Journal of Classification 7:5-31 (1990)

Distribution of Some Similarity Coefficients for Dyadic Binary Data in the Case of Associated Attributes

Tom A. B. Snijders
University of Groningen

Maarten Dormaar
University of Limburg

Wijbrandt H. van Schuur
University of Groningen

Chantal Dijkman-Caes
University of Limburg

Ger Driessen
University of Limburg

Acknowledgement. The authors are grateful to the editor, the referees, and Melissa Bowerman for suggestions leading to an improved presentation.

Authors' Addresses: Tom A. B. Snijders and Wijbrandt H. van Schuur, Department of Statistics and Measurement Theory, FPPSW, University of Groningen, Oude Boteringestraat 23, 9712 GC Groningen, The Netherlands. Maarten Dormaar, Chantal Dijkman-Caes and Ger Driessen, Department of Social Psychiatry, University of Limburg, P.O. Box 616, 6200 MD Maastricht, The Netherlands.

Abstract: Parameters are derived of distributions of three coefficients of similarity between pairs (dyads) of operational taxonomic units for multivariate binary data (presence/absence of attributes) under statistical independence. These are applied to test independence for dyadic data. Association among attributes within operational taxonomic units is allowed. It is also permissible for the two units in the dyad to be drawn from different populations having different presence probabilities of attributes. The variance of the distribution of the similarity coefficients under statistical independence is shown to be relatively large in many empirical situations. This result implies that the practical interpretation of these coefficients requires much care. An application using the Jaccard index is given for the assessment of consensus between psychotherapists and their clients.

Résumé: La distribution des coefficients de similarité pour les données binaires et les attributs associés. Les paramètres de la distribution de trois coefficients de similarité entre paires d'éléments taxinomiques opérationels de données multivariables binaires (présence/absence) ont été dérivés dans l'hypothèse d'indépendance statistique. Ces paramètres sont utilisés dans un test d'indépendance pour les données dyadiques. L'existence est autorisée, dans la population d'éléments, d'une association entre plusieurs attributs. Il est également permis que les deux éléments de la dyade soient tirés de deux populations différentes, ayant différentes probabilités quant à la présence des attributs. Dans beaucoup de situations empiriques, la variance des coefficients de similarité peut être relativement élevée dans le cas d'indépendance statistique. Par conséquence, ces coefficients doivent être interprétés avec précaution. Un exemple est donné pour le coefficient de Jaccard, qui a été employé dans une recherche sur la concordance entre des psychothérapeutes et leurs clients.

Keywords: Consensus; Dice coefficient; Jaccard coefficient; Simple Matching coefficient; Multivariate binary data; Observer agreement; Similarity coefficients; Beta distribution.

1. Introduction: Some Similarity Coefficients for Binary Data

Similarity between operational taxonomic units (or *units*, for short) can be defined on the basis of the common presence and absence of attributes. Coefficients of similarity, or association, based on multivariate binary data are widely used in the field of taxonomy (Sneath and Sokal 1973; Everitt 1980; Anderberg 1976), in biology (Washington 1984), and in other disciplines (Austin and Colwell 1977; Hubalek 1982). These coefficients are often used to convert a two-mode (e.g., units by attributes) matrix to a one-mode (units by units) matrix of proximities between pairs of units. In this paper we treat the case where data arise in pairs of units, or *dyads*, so that representation in a two-mode matrix would imply a loss of information. The dependence between the units in the dyad will be investigated; in other words, the paper is concerned with the Q mode of analysis (relations among units) rather than the

R mode (relations among variables). We restrict attention to three of the better-known similarity coefficients: the Simple Matching coefficient, M (Sokal and Michener 1958), the Jaccard index, J (Jaccard 1900, 1908), and the Dice coefficient, D (Dice 1945; see also Dice 1952; also called Czekanowski-Dice coefficient by Wishart 1978). These are defined by

$$M = \frac{a+d}{a+b+c+d} \quad J = \frac{a}{a+b+c} \quad D = \frac{2a}{2a+b+c}, \tag{1}$$

where a, b, c, and d refer to entries in the contingency table for two units based on attributes present (1) or absent (0) among n attributes:

(2)

		Unit B		
		1	0	
Unit A	1	a	b	$a+b=N_A$
	0	c	d	$c+d$
		$a+c=N_B$	$b+d$	$a+b+c+d=n$

The similarity coeficients J and D are increasing functions of each other:

$$J = D/(2-D), \; D = 2J/(1+J) \; . \tag{3}$$

Thus, J and D express the same information in numerically different ways. Probability statements about J and about D can be transformed into each other. Our interest is more in J than in D; the coefficient D is discussed because, as will be shown in Section 6, the mean and variance of D can be calculated more easily than those of J, so that D can be used as an intermediary to study the statistical properties of J. Our discussion of similarity coefficients is concerned with testing their values for dyads under a null model of absence of association between units A and B; we do not consider the application of such coefficients in cluster analysis.

According to each coefficient M, J, and D, two units are more similar to the extent that a larger number of attributes is jointly present in both. Each coefficient ranges between 0 and 1. The difference between J and D on the one hand and M on the other lies in the treatment of joint absence of attributes in both units. These "negative matches" are ostensibly ignored in J and D, but used in M. The Jaccard index was proposed by Jaccard (1900, 1908) in order to remove the artifact that two units (in his study: ecological sites) would be considered similar because *neither* contains many of the attributes (in his study: plant species under consideration).

2. Comparing Observed Values of Association Coefficients with Expected Values Under Statistical Independence: Formulation of a Null Hypothesis and Tests of Significance

In applications of similarity coefficients, substantive theory often predicts similarity between specific units (individuals, species, sites) to be higher than between arbitrary units. This situation can lead the researcher to test the similarity coefficients for specific dyads against a null hypothesis of independence (e.g., Coquin-Viennot 1975). However, knowing that a dyad's similarity coefficient is higher than expected under the null hypothesis is not enough; we want to know whether it is *significantly* higher. In the absence of criteria for determining significantly high values of similarity, some researchers (e.g., Quesada, Ventosa, Rodriguez-Valera, Megias, and Ramos-Cormenzana 1983) resort to using arbitrary levels (e.g., 0.65) as a lower boundary for an interpretable level of similarity. The present paper is devoted to a presentation of procedures for a significance test.

Such coefficients as M, D, or J are not only sensitive to similarity between the two units, but also to other characteristics of the data. Since the aim of using these similarity coefficients is to study similarity within dyads, it is necessary to control suitably for factors which could influence these coefficients, and thereby may cause problems in assessing their observed values. Two of these factors explicitly to be taken into account are discussed first.

The first is that some attributes may be present more frequently than others. If so, average values of similarity coefficients will be higher than if all attributes have equal relative frequencies. When units A and B in (2) are drawn from two different populations (e.g., from populations of therapists and clients, as in Section 3) a further complication arises. If the frequency orders of attributes are the same in both populations, the average values of similarity coefficients will be higher than if the frequency orders are different. Differential frequency of attributes, and different populations of units A and B, do not constitute serious problems, but they do need to be taken into account.

A second factor results from the association *within* units among the given attributes. The attributes used in applications of M, D, and J may be related in a meaningful way; in some applications they can be regarded as different indicators of one or more underlying latent concepts. When the attributes are associated from a conceptual point of view, they will be mutually statistically dependent in the population(s) of units. This dependence affects the probability distributions of M, D, and J. We may expect that association among attributes will affect the standard deviations of the similarity coefficients, since it affects the standard deviations of the numbers N_A and N_B of attributes present. This problem is more serious than the first one and has

hardly been addressed in the literature we know, which mostly assumes, explicitly or — more often — implicitly, that under the null hypothesis the attributes are mutually independent within units. (An exception is Heltshe (1988), who gave jackknife estimates for the Simple Matching and Jaccard coefficients in the situation in which a quadrat sample is taken to estimate the population value of these similarity coefficients. This situation is essentially different from the one we have studied.) We propose to take these problems into account, first for the Simple Matching coefficient, and subsequently for the Jaccard index, using the Dice index as an intermediate step.

