PERCENT SIMILARITY: THE PREDICTION OF BIAS

E. L. VENRICK[1]

ABSTRACT

An equation is developed which predicts the percent similarity index between replicate samples from an association with specified structure and heterogeneity. A second equation gives a first approximation of the variance between replicate indices. The magnitude of the expected index depends not only upon the heterogeneity of the species but also upon the number of species, their abundance, and their diversity. Because of these dependencies, care must be used in interpreting the percent similarity index.

Many community ecologists use the percent similarity index (PSI; here symbolized by I) to compare the species composition of different communities or community subsets (Whittaker and Fairbanks 1958; Miller 1970; Murdoch et al. 1972; Hicks and Tahvanainen 1974; Donaldson 1975; Haedrich et al. 1975; Silver 1975; Haedrich and Krefft 1978; Reid et al. 1978; Silver et al. 1978; Abramsky et al. 1979). This index, derived from the Bray-Curtis similarity coefficient (Boesch 1977) was proposed by Whittaker (1952) and may be expressed as

$$E(I) = \sum_{i=1}^{n} \min [E(p_{i,1})\, E(p_{i,2})]$$
$$= 1 - 0.5 \sum_{i=1}^{n} |E(p_{i,1}) - E(p_{i,2})|,$$

where I is the similarity index between two communities (1 and 2), n is the total number of species in the combined species list, and $p_{i,1}$ and $p_{i,2}$ are the proportions of species i in the two associations such that, within each association,

$$\sum_{i=1}^{n} p_{i,1} \text{ and } \sum_{i=1}^{n} p_{i,2} = 1.00.$$

A variant of this index is based upon the percent composition instead of proportions and equals $I \times 100\%$. From this variant comes the common designation "percent similarity index." The present study is developed in terms of proportions but the familiar name is retained. All conclusions in this paper are applicable to both forms of the index, although the formulae must be scaled accordingly.

The theoretical range of the percent similarity index is from 0.0 for two associations with no species in common to 1.0 for two identical associations. In actuality, a value of 1.0 is unlikely to be observed even between replicate samples of the same association[2] because species abundance fluctuations in the field, often augmented by sampling errors in the laboratory, reduce the index below 1.0. At present, the only means of estimating the magnitude of this bias is to count replicate samples within each of the two (or more) associations being compared, or to obtain the index between replicate samples by means of computer simulation. Both are time consuming. Recognition of this bias has led to the development of several different similarity indices in which certain types of bias are reduced (Morisita 1959; Lance and Williams 1966; Horn 1966; Grassle and Smith 1976; Wolda 1981). Nevertheless, the percent similarity index remains popular because of its simplicity.

The following paper develops the mathematical formulae relating the percent similarity index expected between replicate samples and its variance to the abundances of the component species and the variances and covariances of the abundance estimates. Equations are developed for the specific case of bias introduced by subsampling error in the laboratory where the magnitudes of the variances and covariances may be controlled. However, when estimates of these parameters are available for field populations, the general equations may be applicable to the estimation of I between replicate field samples. The equations not only offer a method of evaluating I, but provide insight into the influence of changes in community structure (i.e., the number of component species, and their abundances, variances, and diversity) on the bias of the similarity index.

[1]Marine Life Research Group, Scripps Institution of Oceanography, La Jolla, CA 92093.

[2]The precise definition of "association" may vary considerably from study to study. It will generally have spatial dimensions and may have a temporal dimension as well.

Manuscript accepted August 1982.
FISHERY BULLETIN: VOL. 81, NO. 2, 1983.

METHODS

The diversity index used in this paper is the standardized Shannon-Wiener index (Fager 1972):

$$H' = (H - H_{\min})/(H_{\max} - H_{\min})$$

where $H = -\sum_{i=1}^{n} p_i \ln p_i$

$$H_{\max} = \ln n$$

$$H_{\min} = \ln T - \left[\frac{(T - n + 1)}{T}\right] [\ln(T - n + 1)]$$

p_i = proportion of species i
T = total number of individuals in the sample
n = total number of species in the sample.

Use of 1 − Simpson's diversity index (Fager 1972) gave similar results.

Development of the theoretical equations for I and its variance ($s^2(I)$) was accompanied by computerized simulation modeling to examine the accuracy of the equations; values predicted by the equations were compared with those observed in the simulation studies. Two measures of accuracy were used:

$$\text{relative error} = [|\text{predicted} - \text{observed}|/\text{predicted}] \times 100\%$$

$$\text{relative bias} = [(\text{predicted} - \text{observed})/\text{predicted}] \times 100\%.$$

Species distributions sampled in the simulation studies were independent and normal. The consequences of these two assumptions are evaluated in detail in a later section. In each simulation the relationship between the mean and variance ($\sigma_i^2/\mu_i = q$) was held constant for all species in an association. This was a convenience, not a necessary condition.

