, that is, $S_{\mathrm{Sim}} = a/\min(p_1, p_2)$, it holds that

$$[S_{\mathrm{Sim}}]_{\max} = \lim_{\theta \to \infty} M_\theta \left(1, \frac{\min(p_1, p_2)}{\max(p_1, p_2)} \right) = \max \left(1, \frac{\min(p_1, p_2)}{\max(p_1, p_2)} \right) = 1.$$

Next, we consider the maximum value of the covariance $(ad - bc)$ of two binary variables. The maximum covariance given the marginals p_1 and p_2, denoted $(ad - bc)_{\max}$, is given by

$$(ad - bc)_{\max} = \begin{cases} p_1 q_2 & \text{if} \quad b = 0 \\ p_2 q_1 & \text{if} \quad c = 0 \end{cases} \qquad \text{or} \quad (ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1).$$

We may obtain the maximum value of measures for binary variables that use the covariance in the numerator by replacing covariance $(ad-bc)$ by $(ad-bc)_{\max}$. With respect to the power mean of the quantities

$$S_{\text{Cole1}} = \frac{ad-bc}{p_1q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad-bc}{p_2q_1} \quad \text{(Cole, 1949)}$$

the equality $(ad-bc)_{\max} = \min(p_1q_2, p_2q_1)$ leads to

$$\left[M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_1q_2}\right)\right]_{\max} = M_\theta\left(1, \frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}\right).$$

Thus, the maximum value of a coefficient that is a special case of the power mean of S_{Cole1} and S_{Cole2}, is equal to the coefficient corresponding to the same θ of the value 1 and the quantity

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

Hence, only for the maximum function, that is, S_{Loe}, it holds that

$$[S_{\text{Loe}}]_{\max} = \lim_{\theta\to\infty} M_\theta\left(1, \frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}\right) = 1.$$

5.2 Correction for maximum value

Let x/y and x/z be two real positive values, of which the maximum depends on x only, that is

$$\left[\frac{x}{y}\right]_{\max} = \frac{x_{\max}}{y} \quad \text{and} \quad \left[\frac{x}{z}\right]_{\max} = \frac{x_{\max}}{z}.$$

Examples of x/y and x/z are S_{Dice1} and S_{Dice2}. For example, $x = a$ or $x = ad-bc$ and y and z are functions of p_1 and p_2 only. It turns out that division of the power mean of x/y and x/z by its maximum value given quantities y and z, does not depend on the choice of θ. Moreover, the outcome of the division does not depend on the definitions of y and z.

Proposition 5.1. *Let x/y and x/z be two real positive values defined as above. Then*

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) \Big/ \left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\max} = \frac{x}{x_{\max}}.$$

Proof:

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) = \left[\frac{1}{2}\left(\frac{x}{y}\right)^\theta + \frac{1}{2}\left(\frac{x}{z}\right)^\theta\right]^{1/\theta} = \left[\frac{x^\theta(y^\theta+z^\theta)}{2y^\theta z^\theta}\right]^{1/\theta} = \frac{x}{yz}\left[\frac{y^\theta+z^\theta}{2}\right]^{1/\theta}$$

and

$$\left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\max} = \frac{x_{\max}}{yz}\left[\frac{y^\theta+z^\theta}{2}\right]^{1/\theta}. \quad \square$$

An interesting consequence of Proposition 5.1 is the following property. Dividing the power mean of x/y and x/z by its maximum value gives the maximum function of x/y and x/z.

Corollary 5.1. *Let x/y and x/z be defined as above. If $x_{\max} = \min(y, z)$, then*

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) / \left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\max} = \lim_{\theta\to\infty} M_\theta\left(\frac{x}{y}, \frac{x}{z}\right).$$

As a first example, consider the power mean of

$$x = \frac{a}{p_1} \quad \text{and} \quad y = \frac{a}{p_2}.$$

Because $a_{\max} = \min(p_1, p_2)$, we have

$$\frac{M_\theta(x, y)}{[M_\theta(x, y)]_{\max}} = \lim_{\theta\to\infty} M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) = \frac{a}{\min(p_1, p_2)} = S_{\text{Sim}}.$$