These considerations lead to the following formulation of the null hypothesis that in our opinion is most relevant for testing an observed value of a similarity coefficient. A population of dyads is postulated. Either the two units in each dyad are *a priori* indistinguishable, or they can be distinguished as one unit of type A and the other of type B, drawn, respectively, from a population A and a population B of units. For a given set of n attributes, the presence (denoted by 1) or absence (denoted by 0) of each attribute can be observed for both members of the dyads investigated. There may be association (i.e., statistical dependence) among attributes within units (see Coleman, Mares, Willig, and Hsieh 1982, for a formulation of independence in which association is assumed to be absent). If units of types A and B are distinguished, then the joint probability distribution of attributes present may differ between population A and population B. The null hypothesis is *independence between units* in a single dyad under investigation.

It is assumed in this paper that sufficiently large samples of units A and B are available that the joint probability distribution of presence of the attributes within either population of units are known. Parameters of the probability distribution of the similarity coefficients M, D, and J will be derived under the null hypothesis. The null distribution of many other similarity coefficients can be investigated using the same approach.

The null distribution of similarity coefficients in the situation where no distinction between units A and B is made is identical to the null distribution in the situation in which such a distinction *is* made, but the joint probability distribution in population A is the same as in population B. Therefore, no generality is lost by formulating all results for the case where the distinction between the two types is made.

3. Using the Jaccard Index for the Measurement of Consensus

Agreement may have cultural or interpersonal origins. In the first case, two persons hold the same opinions quite independently: they share some beliefs or attributes without having been in contact with each other. In the second case, they agree because they have exchanged views on the subject-

matter and have thus influenced each other's previously-held opinions. The latter situation is called "co-orientation through communication" in Scheff's consensus model, where this process is postulated as the basis of interpersonal consensus. If psychotherapist and client have discussed the nature of the latter's problem, as is generally the case in the first therapeutic contacts, communication has explicitly taken place.

We have been using the Jaccard index to measure consensus between a psychotherapist and his client with respect to the definition of the client's problem(s). Attributes are here specific formulations of possible problems. If both therapist and client agree that certain attributes (problem formulations) apply to the client, then they exhibit agreement or "zero-level of co-orientation" (Scheff 1967). If, on the other hand, both therapist and client agree that certain attributes do not apply to the client, then this agreement should not be taken to contribute to consensus between therapist and client. Therefore, the J index is appropriate here; it avoids the "common deficiencies" problem present in many other coefficients (Gregson 1975, p. 48).

In our research on consensus (Dormaar, Dijkman-Caes, and De Vries, 1989), we used data based on the responses of psychiatric clients and their therapists to a list with eight problem formulations. Clients and therapists were asked to state for each problem formulation whether they considered it valid for the client. The 162 clients filled out the list after the second therapy session and so did their therapists. There were 22 different therapists who participated between 2 and 12 times in this research and who completed the list of problem formulations for a total of 115 of the 162 clients. For this sample of 115 client-therapist dyads the mean J-value was 0.49 and the standard deviation 0.21.

We hypothesize that a distribution of consensus values in a population of dyads, each consisting of "client and client's *own* therapist," will have a higher mean value than the distribution of these values in a population of "client - *arbitrary* therapist" dyads. We expect that within the first type of dyads co-orientation will have occurred, leading to more agreement and higher levels of consensus than within dyads of the second type. The distribution for dyads of the second type corresponds to the null hypothesis formulated above that clients and therapists respond to the problem formulations independently.

Empirical rejection of this null hypothesis would provide support for the model of interpersonal consensus outlined above. Moreover, the availability of the probability distribution under the null hypothesis of statistical independence would enable the researcher to select those dyads which exhibit evidence of having gone successfully through the process of co-orientation. These dyads could be defined as those reaching a degree of

consensus that is highly improbable under the null hypothesis of mutual independence, operationalized, for instance, as J-values with p-values under statistical independence lower than 5%.

As a first way of answering the question whether this value can be interpreted as a high level of consensus, a simulation study was carried out, analogous to Johnson and Millie's (1982) procedure to find confidence intervals for Stander's similarity index. A dataset of 10,000 independent dyads was randomly simulated on the basis of the marginal popularities of the eight problem formulations for clients and therapists separately. The simulation model assumed no association among the attributes (problem formulations). In this case the mean J-value was 0.38 and the standard deviation 0.19.

In a second simulation approach all 115 response vectors of clients to the eight problem formulations were combined with all 115 therapist response vectors. This resulted in a total of $115 \times 115 = 13{,}225$ simulated J-values, among which 115 were genuinely observed and the other 13,110 were hypothetical. The entire set of 13,225 simulated dyads had a mean J-value of 0.39 with a standard deviation of 0.22. The difference between this value and that of the genuine dyads is considerable, but it cannot be legitimately tested with the t-test because the assumption of independent observations is not valid.

These simulation results demonstrate that the standard deviation of J is affected by the presence of association among attributes within units. Rather than using simulation studies to assess the deviation from our observed J-values to values expected under statistical independence, we now turn to a more formal approach to this problem.

4. Mean and Variance of the Simple Matching, Dice, and Jaccard Coefficients Under Absence of Association

Goodall (1967) considered the Simple Matching coefficient in the situation in which attributes are independent within units, and in which there is no distinction between types of units A and B. Denoting the probability of presence of attribute j by p_j, Goodall derived under the null hypothesis that

$$E(M) = \mu = n^{-1} \sum_{j=1}^{n} (p_j^2 + (1-p_j)^2) \tag{4}$$

$$\text{var}(M) = n^{-1}(\mu(1-\mu) - n^{-1} \sum_{j=1}^{n} (p_j^2 + (1-p_j)^2 - \mu)^2) \ . \tag{5}$$

These formulae can be derived by expressing M as a mean (over attributes) of agreement indicators. They can simply be extended to the situation in which there is a distinction between units of types A and B.

A similar straightforward approach to the mean and variance of J or D is not possible, because the denominators of J and D are random whereas the denominator of M is fixed. In this section we consider some previously published formulae for the mean and variance of J and D, as well as one new formula, which all assume absence of association of attributes within units. The formulae in the literature deal with the randomness of the denominators by using the delta method (see, e.g., Bishop, Fienberg and Holland 1975, Section 14.6). This is an approximation valid for large values of n, the number of attributes. The practical validity of the formulae is doubtful in many cases because similarity coefficients are often applied when the value of n is rather small.

Goodall (1978, expression (46) on p. 126) gives the expression

$$E(J) = (\sum_{j=1}^{n} p_j^2) / (\sum_{j=1}^{n} p_j (2-p_j)) \qquad (6)$$

(mistakenly omitting the division slash), which is obtained by replacing both the numerator and the denominator of J by their expected values. For large n, this approximation is valid. A related exact approach is the following. Under the condition that attribute j is present for at least one of the units, the conditional probability that it is present for both units is given by $p_j / (2-p_j)$. Defining

$$I_j = \begin{cases} 1 & \text{if attribute } j \text{ is present for at least one of the units} \\ 0 & \text{if attribute } j \text{ is present for neither unit,} \end{cases}$$

the vector $I = (I_1, \ldots, I_n)$ indicates those attributes which contribute to the denominator of J. This denominator can be expressed as

$$a + b + c = \sum_{j=1}^{n} I_j .$$

Conditional on the outcome of I, the expected value of J is

$$\begin{aligned} E\{J \mid I_1, I_2, \ldots, I_n\} &= \{\sum_{j=1}^{n} I_j p_j / (2-p_j)\} / \{a + b + c\} \\ &= \sum_{j=1}^{n} \{I_j / (a + b + c)\}\{p_j / (2-p_j)\} . \end{aligned}$$

To calculate the expected value of J from this formula, the random variables $I_j / (a + b + c)$ must be replaced by their expected values. Note that the sum over j of these random variables (and hence of their expected values) is necessarily equal to 1. If $p_j = p$ independently of j, the result is

$$E(J) = p / (2 - p) . \tag{7}$$

In other cases, this approach seems to be intractable: again because of the random nature of the denominator $(a + b + c)$, the expected value of $I_j / (a + b + c)$ cannot be computed. However, this approach does allow the conclusion, that $E(J)$ is a weighted average of the values $p_j / (2 - p_j)$, where the weights depend on $p_1, \ldots, p_n$.

