Binary (Presence–Absence) Similarity Coefficients
Author(s): Alan H. Cheetham and Joseph E. Hazel
Reviewed work(s):
Source: Journal of Paleontology, Vol. 43, No. 5 (Sep., 1969), pp. 1130-1136
Published by: SEPM Society for Sedimentary Geology
Stable URL: http://www.jstor.org/stable/1302424
Accessed: 14/04/2012 21:54

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

SEPM Society for Sedimentary Geology is collaborating with JSTOR to digitize, preserve and extend access to Journal of Paleontology.
ABSTRACT—Various numerical coefficients have been employed in comparisons of taxa or bioassociational units, especially in studies involving large arrays of multivariate data. Nomenclatural and conceptual difficulties have arisen from the translation of coefficients from variable-oriented (R-mode) to case-oriented (Q-mode) use. This paper points out the conceptual relationships and some of the properties of coefficients expressing similarity or difference between units defined by binary data.

INTRODUCTION

In recent years there has been increased use in systematic biology of large arrays of multivariate data for both taxonomic and bioassociational (including biogeographic, ecologic, paleoecologic, and biostratigraphic) interpretations. Many taxa and bioassociational units are polythetic in concept, i.e., distinctive in the aggregate of their attributes, one or more of which they share with other such units. Where either the number of examples or cases composing the units or the number of attributes describing them is large, adequate comparison among units becomes more and more dependent upon numerical coefficients of similarity or difference. Various numerical indices have been employed in systematic analyses of quantitative, multistate qualitative, and binary (presence-absence) data. The increasing availability of digital computers, and of programs for them which calculate these indices, doubtless will result in continued acceleration of the use of numerical methods of comparison.

Coefficients of association in studies of relationships between variables or attributes (R-mode) are a well-known part of standard statistical procedure. Translation of these measures to express relationships between examples or cases (Q-mode) of taxa or bioassociational units has resulted in both conceptual and nomenclatural complications. Our purpose is not to justify or criticize the use of any particular coefficient or coefficients, but rather to point out some of the conceptual and nomenclatural relationships among them.

Coefficients which express relationships of either similarity or difference between units defined by binary data are especially important in systematic biology because of the wide availability of data of this type, such as features present or absent in specimens or species present or absent in samples. This discussion is limited to binary coefficients. Some binary coefficients, however, are simply special cases of those, such as the Correlation Coefficient, that express similarity or difference between units based upon quantitative or multistate qualitative data. These relationships are included in the discussions of the binary coefficients to which they are applicable.

Most of the binary coefficients discussed here appear to have been developed intuitively and tested empirically. Those that are special cases of quantitative or multistate coefficients, however, share the mathematical derivation of their non-binary equivalents. Used in R-mode studies, some of these mathematically derived binary coefficients, such as the Phi Coefficient, have known probability distributions on which statistical inferences can be based. In Q-mode analysis however, coefficients do not estimate population parameters in an exactly analogous way, because the defining attributes are more than just an unbiased sample of a universe of possible attributes. Consequently, numerical comparison of taxa or bioassociational units is not, strictly speaking, a statistical procedure.

The properties of some of the coefficients listed here have been discussed in greater detail by Sokal & Sneath (1963, p. 121-157) for taxonomic applications and by Simpson (1960) for bioassociations. Not all coefficients are logically applied to both areas. Those incorporating negative matches (mutual absences) in the numerator are used commonly in taxonomic studies and rarely in bioassociational ones. Those excluding some mismatches (presences in one sample but not the other) from the denominator are common in bioassociational studies, but, as far as we know, have not been used in taxonomy.

