CHAPTER 6

The Estimation of Taxonomic Resemblance

This chapter presents a detailed exposition and evaluation of the various numerical methods which have been advocated for expressing the similarity between taxa. We would like to remind the reader before proceeding that the terms "resemblance," "similarity," and "affinity" are used interchangeably throughout this book, and that unless specifically qualified they imply a solely phenetic relationship.

6.1. METHODS OF ESTIMATION IN VARIOUS SCIENCES

The problem of finding measures of the resemblance between pairs of entities based on a number of characteristics is not a new one; it has been encountered in a number of sciences whenever classificatory tasks have been undertaken. Besides taxonomy, the fields most frequently involved are psychology and ecology (see Section 10.6).

We adopt the convention used in psychology of arranging data for such an undertaking in the form of an $n \times t$ matrix whose t columns represent the t fundamental entities to be grouped on the basis of resemblances and whose n rows are n unit characters. Each entry X_{ij} in such a matrix is the score of operational taxonomic unit j for character i. The scale in which this score is expressed may simply allow for two states, such as present or absent. This is often conveniently symbolized by +

and −; however, the use of 1 and 0 facilitates the numerical treatment of such data. Other scales may be ordinal, continuous, or percentages of the maximum expression of a given character. The consequences of the use of the various scales are discussed later.

Cattell (1952) has pointed out that most matrices of this sort can be examined from two points of view. The association of pairs of characters (rows) can be examined over all OTU's (columns). This is called the R-technique. The converse practice, the association of pairs of OTU's (columns) over all characters (rows), has been called the Q-technique.

In psychology the R-technique is used to quantify relations among various kinds of psychological tests computed from a number of subjects (individuals). It will yield groups of related tests. The Q-technique, on the other hand, will evaluate relationships among persons based on their performance on mental or personality tests. It should result in identifying and categorizing types of persons.

In ecology the application of similar techniques has been particularly fruitful in community studies. Rows in such studies would represent species (the floral or faunal composition), while columns would be stands or other sampled plots. In this field of research the R-technique aims at the quantification of the relation between the species sampled from a number of presumably homogeneous plots, which may be quadrats or entire stands, as the case may be. These methods have been helpful in: (1) discovering ecological relationships among different species, such as common environmental requirements or dependence of one species upon the other, and (2) the use of combinations of species in classifying and identifying ecological associations and communities. Difficulties have been encountered, since plant ecologists have been unable to extend this type of treatment to several species simultaneously. Recently Dagnelie (1960) has shown the way to a multivariate analysis of such data.

The application of the Q-technique in ecology is the computation of coefficients of relationship between stands or other sampling units based on lists of species common to such units. A great variety of coefficients have been developed and applied to this task, which has as its aim the grouping of similar stands into associations and communities. The analogies with taxonomic procedures are quite striking.

In taxonomy both techniques have been employed. In this field the R-technique refers to correlations among characters based on operational taxonomic units. These could be from the lowest possible rank (individual organisms) up through local populations, varieties, subspecies, species, and supraspecific taxa. At the infraspecific level, much R-type

work has been performed (for example, Clark, 1941; Jolicoeur, 1959; Olson and Miller, 1958; Sokal, 1952, 1959, 1962a; Sokal and Hunter, 1955; Sokal and Rinkel, 1963). Such work contributes mainly to an understanding of ontogenetic processes and of minor genetic differentiation. At the higher taxonomic levels, analysis of R-type matrices should lead to information on phylogenetic factors at work within the group studied. So far as we are aware, the only taxonomic study of this nature has been that of Stroud (1953) on 43 species of the termite genus *Kalotermes*. We feel that much information of systematic value can come from such studies and shall discuss them in Section 7.6.

However, our main emphasis in numerical taxonomy is on Q-type studies. They refer to the quantifications of relations between pairs of taxa, frequently species, based on a preferably large number of characters. The history of this approach has been covered in Section 3.1. The resulting estimates of resemblance form the bases for the various procedures of numerical taxonomy. Criticisms of the Q-technique applied to problems in psychology (Cattell, 1952) have been shown not to apply to work in taxonomy (Sokal and Michener, 1958).

6.2. ESTIMATES OF RESEMBLANCE PROPOSED IN TAXONOMY

In a discussion of the various methods used for assessing taxonomic similarities, consideration must be given to the important problems of scaling and coding of characters and also to those of sampling (the reliability of the data used in a computation). However, these cannot be suitably discussed until the methods themselves have been presented. We shall therefore defer all consideration of these problems until the methods have been introduced and evaluated.

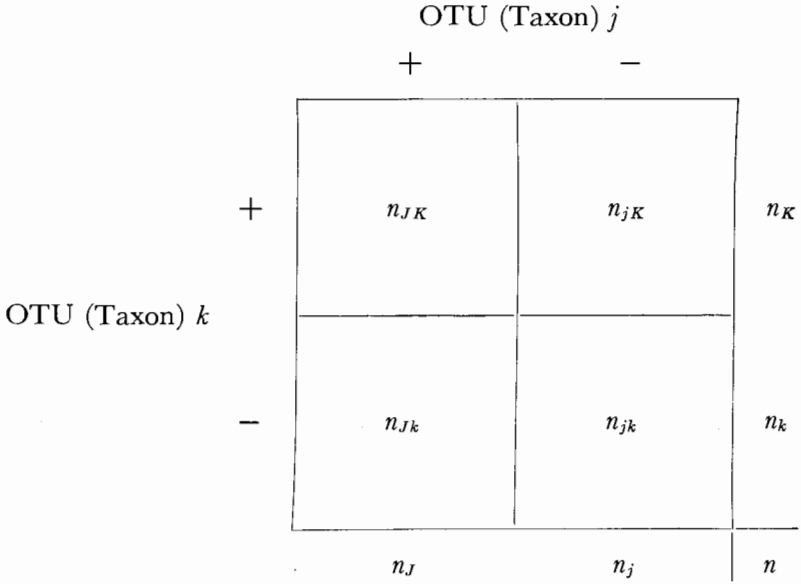
The various techniques for computing resemblances between taxa can be conveniently grouped into three types of coefficients—those of association, of correlation, and of distance. We would like to refer to them collectively as coefficients of resemblance or similarity.

6.2.1. Coefficients of association

These coefficients—also known as coefficients of similarity or relationship, or matching coefficients—have been widely used, particularly in ecology. Reviews of the different types of coefficients which have been proposed and used can be found in Cole (1949, 1957) and Dagnelie

(1960). We should point out at this time that all the terms employed here for these coefficients (such as association, similarity, relationship) have been used in a variety of meanings in English and in other languages. The designations adopted here are therefore arbitrary.

The basic arrangement of data for computation of coefficients of association is the familiar 2 × 2 table. An example of such an arrangement is shown here:



The data consist of n characters scored for two OTU's labeled j and k. They are subdivided into positive and negative classes for each of the two operational taxonomic units. We shall use capital letter subscripts to indicate positive or "1" states and lower-case letter subscripts to show negative or "0" states. The number of characters in which both OTU's are positive is labeled n_{JK}, the number in which both are negative n_{jk}. The number of characters positive for one OTU and negative for the other are n_{Jk} and n_{jK}, respectively. Marginal totals are n_J and n_K for positive characters of OTU j and k, respectively; similarly, n_j and n_k are the marginal totals for the negative characters. For convenience in writing formulas and in thinking about the coefficients, we establish the following symbolism: the number of characters in "matched" cells (of the 2 × 2 table), or m, will be $m = n_{JK} + n_{jk}$, and the number of characters in "unmatched" cells, or u, will be $u = n_{Jk} + n_{jK}$. The total number of characters, or n, is

$$n = m + u = n_{JK} + n_{jk} + n_{Jk} + n_{jK}$$
$$= n_J + n_j = n_K + n_k.$$

The 2×2 setup of the data in numerical taxonomy is primarily a convenient arrangement and must not be confused with the conventional 2×2 tables used for tests of independence in statistics. To elaborate this very important point, let us compare a typical example of data suitable for a test of independence in a 2×2 table with a Q-type arrangement between two OTU's. Suppose we wish to investigate the relation between smoking and heart disease in English males of a given age group. We would attempt to obtain a random sample of, say, 500 men of the specified age from different parts of England and various walks of life. We would then investigate each man to learn whether he smoked more than a specified amount and on this basis classify him as a smoker or nonsmoker (positive or negative). A thorough medical examination would establish whether each man suffered from heart disease according to a specified set of criteria; again the sample would be divided into positive and negative classes on the basis of the cardiac findings. The data would be arrayed in a 2×2 table. We would assume that the proportion of smokers and the incidence of heart disease in the sample are representative of their population statistics for the age group. The null hypothesis would suppose independence between the two properties, smoking and heart disease, and the test of this hypothesis could be carried out in the customary way.

In contrast, while we can legitimately assume that the characters based on a Q-type study in numerical taxonomy are a random sample from the infinity of characters that could be discovered, we cannot consider the proportions of positive scores in either taxonomic unit as representative of any parametric value. The meaning of "positive" and "negative" in connection with a 2×2 table for a coefficient of association can vary from the presence or absence of a structure or a chemical reaction to two alternative states of a dimorphic character, without the implication of absence or retrogression possibly carried by the term "negative." Thus in the latter instances the choice of which of the two states is to be called "positive" is quite arbitrary. When a strain of microorganisms becomes drug-resistant, should we call the resistant or the susceptible strain positive? We must not imply that there is a parametric value of positiveness which is estimated by the proportion in our sample. This proportion may change in quite an unpredictable manner when we add new characters and information. This difficulty is common to Q-type studies of association and correlation; we refer to it

as heterogeneity of column vectors (of the original data matrix) and shall have more to say about it in connection with correlational studies. For the reasons stated, it would not be proper from the point of view of statistical theory to test these 2 × 2 tables with the customary chi-square test for independence (nor would we wish to do so, since our aim is to estimate resemblance).

An appropriate mode of thinking about Q-type data in a 2 × 2 table is to consider them a sample proportion of matches from an infinitely large population of character matches which could be attempted (see Section 5.6.2). If there is a parametric resemblance value between the two OTU's, this can be expressed as a percentage of matching characters. It is this percentage that we are estimating by our sample. In this manner the resemblance can be expressed as a single proportion, and the sampling error can be computed by conventional formulas based on the binomial distribution.

6.2.1.1. Possible kinds of coefficients

Much confusion exists in the literature regarding the terminology of coefficients of association. Since they have been used both for R- and Q-type studies, identical coefficients have often been proposed under different names for the two types of matrices. We have adopted the convention of calling the coefficient by the name of its originator (or the first we know to propose it), regardless of the type of study for which the coefficient was proposed.

The fundamental formula consists of the number of matches divided by a term implying the possible number of comparisons but varying in its detailed composition. Table 6-1 presents a scheme for classifying most coefficients of association, based on the nature of the numerator and the denominator. Of the fourteen possible combinations, only three have so far been proposed for use in numerical taxonomy. The others have either been suggested for ecological work or have not yet been used. The coefficients have been separated (vertically) into two groups in Table 6-1 by whether the numerator includes the "negative matches" or not. These are matches for the negative states of characters between the two OTU's.

Whether negative matches should be incorporated into a coefficient of association may occasion some doubts. 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 camel, louse, and nema-

Table 6-1. *Coefficients of Association*

The coefficients of association are divided in two ways. The two vertical columns represent omission (left) and inclusion (right) of negative matches. The horizontal divisions of the table are based on the nature of the denominator indicated in the left margin of the table. Each coefficient is arranged in its box as follows. First line: author of coefficient and reference; second line: formula for coefficient; third and fourth lines: limits of coefficient and conditions necessary for coefficients to assume these limits; fifth line: expected value of the coefficient on the assumption of independence of positive and negative states in the two taxa being compared—that is, $\mathcal{E}(n_{JK}) = n_J n_K / n$ and similarly for the other three cells of the 2×2 table.