Janson and Vegelius (1981), and Elston, Schroeder, and Rojahn (1982) considered several similarity coefficients, among them the Jaccard coefficient, in the situation in which attributes are independent within units and each attribute has the same probability of being present. These assumptions are very restrictive and likely to be unrealistic in most applications. However, these authors did distinguish between units of type A and B, and they derived formulae which are valid also for the alternative hypothesis that there is dependence among units within dyads. We will discuss their approach briefly.

Denote the probabilities of presence of an attribute j for units of types A and B by p_A and p_B, respectively. The probability of joint presence of an attribute for both units is denoted by p_{AB}. Under the null hypothesis of independence between units within dyads, $p_{AB} = p_A p_B$. Both Janson and Vegelius and Elston et al. derived the approximate formulae

$$E(J) \simeq p_{AB} / (p_A + p_B - p_{AB}) \tag{8}$$

$$\text{var}(J) \simeq \{p_{AB} (p_A + p_B - 2p_{AB})\} / \{n(p_A + p_B - p_{AB})^3\} . \tag{9}$$

By conditioning on $I_1, \ldots, I_n$ as in the derivation of (7), it can be concluded that (8) is an exact expression for $E(J)$, provided that the strong assumptions made do indeed hold. If units A and B are not distinguished, and the units are independent within dyads, (8) reduces to (7). Janson and Vegelius also give

$$E(D) \simeq 2p_{AB} / (p_A + p_B) \tag{10}$$

$$\text{var}(D) \simeq 4p_{AB}\{(p_A + p_B - p_{AB})(p_A + p_B - 2p_{AB})\} / \{n(p_A + p_B)^4\} . \tag{11}$$

Baroni-Urbani and Buser (1976) and Baroni-Urbani (1980) considered the Jaccard coefficient when all attributes are independent and have a common probability of 1/2 of being present. This situation is a special case of the

one considered by Janson and Vegelius and by Elston et al., and it presumably has little practical significance. Connor and Simberloff (1978) also mentioned a null hypothesis in which each attribute has its own mean and variance. However, since they considered it extremely difficult to determine these values, they did not elaborate.

5. Mean and Variance of M for Two Sets of Units and Associated Attributes

In this section we give formulae for the mean and variance of the Simple Matching coefficient under the null hypothesis of independence between units within dyads, taking account of the association among attributes within units. The following definitions are needed. Define the random variables

$$X_{Aj} = \begin{cases} 1 & \text{if attribute } j \text{ is present for unit } A \\ 0 & \text{if attribute } j \text{ is not present for unit } A \end{cases}$$

and X_{Bj} similarly. It will be convenient not to work with a, b, c, and d as in Table 2 but with

$$N_A = a + b, \quad N_B = a + c, \quad N_{11} = a \tag{12}$$

which can be expressed by

$$N_A = \sum_{j=1}^{n} X_{Aj}, \quad N_B = \sum_{j=1}^{n} X_{Bj}, \quad N_{11} = \sum_{j=1}^{n} X_{Aj} X_{Bj} .$$

The Simple Matching coefficient is then given by

$$M = (2N_{11} - N_A - N_B + n) / n . \tag{13}$$

Define

$p_{Aj} = P\{X_{Aj} = 1\} =$ probability of presence of attribute j for a unit of type A ($j = 1, \ldots, n$)

$p_{Ajh} = P\{X_{Aj} = X_{Ah} = 1\} =$ probability of simultaneous presence of attributes j and h for a unit of type A ($j, h = 1, \ldots, n$),

and define p_{Bj} and p_{Bjh} similarly. Note that $p_{Ajh} = p_{Ahj}$ for all (j,h), and that $p_{Ajj} = p_{Aj}$. Replacing an index j and/or h by a + sign will denote summation over that index; e.g.,

$$p_{A++} = \sum_{j=1}^{n} \sum_{h=1}^{n} p_{Ajh} .$$

It follows from these definitions that

$$p_{Aj} = E(X_{Aj}),\ p_{A+} = E(N_A),\ p_{B+} = E(N_B) . \tag{14}$$

E.g., p_{A+} is the mean number of attributes present for a unit of type A. Now define

$$m_{11} = \sum_{j=1}^{n} p_{Aj}\, p_{Bj} . \tag{15}$$

Then the number m_{11} is the mean number of attributes simultaneously present for both units within an independent dyad:

$$m_{11} = E(N_{11}) . \tag{16}$$

It follows from (13), (14), and (16) that the mean of the Matching Coefficient under the null hypothesis is given by

$$E(M) = (2m_{11} - p_{A+} - p_{B+} + n) / n . \tag{17}$$

For an expression for the variance of M, we also need to define

$$\begin{aligned} m_{12} &= \sum_{j=1}^{n} \sum_{h=1}^{n} p_{Aj}\, p_{Bjh} , \\ m_{21} &= \sum_{j=1}^{n} \sum_{h=1}^{n} p_{Ajh}\, p_{Bj} , \\ m_{22} &= \sum_{j=1}^{n} \sum_{h=1}^{n} p_{Ajh}\, p_{Bjh} . \end{aligned} \tag{18}$$

For these quantities there is no simple interpretation analogous to (16). We derive in the Appendix that

$$\begin{aligned} \operatorname{var}(M) = n^{-2}(4\{&m_{22} + m_{11}(p_{A+} + p_{B+}) \\ &- m_{21} - m_{12} - m_{11}^2\} + p_{A++} + p_{B++} - p_{A+}^2 - p_{B+}^2) . \end{aligned} \tag{19}$$

These expressions will be used in Section 7 to derive an approximation to the cumulative distribution function of M, which can be used to test observed values of M.

6. Mean and Variance of D and J

For the Jaccard and Dice coefficients, the situation is more complicated because of the randomness of the denominators. Using the definitions in (12), the Jaccard coefficient can be expressed as

$$J = \frac{N_{11}}{N_A + N_B - N_{11}} . \tag{20}$$

Useful exact expressions for $E(J)$ and $\mathrm{var}(J)$ cannot be derived for the general case in which associated attributes are allowed. Two approximations for $E(J)$ and $\mathrm{var}(J)$ will be presented. The first yields simpler formulae but is much less precise; however, it may be applied in situations in which the data needed for the second approach are not available. For the second approach we use the Dice coefficient to obtain helpful intermediate results.

The first approach uses the delta method (cf. Section 4), which is an approximation valid for large values of n, the number of attributes. Expression (20) is rewritten as

$$J = \frac{P_{11}}{P_A + P_B - P_{11}} \tag{21}$$

where $P_{11} = N_{11} / n$, $P_A = N_A / n$, $P_B = N_B / n$; this non-linear function is then approximated by a function which is linear in P_A, P_B and P_{11}. Proofs are given in the Appendix. The results are

$$E(J) \approx \frac{m_{11}}{p_{A+} + p_{B+} - m_{11}} \tag{22}$$

$$\begin{aligned} \mathrm{var}(J) \approx\ & n^{-1} (p_{A+} + p_{B+} - m_{11})^{-4} \\ & \times \{2m_{11}^2 (p_{A++} + p_{B++} + p_{A+} p_{B+}) \\ & - 2m_{11}(m_{12} + m_{21})(p_{A+} + p_{B+}) + m_{22}(p_{A+} + p_{B+})^2\} . \end{aligned} \tag{23}$$

Expressions (6) - (9) are special cases of (22) and (23). In the continuation of our example presented in Section 8, we show that these approximations are much less precise than those obtained in the second approach. This result is not surprising, since $n = 8$ in this example whereas the delta method assumes that n is large.