As a second example, consider the power mean of

$$x = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_1}.$$

Since $(ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1)$, we have

$$\frac{M_\theta(x, y)}{[M_\theta(x, y)]_{\max}} = \lim_{\theta\to\infty} M_\theta\left(\frac{ad - bc}{p_1 q_2}, \frac{ad - bc}{p_2 q_1}\right) = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)} = S_{\text{Loe}}.$$

As a third example, consider the power mean of the quantities

$$x = \frac{ad - bc}{p_1 q_1} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_2} \qquad \text{(see Peirce, 1884).}$$

Then

$$\begin{aligned}
M_{-1}(x, y) &= \frac{2(ad - bc)}{p_1 q_1 + p_2 q_2} = S_{\text{MP}} && \text{(harmonic mean)} \\
\lim_{\theta\to 0} M_\theta(x, y) &= \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}} = S_{\text{Phi}} && \text{(geometric mean)} \\
M_1(x, y) &= \frac{(ad - bc)(p_1 q_1 + p_2 q_2)}{2 p_1 q_2 p_2 q_1} = S_{\text{Fleiss}} && \text{(arithmetic mean).}
\end{aligned}$$

In light of Corollary 5.1, because $(ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1)$, which is different from $\min(p_1 q_1, p_2 q_2)$, we have

$$\begin{aligned}
\frac{M_\theta(x, y)}{[M_\theta(x, y)]_{\max}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)} &\neq \lim_{\theta\to\infty} M_h\left(\frac{ad - bc}{p_1 q_1}, \frac{ad - bc}{p_2 q_2}\right) \\
&= \frac{ad - bc}{\min(p_1 q_1, p_2 q_2)}.
\end{aligned}$$

Thus, the power mean of these x and y becomes S_{Loe}, although the latter coefficient is not a special case of the power mean.

Instead of considering power means, correction (5.1) can also be approached from a different angle. Below, two assertions are presented with respect to coefficients S_{Sim} and S_{Loe}.

Proposition 5.2. *Let $S = a/x$ with x a function of p_1 and p_2. Then*

$$S/\left[S\right]_{\max} = \frac{a}{\min(p_1, p_2)} = S_{\text{Sim}}.$$

Proof:

$$\left[S\right]_{\max} = \left[\frac{a}{x}\right]_{\max} = \frac{a_{\max}}{x} = \frac{\min(p_1, p_2)}{x}. \quad \text{Hence} \quad S/\left[S\right]_{\max} = S_{\text{Sim}}. \quad \square$$

Proposition 5.3. *Let $S = (ad - bc)/x$ with x a function of p_1 and p_2. Then $S/\left[S\right]_{\max} = S_{Loe}$.*
Proof:

$$\left[S\right]_{\max} = \left[\frac{ad - bc}{x}\right]_{\max} = \frac{(ad - bc)_{\max}}{x} = \frac{\min(p_1 q_2, p_2 q_1)}{x}.$$

Hence $S/\left[S\right]_{\max} = S_{\text{Loe}}$. $\square$

5.3 Correction for minimum value

In addition to the maximum value $\left[S\right]_{\max}$ of a coefficient S, one may study the minimum value $\left[S\right]_{\min}$. For coefficients that are special cases of the power mean of the quantities

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

the minimum value 0 is obtained if $a = 0$. Similarly, coefficients of the form a/x where x is a function of p_1 and p_2, equal 0 whenever $a = 0$. Thus, for this type of coefficients the minimum value is not constrained by the marginals. The section is therefore restricted to the minimum value of coefficients with the covariance $(ad-bc)$ in the numerator. For this class of coefficients the minimum value is obtained if either quantity a, d, or both equal zero. Hence, with unequal marginals $p_1 \neq q_1$, the 2×2 contingency table has the form

0	b	p_1
c	d	q_1
p_2	q_2	1

for example $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

if $a = 0$,

or

a	b	p_1
c	0	q_1
p_2	q_2	1

for example $\begin{bmatrix}1\\1\\1\\1\\0\\0\end{bmatrix}$ and $\begin{bmatrix}0\\0\\0\\1\\1\\1\end{bmatrix}$

if $d = 0$. The minimum covariance of two binary variables given marginal proportions p_1 and p_2, denoted $(ad - bc)_{\min}$, is thus given by

$$(ad - bc)_{\min} = \begin{cases} -p_1 p_2 & \text{if} \quad a = 0 \\ -q_1 q_2 & \text{if} \quad d = 0 \end{cases}$$

which equals

$$(ad - bc)_{\min} = \max(-p_1 p_2, -q_1 q_2) = -\min(p_1 p_2, q_1 q_2).$$

Thus, the minimum value of the covariance can only be obtained if $p_1 p_2 = q_1 q_2$ if and only if $p_1 + p_2 = 1$.