Denominator	Negative Matches in Numerator: Excluded	Negative Matches in Numerator: Included
Matched and unmatched pairs are equally weighted	Jaccard, 1908 (Sneath, 1957b) $S_J = n_{JK}/(n_{JK} + u)$ $S_J \longrightarrow 0$ as $n_{JK}/u \longrightarrow 0$ $S_J \longrightarrow 1$ as $u/n_{JK} \longrightarrow 0$ $\mathcal{E}(S_J) = n_J n_K/[n(n_J + n_K) - n_J n_K]$	Simple Matching (Sokal and Michener, 1958) $S_{SM} = m/(m + u) = m/n$ $S_{SM} \longrightarrow 0$ as $m/u \longrightarrow 0$ $S_{SM} \longrightarrow 1$ as $u/m \longrightarrow 0$ $\mathcal{E}(S_{SM}) = (n_J n_K + n_j n_k)/n^2$
	Russell and Rao, 1940 $S_{RR} = n_{JK}/n$ $S_{RR} \longrightarrow 0$ as $n_{JK}/u \longrightarrow 0$ $S_{RR} \longrightarrow 1$ as $(n_{jk} + u)/n_{JK} \longrightarrow 0$ $\mathcal{E}(S_{RR}) = n_J n_K/n^2$	(as above)
Matched pairs carry twice the weight of unmatched pairs	Dice, 1945 (Sørensen, 1948) $S_D = 2n_{JK}/(2n_{JK} + u)$ $S_D \longrightarrow 0$ as $n_{JK}/u \longrightarrow 0$ $S_D \longrightarrow 1$ as $u/n_{JK} \longrightarrow 0$ $\mathcal{E}(S_D) = 2/[(n/n_K) + (n/n_J)]$	Unnamed coefficient $S = 2m/(2m + u) = 2m/(n + m)$ $S \longrightarrow 0$ as $m/u \longrightarrow 0$ $S \longrightarrow 1$ as $u/m \longrightarrow 0$ $\mathcal{E}(S) = \dfrac{[2(n_J n_K + n_j n_k)]}{[n(n_J + n_K) + 2n_j n_k]}$
Unmatched pairs carry twice the weight of matched pairs	Unnamed coefficient $S = n_{JK}/(n_{JK} + 2u)$ $S \longrightarrow 0$ as $n_{JK}/u \longrightarrow 0$ $S \longrightarrow 1$ as $u/n_{JK} \longrightarrow 0$ $\mathcal{E}(S) = 1/[(2n/n_K) + (2n/n_J) - 3]$	Rogers and Tanimoto, 1960 $S_{RT} = m/(m + 2u) = m/(n + u)$ $S_{RT} \longrightarrow 0$ as $m/u \longrightarrow 0$ $S_{RT} \longrightarrow 1$ as $u/m \longrightarrow 0$ $\mathcal{E}(S_{RT}) = \dfrac{(n_J n_K + n_j n_k)}{(n_{Jk} + n_{jK} + n^2)}$
Unmatched pairs only	Kulczynski, 1927 $S_{K1} = n_{JK}/(n_J + n_K - 2n_{JK})$ $= n_{JK}/u$ $0 \leqq S_{K1} \leqq \infty$ $\mathcal{E}(S_{K1}) = 1/[(n/n_K) + (n/n_J) - 2]$	Unnamed coefficient $S = m/u$ $0 \leqq S \leqq \infty$ $\mathcal{E}(S) = \dfrac{[n_J n_K + (n - n_J)(n - n_K)]}{[n(n_J + n_K) - 2n_J n_K]}$

continued on following page

Table 6-1. *(continued)*

Denominator	Negative Matches in Numerator	
	Excluded	Included
Marginal totals	Kulczynski, 1927 $S_{K2} = \frac{1}{2}[(n_{JK}/n_J) + (n_{JK}/n_K)]$	Unnamed coefficient $S = \frac{1}{4}[(n_{JK}/n_J) + (n_{JK}/n_K) + (n_{jk}/n_j) + (n_{jk}/n_k)]$
	$S_{K2} \longrightarrow 0$ as n_{JK}/n_{Jk} and n_{JK}/n_{jK} both $\longrightarrow 0$	$S \longrightarrow 0$ as n_{JK}/n_{Jk}, n_{JK}/n_{jK}, n_{jk}/n_{Jk}, n_{jk}/n_{jK} all $\longrightarrow 0$
	$S_{K2} \longrightarrow 1$ as n_{Jk}/n_{JK} and n_{jK}/n_{JK} both $\longrightarrow 0$	$S \longrightarrow 1$ as above ratios all $\longrightarrow \infty$
	$\mathcal{E}(S_{K2}) = (n_K + n_J)/2n$	$\mathcal{E}(S) = \frac{1}{2}$
Marginal totals	Ochiai, 1957 $S_O = n_{JK}/\sqrt{n_J n_K}$	Unnamed coefficient $S = n_{JK}n_{jk}/\sqrt{n_J n_K n_j n_k}$
	$S_O \longrightarrow 0$ as $n_{JK}/n_{Jk}n_{jK} \longrightarrow 0$	$S \longrightarrow 0$ as $n_{JK}n_{jk}/n_{Jk}n_{jK} \longrightarrow 0$
	$S_O \longrightarrow 1$ as $n_{Jk}n_{jK}/n_{JK} \longrightarrow 0$	$S \longrightarrow 1$ as $n_{Jk}n_{jK}/n_{JK}n_{jk} \longrightarrow 0$
	$\mathcal{E}(S_O) = \sqrt{n_J n_K}/n$	$\mathcal{E}(S) = \sqrt{n_J n_K n_j n_k}/n^2$

tode), 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 heterogenous assemblage (for example, bat, heron, and dragonfly). Neither can we argue that absence of a character may be due to a multitude of causes and that matched absence in a pair of OTU's is therefore not "true resemblance," for, after all, we know little more about the origins of matched positive characters.

Sneath (1957b) excluded negative matches from consideration in his similarity coefficient. He felt that it was difficult to decide which negative features to include in a study and which to exclude. He stated that it is not pertinent to count "absence of feathers" when comparing two bacteria, but that this feature is applicable in comparing bacteria and birds. It is true that through reductio ad absurdum we can arrive at a universe of negative character matches purporting to establish the similarity between two entities. A similarly absurd procedure would be the introduction of positive matches for characters that are invariant over the group under study. The rules described in the section on inadmissible characters (Section 5.3.3) should forestall such improper procedures. In many two-state characters the two states in which they are expressed do not signify presence and absence of the character. Matches for "negative" states are thus of equal value and interest to those for "positive" states. A reasonable and logically defensible position appears to be the

inclusion of positive as well as negative matches for those characters that vary within the group under study. In this respect our conclusion differs from that of Sneath (1957b) and Sørensen (1948) but agrees with Rogers and Tanimoto (1960) and by implication with Sokal and Michener 1958). Most of the applications of association coefficients since 1957 largely in the field of bacteriology) have included negative matches in their coefficients.

In bacteriological work the problem may be slightly different because of the practice of applying a standard series of tests to a group of bacteria. In such cases appreciable blocks of invariant negative characters may result, which would artificially increase resemblance values between OTU's. Exclusion of negative matches from the computation of a coefficient of association may be the safe procedure here, especially since a large group of negatives may on occasion be due to an unrecognized metabolic block preventing the expression of many other characters. This case is similar to that of the missing organs discussed below (Section 6.5.3). The difficulty particularly with microorganisms is to know what characters are missing. In morphological characters this is determined by position—an insect, for example, cannot have wing veins in an absent wing. Thus we must code the veins NC (see Section 5.3.5). But our knowledge of metabolic characters is more limited. It may be impossible to decide if enzymes **A**, **B**, **C**, **D**, $\cdots$ are present but not expressed because of lack of **Z**, which is necessary for activity. This is analogous in a morphological context to being unable to decide how to score subsidiary characters because we do not know whether the organ is present. The coefficient of Jaccard (as employed by Sneath, 1957b) is appropriate when negative matches are to be excluded.

The horizontal subdivision of the coefficients in Table 6-1 is on the basis of the denominator of the fraction. The first row consists of coefficients whose denominator is the total possible number of matches. Matched and unmatched pairs are here given equal weight. Thus the ratio expresses the proportion of actual out of all potential matches. It appears to us to be the simplest of the various ratios proposed. The second row shows Russell and Rao's coefficient, which is "hybrid" in nature, excluding negative matches from the numerator but not from the denominator. This appears to be of questionable utility.

The third row shows coefficients with denominators in which the matched pairs carry twice the weight of the unmatched pairs. While the limits of these coefficients are identical to those of the values in the first row (similarity coefficient $0 \leqq S \leqq 1$), the intermediate values are

bound to be larger. Thus, comparing the coefficients of Jaccard and of Dice, we find the latter greater at all times except when $u = 0$, at which time $S_J = S_D$.

The fourth row of coefficients uses a denominator in which the unmatched pairs carry twice the weight of the matched pairs. Again the limits remain as before, but now the coefficients are less than those in the first row except when $u = 0$—that is, when there are no unmatched pairs of OTU's. The coefficients in the fifth row differ from those in the first three rows in that they have infinity rather than unity as their upper limit. They represent the ratio of the number of matches to that of the nonmatches.

The sixth row represents the proportion of matches as the average of the proportion of matches in j and k, while the seventh row represents the ratio of matches to the geometric mean of the marginal totals.

Next we consider three coefficients of association which balance the

Table 6-2. *Three Coefficients of Association which Balance Matched Pairs against Unmatched Pairs in the Numerator**

Hamann, 1961
$S_H = (m - u)/n$
$S_H \longrightarrow 0$ as $m \longrightarrow u$
$S_H \longrightarrow +1$ as $u/m \longrightarrow 0$
$S_H \longrightarrow -1$ as $m/u \longrightarrow 0$
$\mathcal{E}(S_H) = \frac{1}{n^2}[(n_J - n_j)(n_K - n_k)]$
Yule, 1911 (refer to Yule and Kendall, 1950)
$S_Y = (n_{JK}n_{jk} - n_{Jk}n_{jK})/(n_{JK}n_{jk} + n_{Jk}n_{jK})$
$S_Y \longrightarrow 0$ as $n_{JK}n_{jk} \longrightarrow n_{Jk}n_{jK}$
$S_Y \longrightarrow +1$ as $n_{Jk}n_{jK}/n_{JK}n_{jk} \longrightarrow 0$
$S_Y \longrightarrow -1$ as $n_{JK}n_{jk}/n_{Jk}n_{jK} \longrightarrow 0$
$\mathcal{E}(S_Y) = 0$
phi coefficient (Pearson; refer to Guilford, 1942)
$S_\phi = (n_{JK}n_{jk} - n_{jK}n_{Jk})/(n_J n_K n_j n_k)^{1/2}$
$S_\phi \longrightarrow 0$ as $n_{JK}n_{jk} \longrightarrow n_{Jk}n_{jK}$
$S_\phi \longrightarrow +1$ as $n_{Jk}n_{jK}/n_{JK}n_{jk} \longrightarrow 0$
$S_\phi \longrightarrow -1$ as $n_{JK}n_{jk}/n_{Jk}n_{jK} \longrightarrow 0$
$\mathcal{E}(S_\phi) = 0$

* The arrangement in the boxes for these three coefficients is the same as in Table 6-1.

number of matched and unmatched pairs in the numerator. These range from -1 to $+1$ and are shown in Table 6-2. All the coefficients of association which have so far been employed in numerical taxonomy are discussed below. The computation of the various coefficients of association is shown in Appendix A.2.