The second approach makes use of the fact that the exact mean and variance of the Dice coefficient can be derived. It must be assumed that $P\{N_A = 0\} = P\{N_B = 0\} = 0$. Otherwise, it would be possible for the

denominators of D and J to be 0, thus leaving D and J undefined. In practice, J will be computed only for those cases in which $N_A > 0$ and $N_B > 0$, so that this assumption is not restrictive from a practical point of view.

Define

$p_A(n_A) =$ $P\{N_A = n_A\}$, the probability that a total of n_A attributes is present for unit A;

$p_{Aj}(n_A) =$ $P\{N_A = n_A \text{ and } X_{Aj} = 1\}$, the probability that a total of n_A attributes is present for unit A, among which is attribute j;

$p_{Ajh}(n_A) =$ $P\{N_A = n_A \text{ and } X_{AJ} = X_{Ah} = 1\}$, the probability that a total of n_A attributes is present for unit A, among which are attributes j and h;

similarly, define $p_B(n_B)$, $p_{Bj}(n_B)$, and $p_{Bjh}(n_B)$. In the Appendix, the following expressions for the mean and variance of D are derived:

$$E(D) = 2 \sum_{n_A=1}^{n} \sum_{n_B=1}^{n} ((n_A + n_B)^{-1} \sum_{j=1}^{n} p_{Aj}(n_A) p_{Bj}(n_B)) \tag{24}$$

$$= 2 \sum_{n_A=1}^{n} \sum_{n_B=1}^{n} ((n_A + n_B)^{-1} \{c(n_A, n_B) + n_A n_B p_A(n_A) p_B(n_B) / n\}) \tag{25}$$

where

$$c(n_A, n_B) = \sum_{j=1}^{n} (\{p_{Aj}(n_A) - n_A p_A(n_A) / n\} \{p_{Bj}(n_B) - n_B p_B(n_B) / n\});$$

$$\text{var}(D) = E(D^2) - (E(D))^2$$

$$= 4 \sum_{n_A=1}^{n} \sum_{n_B=1}^{n} ((n_A + n_B)^{-2} \sum_{j=1}^{n} \sum_{h=1}^{n} p_{Ajh}(n_A) p_{Bjh}(n_B)) - (E(D))^2 \tag{26}$$

$$= 4 \sum_{n_A=1}^{n} \sum_{n_B=1}^{n} ((n_A + n_B)^{-2} \{d(n_A, n_B) \tag{27}$$

$$+ n_A^2 n_B^2 p_A(n_A) p_B(n_B) / n^2\}) - (E(D))^2$$

where

$$d(n_A, n_B) = \sum_{j=1}^{n} \sum_{h=1}^{n} \{p_{Ajh}(n_A) - n_A^2 p_A(n_A) / n^2\}$$

$$\times \{p_{Bjh}(n_B) - n_B^2 p_B(n_B) / n^2\}.$$

The total contribution to (25) of the terms following the + sign is the expectation of the random variable

$$\frac{N_A N_B}{n(N_A + N_B)}$$

which is an increasing function of N_A and N_B. This contribution increases as the popularity of the attributes for units A and/or B increases. The total contribution of the terms $c(n_A,n_B)$ in (25) is higher when the frequency of particular attributes is similar in populations A and B than when it is dissimilar. This point was already mentioned in a qualitative sense in Section 2.

The expressions for $E(D)$ and $\mathrm{var}(D)$ derived above can be employed to derive approximate expressions for $E(J)$ and $\mathrm{var}(J)$ by a variant of the delta method. For approximating $E(J)$ we use the second-order Taylor expansion

$$J = D/(2-D) \approx \{E(D)/(2-E(D))\} + 2(D-E(D))(2-E(D))^{-2} + 2(D-E(D))^2(2-E(D))^{-3};$$

the quadratic term employed here is a refinement of the usual implementation of the delta method, and yields a considerably better approximation to $E(J)$. Note that, in the right hand side of this approximate equality, $E(D)$ is constant. Taking expectations with respect to the random variable D yields the approximation

$$E(J) \approx E(D)/(2-E(D)) + 2(2-E(D))^{-3}\,\mathrm{var}(D). \qquad (28)$$

For approximating $\mathrm{var}(J)$ we use the first-order Taylor expansion

$$J = D/(2-D) \approx \{E(D)/(2-E(D))\} + 2(D-E(D))(2-E(D))^{-2};$$

using the second-order Taylor expansion here would necessitate expressions for the third and fourth moments of D, which we have not wished to derive and which presumably would not contribute much to the accuracy of the resulting expression for $\mathrm{var}(J)$. Taking the variance of the right hand side yields the approximation

$$\mathrm{var}(J) \approx 4(2-E(D))^{-4}\mathrm{var}(D). \qquad (29)$$

The approximation of $E(J)$ and $\mathrm{var}(J)$ according to the first approach will usually not be very reliable, since the non-linear nature of J as a function

of (P_A,P_B,P_{11}) will often be rather pronounced in the region where most of the probability mass of (P_A,P_B,P_{11}) is concentrated. Note that approximation (22) neglects the fact that the denominator of (21) is random. Since the non-linear nature of (3) is less pronounced than that of function (21), and a second-order Taylor series for (3) is used for approximating $E(J)$, the second approach may be expected to yield better approximations than the first. On the other hand, although approximations (22)-(23) are rougher than (28)-(29), the former can be useful because less information on the joint probability distributions is needed to evaluate them. There may be situations in which the values of (22)-(23) can be estimated from available data (or guessed), while (28)-(29) cannot be reliably estimated.

7. Approximations to the Distribution Functions of M and J

In order to make probability statements for assessing the significance of a given observed value of M, D, or J, it is necessary to evaluate the cumulative distribution function (*cdf*), or at least to approximate it. In Section 6 exact expressions were given for the mean and variance of D, but only approximate ones for the mean and variance of J. Therefore it seems sensible in approximating the *cdf* of J to use the exact expressions for $E(D)$ and $\mathrm{var}(D)$, to approximate the *cdf* of D, and then to use relation

$$P\{J \le t\} = P\{D \le 2t/(1+t)\}, \quad (0 \le t \le 1) \tag{30}$$

implied by (3), rather than use the approximate expressions for $E(J)$ and $\mathrm{var}(J)$.

To approximate the *cdf* of M or D, one could use either a normal approximating distribution or a non-normal one that takes into account the boundedness of the interval [0,1] of possible values of M and D. The fact that M and D are restricted to the interval [0,1] causes problems in a normal approximation, because the standard deviation in many practical cases is rather large. It seems therefore preferable to consider the approximation using a Beta distribution, which is a distribution on the interval (0,1) having probability density function

$$f(x) = \{B(p,q)\}^{-1} x^{p-1} (1-x)^{q-1}, \quad (0 < x < 1)$$

where p and q are positive parameters and $B(p,q)$ is the beta function defined by

$$B(p,q) = \int_0^1 x^{p-1} (1-x)^{q-1} \, dx \, .$$

The *cdf* of the Beta distribution will be denoted by $B(x;p,q)$. The mean and variance of this distribution are, respectively,

$$\mu = p / (p+q) \quad \sigma^2 = pq / \{(p+q)^2 (p+q+1)\} \, ; \tag{31}$$

see, e.g., Johnson and Kotz (1970, p.40). The distributions of M and D can be approximated by Beta distributions. There exists only one Beta distribution with a given mean μ and variance σ^2. The parameters are determined by inverting relations (31), which results in

$$p = \mu^2(1-\mu) / \sigma^2 - \mu, \; q = \mu(1-\mu)^2 / \sigma^2 - 1 + \mu \tag{32}$$

(cf. Johnson and Kotz 1970, p. 44). The procedure for approximating the *cdf* of M, D or J now is the following:

1. Calculate ($E(M)$ and $\mathrm{var}(M)$) or ($E(D)$ and $\mathrm{var}(D)$);
2. Set ($\mu = E(M), \sigma^2 = \mathrm{var}(M)$) or ($\mu = E(D), \sigma^2 = \mathrm{var}(D)$);
3. Calculate the parameters p and q of the Beta distribution from (32);
4. Employ a numerical procedure to calculate the *cdf* of the Beta distribution with these values of p and q, e.g., by using subroutine MDBETA in the IMSL library or function BETAI given by Press, Flannery, Teukolsky, and Vetterling (1986).
5. Procedure (1)-(4) yields approximations to the *cdf* of M or D, respectively. Relation (30) can be used for transforming the *cdf* of D to that of J.