With correction for the minimum value the following issue must be taken into consideration. Because the quantity $(ad - bc)_{\min}$ is negative, division of a coefficient by $(ad - bc)_{\min}$ results in a change of sign. However, the minimum value of -1 can be obtained if the quantity $\min(p_1 p_2, q_1 q_2)$ is used instead of $-\min(p_1 p_2, q_1 q_2)$.

Similar as in the previous section, let x/y and x/z be two real positive values, of which the minimum depends on x only, that is

$$\left[\frac{x}{y}\right]_{\min} = \frac{x_{\min}}{y} \quad \text{and} \quad \left[\frac{x}{z}\right]_{\min} = \frac{x_{\min}}{z}.$$

Similar to $S/\left[S\right]_{\max}$, the outcome of $S/\left[S\right]_{\min}$ does not depend on the definitions of y and z with respect to power means. The proof of the next result is similar to the proof of Proposition 5.1.

Proposition 5.4. *Let x/y and x/z be two real positive values defined as above. Then*

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) / \left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\min} = \frac{x}{|x_{\min}|}.$$

As a first example, consider the power mean of

$$x = \frac{ad - bc}{p_1 q_1} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_2}.$$

We have

$$\frac{M_\theta(x, y)}{|[M_\theta(x, y)]_{\min}|} = \lim_{\theta \to \infty} M_\theta \left(\frac{ad - bc}{p_1 q_1}, \frac{ad - bc}{p_2 q_2} \right) = \frac{ad - bc}{\min(p_1 p_2, q_1 q_2)}$$

which is a special case of the power mean. As a second example, consider the power mean of

$$x = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_1}.$$

Again, we obtain

$$\frac{M_\theta(x, y)}{|[M_\theta(x, y)]_{\min}|} = \lim_{\theta \to \infty} M_\theta \left(\frac{ad - bc}{p_1 q_2}, \frac{ad - bc}{p_2 q_1} \right) = \frac{ad - bc}{\min(p_1 p_2, q_1 q_2)}$$

which is not a special case of this power mean.

We end this chapter with an argument made in Davenport and El-Sanhurry (1991). These authors argue that studying the minimum of $(ad - bc)$ is somewhat trivial. The minimum problem can be turned into a maximum problem at any time, simply by recoding the values of one of the binary variables. Maximum and minimum of $(ad - bc)$ are given by

$$(ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1) \quad \text{and} \quad (ad - bc)_{\min} = -\min(p_1 p_2, q_1 q_2).$$

Suppose that the observations of the second variable are recoded, $1 \to 0$ and $0 \to 1$, for example

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \to \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Note that the recoding changes the sign of the covariance $(ad - bc)$ between the two binary vectors. Furthermore, for the second vector $p_2 \to q_2$ and $q_2 \to p_2$. Multiplying $(ad - bc)_{\min}$ by -1 and changing the roles of p_2 and q_2 in $(ad - bc)_{\min}$, we obtain $(ad - bc)_{\max}$.

5.4 Epilogue

In this chapter it was shown that various coefficients become equivalent if they are divided by their maximum value given fixed marginal probabilities p_1 and p_2. For example, the power mean of the quantities

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

has as special cases

$$\begin{aligned} S_{\text{BB}} &= \frac{a}{\max(p_1, p_2)} \\ S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} \\ S_{\text{DK}} &= \frac{a}{\sqrt{p_1, p_2}} \\ \text{and} \quad S_{\text{Kul}} &= \frac{1}{2}\left[\frac{a}{p_1} + \frac{a}{p_2}\right]. \end{aligned}$$

By Proposition 5.1, S_{BB}, S_{Gleas}, S_{DK} and S_{Kul} coincide after correction for maximum value. Furthermore, by Corollary 5.1 all special cases of the power mean become equivalent to the maximum function (also a special case) of the two quantities. For example, S_{BB}, S_{Gleas}, S_{DK} and S_{Kul} become