6.2.1.2. The coefficient of Jaccard (Sneath):

$$S_J = n_{JK}/(n_{JK} + u)$$

Sneath (1957a) used a coefficient he called the *similarity*, which has had a considerable history of application in R-type and Q-type studies in ecology. The earliest record of its employment we have found is by Jaccard (1908), and we shall therefore refer to it as the coefficient of Jaccard, S_J. It is clear that $S_J \to 0$ as $n_{JK}/u \to 0$, and that as $u \to 0$, $S_J \to 1$. In the latter case $n_J = n_K = n_{JK}$. The coefficient of Jaccard omits consideration of negative matches. In its class it is the simplest of the coefficients.

6.2.1.3. The simple matching coefficient:

$$S_{SM} = m/n = m/(m + u)$$

Sokal and Michener (1958), in a paper dealing with their numerical method for evaluating taxonomic relationships, introduced but did not employ a so-called "matching coefficient." This coefficient is the equivalent of the coefficient of Jaccard just discussed but includes negative matches. Because of its simple nature it must have been thought of and applied repeatedly; see, for example, du Mas (1955). In psychology Zubin (1938) proposed such a coefficient. Without tracing its history, we have called it the simple matching coefficient. This is S_S in Sneath (1962). It is the affinity index of Brisbane and Rovira (1961). From the formula it follows that $S_{SM} \to 0$ as $m/u \to 0$, and that $S_{SM} \to 1$ as $u/m \to 0$, in which case $n_J = n_K = n_{JK}$ and $n_j = n_k = n_{jk}$. When first suggested by Sokal and Michener (1958), the coefficient was not restricted to characters with only two states. However, since we are here considering coefficients of association for dichotomous characters only, we are so restricting it now.

6.2.1.4. The coefficient of Rogers and Tanimoto:

$$S_{RT} = m/(m + 2u) = m/(n + u)$$

Rogers and Tanimoto (1960) developed a *similarity ratio* with flexibility to include characters with more than two states and also to take

missing information into account. For purposes of this section we shall limit their coefficient, S_{RT}, to the case of two states per character and with complete information. Its mathematical formulation is then as given above. It follows that $S_{RT} \to 0$ when $m/u \to 0$ and that $S_{RT} \to 1$ when $u/m \to 0$, in which case $n_J = n_K = n_{JK}$ and $n_j = n_k = n_{jk}$. The coefficient of Rogers and Tanimoto includes negative matches, if we wish to call the second of the two states negative. It is more elaborate than the simple matching coefficient but is functionally related to it.

6.2.1.5. The coefficient of Hamann:

$$S_H = (m - u)/n$$

This coefficient has been employed by Hamann (1961) in a study of some families of monocotyledonous plants. It employs the difference between the matched and unmatched pairs as a criterion of association. Thus when the number of matched and unmatched pairs is equal, $m = u$ and $S_H = 0$. This coefficient can range from -1 to $+1$ when $u \to 0$ and $m \to 0$, respectively. It shares this property with the two coefficients listed next. These latter coefficients carry the determinant of the 2×2 table as their numerator. When two OTU's are independent on the basis of their characters, the determinant will be zero, yet the term $(m - u)$ is not necessarily so. As can be seen in Table 6-2, the expected value of S_H will be zero only when $n_J = n_j$ or $n_K = n_k$ or both of these relations hold. This might appear an undesirable property of S_H, yet independence probably does not have a clear meaning in 2×2 tables in numerical taxonomy, as discussed above.

6.2.1.6. The coefficient of Yule:

$$S_Y = (n_{JK}n_{jk} - n_{jK}n_{Jk})/(n_{JK}n_{jk} + n_{jK}n_{Jk})$$

This coefficient, described as Q in Yule and Kendall (1950), is symbolized by us as S_Y in order to conform to the pattern adopted for designating coefficients of association. Its limits are $+1$ when n_{jK} or $n_{Jk} \to 0$ and -1 when n_{JK} or $n_{jk} \to 0$. Yule and Kendall describe it as the simplest possible (of its kind), although not necessarily the most advantageous that may be devised. It has been employed also by Brisbane and Rovira (1961).

6.2.1.7. The phi coefficient:

$$S_\phi = (n_{JK}n_{jk} - n_{jK}n_{Jk})/(n_J n_j n_K n_k)^{1/2}$$

This well-known coefficient can be found in many statistics books, as for example in Yule and Kendall (1950). It is also known as the fourfold

point correlation coefficient. Its customary symbol is ϕ, used by us in subscript form. The limits of S_ϕ are the same as those of S_Y, whose denominator it shares. The phi coefficient is frequently used in statistics and is important because of its relation to χ^2; that is, $\chi^2 = \phi^2 n$. This permits a test of significance; however, because of the problem of heterogeneity of column vectors (Section 6.2.1), it is doubtful whether any meaning can be applied to such a test.

6.2.1.8. Smirnov's coefficient of similarity and the generalized coefficient of Rogers and Tanimoto

Two coefficients of association remain which we have not been able to bring into our general classification. These deal with characters whose number of states are not restricted to two. Such characters may or may not be linearly ordered. Rogers and Tanimoto (1960) treat these by a general formula, from which the formula given in Table 6-1 for two-state characters has been abstracted. The treatment by Smirnov (1960) approaches the problem in quite a different manner. Although these coefficients are quite different, the basic arrangement of data for both is identical.

Table 6-3 illustrates two OTU's scored for four characters *A*, *B*, *C*, and *D*. These characters have states ranging from 2 for character *B* to 5 for character *D*. The two OTU's are scored positively if they possess a given state for a character and negatively if they do not. Thus we see that OTU's 1 and 2 agree in both exhibiting state 2 for character *A* and state 3 for character *C*. They disagree in character states for characters *B* and *D*. If a simple matching coefficient (S_{SM}) were applied to these data, this would imply that the mismatch in character *D* is

Table 6-3. *Four Multistate Characters to Illustrate the Computation of the Coefficients of Rogers and Tanimoto* (1960) *and Smirnov* (1960).

	Character														
	A				B		C			D					
State	1	2	3	4	1	2	1	2	3	1	2	3	4	5	
OTU 1	−	+	−	−	−	+	−	−	+	−	+	−	−	−	
OTU 2	−	+	−	−	+	−	−	−	+	−	−	−	−	+	
	States expressed by Smirnov's system														
OTU 1	a_1	A_2	a_3	a_4	b_1	B_2	c_1	c_2	C_3	d_1	D_2	d_3	d_4	d_5	
OTU 2	a_1	A_2	a_3	a_4	B_1	b_2	c_1	c_2	C_3	d_1	d_2	d_3	d_4	D_5	

equivalent to the mismatch in character B. If the character states are scored entirely qualitatively, then this may well be a legitimate assumption; in such a case the simple matching coefficient might be indicated. The small fictitious example here would give a matching coefficient of two matches divided by four possible matches ($= 0.5$). When we give more weight to the mismatch in character D than to the mismatch in character B we presumably imply dimensionality in the character states. In such instances the coefficients of distance and correlation mentioned below would be more appropriate.

Rogers and Tanimoto's coefficient of similarity calculates the number of agreements in character states divided by the number of character states represented by at least one OTU. Thus the numerator of their coefficient would be 2 (one for the agreement on character state A_2, the other one for the agreement on character state C_3), and the denominator would be 6 (character states A_2, B_1, B_2, C_3, D_2, and D_5 all have at least one positive representative in the comparison). Thus Rogers and Tanimoto's similarity coefficient, S_{RT}, would equal $\frac{2}{6} = 0.33$.

In order to understand Smirnov's coefficient we first have to become familiar with his terminology. In Smirnov's coefficient the number of species (OTU's) involved in any given taxonomic study is of importance; this number is called s. Of importance also is the distribution of the states (modalities) for a given character among the s species of the study. Thus Smirnov will write such a distribution as

$$(E_1) + (E_2) + (E_3) + \cdots + (E_{e-1}) + (E_e) = s,$$

where (E_1) is the number of species exhibiting character state 1 for character E. The number of character states in character E is symbolized by e. Thus character A has a states (four in the example in Table 6.3), while character B has b states (two in the same example). Since all the s species in a study must exhibit one or the other character state, $\sum_{i=1}^{e} (E_i) = s$. Smirnov designates (e_1) as the number of species not possessing character state 1 of character E. Obviously the following relation must hold:

$$(E_1) + (e_1) = s.$$

In an actual example we might have the following distribution of positive character states in one hundred species:

$$30E_1 + 20E_2 + 5E_3 + 45E_4 = 100.$$

Since 30 species possess character state 1, it is clear that 70 of the species

will not possess that character state. We can therefore write the above relation as

$$\begin{aligned} 30E_1 + 70e_1 &= 100, \\ 20E_2 + 80e_2 &= 100, \\ 5E_3 + 95e_3 &= 100, \\ 45E_4 + 55e_4 &= 100. \end{aligned}$$

The key to understanding Smirnov's method is that the similarity based on any one character is weighted as a function of the probability of the simultaneous occurrence of such a character state in two separate OTU's. If two forms share a rare character state, this is given much weight; if they share a commonly occurring character state, this is given little weight. Smirnov argues that when a character rare in the larger taxon under study (genus) occurs in two OTU's (species), then this is more important for determining similarity than if the concurrence is in a character state that is widely distributed. Thus, if in the fictitious example cited above an agreement occurred between two OTU's matching for character state E_1, the weight to be calculated would be as follows:

weight for character state match E_1E_1,

$$w_{E_1E_1} = \frac{(e_1)}{(E_1)} = \frac{70}{30} = 2.33.$$

However, an agreement in character state E_3 results in the following weight,

$$w_{E_3E_3} = \frac{(e_3)}{(E_3)} = \frac{95}{5} = 19.$$

Conversely, an agreement with respect to the absence of character state E_3 is weighted only very slightly:

$$w_{e_3e_3} = \frac{(E_3)}{(e_3)} = \frac{5}{95} = .053.$$

The minimum and maximum possible weights to be applied to an agreement are

$$(w_{EE})_{\min} = \frac{1}{s-1},$$

$$(w_{EE})_{\max} = \frac{s}{2} - 1.$$

Thus we see that weights are functions of s, the number of OTU's in the

study. For any one character E, an average weight is calculated, representing the average of the weights for all the states of the character:

$$\overline{w}_E = \frac{1}{e} \sum w_E = \frac{1}{e} (w_{1,1} + w_{2,2} + \cdots + w_{e,e}),$$

where $w_{1,1}$ represents the weight for the positive match, negative match, or mismatch for character state 1 of character E, there being a total of e states for this character. Mismatches are given a weight of -1. If we had two OTU's exhibiting different states of the fictitious character E above, thus:

OTU 1 shows character state E_1
OTU 2 shows character state E_3,

then

$$\overline{w}_E = \frac{1}{e}\left[-1 + \frac{(E_2)}{(e_2)} - 1 + \frac{(E_4)}{(e_4)}\right]$$

$$= \frac{1}{4}\left[-1 + \frac{20}{80} - 1 + \frac{45}{55}\right] = -.233.$$

The similarity between any two OTU's—which Smirnov calls $t_{f,g}$ for OTU's f and g, respectively—is calculated by summing the weights for the e states of *all* characters and dividing by n, the sum of all the numbers of character states; that is

$$n = \sum_{i=1}^{m} e_i,$$

where m is the number of characters employed and e is the number of states per character. Finally,

$$t_{f,g} = \frac{1}{n} \sum_{i=1}^{m} (\Sigma\, w_{Ei}).$$

Thus, if we were to assume that OTU's 1 and 2 in Table 6-3 had been taken from 120 OTU's and that the frequencies of the states of characters A, B, C, and D were equal in the sample (a highly artificial assumption), we would arrive at the following Smirnov similarity coefficient:

$$t_{1,2} = \frac{1}{14}\left[\left(\frac{30}{90} + \frac{90}{30} + \frac{30}{90} + \frac{30}{90}\right) + (-1 - 1)\right.$$

$$\left. + \left(\frac{40}{80} + \frac{40}{80} + \frac{80}{40}\right) + \left(\frac{24}{96} - 1 + \frac{24}{96} + \frac{24}{96} - 1\right)\right]$$

$$= 0.268.$$

The validity of Smirnov's coefficient must rest on the contention that

a similarity coefficient should be placed on a probabilistic basis. Should similarity in rare structures be made more important than similarity in commonly occurring structures? We would hesitate to take such a step, since this would make the magnitude of the similarity coefficient much too dependent on the size and nature of the group investigated.