An approximation to the distribution of M and to that of J (possibly via that of D) by a normal distribution is also possible, but such an approximation will in many cases give a rather high probability of values smaller than 0 and/or greater than 1, which is obviously undesirable. In those cases where a normal approximation would yield a good result, this result would presumably not be very different from the approximation by a Beta distribution. The reason is that the variance of M or J, respectively, will be relatively small in such cases. This situation leads to large parameters p and q in the Beta distribution. For such parameter values, the Beta distribution itself approaches a normal distribution. In conclusion, there seems to be no strong reason for using a normal approximation given that a Beta approximation is also available.

Goodall (1967) approximates the *cdf* of M by using a normal distribution for $\arcsin(M^{1/2})$, which also takes into account that $0 \leq M \leq 1$. He suggests this transformation because it stabilizes the variance for a binomial proportion; if all attributes are independent within units and have the same probability of being present, then indeed M is simply a binomial proportion. In the general case, however, there are no special arguments for use of this transformation. A helpful property of the Beta approximation proposed here is the fact that the mean and variance of the approximating Beta distribution are exactly equal to the mean and variance of M or D, respectively.

The approximation by a Beta distribution suffers from the usual problems that occur when a discrete distribution is approximated by a continuous one. A continuity correction will usually improve the approximation and can be carried out in the following way for, e.g., the J index. The possible values for J are the rational numbers a/f where a and f are integers with $0 \leq a \leq f \leq n, 0 < f$. Let t_1, t_2, and t_3 be three consecutive possible values for J: so $0 \leq t_1 < t_2 < t_3 \leq 1$, and there are no possible values for J between t_1 and t_2, nor between t_2 and t_3. The outcome t_2 for J is then identified with the interval of continuous values from $(t_1 + t_2)/2$ to $(t_2 + t_3)/2$. To carry out a continuity correction for approximating $P\{J \geq t_2\}$, define $t_4 = (t_1 + t_2)/2$. Then (30) with the continuity correction leads to

$$P\{J \geq t_2\} = P\{D \geq 2t_4/(1 + t_4)\}. \tag{33}$$

For the right side of (33), the Beta approximation can be used.

In the examples we have calculated, it has turned out that a special problem in approximating the *cdf* of J is caused by the relatively large probability that $J = 0$; see Section 8. From expression (20), note that $P\{J = 0\} = P\{N_{11} = 0\}$; in contrast, the probabilities of strictly positive outcomes of N_{11} are spread over several values for J (depending on the denominator $N_A + N_B - N_{11}$), thereby leading to smaller probabilities for the various positive outcomes of J. A somewhat ad hoc procedure for dealing with this problem, which resulted in markedly increased precision in our examples, is the following. The expected value of N_{11} is, from (16), given by m_{11}; possible values for N_{11} range from 0 to n. Approximating the distribution of N_{11} by a binomial distribution yields

$$P\{J = 0\} = P\{D = 0\} = P\{N_{11} = 0\} \simeq (1 - m_{11}/n)^n. \tag{34}$$

Denote the latter value by π. The distribution of D now can be approximated by a mixed distribution with a probability π for the discrete outcome $D = 0$, and a probability $(1 - \pi)$ for a beta distribution with parameters p_0 and q_0. This distribution has *cdf*

$$\pi + (1-\pi)B(x;p_0,q_0) \tag{35}$$

for $0 \leq x \leq 1$. Elementary probability calculations show that this distribution has mean and variance

$$\mu = \mu_0(1-\pi), \quad \sigma^2 = (1-\pi)(\sigma_0^2 + \pi\mu_0^2), \tag{36}$$

where μ_0 and σ_0^2 correspond to p_0 and q_0 as in (31). The distribution of D is now approximated by equating mean and variance μ and σ^2 in (36) to the exact mean and variance of D. Solving (36) for μ_0 and σ_0^2 shows that p_0 and q_0 can be calculated from (32), substituting

$$\begin{aligned} \mu_0 &= E(D) / (1-\pi), \\ \sigma_0^2 &= (\mathrm{var}(D)) / (1-\pi) - (E(D))^2\, \pi(1-\pi)^{-2} \end{aligned} \tag{37}$$

for μ and σ^2 in those equations. Approximate cumulative probabilities for J can be calculated using *cdf* (35) for D, and using (33) for the correspondence between the *cdf*'s of J and D.

8. Elaboration of the Example of Consensus Assessment

The values found for J in the study of consensus between psychotherapist and client, discussed briefly in Section 3 and more fully in Dormaar, Dijkman-Caes, and De Vries (1989), were investigated under the null model defined in Section 2. The parameters p_{Aj}, etc., were estimated from the sample of 115 psychotherapist-client dyads for which complete data were available. Thus, the estimated null distribution of J coincides with the distribution of the 13,225 simulated J-values obtained by taking all pairs of any psychotherapist and any client, which was referred to in Section 3. This distribution can be regarded as the permutation distribution; consequently, given that estimated probabilities are used, the probability distribution of J for which approximations are given in Sections 6 and 7 is simply the permutation distribution. The availability in this case of the permutation distribution can be used to study the values of the various approximations in this particular case.

Under the permutation distribution, results for the Dice coefficient were $E(D) = 0.521$ and $\mathrm{var}(D) = 0.0555$, while for the Jaccard coefficient $E(J) = 0.386$ and $\mathrm{var}(J) = 0.0476$. The delta method approximations (22) and (23) to the moments of J were 0.491 and 0.0347, while approximations (28) and (29) yielded 0.386 and 0.0463, respectively. The poor performance of the first approximation and the good performance of the second one are striking. With (32), we obtain $p = 1.82$ and $q = 1.68$ as parameters for the