Because of separate development of coefficients in the two areas of application, similar or identical coefficients have been proposed under different names, and the same name has been used for different coefficients. Unlike statistical coefficients, these similarity or difference measures have not been assigned standard symbols, and even the symbols used for their computation
are different in the two areas (see, for example, Simpson, 1960; Sokal & Sneath, 1963, p. 126). The following list presents these computational symbols, bioassociational on the left, taxonomic on the right. Each symbol refers to a number of taxa in bioassociational application or a number of characters in taxonomic application. To emphasize the equivalency of coefficients in the two areas of application, we have used both sets of symbols in the expressions that follow the list.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Bioassociational</th>
<th>Taxonomic</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Present in both units compared</td>
<td>njJK</td>
</tr>
<tr>
<td>E₁</td>
<td>Present in first but not second</td>
<td>njK</td>
</tr>
<tr>
<td>E₂</td>
<td>Present in second but not first</td>
<td>njK</td>
</tr>
<tr>
<td>A</td>
<td>Absent in both (but present in others)</td>
<td>njK</td>
</tr>
<tr>
<td>N₁(=C+E₁)</td>
<td>Total present in first</td>
<td>njJK + njK</td>
</tr>
<tr>
<td>N₂(=C+E₂)</td>
<td>Total present in second</td>
<td>njJK + njK</td>
</tr>
<tr>
<td>N₄(=N₁+N₂-C)</td>
<td>Total present in both units</td>
<td>njJK + njK</td>
</tr>
<tr>
<td>N₅(=N₁+N₂)</td>
<td>Total present in all units (where there are more than two)</td>
<td>n</td>
</tr>
</tbody>
</table>

In the following discussion we do not differentiate between coefficients in which the numerator is multiplied by 100 and those in which it is not. The former, expressed as percentages, are generally used for bioassociational indices, but not in taxonomic work.

**COEFFICIENTS**

The most widely used measure in both taxonomic and bioassociational studies is

\[ \frac{C}{N₁ + N₂ - 2C} = \frac{njJK}{njJK + njK + njK}, \]  

which was introduced by Jaccard (1908, p. 226-227) as the Coefficient of Community (see also Hagmeier & Stults, 1964; Peters, 1968). In numerical taxonomy, this coefficient is well known as the Jaccard Coefficient (Sokal & Sneath, 1963, p. 129), and under that name has also been used in bioassociational studies such as those of Kaeasler (1966), Maddocks (1966), Valentine (1966), and Mello & Buzas (1968).

The Simple Matching Coefficient,

\[ \frac{C + A}{N₁ + A} = \frac{njJK + njK}{n}, \]  

was proposed by Sokal & Michener (1958) for numerical taxonomy, but it has also been used by Kaeasler (1966) in comparing bioassemblies. It can be used, of course, only where three or more assemblages are being compared; otherwise, A (njA)=0, and expression (2) becomes identical with (1).

Another commonly used index in both taxonomy and bioassociational studies is

\[ \frac{2C}{N₁ + N₂} = \frac{2njJK}{2njJK + njK + njK}, \]  

called the Dice Coefficient by Sokal & Sneath (1963, p. 128) and the Burt Coefficient or Pirlot Index by Peters (1968).

Other coefficients proposed for use in both taxonomy and bioassociational studies include the First Kulczynski Coefficient (Sokal & Sneath, 1963, p. 129; Peters, 1968),

\[ \frac{C}{N₁ + N₂ - 2C} = \frac{njJK}{njJK + njK + njK}; \]  

and the measure attributed to Otsuka by Ochiai (1957) and called the Otsuka Coefficient by Sokal & Sneath (1963, p. 130) and the Otsuka Coefficient by Peters (1968),

\[ \frac{C}{\sqrt{N₁N₂}} = \frac{njJK}{\sqrt{njJK}}. \]  

Expression (6), the Otsuka Coefficient, is a special case of the Coefficient of Proportional Similarity (Imbrie & Purdy, 1962, p. 355),

\[ \cos \theta = \frac{\sum (Xij)(Xik)}{\sqrt{\sum (Xij)^2}\sqrt{\sum (Xik)^2}}, \]  

in which \(X_i\) (the character state in numerical taxonomy) may take on only the binary values 0 and 1. In the binary case, the product (\(X_{ij}\)) (\(X_{ik}\)) is equal to 1 for positive matches (mutual presences) and to 0 for mismatches and negative matches (mutual absences), and thus the numerator is equal to the sum of the positive (\(C=njJK\)). Each of the quantities \(\sum (Xij)^2\) and \(\sum (Xik)^2\) is in like manner equal to the total number of presences in each unit compared, that is, to \(N₁ (=nj)\) and \(N₂ (=njK)\), respectively; the denominator thus becomes \(\sqrt{N₁N₂} (=\sqrt{njJK})\).