Superficially, the idea of weighting characters on the basis of the rareness of their occurrence has a certain attraction, particularly if the rarer structure or character is a complicated one. As it seems quite unlikely that independent evolution had produced this same structure, people tend to give more weight to such similarities, especially if they are deducing phyletic relationships. However, as we have stated elsewhere, we believe that in such cases the importance of the character would be shown by the presence of numerous correlated characters, which together automatically weight the character. We feel that such "built-in" weighting is preferable to Smirnov's system.

Smirnov's coefficient has one other disadvantage. It does not result in unity when comparing OTU's with themselves, but gives different coefficients for different OTU's. Smirnov interprets the magnitude of $t_{f,f}$ as a measure of the uniqueness of species f with respect to the others with which it is being compared. A similar measure would be found in the *uniqueness* of factor analysis (see Section 7.3.3).

In view of the above drawbacks and the fact that multistate characters can usually be handled preferably by distance or correlation analysis, we cannot recommend Smirnov's coefficient for use in numerical taxonomy.

6.2.1.9. Comparison of coefficients

When we try to evaluate the relative merits of the six coefficients adapted to two-state characters, we must first consider whether inclusion of the negative matches is justified. If not, then the coefficient of Jaccard (or Sneath) appears appropriate as being the simplest of the coefficients in its class. There may be cases, particularly when there are many biochemical characters, in which sufficient grounds for rejecting negative matches can be found. But we have already stated that in most cases it would appear that all matches should be considered. Under those circumstances we would prefer the simple matching coefficient over that of Rogers and Tanimoto. The former is a simpler quantity and is easier to interpret, for S_{SM} is functionally related to S_{RT}. The relation can be expressed as

$$\frac{S_{SM}}{S_{RT}} = \frac{m + 2u}{m + u}.$$

From this ratio it can be seen that in most cases $S_{SM} > S_{RT}$, that $S_{SM} \to S_{RT}$ when $u \to 0$ and that $S_{SM} \to 2S_{RT}$ when $m \to 0$. We can interpret S_{SM} as the probability that two OTU's j and k will match for a given character selected at random. Such an interpretation leads directly to the concept that the probability of matching is an expression of the phenetic relationship between the two taxa (see Section 5.6.2).

Although ordinarily we would employ neither S_J nor S_{RT}, it may be useful to describe the mutual relations between these two coefficients and S_{SM}. The relation between S_J and S_{SM} can be expressed as

$$\frac{S_J}{S_{SM}} = \frac{n_{JK}^2 + n_{JK}n_{jk} + n_{JK}u}{n_{JK}^2 + n_{JK}n_{jk} + (n_{JK} + n_{jk})u},$$

from which we learn that

$$S_J < S_{SM} \quad \text{when} \quad n_{jk} > 0,$$
$$S_J \to S_{SM} \quad \text{when} \quad n_{jk} \to 0 \quad \text{and} \quad u > 0,$$

when $u \to 0$ and n_{JK}, n_{jk} are both >0, then $S_J \doteq S_{SM} \to 1$. By comparison,

$$\frac{S_J}{S_{RT}} = \frac{n_{JK}^2 + n_{JK}n_{jk} + 2n_{JK}u}{n_{JK}^2 + n_{JK}n_{jk} + n_{JK}u + n_{jk}u}.$$

Hence

$$S_J > S_{RT} \quad \text{when} \quad n_{JK} > n_{jk},$$
$$S_J = S_{RT} \quad \text{when} \quad n_{JK} = n_{jk},$$
$$S_J < S_{RT} \quad \text{when} \quad n_{JK} < n_{jk},$$

when $u \to 0$ and n_{JK}, n_{jk} are both >0, then $S_J \doteq S_{RT} \to 1$.

Cole (1949) has discussed the relations among coefficients of association which range from -1 to $+1$. His preferred coefficient, C_7, would not be applicable to numerical taxonomy as it ignores negative matches. We have some reservations about using coefficients with this range. Should organisms be related quantitatively along a scale which permits them to be negatively correlated? Positive resemblance between organisms and absence of resemblance can be easily understood. What, however, is negative resemblance between organisms? In perfect negative correlation between OTU's **A** and **B**, every character on the basis of which the OTU's are being compared must have opposite character states in the two taxa. As a consequence, perfect negative association of this sort would result in an "anti-organism" whose organic feasibility and viability would be somewhat in doubt. On the whole, it seems to us that a similarity value scale ranging from 0 to 1 is to be preferred. On

the other hand, we should point out that some coefficients such as S_{SM} indicate a completely negative correlation between two OTU's as zero.

6.2.2. Coefficients of correlation

Coefficients of correlation have been repeatedly used in Q-type studies in both psychology and ecology. In the former science, Stephenson (1936) originated the Q-technique (under the name of inverted factor technique). The use of correlation coefficients in ecology is reviewed by Dagnelie (1960). These coefficients have been employed in numerical taxonomy by Michener and Sokal (1957), Sokal and Michener (1958), Sokal (1958), Morishima and Oka (1960), Ehrlich (1961c), Soria and Heiser (1961), and Rohlf (1962). Only the product-moment correlation coefficient has been used to date, and this on data where most if not all of the characters were present in more than two states. This coefficient, computed between taxa j and k, is

$$r_{jk} = \frac{\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)}{\sqrt{\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2 \sum_{i=1}^{n} (X_{ik} - \overline{X}_k)^2}},$$

where X_{ij} stands for the character state value of character i in OTU j, $\overline{X}_j$ is the mean of all state values for OTU j, and n is the number of characters sampled. Since this formula is based on moments around the mean, it takes into account the magnitudes of mismatches between taxa for characters with more than two states. In this respect correlation coefficients are superior to the coefficients of association described in Section 6.2.1. They resemble the three coefficients of association in Table 6-2 in that their limits range from -1 to $+1$. Thus negative correlation between taxa is at least theoretically possible.

Among the studies published to date, only that of Morishima and Oka (1960) shows high and significant negative correlations. We explain the generally positive nature of taxonomic Q-correlation matrices by saying that we are unlikely to find a pair of OTU's antithetical for an appreciable number of characters (the improbability of an "anti-organism"). Another explanation for the same phenomenon put forward by Michener and Sokal (1957) and Sokal and Michener (1958) appears on re-examination to be in error.

As a measure of phenetic resemblance, r_{jk} has undisputed merit, but

doubt must prevail about the significance of coefficients computed for a Q-type study. The heterogeneity of column vectors, noted in the previous section, is an equally irksome problem here. When the data are arbitrarily coded and the number of states varies for different characters, the correlations cannot meet the basic assumptions of the bivariate normal frequency distribution. This problem has already been faced by the psychologists, who have suggested standardizing rows of the basic data matrix. By this is meant calculating the mean and standard deviation of each character and transforming each character score into a standard deviate—that is, dividing its deviation from its mean by the standard deviation. This will create a mean of zero and a variance of one for every character. We can therefore postulate that the variates for each OTU (each of the column vectors) are sampled from n populations having a common mean (zero) and standard deviation (unity).

Studies of the effect of standardization of characters have been carried out by Rohlf and Sokal (1963) on the 97 species of bees of the *Hoplitis* complex first analyzed by Michener and Sokal (1957) and on the 48 species of the mosquito genus *Aedes* analyzed by Rohlf (1962). The above investigations have shown that standardization of characters reduces the average correlation within a matrix to approximately zero from the previous positive value. The standard deviations of the correlation coefficients based on standardized characters are larger than expected, and the coefficients are skewed to the right. Among correlations based on standardized characters, few negative correlations lower than -0.3 have been observed, while positive correlations can range almost up to unity. The phenetic relationships obtained from standardized correlation matrices are quite similar to those based on unstandardized correlation matrices. Standardization can also be achieved in a general way by various devices designed to equalize the variances of the different characters. Use of a percentage scale (Cain and Harrison, 1958) or a ratio of the variable against a standard (Haltenorth, 1937) are cases in point. In such instances negative correlations could occur.

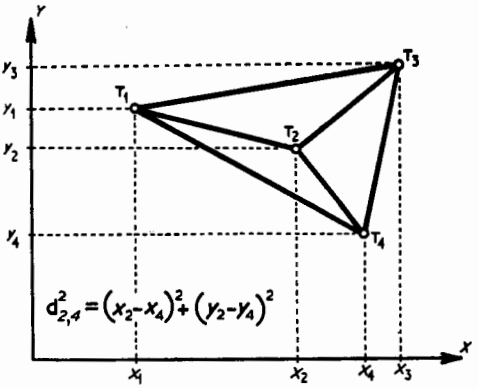
Why did Morishima and Oka's (1960) study show appreciable negative correlations? These authors, who analyzed 16 strains and species of rice, may have initially coded their data in such a manner that means and variances of character state codes within a character were approximately identical. Such a view is supported by the fact that standardization of approximately half of their characters did not appreciably lower the mean of their correlation coefficients, which already was near zero.

6.2.3. Measures of distance

These statistics of resemblance between OTU's are based on a geometrical model. As we shall see later, we can relate them in special cases to the other types of estimates of resemblance. However, at face value they appear quite different. Since the techniques involved may be less familiar, we shall go into them in some detail.

Let us assume we have two OTU's for which n characters have been studied. The states for each character have been assigned values along a scale ranging for convenience from zero to unity. We now draw a conventional pair of rectangular coordinates in which the abscissa

FIGURE 6-1

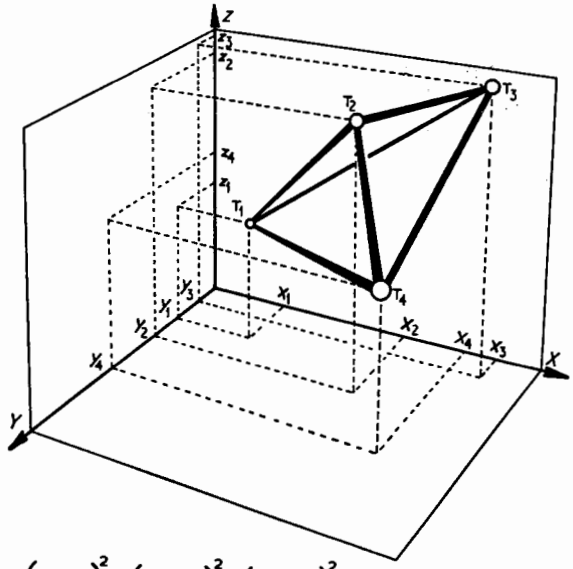


Representation of four OTU's (T_1, T_2, T_3, T_4) *as points on a plane determined by their character states for two characters,* X *and* Y. *Each character is represented by a dimension—in this case two. The quantity labeled* $d^2_{2,4}$ *in the figure is referred to as* $\Delta^2_{2,4}$ *in the text. To obtain the taxonomic distance,* $d_{2,4}$, *between OTU's* T_2 *and* T_4 *as defined in the text we must divide the quantity on the right side of the equation by* n, *the number of characters, and take the square root of the quotient (see Section 6.2.3.2).*

represents character X and the ordinate character Y. Next we plot the position of the two OTU's with respect to these coordinate axes. A hypothetical case is shown in Figure 6-1. If the two OTU's are identical as regards the two characters under consideration, their positions will coincide and the distance between them will be zero. The greater the disparity between them, the greater will be the distance. Thus distance is seen to be the complement of similarity. From elementary analytical geometry we can show that the maximum distance possible between the OTU's would be $\sqrt{2}$ when they occupy respectively the tips of the two coordinate axes.