To determine empirically the values of I and $s^2(I)$ for an association, 100 pairs of replicate samples were drawn; the value of I was calculated for each pair and the mean and variance were determined over the 100 pairs. These values, I and $s^2(I)$, were compared with the values $\hat{I}$ and $\hat{s}^2(I)$ estimated from the statistics observed in each sample of an independent set of 100 single samples drawn from the same association. The comparison allowed determination and correction of the bias of the predictive formulae for mean and variance and the determination of the variance of the estimate. The number of species in the association, their abundances, variances, and diversity were varied independently to examine their influence on the value of I and $s^2(I)$ and on the accuracy of the values estimated by the formulae.

To examine any errors introduced by use of the normal distribution in the simulations, a second series of simulations was run to sample species distributed independently according to a negative binomial distribution (Bliss and Fisher 1953). The negative binomial distribution is generally characterized by the parameters μ and $k = \mu^2/(\sigma^2 - \mu)$. However, an alternative parameter $q = (\mu/k) + 1 = (\sigma^2/\mu)$ is identical to the parameter q used throughout this study to express population heterogeneity. Thus, I have chosen to define negative binomial distributions by q rather than k. In these simulations, the parameters used in the formulae for the expected similarity index and its variance were not estimated from single samples but were the given parameters of the distribution.

RESULTS

Percent Similarity Index

An equation for predicting the similarity index between replicate samples from one association is

$$\hat{I} = 1 - \frac{0.5642}{\tau^2} \sum_{i=1}^{n} \{[\tau^2\sigma^2(x_i) - 2\mu_i\tau\sigma^2(x_i,T) + \mu_i^2\sigma^2(T)]\}^{1/2},$$

where n is the total number of species, μ_i and $\sigma^2(x_i)$ are the mean and variance of the estimate of abundance of the ith species, τ and σ^2 (T) are the mean and variance of the estimate of abundance of the total number of individuals, and $\sigma^2(x_i,T)$ is the covariance between x_i and T (Appendix Equation (5)).[3] The goal of this study is to estimate, from a single sample of an association, the value of $\hat{I}$ expected between replicate samples. Thus, the parameters necessary for Appendix Equation (5) must be obtained from one sample or must be independently known. The observed abundances, x_i, and T are unbiased estimators of the true mean abundances. To simplify the estimation of the variance and covariance components in the present study, two assumptions have been made: 1) The component species are independently distributed, which may be strictly true only under controlled laboratory conditions, as when a subsample is drawn

[3]These statistics must be applicable to the association represented by $\hat{I}$. Thus, if the association has a temporal dimension, this must be represented by the means and variances.

from a sample; and 2) the variance of a single species may be obtained from a predetermined relationship between the mean and the variance: $\sigma^2(x_i) \approx q\mu_i \approx qx_i$. A relationship between mean and variance has been demonstrated for phytoplankton subsampled in the laboratory (Venrick et al. 1977; Venrick 1978), although the validity of this approximation in field populations remains to be investigated.

Using these simplifying relationships and correcting for biases, Appendix Equation (5) becomes

$$\hat{I} = 1 - 0.5765(q/T^3)^{1/4} \sum_{i=1}^{n} (Tx_i - x_i^2)^{1/2}.$$

(Appendix Equation (7)).

It is evident from Appendix Equations (5) and (7) that the expected similarity index between replicate samples is a function of many of the parameters of the association: total number of species, their abundance and heterogeneity, and diversity. These relationships are interactive. The relationship between $\hat{I}$ and the number of species when T is held constant (Fig. 1) is nonlinear, with $\hat{I}$ approaching 1.0 as n approaches 1. Increasing the heterogeneity (q) or decreasing the total number of individuals (T) decreases the expected similarity and increases the dependence of $\hat{I}$ on n. When abundances of component species, rather than T, are held constant, the value of $\hat{I}$ is essentially independent of n, except at very low species numbers (Fig. 2). The relationship between $\hat{I}$ and diversity is approximately linear for values of $H' > 0.2$, the value of $\hat{I}$ decreasing as diversity increases (Fig. 3), but the slope of the relationship depends upon the other parameters. Although $\hat{I}$ is related to total abundance (T), scaling the abundance data by some factor (as when counts per sample are standardized to some different sample area or volume) does not alter the expected similarity index, since the values of T and q are automatically scaled by the same factor while $\sigma^2(x_i)$, $\sigma^2(T)$, and $\sigma^2(x_i,T)$ are scaled by the square of that factor and the factor cancels out in both Appendix Equations (5) and (7).