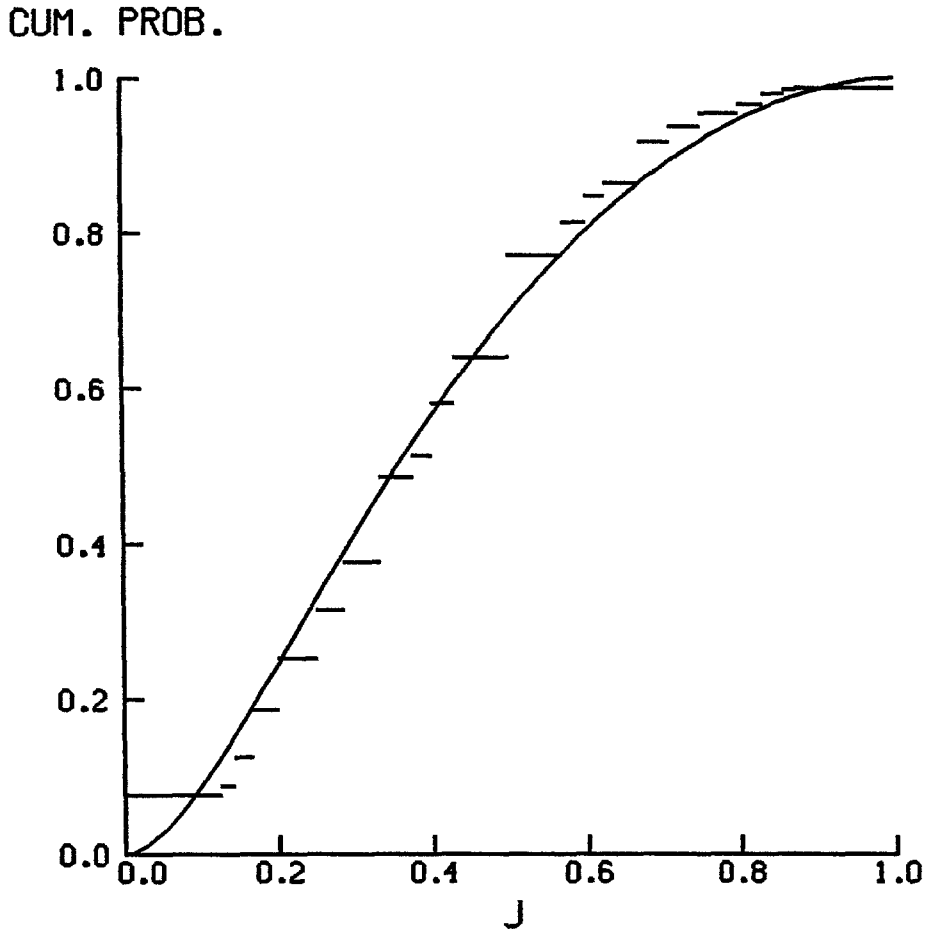

Figure 1. Cumulative distribution functions of J: permutation distribution (step function) and approximating continuous distribution function using the Beta distribution for D.

approximating Beta distribution for D. Figure 1 gives the *cdf* of J under the permutation distribution and the approximating *cdf* obtained by transforming the Beta(1.82, 1.68) distribution according to (30).

In accordance with the last paragraph of Section 7, the probability that $J = 0$ is quite large under the permutation distribution and not so large under the continuity-corrected Beta approximation. A further consequence of this poor approximation to $P\{J = 0\}$ is that the remainder of the approximating *cdf* is less steep than the permutation *cdf*. The approximation with an extra discrete probability for $J = 0$ gives better results. From (37) we obtain $\mu_0 = 0.561$, $\sigma_0^2 = 0.0373$, which leads to $p_0 = 3.138$, $q_0 = 2.456$. Figure 2 gives the corresponding *cdf* (35). The approximation to the permutation *cdf* is much better, especially in the right tail.

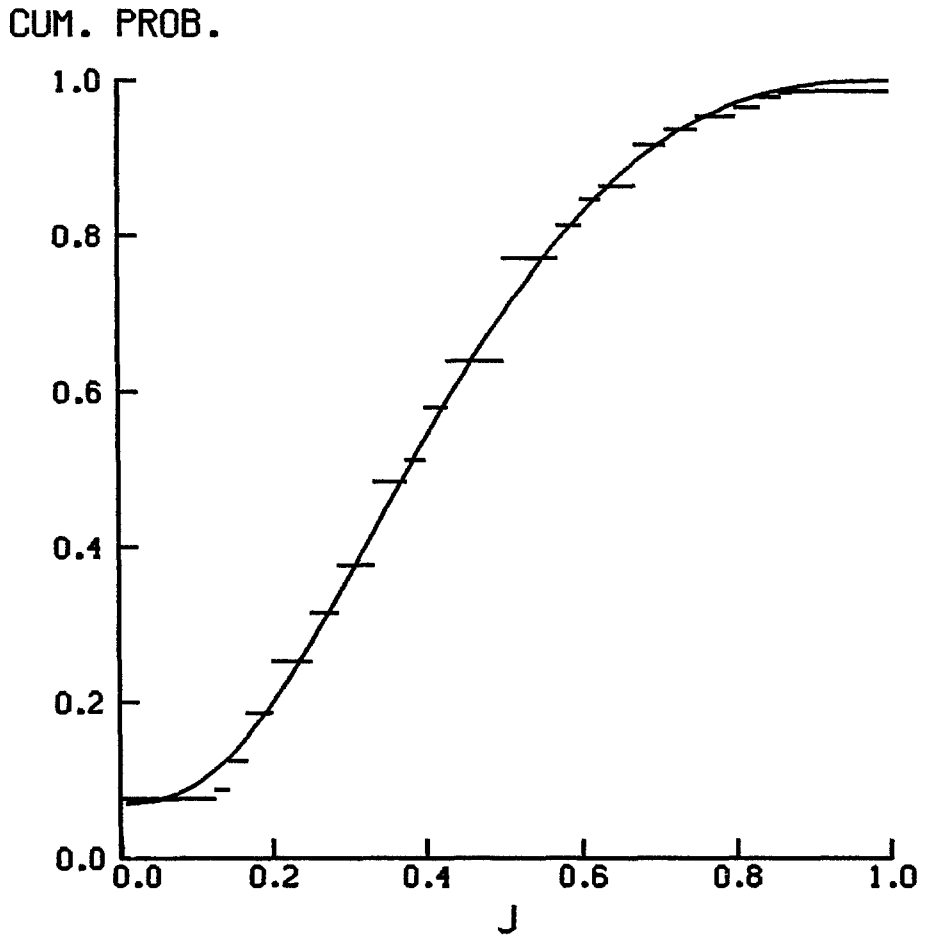

Figure 2. Cumulative distribution functions of J: permutation distribution (step function) and approximating continuous distribution function using the Beta distribution for D with an extra positive probability for $J = 0$.

When particular J-values are to be tested under the null model of absence of consensus between psychotherapist and client (formulated statistically as independence between the attributes chosen by the client and the psychotherapist while allowing for association of the attributes chosen by each individual), the upper 5% level of the null distribution would be a reasonable threshold value. In the permutation distribution of the J index, this is the value $J = 0.75$. Relevant null probabilities are $P\{J \geq 4/5\} = 0.046$, $P\{J \geq 3/4\} = 0.063$, and $P\{J \geq 5/7\} = 0.084$; there are no possible outcomes for J between $5/7 = 0.71$, $3/4 = 0.75$, and $4/5 = 0.8$. Therefore, in order to apply the continuity correction discussed in the paragraph leading to (33), we take $t_4 = 0.775$ for $t_2 = 0.8$, and $t_4 = 0.73$ for $t_2 = 0.75$. The resulting values

of the approximation for (33) obtained by using *cdf* (35) are

$$P\{J \geq 4/5\} \approx 0.040, P\{J \geq 3/4\} \approx 0.064 .$$

The correspondence is quite satisfactory.

Of the 115 true client-psychotherapist dyads, only 16 had a J-value at least equal to .75, while 27 more had a J-value at least equal to .71. From a statistical point of view, these frequencies are considerably higher than the 6.3% and 8.4% expected under independence. From the clinical point of view, however, it is rather disappointing that there are relatively few dyads for which the obtained value of J suggests consensus-after-communication. Tentatively, we conclude that the Jaccard index on 8 attributes is not a powerful method to establish consensus.

9. Discussion

A method has been proposed for testing the significance of observed values of similarity coefficients such as the indices of Jaccard, Dice, and the Simple Matching coefficient. This approach requires the computation of an (estimated) mean and variance of the similarity coefficient, which is feasible when a sufficiently large sample of units is available. In our example, samples of more than 100 units were available, which certainly seems sufficient. The null distributions of the M and D coefficients can be quite well approximated by Beta distributions, provided that a continuity correction is used. For the distributions of J and D, the approximating distribution should contain an extra, discrete probability for the outcome 0.