When we wish to estimate taxonomic distance on the basis of three characters, we must add a third coordinate (Z) to our diagram. On paper such a three-dimensional model can only be shown as a two-dimensional projection (Figure 6-2). The maximal distance is now $\sqrt{3}$.

We cannot visualize the geometry of adding a fourth and subsequent



$$d^2_{2,4} = (x_2 - x_4)^2 + (y_2 - y_4)^2 + (z_2 - z_4)^2$$

FIGURE 6-2

Representation of four OTU's in a three-dimensional space, obtained from Figure 6-1 by adding a third character, Z. (*For the meaning of* $d^2_{2,4}$, *see Figure 6-1.*)

characters. The requirements of each new coordinate axis are that it be at right angles with all previous ones. Although we cannot depict such an axis graphically, we can postulate its existence and demonstrate algebraically that most of the geometric theorems of conventional three-dimensional space can be extended to n dimensions in so-called Euclidean hyperspace. Thus we are at liberty to postulate n dimensions for n characters. We can compute the distance between the two OTU's in hyperspace. The maximum distance will be $\sqrt{n}$ in an n-space, a hyperspace of n dimensions. (It should be re-emphasized here that these maximal distances are based on characters with maximal values of unity.)

Before we describe the various techniques which have been suggested to establish measures of distance, we have to discuss the coding of the character states for distance analysis. Coding them between the limits of 0 to 1 (as is suggested by Cain and Harrison, 1958) seems a logical standardization, permitting maximum distances to be calculated. While this does not equalize the variances of the character states (rows) it does constrain them somewhat. A drawback of this scheme of coding would be its inability to admit OTU's with more extreme character states without recoding all characters.

Another technique is to standardize rows as described above for r—that is, to compute the mean and standard deviation of each row (the states of each character) and to express each state as a deviation from the mean in standard deviation units (Sokal, 1961). Since this would result in negative as well as positive values, adding 5 to each value would leave the variance unchanged yet would make computation easier when desk calculators are employed. When working with a computer, coding by the addition of 5 is generally unnecessary.

The standardization of the character states would make all character variances equal to unity. However, we are still faced with the problem of what to do when we wish to add a new species with one or more character states beyond the previous limits. Our coding system will now not have limits of 0 and 1 as before, and therefore it should not be difficult to express the new variate in standardized form. To give a concrete example, if the range of states for a character has been from 1 to 6, its mean 4.1 and its standard deviation 0.8, then the previous limits of the range must have been coded $(1 - 4.1)/0.8 = -3.875$ and $(6 - 4.1)/0.8 = 2.375$, respectively. A new state of 7 is now to be coded. It simply becomes $(7 - 4.1)/0.8 = 3.625$. This value is not really correct, however, since both mean and standard deviation of the character states have changed with the addition of the new state. We believe that when only few new OTU's and few cases of new extremes are involved this is not a serious problem, as the variance would be inappreciably altered. When a larger number of new states is involved (a case which we consider quite unlikely if fairly exhaustive comparative study has preceded the analysis), a fresh standardization of the affected characters will be necessary.

If desired, a normalization rather than a standardization can be carried out by means of rankits. (A rankit is the average deviate of the rth largest in a sample of n observations drawn at random from a normal distribution with a mean of zero and a variance of unity; Bliss and Calhoun, 1954.) However, it seems to us that in the procedures outlined below this will not be necessary. For the purposes of the discussion which follows we shall consider our data to be either coded from 0 to 1 or to have been standardized.

We shall establish the following symbolism to deal with the examples in Sections 6.2.3.1, 6.2.3.2, and 6.2.3.3. Let X_{ij} be the value of the state of character i in OTU j, where i varies from 1 to n and j varies from 1 to t. Thus, for example, the difference between OTU's 6 and 8 for

character 3 would be written as $(X_{3,6} - X_{3,8})$. (The comma is introduced between the subscripts when dealing with numerals or triple letters in order to prevent confusion.)

6.2.3.1. Average differences

The expression

$$\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - X_{ik})$$

is not particularly suitable for measuring the distance between OTU's j and k, since the differences could be negative as well as positive. In a random, symmetrical model the expected value of this expression would be zero. The obvious correction would be to use

$$\frac{1}{n}\sum_{i=1}^{n}|X_{ij} - X_{ik}|,$$

the absolute (positive) values of the differences between the OTU's for each character. This is the *mean character difference* (*M.C.D.*), which has been proposed by Cain and Harrison (1958) as a measure of taxonomic resemblance. It had previously been used in anthropology by Czekanowski (1932), who called it *durchschnittliche Differenz*. Haltenorth (1937) employed this coefficient in an extensive study of 86 characters of eight species of the large cats. Each character was a mean based on a large number of specimens. In computing it Haltenorth counted as zero all differences which were not statistically significant. We consider this to be an undesirable feature of his system, as is explained in Section 6.2.4. However, a reanalysis of his original data by Sokal—employing d, the coefficient of taxonomic distance (Section 6.2.3.2)—resembled Haltenorth's results closely.

The simplicity of this statistic is in its favor; however, it does suffer several major disadvantages. It will always underestimate the true Euclidean distance between the taxa in space, and when some character differences are small while others are large, it will underestimate the actual distance considerably. It also lacks some of the desirable attributes of the alternative measure, the Euclidean distance or its square, described below. It cannot be partitioned into components. In general, it stands in the same relation to the distance as the average deviation to the standard deviation and suffers from similar disabilities as the former. In considering which of the methods to apply it might be argued that the mean character difference is simpler and hence to be preferred;

however, it is reasonable to expect that any worthwhile study in numerical taxonomy will depend on machine computation. As simple a study as 15 OTU's of 60 characters each will require 105 comparisons and consequently 6300 subtractions. Calculating the taxonomic distance on a good desk calculator would not be prohibitively more time-consuming than obtaining Cain and Harrison's mean character difference. Most numerical taxonomic studies will probably employ electronic computers, however. The limitations of these machines in relation to our work will be shortage of storage and slowness of input rather than computation time, which would not differ appreciably among correlational, distance, or mean character difference techniques.

6.2.3.2. Taxonomic distance

The distance between two OTU's in two- and three-dimensional spaces has been illustrated in Figures 6-1 and 6-2. We can generalize this concept of the Euclidean distance between two points in an n-dimensional space. The formula for such a distance, Δ_{jk}, between OTU's j and k, using the symbolism of the last section, is

$$\Delta_{jk} = \left[\sum_{i=1}^{n} (X_{ij} - X_{ik})^2\right]^{1/2}.$$

The symbol Δ_{jk} has been adopted in place of δ_{jk}, suggested by Sokal 1961), in order to conform to conventional statistical usage, in which sample statistics are identified by Roman letters and parameters are symbolized by lower-case Greek letters. Since we are reserving the symbol d for average distance (as specified below) and since D is used for Mahalanobis' generalized distance, we assigned the symbol Δ for the above quantity. This convention has also been employed by Rohlf and Sokal (1963). It will sometimes be found useful to employ the square of the distance; the formula then becomes

$$\Delta_{jk}^2 = \sum_{i=1}^{n} (X_{ij} - X_{ik})^2.$$

Since Δ_{jk}^2 increases with the number of characters used in the comparison, an average distance is commonly computed. This is

$$d_{jk} = \sqrt{\Delta_{jk}^2/n} \quad \text{or} \quad d_{jk}^2 = \Delta_{jk}^2/n$$

in square root and square form, respectively. The average squared distance, d_{jk}^2, was employed by Sokal (1961) and called $\overline{\delta_{jk}^2}$ by him.

The idea for such a coefficient has come to many people. So far as we

can learn, this form of a measure of distance was first employed by Heincke as early as 1898. Schilder and Schilder (1951) demonstrated such a coefficient without standardizing characters. Clark (1952) has employed the same coefficient in comparing several populations of snakes, some only subspecifically distinct, others in different genera. His distances for each character are computed as a ratio varying between zero and unity. Thus his distance, called by him the *coefficient of divergence*, is computed as

$$CD_{jk} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_{ij} - X_{ik}}{X_{ij} + X_{ik}}\right)^2\right]^{1/2}.$$

In Clark's original formula each of the X terms actually represented a mean, $\overline{X}$, since he used a number of specimens to represent each OTU. For numerical taxonomy we have seen that single values will frequently suffice.

Bielicki (1962) describes a coefficient of distance for use in anthropology. Based upon an earlier statistic of Wanke (1953), it is similar to the coefficient of Clark (1952).

Sokal (1961) brought taxonomic distance to the attention of numerical taxonomists, employing the formulation of d_{jk}^2 and suggesting standardization of character state codes for each character.

Zarapkin (1934, 1939, 1943) employed a statistical approach related to distance although not identified as such. He employed a so-called standard population for which he would compute the mean and standard deviation of each character considered. Then for each character i and for each OTU j he computed $(X_{ij} - \overline{X}_{i,st})/s_{i,st}$, where $\overline{X}_{i,st}$ represents the mean of a standard population for character i, while $s_{i,st}$ is its standard deviation. This quantity is essentially a standardized deviation, except that the standardization is based on the s of the individuals within the standard population rather than the s of the character in question across all the OTU's of the study, which would seem to us to be a more reasonable scale of reference. When a standard population was unavailable and a single individual had to serve as an OTU, then the deviation was calculated as a percentage $(X_{ij} - X_{i,st})/X_{i,st}$. Zarapkin studied the frequency distributions of these deviations for many characters in a variety of populations, finding the shape and spread of these frequency distributions to be related to the taxonomic differences between the OTU's under consideration and the standard. To summarize his findings Zarapkin computed the standard deviations of all the deviations between each OTU and the standard population (that

is, over all the characters). He called this quantity $\mathfrak{S}$. It is easily seen that, to the extent that the mean of the deviations between any OTU and the standard population approaches zero, the quantity is nothing but the distance between the OTU and the standard population expressed in the standardized units of the latter. The difference between $\mathfrak{S}_{jk}^2$ and d_{jk}^2 is the size coefficient of Penrose, C_Q^2 (see Section 6.2.3.3). In this manner Zarapkin was able to obtain distances of a variety of populations from a standard but not of the populations from each other.

We would not recommend this method. (1) It sets up an arbitrary standard population with respect to which the others are viewed, rather than considering them simultaneously (and thus from a multivariate point of view). (2) By standardizing on the basis of the standard population we may be introducing considerable bias into the interpretation of our findings. Worse yet, we may standardize interpopulation deviations on the basis of intrapopulation standard deviations. (3) Depending on the size of the standard population, $\mathfrak{S}$ may or may not represent taxonomic distance. An unequivocal statistic is preferable in this regard.

Related to taxonomic distance is the *coefficient of racial likeness*, developed by Karl Pearson (1926) for measuring resemblances between samples of skulls of various origins. The problem here differs in that continuous characters (lengths, ratios, and others) are measured, which vary from specimen to specimen and can only be expressed as means. This coefficient is computed as

$$\text{C.R.L.} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(\bar{X}_{ij} - \bar{X}_{ik})^2}{\frac{s_{ij}^2}{n_j} + \frac{s_{ik}^2}{n_k}} \right] \right\}^{1/2} - \frac{2}{n},$$

where $\bar{X}_{ij}$ stands for the sample mean of the ith character for entity j, s_{ij}^2 for the variance of the same, and n_j for the sample size of entity j.