Measures of similarity proposed for use in bioassociational but not taxonomic studies include the Correlation Ratio (Sorgenfrei, 1959, p. 403),

\[ \frac{C^2}{N₁N₂} = \frac{(njJK)^2}{njJK}; \]  

which is the square of expression (6); the Simp-
<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Column</th>
<th>1. As $N_1 \rightarrow N_2$</th>
<th>2. As $C \rightarrow 0$</th>
<th>3. As $C \rightarrow N_t$</th>
<th>4. As $A \rightarrow 0$</th>
<th>5. As $\frac{N_1}{N_2} = \frac{1}{2}$ and $\frac{C}{N_t} = \frac{1}{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Jaccard</td>
<td>$\frac{C}{N_1+N_2-C}$</td>
<td>$\frac{C}{2E_1+C}$</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>2 Simple M.</td>
<td>$\frac{C+A}{N_1+N}$</td>
<td>$\frac{C+A}{2E_1+C+A}$</td>
<td>$\frac{A}{E_1+E_2+A}$</td>
<td>$\rightarrow 1$</td>
<td>$\rightarrow \frac{C}{N_t}$ \quad \text{[Jaccard (1)]}</td>
<td>$\frac{C+A}{5C_2}$ \quad \text{[if $A=C$, then $=1/3$]}</td>
</tr>
<tr>
<td>3 Dice</td>
<td>$\frac{2C}{N_1+N_2}$</td>
<td>$\frac{C}{N_1}$ \quad \text{[Simpson (8)]}</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>4 1st Kulcz.</td>
<td>$\frac{C}{N_1+N_2-2C}$</td>
<td>$\frac{C}{2E_1}$</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow \infty$</td>
<td>unaffected</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>5 2nd Kulcz.</td>
<td>$\frac{C(N_1+N_2)}{2(N_1N_2)}$</td>
<td>$\frac{C}{N_1}$ \quad \text{[Simpson (8)]}</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{3}{8}$</td>
</tr>
<tr>
<td>6 Otsuka</td>
<td>$\frac{C}{\sqrt{N_1N_2}}$</td>
<td>$\frac{C}{N_1}$ \quad \text{[Simpson (8)]}</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{1}{\sqrt{8}}$</td>
</tr>
<tr>
<td>8 Corr. Ratio</td>
<td>$\frac{C^2}{N_1N_2}$</td>
<td>$\frac{C^2}{(N_1)^2}$ \quad \text{[(Simpson (8))^2]}</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{1}{8}$</td>
</tr>
<tr>
<td>9 Simpson</td>
<td>$\frac{C}{N_t}$</td>
<td>$\frac{C}{N_t}$</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>10 Braun-Bl.</td>
<td>$\frac{C}{N_2}$</td>
<td>$\frac{C}{N_t}$ \quad \text{[Simpson (8)]}</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 1$</td>
<td>unaffected</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>11 Fager</td>
<td>$\frac{C}{\sqrt{N_1N_2}}$ \quad $\frac{1}{2\sqrt{N_2}}$</td>
<td>$\frac{C}{N_1}$ \quad $\frac{1}{2\sqrt{N_2}}$</td>
<td>$\rightarrow \frac{1}{2\sqrt{N_2}}$</td>
<td>$\rightarrow 1 - \frac{1}{2\sqrt{N_2}}$</td>
<td>unaffected</td>
<td>$\frac{1}{\sqrt{8}}$ \quad $\frac{1}{4\sqrt{C}}$</td>
</tr>
</tbody>
</table>