The methodological message of this paper is twofold. First, the null mean and standard deviation of the similarity coefficients will in practice often be large, unless the number of attributes is great. As a consequence, observed values of the coefficients can be interpreted as contradicting the null model only if they are extremely high: in our example with 8 attributes, the 5% threshold value for the Jaccard index was as high as 0.75. This result suggests that it is difficult to draw statistically significant conclusions if one wants to establish "consensus" on the basis of a small number of attributes (e.g., only 8). Increasing the number of attributes with a given average presence probability will decrease the variances of M and J. However, the number of attributes useful in this context is limited for both theoretical and practical reasons. Users of these measures of association, which are so prevalent in methods of classification, ought to be aware of these limitations.

Second, one should take account of the possible association of attributes within taxonomic units. Such an association may strongly affect the null distribution of similarity coefficients.

The method used here for determining the mean and variance of M, J, and D can easily be extended to similar coefficients, although we did not have encyclopaedic ambitions in this respect when writing this paper. Some related coefficients (see, e.g., Austin and Colwell 1977, p. 206; Wishart 1978, p. 112; Gower and Legendre 1986, p. 13) are:

Total difference	$(b+c)/(a+b+c+d) = 1-M$
Error sum of squares	$(b+c)/2(a+b+c+d) = (1-M)/2$
Variance	$(b+c)/4(a+b+c+d) = (1-M)/4$
Unnamed	$2(a+d)/(a+b+c+d) = 2M$
Rogers and Tanimoto	$(a+d)/(a+d+2(b+c)) = M/(2-M)$
Hamann	$(a+d-b-c)/(a+b+c+d) = 2M-1$
Non-metric	$(b+c)/(2a+b+c) = 1-D$.

These measures are functions of M or D, which means that the results of this paper can be directly applied to them. It may also be possible to apply the same methods to additional measures that are not expressible as functions of M or D.

Quite a different approach to testing the significance of a set of similarity coefficients can be taken by considering the original two-mode units/attributes 0-1 matrix of presence - absence data, and conditioning on the marginal totals. This strategy implies a null distribution which fixes the number of attributes for each of the units, and the number of units having each of the attributes. This approach is worked out in Snijders (1989).

A computer program in Fortran containing the calculations of the means and variances treated in this paper can be obtained from one of the authors, G. Driessen.

Appendix. Derivations

1. Derivation of (19)

The definitions in Section 5 imply

$$E(X_{Aj}) = p_{Aj}\,,\ \operatorname{var}(X_{Aj}) = p_{Aj}(1-p_{Aj})\,,$$

$$E(X_{Aj}X_{Ah}) = p_{Ajh}\,,\ \operatorname{cov}(X_{Aj},X_{Ah}) = p_{Ajh} - p_{Aj}p_{Ah} \quad (\text{all } j,h)$$

and similarly for unit B. This result implies

$$E(N_A) = p_{A+}\,,\quad E(N_B) = p_{B+} \tag{A.1}$$

$$E(N_A^2) = E\Big(\sum_{j=1}^{n}\sum_{h=1}^{n} X_{Aj}X_{Ah}\Big) = \sum_{j=1}^{n}\sum_{h=1}^{n} p_{Ajh} = p_{A++}$$

$$\mathrm{var}\,(N_A) = E(N_A^2) - (E(N_A))^2 = p_{A++} - p_{A+}^2\,. \tag{A.2}$$

The independence of units A and B implies that $(X_{A1},\ldots,X_{An})$ and $(X_{B1},\ldots,X_{Bn})$ are independent; N_A and N_B are then also independent. Hence

$$\mathrm{cov}\,(N_A,N_B) = 0$$

$$E(N_{11}) = \sum_{j=1}^{n} p_{Aj}p_{Bj} = m_{11} \tag{A.3}$$

$$E(N_{11}^2) = E\Big(\sum_{j=1}^{n}\sum_{h=1}^{n} X_{Aj}X_{Bj}X_{Ah}X_{Bh}\Big) = \sum_{j=1}^{n}\sum_{h=1}^{n} p_{Ajh}p_{Bjh} = m_{22}$$

$$\mathrm{var}\,(N_{11}) = E(N_{11}^2) - (E(N_{11}))^2 = m_{22} - m_{11}^2 \tag{A.4}$$

$$\mathrm{cov}\,(X_{Aj}X_{Bj},X_{Ah}) = E(X_{Aj}X_{Bj}X_{Ah}) - (E(X_{Aj}X_{Bj}))(E(X_{Ah}))$$

$$= p_{Ajh}p_{Bj} - p_{Aj}p_{Ah}p_{Bj}$$

$$\mathrm{cov}\,(N_{11},N_A) = \sum_{j=1}^{n}\sum_{h=1}^{n} \mathrm{cov}\,(X_{Aj}X_{Bj},X_{Ah}) \tag{A.5}$$

$$= \sum_{j=1}^{n}\sum_{h=1}^{n} (p_{Ajh}p_{Bj} - p_{Aj}p_{Ah}p_{Bj}) = m_{21} - p_{A+}m_{11}\,.$$

Similarly,

$$\mathrm{cov}\,(N_{11},N_B) = m_{12} - p_{B+}m_{11}\,. \tag{A.6}$$

Expression (13) for the Simple Matching coefficient implies

$$\mathrm{var}\,(M) = 4\{\mathrm{var}\,(N_{11}) - \mathrm{cov}\,(N_{11},N_A) - \mathrm{cov}\,(N_{11},N_B)\}$$
$$+ \mathrm{var}\,(N_A) + \mathrm{var}\,(N_B) + 2\mathrm{cov}\,(N_A,N_B)\,.$$

The latter expression, together with (A.1) to (A.6), yields formula (19).

2. Derivation of (22) and (23)

For the delta method, the partial derivatives of J, as given by (21), with respect to P_A, P_B, and P_{11} are needed. These are

$$\partial J / \partial P_{11} = (P_A + P_B) / (P_A + P_B - P_{11})^2 \tag{A.7}$$

$$\partial J / \partial P_A = \partial J / \partial P_B = -P_{11} / (P_A + P_B - P_{11})^2 . \tag{A.8}$$

The mean $E(J)$ is approximated by replacing random variables in (21) with their mean values, yielding (22). The approximation to var (J) as given by the delta method is

$$\sum_Y \sum_Z (\partial J / \partial Y)(\partial J / \partial Z) \operatorname{cov}(Y,Z) ,$$

where the sums extend over Y and Z each taking each of the values P_A, P_B, and P_{11}, and where the partial derivatives (A.7), (A.8) are evaluated at the expectations (A.1), (A.2). Substituting the expressions for the variances, covariances, and partial derivatives yields (23).

3. Derivation of (24) to (27).

The expressions for $E(D)$ and var (D) are derived by conditioning on the outcomes of N_A and N_B. It is clear that

$$E\{X_{Aj} \mid N_A = n_A\} = p_{Aj}(n_A) / p_A(n_A) ,$$

and similarly for X_{Bj}. Using this result with the independence of X_A and X_B, we obtain

$$\begin{aligned} E\{N_{11} \mid N_A = n_A, N_B = n_B\} &= E\{\sum_{j=1}^{n} X_{Aj} X_{Bj} \mid N_A = n_A, N_B = n_B\} \\ &= \sum_{j=1}^{n} \{p_{Aj}(n_A) p_{Bj}(n_B)\} / \{p_A(n_A) p_B(n_B)\} . \end{aligned}$$

Denote this quantity by $f(n_A, n_B)$. Note that the independence of the dyad implies that

$$P\{N_A = n_A, N_B = n_B\} = p_A(n_A) p_B(n_B) .$$

The definition of D can be expressed as

$$D = 2N_{11} / (N_A + N_B) ,$$

which implies

$$E(D) = 2E\{f(N_A,N_B) / (N_B + N_B)\} \tag{A.9}$$
$$= 2 \sum_{n_A=1}^{n} \sum_{n_B=1}^{n} p_A(n_A)p_B(n_B)f(n_A,n_B) / (n_A + n_B),$$

which is equal to (24). In order to derive (25) and (27), the following relations will be used; $I\{E\}$ denotes the indicator function of the event E, equal to 1 or 0, respectively, depending on whether E occurs or not.