In numerical taxonomy we can develop the following formula for the C.R.L., with Q-type data which have been standardized by rows. Since we have no means but only single values (with a variance of one) representing the OTU's, we obtain

$$\text{C.R.L.} = \left[\frac{1}{2n} \sum_{i=1}^{n} (X_{ij} - X_{ik})^2 \right]^{1/2} - \frac{2}{n},$$

which approximates $\sqrt{2d_{jk}}$ for any of the cases in which we are interested ($n \geqq 60$). The quantity d_{jk}^2 has therefore been called the reduced C.R.L. (Morant, 1936). The standard error of the C.R.L. approaches $1/\sqrt{2n}$ when n is sizable. The C.R.L. is sometimes given in squared form (as in

Sokal, 1961). We prefer the plain form here, to be able to analogize it with d. The expected value of the C.R.L. for independent characters is $1 - (2/n)$. It should approach 1 for a large number of characters if the states are normally distributed.

On the assumptions that the observations (the n standardized characters) were independent and normally distributed with a mean of zero and a variance of unity, Rohlf (1962) computed the expected value of d for even values of n as

$$\mathcal{E}(d) = \frac{(n-1)!}{\left[\left(\frac{n}{2}-1\right)!\right]^2 2^{n-2}} \left(\frac{\pi}{n}\right)^{1/2}.$$

Using Stirling's formula this reduces to

$$\mathcal{E}(d) \doteq \sqrt{2}\left(1 - \frac{1}{n}\right)^{1/2}\left(1 + \frac{1}{n-2}\right)^{n-1}\frac{1}{e}.$$

Thus for the large values of n generally employed in numerical taxonomy the expected value of d approaches $\sqrt{2}$ very closely. The expected variance of d is

$$\mathcal{E}(\sigma_d^2) = 2 - [\mathcal{E}(d)]^2,$$

which approaches zero as n tends to infinity. Figure 5-9 gives the 95% confidence limits to the expected value of d at various sample sizes, n. We see that after n approximately equals 75, the width of the band decreases very slowly.

Generalized distance is a statistic related to the coefficient of racial likeness. Developed by Mahalanobis (1936) and Rao (1948), it is a weighted coefficient similar to a squared distance in which both the variance of separate characters and the correlations among characters are taken into account. Generalized distance procedures are not appropriate for the usual type of work in numerical taxonomy, since in the latter field we recommend a single code value to represent the state of a given character in a given taxon. We do not, therefore, consider variation for a given character within a taxon. But the osteological material with which Pearson and the Indian statisticians were dealing had characters which were largely continuous, varying with the population; in their attempts to compute distance coefficients for such material they had to take into account the mean and variance of each of the characters considered. Since the material with which we are working consists of an appreciable proportion of discontinuous characters and since there is in many cases no variation of the particular state code within the taxon,

the problem of intrataxon variation in state codes does not usually arise in our data, and we may assume their values to be fixed. Furthermore, as a matter of simple practicality we cannot consider using generalized distance in the study of numerical taxonomy, for these methods require inversions or similarly involved operations on matrices of the order of the number of characters considered. As we are often employing as many as a hundred or more characters, inversions of such matrices would be entirely impractical. No claim is made here that the characters which we are considering in numerical taxonomy are always invariable within taxa. On the contrary, in low-ranking taxonomic units (such as subspecies, races, or varieties) as well as high-ranking taxa, many character states are not constant for the entire taxon. The several methods suggested for treating such data are discussed in Section 6.4. But generalized distance allows for correlation among characters which the methods proposed so far in numerical taxonomy do not. Until R-type correlations are studied in numerical taxonomy the importance of this aspect of generalized distance for this work cannot be evaluated.

We might point out here that the simple matching coefficient S_{SM} is related to distance in the following way. If we consider character axes ranging from zero to one so that the maximum difference between any two OTU's for any one character is unity, then the maximum distance between a pair of OTU's based on n characters is $\sqrt{n}$. It can easily be shown that the average squared distance $d_{jk}^2 = u/n$; therefore $d_{jk}^2 = 1 - S_{SM}$.

Rogers and Tanimoto (1960) have also defined a distance coefficient, d_{ij}. This coefficient is defined as $d_{ij} = -\log_2 S_{ij}$, where S_{ij} is an association coefficient between OTU's i and j and ranges from zero to unity. These distances define a semimetric space, rather than a metric one, such as is defined by the distances discussed above. It seems most appropriate to discuss the distance of Rogers and Tanimoto in connection with their method of clustering, presented in Section 7.3.2.5.

6.2.3.3. Size and shape

Penrose (1954) has suggested dividing the reduced C.R.L. (d_{jk}^2, or C_H^2 in his symbolism) into two parts: a coefficient of "size,"

$$C_Q^2 = \frac{1}{n^2}\left[\sum_{i=1}^{n}(X_{ij} - X_{ik})\right]^2,$$

and a coefficient of "shape,"

$$\frac{n-1}{n} C_Z^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - X_{ik})^2 - \frac{1}{n^2} \left[\sum_{i=1}^{n} (X_{ij} - X_{ik}) \right]^2$$

$$= C_H^2 - C_Q^2 = d_{jk}^2 - C_Q^2.$$

The "shape" coefficient, which is proportional to the square of Zarapkin's coefficient ($\mathfrak{S}^2$ of the previous section, identified as C_Z^2 by Penrose), represents the variance of differences between the character states of the OTU's being compared. The squared coefficient of Zarapkin is indeed nothing but the estimated variance (i.e., corrected for degrees of freedom). It is likely to be sizable when considerable discrepancy in the magnitude of the differences occurs, including a mixture of positive, negative, and negligible terms.

The C_Q^2 is identical to the correction term used in calculating the variance of differences between the character states. It represents the magnitude and direction of the differences. When C_Q^2 is large, the character states of the two OTU's being compared are quite different in magnitude, and the differences that exist are largely in one direction. A large C_Q^2 would appear, for example, if one OTU were very similar to another but much larger along most of the character scales. Thus where size is an important factor this should be revealed by the magnitude of C_Q^2.

In many studies of numerical taxonomy this partition may not be of too much significance, since C_Q^2 is not likely to be appreciable. This is so because coding is not along a uniform scale for all characters, and thus the sum of the differences is likely to be close to zero. However, when we are comparing organisms of different sizes, many of whose characters are size-conditioned, C_Q^2 is likely to yield valuable information. The magnitude of the shape coefficient will indicate whether the differences between characters are consistent or variable. Again, in much of our work we expect them to be variable. Although we still know too little about the relative importance of the size and shape factors to be dogmatic, we may note that the shape component may be the better estimate of affinity since we tend to reduce the size factor by our scoring procedures.

Rohlf and Sokal (1963) have pointed out that the shape coefficient is not a measure of similarity in proportions, as the name might imply. It is zero—indicating identity in "shape"—only if the difference between two OTU's is constant for all of the characters. These authors found that the product-moment correlation coefficient, r, is a better measure of similarity in shape between two OTU's.

If enough characters were available in a given study, one might segregate all those characters which manifest a size trend (such as skeletal dimensions in a vertebrate group) and recompute distances between OTU's on the basis of these characters only. One could then partition these distances into shape and size components, which should give some insight into the pattern of evolutionary change within the groups. It is not likely that the size component will be very large. The rules for treating empirical correlations (Section 5.3.3.6) would probably have reduced the importance of a general size factor even if many characters had shown its effects. To illustrate this problem, suppose we included in a study a number of different but approximately equally sized species of flies whose resemblances were computed from, say, 100 characters. If we add a species **A′**, which is apparently identical to a species **A** already in the study but several times as large, we shall have to add a number of new characters expressing size differences of various parts of the body. For all these new characters all the old species (including **A**) will exhibit the small state, while **A′** will show the large state. Since these new characters will therefore covary identically with each other, we must show cause, according to the procedure described in the section on empirical correlations, why more than one character should be employed. If, as we are likely to feel, a single size factor is responsible for the enlargement of the various parts of the body in species **A′**, we reduce the new characters to a single one and general size is *ab initio* an unimportant component of the distance coefficient.

6.2.4. Statistical significance of similarity coefficients

The reader may have wondered why the foregoing sections on statistics measuring similarity among OTU's have not made more than passing references to the significance of the coefficients being calculated. The computation of the separate significance of each individual similarity coefficient is not too important in numerical taxonomy. This is so for two reasons. First, we are concerned with the general significance of the similarity coefficient matrix among all the OTU's and not so much with the separate coefficients. Since we are concerned with the joint significance of the entire matrix, to omit values which are not significant, as was done for instance by Haltenorth (1937), is probably an improper procedure. Even values which individually are not significant are the best obtainable estimates of the relations between the two OTU's. This should of course not be interpreted to mean that statistical significance

is to be ignored entirely. A matrix without a single significant coefficient (or with only a few such) is clearly not worth investigation. On the other hand, if a sufficient number of significant similarity coefficients are present, it is not necessary to demonstrate the individual significance of every coefficient.

The second consideration is that the significance levels of individual similarity coefficients based on n characters are not likely to be those customarily given in textbooks of statistics. This problem has been previously referred to as the heterogeneity of column vectors. We mean by this that variables j and k representing OTU's j and k are not taken as random samples from a common population, as is required by statistical theory, but are really taken from a heterogeneous sample wherein each variate estimates a different character. This problem is somewhat alleviated by standardizing characters—that is, standardizing the rows of the data matrix. Such a procedure results in all characters having a mean of zero and a variance of one. However, there is obvious correlation among the rows or the characters. This means that when a correlation between two OTU's is based on n characters it is not really based on n independent dimensions of variation, and the number of degrees of freedom on the basis of which its significance is to be computed is likely to be less than n. Since we have as yet no way of approaching this problem, we have to put general faith in the validity of the matrices, just as persons working on multiple factor analysis, a closely related field, have had general confidence in the validity of factor loadings without being able to assign standard errors and confidence limits to their estimates. The conventional standard errors can therefore serve as guide lines toward the significance level of the similarity statistics. Type I errors, however, are likely to be greater than the standard statistical tests indicate.

6.3. SCALING AND CODING CHARACTERS

The logical basis for coding characters has already been discussed in Section 5.3. We are therefore concerned here only with the numerical and statistical consequences of the procedure adopted. The nature of the scale in which the character is coded will limit the choice of possible similarity coefficients to be adopted.

When we consider the scaling of each character, we have to distinguish between (1) phenetically discrete and (2) phenetically overlapping characters.

(1) By a *phenetically discrete character* we mean a character that does not vary appreciably within the OTU's; consequently the taxa under study can be easily grouped according to the various states for the character in question, with little or no possibility of misclassification for any given OTU. Meristic characters will often fall into this category. Where the number of antennal, palpal, or tarsal segments, the number of vertebrae of a given region, or the number of petals of a flower is constant for a certain taxon, but varies among taxa, the character may be taken to be phenetically discrete. In an entomological study, if a group of species were either apterous, micropterous, or macropterous but did not vary intraspecifically in this regard, we could again consider the character phenetically discrete. On the other hand, if, as in some Heteroptera, species occurred which were dimorphic or polymorphic with regard to their wingedness, the character could not be so considered. "Presence-absence" characters with perfect penetrance are included among phenetically discrete characters and are likely to constitute a considerable proportion of characters in botanical and zoological studies and the preponderant part in microbiological analyses.

It may be argued that phenetic discreteness does not necessarily imply genetic discontinuity. Indeed, there is good reason to believe that many stepped characters (such as polydactyly in mammals or tolerance to toxicants in insects) are caused by thresholds superimposed upon continuously varying effects (genes or gene products). However, these cases are usually recognizable by the variation in the expression of the character within a given class (taxon). If meristic characters are phenetically discrete characters, we must conclude that the variation ranges widely and is multimodal and that the thresholds are spaced in such a way as to effectively split up the distribution into nonoverlapping segments.