Table 1—Characteristics of coefficients of similarity and difference based on binary (presence-absence) data. Bioassociational symbols are used; see text for numerical taxonomic equivalents. The expressions in column 1 result when two samples have the same number of positive features (equal diversity); those in column 2 when they share no positive features; those in column 3 when they become identical in positive features; and those in column 4 when there are no shared negative features (mutual absences). Column 5 offers a comparison of the coefficients when one sample has twice as many positive features (presences) as the other and the number of features in common (mutual presences) is one half the number in the less positive (less diverse) sample.
<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{C+A}{N_1+A+E_1+E_2} )</td>
<td>Rogers &amp; Tan, (C+A)-(E_1+E_2)</td>
</tr>
<tr>
<td>( \frac{N_t + A}{N_1 + A} )</td>
<td>Hamann, ( C+A )</td>
</tr>
<tr>
<td>( \frac{C-A-E_1E_2}{C+A+E_1E_2} )</td>
<td>Yule, CA-E_1E_2/CA+E_1E_2</td>
</tr>
<tr>
<td>( \frac{C-A-E_1E_2}{\sqrt{N_1N_2(E_1+A)(E_2+A)}} )</td>
<td>Phi, CA-E_1E_2</td>
</tr>
<tr>
<td>( \left( \frac{N_1}{N_2} \right)^{1/2} + \left( \frac{N_2}{N_1} \right)^{1/2} = 1 )</td>
<td>Resem. Eq.</td>
</tr>
<tr>
<td>( \frac{1-C}{N_2} )</td>
<td>Coeff. Diff.</td>
</tr>
<tr>
<td>( \frac{2E_1}{C+A+2E_1} )</td>
<td>No.Feat. Diff. E_1+E_2</td>
</tr>
<tr>
<td>( \sqrt{\frac{E_1+E_2}{N_t+A}} )</td>
<td>Sokal Dist. (Binary Case)</td>
</tr>
<tr>
<td>( \sqrt{1-\frac{C+A}{N_1+A}} )</td>
<td>Sokal Dist. (Binary Case)</td>
</tr>
<tr>
<td>( \frac{C-A}{N_1+E_1+E_2} )</td>
<td>C+A</td>
</tr>
<tr>
<td>( \frac{A-3C}{A+5C} )</td>
<td>C-A-3C</td>
</tr>
<tr>
<td>( \frac{A-3C}{A+3C} )</td>
<td>C-A-3C</td>
</tr>
<tr>
<td>( \frac{A}{N_t} )</td>
<td>Nt-C</td>
</tr>
<tr>
<td>( \frac{A}{A+3C} )</td>
<td>A+3C</td>
</tr>
<tr>
<td>( \frac{A}{A-3C} )</td>
<td>A-3C</td>
</tr>
<tr>
<td>( \sqrt{\frac{4C}{5C+A}} )</td>
<td>( \sqrt{\frac{4C}{5C+A}} )</td>
</tr>
<tr>
<td>( \sqrt{\frac{1-C}{N_1}} )</td>
<td>Jaccard (1)</td>
</tr>
<tr>
<td>( \sqrt{\frac{1-C+A}{5C+A}} )</td>
<td>( \sqrt{\frac{1-C+A}{5C+A}} )</td>
</tr>
</tbody>
</table>

* When the two samples being compared are identical the equation is insoluble.
son Coefficient (Simpson, 1943; 1947; 1960; Cheetham & Deboo, 1963; Hagmeier & Stults, 1964; Peters, 1968),
\[
\frac{C}{N_1} = \frac{n_{jk}}{n_j}, \text{ where } N_1 \leq N_2, \ n_j \leq n_K;
\]
the coefficient introduced by Braun-Blanquet (1932, p. 632) and referred to by Peters (1968) as the Jaccard Coefficient,
\[
\frac{C}{N_2} = \frac{n_{jk}}{n_k}, \text{ where } N_1 \leq N_2, \ n_j \leq n_K;
\]
and the Fager Coefficient,
\[
\frac{C}{\sqrt{N_1 N_2}} - \frac{1}{2 \sqrt{N_2}} = \frac{n_{jk} - 1}{\sqrt{n_{jk} n_k}}, \text{ where } N_1 \leq N_2, \ n_j \leq n_K,
\]
which was introduced by Fager in Fager & McGowan (1963) and subsequently has been used by Elliot (1963) and Valentine & Peddicord (1967).

Measures of similarity which have been proposed for use in taxonomy but not in bioassociational studies include the Rogers and Tanimoto Coefficient (Sokal & Sneath, 1963, p. 129),
\[
\frac{C + A}{N_1 + A + E_1 + E_2} = \frac{n_{jk} + n_k}{n + n_{jk} + n_{jk}};
\]
the Hamann Coefficient (Sokal & Sneath, 1963, p. 132),
\[
\frac{C + A}{n} - \frac{(E_1 + E_2)}{N_1} = \frac{(n_{jk} + n_k)}{n} - \frac{(n_k + n_{jk})}{n};
\]
the Yule Coefficient (Sokal & Sneath, 1963, p. 132),
\[
\frac{C - E_1 E_2}{\sqrt{N_1 N_2 (E_1 + A)(E_2 + A)}} = \frac{n_{jk} - n_k n_{jk}}{\sqrt{n_{jk} n_k (n_{jk} + n_{jk}) (n_{jk} + n_{jk})}};
\]
and the Phi Coefficient (Sokal & Sneath, 1963, p. 132),
\[
\frac{C - E_1 E_2}{\Sigma X_{ij}^2} = \frac{n_{jk} - n_k n_{jk}}{\Sigma X_{ik}^2}.
\]