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

after correction (5.1). As a second example, by Proposition 5.1 and Corollary 5.1,

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

are special cases of the power mean of

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

Coefficient S_{Cohen} and S_{Phi} become

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

after correction for maximum value. Moreover, by Proposition 5.3, S_{Cole1}, S_{Cole2},

$$S_{\text{MP}} = \frac{2(ad - bc)}{p_1 q_1 + p_2 q_2} \quad \text{and} \quad S_{\text{Fleiss}} = \frac{(ad - bc)(p_1 q_1 + p_2 q_2)}{2 p_1 q_2 p_2 q_1}$$

also become equivalent to S_{Loe}, after division by their maximum value given fixed marginals p_1 and p_2.

5.5 Loevinger's coefficient

Correction for chance and correction for maximum value were treated separately in Chapters 4 and 5. This section is used to show two properties of

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

the coefficient by Loevinger (1947, 1948), with respect to correction for chance and correction for maximum value simultaneously. With respect to both properties it is assumed that $E(a)_{\text{Cohen}} = p_1 p_2$ is the appropriate expectation.

First of all, if $E(a) = p_1 p_2$ and $a_{\max} = \min(p_1, p_2)$, then coefficient S_{Loe} can be defined as

$$S_{\text{Loe}} = \frac{a - E(a)}{a_{\max} - E(a)}$$

or dually

$$S_{\text{Loe}} = \frac{d - E(d)}{d_{\max} - E(d)}$$

where $E(d) = q_1 q_2$ and $d_{\max} = \min(q_1, q_2)$. Furthermore, under the same conditions, any coefficient in the $\mathcal{L}$ family (of the form $\lambda + \mu a$) becomes S_{Loe} after correction for maximum value and correction for chance. Moreover, the result does not depend on what correction is considered first.

Proposition 5.5. *A coefficient of the form $\lambda + \mu a$ becomes S_{Loe} after correction* (4.1) *and* (5.1).

Proof: Dividing coefficient $\lambda + \mu a$ by its maximum value given fixed marginals p_1 and p_2, we obtain

$$\frac{\lambda + \mu a}{\lambda + \mu \min(p_1, p_2)}. \tag{5.2}$$

The expectation of (5.2) is given by

$$E\left[\frac{\lambda + \mu a}{\lambda + \mu \min(p_1, p_2)}\right] = \frac{\lambda + \mu E(a)}{\lambda + \mu \min(p_1, p_2)} = \frac{\lambda + \mu p_1 p_2}{\lambda + \mu \min(p_1, p_2)}. \tag{5.3}$$

Using (5.2) and (5.3) in (4.1), and multiplying by $\lambda + \mu \min(p_1, p_2)$, we obtain

$$\frac{\lambda + \mu a - \lambda - \mu p_1 p_2}{\lambda + \mu \min(p_1, p_2) - \lambda - \mu p_1 p_2} = \frac{a - p_1 p_2}{\min(p_1, p_2) - p_1 p_2} = S_{\text{Loe}}.$$

Alternatively, Using $\lambda + \mu a$ and the corresponding expectation

$$\lambda + \mu p_1 p_2$$

in (4.1), we obtain

$$\frac{\lambda + \mu a - \lambda - \mu p_1 p_2}{1 - \lambda - \mu p_1 p_2} = \frac{a - p_1 p_2}{(1 - \lambda)/\mu - p_1 p_2}. \tag{5.4}$$

The maximum value of (5.4) given fixed marginals p_1 and p_2, is given by

$$\frac{\min(p_1, p_2) - p_1 p_2}{(1-\lambda)/\mu - p_1 p_2}. \tag{5.5}$$

Dividing (5.4) by (5.5), we obtain

$$\frac{a - p_1 p_2}{\min(p_1, p_2) - p_1 p_2} = S_{\text{Loe}}.$$

This completes the proof. □

Zero value under statistical independence, and maximum value unity independent of the marginal distributions, are two properties or desiderata that similarity coefficients may have in general. Proposition 5.5 shows that the linear transformations that set the value under independence at zero (4.1) and the maximum value at unity (5.1), transform all coefficients in $\mathcal{L}$ family (of the form $\lambda + \mu a$) into the same underlying coefficient. This coefficient happens to be S_{Loe}.