"Presence-absence" characters, whose states are expressed as 0 and 1, respectively, can be handled by any of the coefficients of association. If a coefficient of correlation is desired, some consideration as to the nature of the underlying variation is necessary. The tetrachoric correlation coefficient, r_t (Treloar, 1942), is to be computed when a continuous distribution of the character is assumed, with the division into two classes established as a convenience for the taxonomist; when, on the other hand, basic dichotomy is believed to exist, the fourfold point correlation coefficient ϕ appears appropriate. The two coefficients are not identical. Hence the correlation between taxa will quite properly depend in part on the assumptions behind their characters. Since it is not very likely that all characters used in a study will be subject to the same assumptions,

it may be difficult to agree on any one set of assumptions. The use of distance as a measure of relationship is possible with presence-absence data. In such a case the taxa are located at some corner of the hyperspatial cube represented by the positive manifold.

The use of phenetically discrete characters divided into more than two states will result in the substitution of the Pearson product-moment r for the correlation coefficients mentioned previously.

Of the coefficients of association so far employed for taxonomic work, only those by Rogers and Tanimoto and by Smirnov make special allowance for characters with more than two states. The probability of matches occurring in such characters becomes greatly reduced. Let us consider for the sake of simplicity a random model with an equal frequency of occurrence of each state. While in the case of two states the probability of a match for a given character (negative matches included) is $1/2$, and $(1/2)^n$ for matches on all of n characters, in the case of 3 states these probabilities are $1/3$ and $(1/3)^n$ respectively and in the general case of c states $1/c$ and $(1/c)^n$. Thus values of the association coefficients would be much reduced. There would also be no opportunity to allow for the magnitude of a mismatch in computing the coefficient. For these reasons association coefficients are not very suitable for multistate characters. In these characters negative matches as such usually do not have any meaning, unless one of the states of the character refers to its absence. It thus becomes difficult to array the data in a 2×2 table; only 2 cells can really be filled in—the number of matches (m) and the number of nonmatches (u). Hence the appropriate coefficients of association are the simple matching coefficient and that of Rogers and Tanimoto.

The distance formulas are equally applicable in the cases of phenetically separate characters of two or more than two states. It is somewhat difficult to predict the effect of dividing a character into several states as compared to a zero-or-one scale. Distances between taxa should in general decrease when a previous two-state case is recoded into more states. However, if the previous example had many matches at zero or one, which on finer classification would be shown to be short distances, the overall distance may well increase somewhat.

(2) *Phenetically overlapping characters* differ in their means from taxon to taxon but exhibit considerable variation within taxa and overlap between them. Characters such as those expressing size, color intensity, and ratios of body measurements would quite likely fall into this group. How-

ever, statistically discontinuous or meristic characters, such as segment or tooth number, may also be phenetically overlapping characters. Numerically they present no special problems in the computation of product-moment correlation coefficients and of distances. The coefficients of association discussed in this paper cannot be computed from such information directly, since matches along the scale of a continuous variable would be quite unlikely. However, with the device of grouping the means into a small number of classes, the approach suggested in the previous section can be utilized.

The employment of phenetically overlapping characters gives rise to problems of a statistical nature. Since the expression of a given character varies within an OTU, the mean used to describe the state of the character for a given taxon is merely an estimate subject to sampling error. No difficulty occurs in setting confidence limits to individual means, but the distribution and hence the validity of coefficients of resemblance based on such measures are difficult to evaluate. This problem arises constantly in physical anthropology. Estimates of distance (particularly Pearson's coefficient of racial likeness and Mahalanobis' generalized distance) therefore take the variance of the estimates into consideration. Coefficients of association do not. This whole problem is treated in greater detail in the next section.

A measure of similarity between the gene pools of freely interbreeding populations conforming to the Hardy-Weinberg equation has been proposed by Sneath (unpublished) for studying the overall similarity between the blood groups of human races. This compares not the phenotypes but the gene frequencies in the populations; the measure is

$$S.B._{1.2} = \frac{1}{L}\,(K_{\alpha 1.2} + K_{\beta 1.2} + \cdots + K_{\lambda 1.2}),$$

where $S.B._{1.2}$ is the overall similarity between populations 1 and 2, L is the number of gene loci, and K_α to K_λ are the proportion of matches at each of these L loci; the value of K for each locus is

$$K_{1.2} = (a_1a_2 + b_1b_2 + c_1c_2 + \cdots + h_1h_2),$$

where a_1 to h_1 are the gene frequencies of the different alleles, a to h, at that locus in population 1, and a_2 to h_2 are the respective gene frequencies in population 2.

This formula in effect excludes negative matches, for reasons discussed in a forthcoming paper by Sneath.

6.4. CHARACTER VARIATION WITHIN TAXA

How should we record characters if they vary within the operational taxonomic units which we employ? In the earliest studies of numerical taxonomy this was deliberately ignored, and Michener and Sokal (1957) were fortunate in not having many characters that varied within the species they studied. Wherever intraspecific variation of a given character occurred, the commonest state was chosen to represent the species. However, in much of the work to be done in the future, and particularly in taxonomic studies involving categories above the species, character variation within taxa is going to figure prominently in the analysis of the data. Whenever intrataxon variation of characters occurs, one necessarily takes the risk that these groups are not valid natural taxa, with the consequent danger that the analysis will give misleading results. In order to understand fully the implications of variation of characters within taxa, we have to digress for a moment into related fields—psychology and ecology.

Workers in psychology and ecology engaged in the computation of coefficients of correlation and of association are much concerned with the question of the reliability of single variates used in their computations. The responses of an individual to a particular psychological test repeatedly administered are rarely the same. Even in controlled situations, where learning and conditioning can be ignored, responses vary, owing to a variety of pyschological and physiological conditions mostly unknown and subsumed under the heading "individual variability" or "error." From this comes the concept of the reliability of a test, its correlation with itself when repeatedly administered. This problem occurs with most biochemical estimations, which usually have rather low accuracy, and is especially troublesome in microbiology. It occurs even when a single strain is tested. If the observational variation is sufficient to make a character seriously unreliable it should of course be omitted.

Sampling problems in ecological work are similar. Let us study, for example, the ecological associations of arthropods found in decaying tree stumps in a mixed hardwood forest. We attempt to compute coefficients of association in a Q-type study between faunas collected from individual stumps. Sampling error may arise in one or both of two ways: (a) species **X** may not be found in tree stump **A** since the sample taken from it did not contain **X** (other areas of stump **A** would have **X**, however); (b) no members of species **X** may be in the tree stump at all, although it may be ecologically quite suitable for it. In the dispersal process of the species **X**,

stump **A** has (at least so far) been missed. This does not mean that species **X** may not invade the stump at a later time.

Error (a) is a relatively straightforward problem of sampling, its magnitude depending among other factors on the density and distribution of species **X** in stump **A** and the size of the sample taken in relation to the size of the stump. Error (b) is complicated by the time dimension and requires an answer to the question of whether the absence of **X** from the fauna of stump **A** should be considered an error at all. Ecologists have debated whether an "archetypal" association including all possible species exists at all. We tend to support the more recent views (Whittaker, 1953) of associations as relatively undefined entities (stands) containing an assemblage of species with varying probabilities of occurrence. An association is therefore a cluster of stands whose species compositions do not fully overlap but which possesses centers of greater density representing the most typical form of the association. Thus absence of species **X** from stump **A** should not be thought of as an error; rather, it is an essential item of information about the latitude in species composition in the association and the cohesion of the cluster of stands. In summary, we would feel that allowance should be made for sampling errors within a stand but that we need not be concerned with "errors" in the faunal or floral composition of various stands.

We have digressed so fully into these problems of ecological research because the question of error in such analyses has analogues in numerical taxonomy. A sampling error of type (a) in numerical taxonomy might occur in the following situation; if we sample one or a few specimens from a polymorphic population, we might observe and record only one character state, while two or more were to be found in the population. An error of type (b) would consist of using a local population (or subspecies) to represent an entire species, ignoring the fact that some characters had different states in different local populations. Such errors can, of course, occur at higher levels, too. Some sampling error of type (a) is unavoidable so long as the OTU is to represent a hierarchic level higher than single individuals. Whether the error of type (b) should be called an error is dubious. As in ecology, the variation itself represents an important property of the system. In numerical taxonomy it would be a measure of the heterogeneity of the OTU under study, worth investigating in its own right. We shall now treat the subject of errors in character state coding in numerical taxonomy in a more systematic fashion. Errors in coding the phenetic value of a character in a given taxon may arise from the following four sources (apart from observational errors).

(1) If, as is not infrequently the case, a species is known and described from a single specimen, we run the danger of employing in our computation data that may not be typical or representative of it. Even in the case of phenetically discrete characters, occasional variants and mutants are bound to occur, and while it is not very likely that a single specimen taken at random from the population will show one of these, the possibility should not be neglected. With one variant per 1000 individuals and a consideration of 100 characters, a specimen picked at random has almost a 10% chance of carrying at least one variant. However, in most studies, particularly in those applications of numerical taxonomy which may be expected in the reasonable future, one would expect that a fairly representative sample of each OTU has been examined and that aberrations have been recognized as such.

(2) Cases of phenetically overlapping characters present more of a problem. Whether they are qualitative (melanic forms in a group of moths), meristic (number of antennal segments in a group of grasshoppers), or continuous (size of leaves in species of elms, rate of sugar fermentation in bacterial substrains), their means are not very representative if they have been derived from a single, reasonably homogeneous population from a limited geographical area. However, even if very complete knowledge of the variation within each taxon were available, it would still be difficult to decide how to compute a mean and its variance for every character of the taxon. One way might be weighting based on frequency in the population. If, for example, a mean for skin color in the human species is to be computed, one could multiply the various color values by the respective frequencies of these colors and thus obtain a mean for the species. It might be felt, however, that the actual frequencies of the various types at present living were not really representative of the common stock from which they presumably originated. Since we do not really know the color of the ancestral stock for *Homo sapiens* and are unlikely to know ancestral character states in most instances, such considerations are not particularly useful. An unweighted mean or midpoint of the range of variation may therefore be preferable to a weighted mean.

We are unable at this early stage in the development of numerical taxonomy to present recommendations for the various alternatives. Experience in comparing several of these will have to be gained. Two alternatives which have not yet been mentioned are (a) to omit the variable character from the analysis and (b) to employ a character state for a taxon postulated for its archetype or ancestral form. We may immedi-

ately dismiss the latter alternative by saying that it will nearly always introduce an unwarrantable element of speculation, prejudice, or vagueness. The first alternative also has disadvantages. In most instances it may not greatly matter, although many features may have to be eliminated on this account, leaving too few for good work. Yet if a feature is rare in one taxon and very common in the next and would be excluded, it could not then contribute to the dissimilarity between the taxa, as it clearly should.

(3) If a numerical taxonomic study of higher categories is to be undertaken, the problem of how to weight different types just discussed reappears in a new guise. Should a higher taxon which is to be used as an OTU be represented as the weighted mean of its constituent taxa and, if so, how should the weighting be done?

One solution to this dilemma is to introduce into the study one representative of each varying constituent of each polymorphic taxon and analyze them all together. If our notions about affinities are correct, the first clusters should represent the various polymorphic taxa; that is, the variants composing them should correlate much more closely with each other than any one of them or their common taxon does with any other taxa. However, introducing many variants adds much labor and expense to a study and may make it prohibitive.

Another solution is to use only a single representative of the polymorphic group. This would be done in the expectation that the variance of the polymorphic forms within their taxon is less than the variance among the taxa of the study. Thus the error introduced by choosing a single representative of a taxon should not be large enough to seriously affect the estimation of the similarity among the taxa of the study. If this were not so it would raise the question of the validity of the represented taxon. We have called this approach the *exemplar method.* The single representatives of the OTU's are exemplars of the taxa they represent.