Expression (15), the Phi Coefficient, is a special case of the Pearson Product-Moment Correlation Coefficient, computed by the formula (Sokal & Sneath, 1963, p. 296)
\[
r = \frac{\Sigma (X_{ij}) (X_{ik}) - \frac{1}{n} (\Sigma X_{ij}) (\Sigma X_{ik})}{\sqrt{\left(\Sigma (X_{ij})^2 - \frac{1}{n} (\Sigma X_{ij})^2\right) \left(\Sigma (X_{ik})^2 - \frac{1}{n} (\Sigma X_{ik})^2\right)}}.
\]

In the binary case, substitution of the quantities 0 and 1, as used above in the discussion of expression (7), demonstrates the equivalence of the Phi and Correlation Coefficients.

In addition to the foregoing measures of similarity, three coefficients of difference have been proposed for computation from binary data. In bioassociational studies, the coefficient \( z \) is calculated from the Resemblance Equation,
\[
\frac{(n_{jk} + n_{jk})^{1/2}}{n - n_{jk}} + \frac{(n_{jk} + n_{jk})^{1/2}}{n - n_{jk}} = 1.
\]

The Resemblance Equation was proposed by Preston (1962), and Hagmeier & Stults (1964) used \( 1 - z \) as a measure of similarity. The Coefficient of Difference (Savage, 1960; Peters, 1968),
\[
1 - \frac{C}{N_2} = \frac{n_{jk}}{n_k}, \text{ where } N_1 \leq N_2, \ n_j = n_K
\]
is simply one minus expression (10).

The Number of Features of Difference (NFD) used by Stephenson, Williams, & Lance (1968, p. 4, 5),
\[
E_1 + E_2 = n_{jk} + n_{jk},
\]
is a special case of the squared Taxonomic Distance of Sokal (1961) and Sokal & Sneath (1963, p. 147),
\[
\sum_{i=1}^{n} (X_{ij} - X_{ik})^2,
\]
in which \( X_i \) may take on only the binary values 0 and 1; in this case, the value \( (X_{ij} - X_{ik})^2 \) is equal to 0 for matches and 1 for mismatches, and the squared Taxonomic Distance is equal to the sum of the mismatches, as given in expression (19). As pointed out by Sokal & Sneath (1963, p. 147), this value increases with the number of characters used in comparison, so that in practice, an average Taxonomic Distance (commonly called the Sokal Distance),
\[
\sqrt{\frac{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2}{n}},
\]
is an average of the squared Taxonomic Distances for all pairs of comparisons.
is used. In the binary case, the Sokal Distance becomes

$$
\sqrt{\frac{E_1 + E_2}{N_1 + A}} = \sqrt{\frac{n_{jk} + n_{jk}}{n}} ,
$$
(22)

which is equivalent to

$$
\sqrt{1 - \frac{C + A}{N_1 + A}} = \sqrt{1 - \frac{n_{jk} + n_{jk}}{n}}
$$
(22a)

that is, the square root of one minus expression (2).

**PROPERTIES OF COEFFICIENTS**

Table 1 lists the coefficients treated above under the names which appear to be the most appropriate for them. Some properties of the coefficients are indicated by variation of certain terms in the numerator and denominator. The expressions in column 1 result when two samples being compared have the same number of positive features (equal diversity). Under this condition, for example, the Dice, 2nd Kulczynski, Otsuka, and Braun-Blanquet Coefficients become identical with the Simpson Coefficient, and the Coefficient of Difference becomes one minus the Simpson Coefficient.

When there are no positive features (mutual presences) shared by two samples being compared (column 2), all similarity coefficients except those that use shared negative features (mutual absences) take on their minimum values. When two samples are identical in positive features (mutual presences), the similarity coefficients assume maximum (and difference coefficients, minimum) values, except for the Fager Coefficient which still has a negative constant, and the Resemblance Equation which becomes insoluble (column 3).

When no two samples share no negative features (no mutual absences), the Simple Matching Coefficient behaves differently as the number of shared negative features (mutual absences) varies.

**ACKNOWLEDGMENTS**

We are indebted to James F. Mello, U. S. Geological Survey, James A. Peters, Smithsonian Institution, John Imbrie, Brown University, and David H. Raup, University of Rochester, for critically reading the manuscript.

Publication is authorized by the Director, U. S. Geological Survey.

**REFERENCES**


Manuscript received December 4, 1968