$$\sum_{h=1}^{n} p_{Ajh}(n_A) = \sum_{h=1}^{n} E(X_{Aj}X_{Ah}I\{N_A = n_A\}) \tag{A.10}$$
$$= E(X_{Aj}(\sum_{h=1}^{n} X_{Ah})I\{N_A = n_A\})$$
$$= E(X_{Aj}N_AI\{N_A = n_A\}) = n_Ap_{Aj}(n_A).$$

Similarly it can be proved that

$$\sum_{j=1}^{n} p_{Aj}(n_A) = n_Ap_A(n_A), \tag{A.11}$$

so that

$$\sum_{j=1}^{n} \sum_{h=1}^{n} p_{Ajh}(n_A) = n_A^2p_A(n_A). \tag{A.12}$$

The same relation holds for B. From (A.11) it follows that

$$\sum_{j=1}^{n} \{p_{Aj}(n_A)p_{Bj}(n_B)\}$$
$$= \sum_{j=1}^{n} \{p_{Aj}(n_A) - n_Ap_A(n_A) / n\} \{p_{Bj}(n_B) - n_Bp_B(n_B) / n\}$$
$$+ n_An_Bp_B(n_A)p_B(n_B) / n,$$

which together with (24) establishes (25).

To derive (26) we must find an expression for $E(D^2)$, which is also derived by conditioning on the outcomes of N_A and N_B. First note that

$$E\{N_{11}^2 \mid N_A = n_A,N_B = n_B\}$$

$$= E\{\sum_{j=1}^{n}\sum_{h=1}^{n} X_{Aj}X_{Bj}X_{Ah}X_{Bh} \mid N_A = n_a, N_B = n_b\}$$

$$= \sum_{j=1}^{n}\sum_{h=1}^{n} p_{Ajh}(n_A)p_{Bjh}(n_B) / \{p_A(n_A)p_B(n_B)\} .$$

Analogous to the derivation of (A.9), this result leads to

$$E(D^2) = 4 \sum_{n_A=1}^{n}\sum_{n_B=1}^{n} ((n_A + n_B)^{-2} \sum_{j=1}^{n}\sum_{h=1}^{n} p_{Ajh}(n_A)p_{Bjh}(n_B))$$

and hence to (26). This leads to (27), because (A.12) implies that

$$\sum_{j=1}^{n}\sum_{h=1}^{n} p_{Ajh}(n_A)p_{Bjh}(n_B)$$

$$= \sum_{j=1}^{n}\sum_{h=1}^{n} \{p_{Ajh}(n_A) - n_A^2 p_A(n_A) / n^2\} \{p_{Bjh}(n_B) - n_B^2 p_B(n_B) / n^2\}$$

$$+ n_A^2 n_B^2 p_A(n_A)p_B(n_B) / n^2 .$$

References

ANDERBERG, M. R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

AUSTIN, B., and Colwell, R. R. (1977), "Evaluation of Some Coefficients for Use in Numerical Taxonomy of Microorganisms," *International Journal of Systematic Bacteriology, 27*, 204-210.

BARONI-URBANI, C., and BUSER, M. W. (1976), "Similarity of Binary Data," *Systematic Zoology, 25*, 251-259.

BARONI-URBANI, C. (1980), "A Statistical Table for the Degree of Coexistence Between Two Species," *Oecologia, 44*, 287-289.

BISHOP, Y. M . M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Mass.: MIT Press.

COLEMAN, B. D., Mares, M. A., Willig, M. R., and Hsieh, Y.-H. (1982), "Randomness, Area and Species Richness," *Ecology, 63*, 1121-1133.

CONNOR, E. F., and Simberloff, D. (1978), "Species Number and Compositional Similarity of the Galapagos Flora and Avifauna," *Ecological Monographs, 48*, 219-248.

COQUIN-VIENNOT, D. (1975), "Recherche d'une organisation mnemonique interne dans un ensemble de donnees," *Annee Psychologique, 75*, 575-597.

DICE, L. R. (1945), "Measures of the Amount of Ecological Association Between Species," *Ecology, 26*, 297-302.

DICE, L. R. (1952), "Measure of the Spacing Between Individuals Within a Population," *Contributions of the Laboratory of Vertebrate Biology of the University of Michigan, 55*, 1-23.

DORMAAR, M., Dijkman-Caes, C., and De Vries, M. W. (1989), "Consensus in Client-Therapist Interactions; A Measure of the Therapeutic Relationship Related to Outcome," Accepted for publication, *Psychotherapy and Psychosomatics*.

ELSTON, R.C., Schroeder, S. R., and Rohjan, J. (1982), "Measures of Observer Agreement When Binomial Data Are Collected in Free Operant Situations," *Journal of Behavioral Assessment, 4*, 299-310.

EVERITT, B. S. (1980), *Cluster Analysis* (2nd ed.), London: Gower.

GOODALL, D. W. (1967), "The Distribution of the Matching Coefficient," *Biometrics, 23*, 647-656.

GOODALL, D. W. (1978), "Sample Similarity and Species Correlation," in *Ordination of Plant Communities*, Ed. R. H. Whittaker, The Hague: Junk, 101-149.

GOWER, J. C., and Legendre, P. (1986), "Metric and Euclidean Properties of Dissimilarity Coefficients," *Journal of Classification, 3*, 5-48.

GREGSON, R. A .M. (1975), *Psychometrics of Similarity*, New York: Academic Press.

HELTSHE, J. F. (1988), "Jackknife Estimate of the Matching Coefficient of Similarity," *Biometrics, 44*, 447-460.

HUBALEK, Z. (1982), "Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation," *Biological Review, 57*, 669-689.

JACCARD, P. (1900), "Contributions au problème de l'immigration post-glaciaire de la flore alpine," *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547-579.

JACCARD, P. (1908), "Nouvelles recherches sur la distribution florale," *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44, 223-270.

JANSON, S., and Vegelius, J. (1981), "Measures of Ecological Association," *Oecologia, 49*, 371-376.

JOHNSON, B. E., and Millie, D. F. (1982), "The Estimation and Applicability of Confidence Intervals for Stander's Similarity Index (SIMI) in Algal Assemblage Comparisons," *Hydrobiologica, 89*, 3-8.

JOHNSON, N. L. and Kotz, S. (1970), *Distributions in Statistics: Continuous Distributions - 2*, New York: Wiley.

PRESS, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986), *Numerical Recipes*, Cambridge: Cambridge University Press.

QUESADA, E., Ventosa, A., Rodgriguez-Valera, F., Megias, L., and Ramos-Cormenzana, A. (1983), "Numerical Taxonomy of Moderate Halophilic Gram-negative Bacteria from Hypersaline Soils," *Journal of General Microbiology, 129*, 2649-2657.

SCHEFF, T. J. (1967), "Toward a Sociological Model of Consensus," *American Sociological Review, 32*, 32-46.

SNEATH, P. H. A., and Sokal, R. R. (1973), *Numerical Taxonomy*, San Francisco: Freeman.

SNIJDERS, T. A .B. (1989), "Enumeration and Simulation Methods for 0-1 Matrices with Given Marginals," Submitted for publication.

SOKAL, R. R., and Michener, C. D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Scientific Bulletin, 38*, 1409-1438.

WASHINGTON, H. G. (1984), "Diversity, Biotic and Similarity Indices. A Review with Special Relevance to Aquatic Ecosystems," *Water Research, 18*, 653-694.

WISHART, D. (1978), *Clustan User Manual* (3d ed.), Edinburgh: Program Library Unit, Edinburgh University.