Thus, to cite an example, if we were studying the relationship of *Homo sapiens* to various anthropoids, we could use a specimen from any of the races of man. The correlation of such an individual with any given anthropoid should be independent of his race, on the average, and would therefore approximate that of some hypothetical average man with the same anthropoid.

The above paragraph was written before the authors became acquainted with a paper by Zarapkin (1943) comparing hands and feet of man and three apes. In this study Zarapkin compared the deviations of single specimens of the animals and found clear distinctions at the racial,

specific, and generic levels which transcended individual variation. In a test of the exemplar method carried out by Sokal (1962b) on Smirnov's (1925) data on genera of syrphid flies, two taxa (genus groups), **A** and **B**, joined at a (coded) similarity level of 850. The average value of the similarities between members of **A** and **B** turned out to be 851, with a standard deviation of 20. The range of individual similarity coefficients is from 900 (upper value) to 813 (lower value). Thus we are able to estimate the amount of possible error involved by taking at random a member of taxon **A** and one of taxon **B** as exemplars of their respective groups. In this particular example, also, the magnitude of error is quite tolerable. Other tests of the exemplar method are currently underway.

A solution of the problem of intrataxon variation of characters will partly depend on what the investigator wishes to study. If he wishes to compare a typical mammal with a typical bird, he must himself decide what he means by typical—whether "central" or commonest. He must also take the consequences of his decision, for it may be that the commonest form is very eccentric. If he is in doubt, he will do well to use several forms to represent the taxon, using, when necessary, single specimens for this. We suggest that in general a combination of the two approaches will prove of most value.

(4) A problem peculiar to microbiology is that many biochemical tests can select for mutations. A single strain may then give different results on two occasions, depending on whether a mutation had occurred. This is likely to affect few of the characters, and it may therefore not matter which state is scored, but if a mutation is regularly observed, this fact (and the mutation rate if measured) is a perfectly valid character of the strain. Such problems in clones of higher organisms, though rare, can be treated similarly.

6.5. UNWARRANTED COMPARISONS

Up to this point we have ignored a major problem which must have occurred to most readers: there are likely to be numerous cases in which, for a certain OTU, no information is available for a particular character, making it impossible to compare this OTU with others. We may distinguish several ways in which this situation may occur.

6.5.1. Missing data

More frequently in some studies than in others, certain items of information may be unobtainable. The only available specimen may be

damaged and have some structures missing; museum regulations may prevent dissection for the study of internal characters; distributional or ecological facts may be unknown; equipment for complex chemical or physical tests may not be available. Yet in many OTU's of the study we may have information on the character in question, and the one obvious and simplest solution—the elimination of the character from consideration—would seem deplorable. Where such missing data occur, the character state should be labeled with some agreed code for "inapplicable," which should be clearly distinguishable from a minus or a zero. Sneath (1957b) labeled such cases NC ("no comparison").

6.5.2. Missing characters

Many instances will arise where a given character present in one OTU is absent in another. In most cases the zero or minus state of the character will be the appropriate code for this condition. In some cases, however, a character may be masked by another so that we cannot score it; for example, black pigment would prevent our scoring for the character "presence of yellow pigment." The latter character would then in a sense be a missing character, and this would be another form of missing data discussed above. When this occurs the character state should be labeled inapplicable.

6.5.3. Missing organs

More frequent are instances where organs or relatively major parts of the body are absent or strongly modified in a given OTU, with the result that logically subordinate characters contained within the missing part cannot be scored. If, in a study of a group of insects, we have included "presence or absence of a wing" as one character and also five wing vein characters, we cannot score the venational characters in a wingless taxon. We cannot be sure whether the difference between the two taxa involves one character or all six, since the wingless form may have maintained its wing vein genes which cannot now be expressed as such. The only consistent procedure is to consider the wing vein characters inapplicable in the wingless taxon. Thus they cannot be compared with the corresponding states in the winged form. They also cannot be compared with the wing vein characters in another wingless taxon, since the fact that they are both not manifested does not provide a basis for comparison. Thus the wingless and the winged forms could be compared only on the basis of

the single character "presence or absence of wings." We can compare the two wingless forms on the basis of the same character if we accept "negative matches" (Section 6.2.1.1). This may seem to be splitting hairs, and we must admit that it is not always easy to decide when a character is inapplicable and when it is negative or absent. However, once the decision has been made the subsequent procedure is logical and consistent. A block in a metabolic pathway poses a similar problem (Section 6.2.1.1).

6.5.4. The estimation of resemblance when some characters are inapplicable

No problem exists in the estimation of resemblance when coefficients of association are used. Any pair of character states including the code for "inapplicable" is simply excluded from the 2×2 table of matches and from the computation. This is an acceptable procedure unless too many of the characters are inapplicable, in which case the inclusion of the responsible OTU is inadvisable (see Section 6.5.5 on relevance).

When the resemblance between two OTU's is calculated by means of a correlation coefficient or a distance coefficient, characters which are inapplicable for either one are omitted in the calculation of the coefficient for the pair. It is obvious that the divisor in calculations of either coefficient has to be adjusted. In the distance coefficient the number of characters has to be reduced by the number of inapplicable comparisons, but in the correlation coefficient it is often necessary to compute separate sums of squares for the denominator of each individual r, since different characters are likely to be inapplicable in different OTU pairs. Such operations can, of course, be programmed but result in slowing down the speed of computations appreciably and also require greater storage (or repeated passes through the computer), since we have to recompute the sum of squares of each variable for each correlation coefficient rather than only once for each matrix. When possible, therefore, inapplicable characters are to be avoided. While a desk calculator operator would not be likely to be slowed down by a few inapplicable values, since he can easily scan the data and use standard procedures when records are complete, prior scanning is not feasible on most electronic computers unless their capacity is such that all the data can be stored before the outset of the computations. A computer program would therefore have to proceed with the time-consuming checking of all input for inapplicable data and the separate computation of sums of squares. When a given table of data

contains inapplicable entries for only a few characters or a few OTU's or for a few of both, we recommend the removal of the responsible rows and columns rather than processing by the slower method.

One final consideration: if the coefficients in a resemblance matrix are based upon different samples of characters, the resulting coefficients are subject to two sources of error. First, they may be different because of the qualitative differences in the characters from which the coefficients are computed. We believe that adequate sample sizes will minimize this error, basing our belief on the hypotheses of nonspecificity and the matches asymptote. A second error is statistical. Confidence limits of the coefficients will vary, and a difference which is statistically significant between two coefficients in a matrix may be nonsignificant between another two coefficients in the same matrix, based upon a smaller sample of characters. We are particularly concerned with this fact when an elaborate statistical treatment of the resemblance matrix such as factor analysis is planned. We therefore hope that the number of character state codes labeled "inapplicable" can be kept at a minimum in a given study.

6.5.5. Relevance

Cain and Harrison (1958) have introduced the useful concept of the relevance of a comparison. They define this as the ratio of "twice the number of applicable characters considered (since these are shown by both forms) to the number of inapplicable ones (each of which will be shown by only one of the forms)." This ratio has the undesirable property of being indeterminate at its upper limit, and it is also not clear whether Cain and Harrison included those characters inapplicable to both forms being compared [such as character (6) for taxa **D** and **E** in their study]. We prefer a simpler coefficient of relevance,

$$R_{jk} = \frac{a_{jk}}{n},$$

where a_{jk} is the number of characters applicable in taxon j which are also applicable in taxon k (or vice versa), and n is the number of characters employed in the study. By this formulation R_{jk} ranges from zero to unity.

We do not yet have enough experience with numerical taxonomy to know what relevance values are likely to be found in comparisons involving OTU's at various hierarchic levels. Nor do we know whether different groups of organisms would exhibit appreciable differences in

this regard. However, it is obvious that low relevance values are undesirable; minimal values of 0.7 in most studies and of 0.5 in those based on many characters would seem indicated. In studies where widely varying relevance values seem unavoidable, the values may be used to indicate the reliability of the similarity coefficient to which they are pertinent.

6.6. COMPARISON AND EVALUATION OF METHODS

Evaluations of the three major approaches to estimates of affinity must at this time remain quite tentative. Although an appreciable number of studies employing numerical taxonomy have by now been published and all three types of coefficient employed, there are only five studies known to us in which different types of coefficients are compared. These are studies by Rohlf and Sokal (1963), Rohlf (1962), Sneath (1961), Gilardi et al. (1960), and Hill et al. (1961).

Rohlf and Sokal (1963) and Rohlf (1962) suggest that a correlation coefficient should be used, rather than a distance, whenever most of the characters used in a study are measurements of various parts of an organism and the OTU's differ much in overall size. When characters are independent of size, distance coefficients seem more meaningful. Until more experience is gained in numerical taxonomy, these authors suggest that both distance and correlation be applied and that taxonomies be erected that take both into consideration.

Rohlf (1962) in his study of the 48 species of *Aedes* mosquitoes found relationships indicated by correlations and by distance to be correlated to the extent of $r = -0.52$. Rohlf decided to employ distances based on standardized characters for establishing his classification because (a) these coefficients gave higher correlation between adults and larvae in his study, and (b) the relationships indicated by distances corresponded more closely to the previous classification of the genus.

Sneath (1961) compared the mean character difference (M.C.D.) and correlation (r) in *Knightia*, which though based on few characters showed quite good concordance. Gilardi et al. (1960) and Hill et al. (1961) compared the S_J and S_{SM} values for a series of bacteria, and again there was good concordance for most values, but one or two results suggested that with S_{SM} inapplicable features were being counted as similarities in some cases.

Thus our discussion of the relative merits of the various methods must

have a tentative aspect and be based mainly upon deductive inferences of the properties of the coefficients and the procedures used to compute them. Our comments will be restricted to the three coefficients (one from each method) which we at present feel have the most justification and the best promise for further work. In association coefficients this is the simple matching coefficient (or that of Jaccard, if negative matches are to be excluded); among correlation coefficients it is Pearson's product-moment r; and among distance measures it is the coefficient of taxonomic distance, d, based on standardized characters. Association and distance are easier to interpret conceptually than correlation. We can think of an association coefficient as the proportion of agreements to be found in unit taxonomic characters between two organisms. The concept of distance in a hyperspace bounded by coordinates representing the characters of the study is self-evident (see Figure 6-2). Of the various conventional interpretations of a correlation coefficient perhaps the most applicable for numerical analysis is through its square, the coefficient of determination, which identifies the proportion of the common variance of the two OTU's. This is not an altogether satisfactory idea. While not likely to be found in practice, we can hypothesize an OTU which has the same numerical code for its state in every character. Since the variance of this OTU would be zero, its correlation with another OTU would be indeterminate. Such a case would not, however, invalidate the coefficients of association or distance.

As regards simplicity of computation the correlation coefficient is the most complicated, followed by that of distance; the association coefficient is the simplest of all. However, such considerations are probably not very important. The computations even for a small numerical taxonomic study are so tedious that they will almost inevitably be processed by electronic computers. The differences in computation time will consequently be negligible. If necessary, correlation and distance computations can be satisfactorily carried out on a desk calculator. The association coefficient involves mostly matching of two-state characters, which can conveniently be performed by inspection or with a tally counter. This method lends itself most easily to mechanical improvisations, which should simplify the counting procedure. Punched strips of paper superimposed one on top of the other, mechanical sorting of punched cards, or specially marked X-ray plates (Sneath, 1957b) are among such devices.

From the point of view of utilizing the coefficients for the classificatory procedures outlined in Chapter 7, cluster analysis can be performed on

all three of them, although the distances would first have to be transformed into a complementary function (or the cluster analysis procedure revised to pick out the lowest rather than the highest coefficients). If we wish to use factor analysis for our classificatory procedure, the coefficients must be correlations.