Variance of I

Appendix Equations (5) and (7) predict the value of $\hat{I}$ likely to be observed between replicate samples from a specified association. This is a mean value which has a variance associated with it. Unfortunately, it was not possible to calculate an exact expression for $\hat{\sigma}^2(I)$. However, in some situations the approximate equation may be useful:

$$\hat{\sigma}^2(I) = \frac{\beta q}{T^3} \sum_{i=1}^{n} (Tx_i - x_i^2)$$

where β is obtained from Figure 4 (Appendix Equation (9)).

Comparison of Appendix Equations (7) and (9) indicates that $\hat{\sigma}^2(I)$ is related to $(1 - \hat{I})$; lower similarity indices have larger associated variances. In general, the relationship between the variance of I

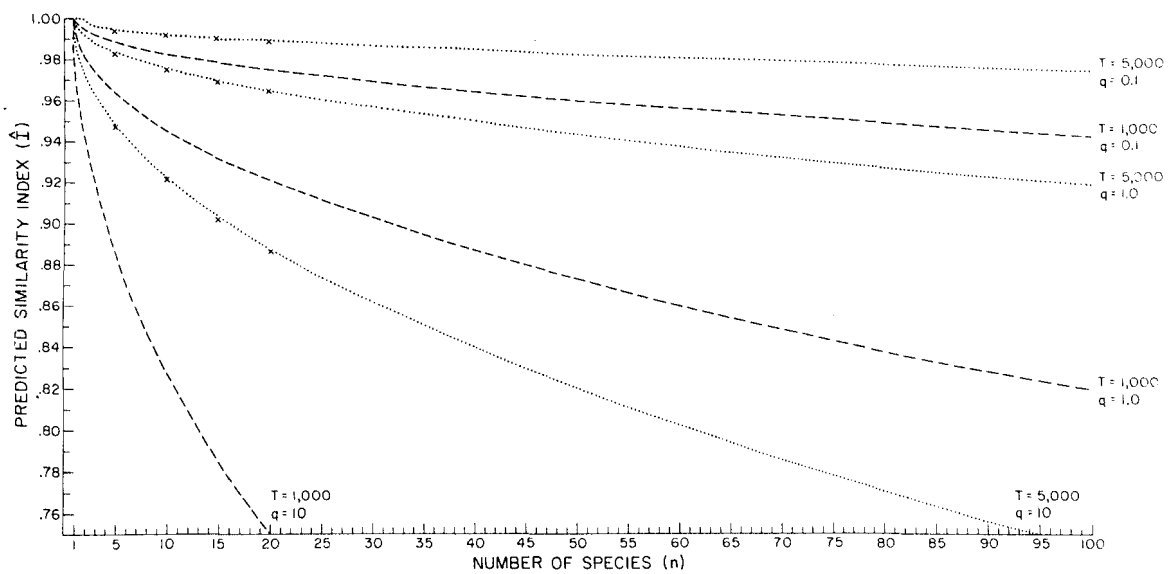

FIGURE 1.—Relationship between $\hat{I}$ and the number of species (n) for associations of different heterogeneity (q) and total number of individuals (T). In all cases, diversity (H') = 1.0. For each curve, abundance (x_i) is a constant. Curves are derived from Appendix Equation (7). X's indicate values of I observed in computer simulation and are included to indicate the accuracy of the equation.

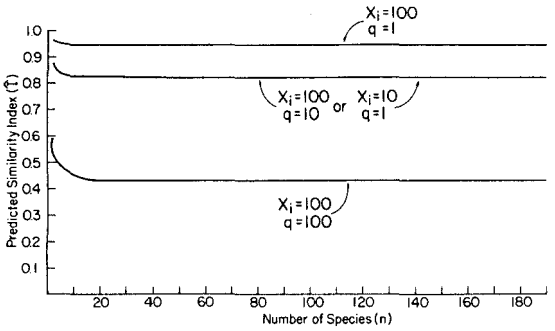

FIGURE 2.—Relationship between $\hat{I}$ and the number of species (n) for associations with different abundances (x_i) and heterogeneity (q). In all cases, diversity (H') = 1.0. For each curve, total number of individuals (T) is a constant. Curves are derived from Appendix Equation (7).

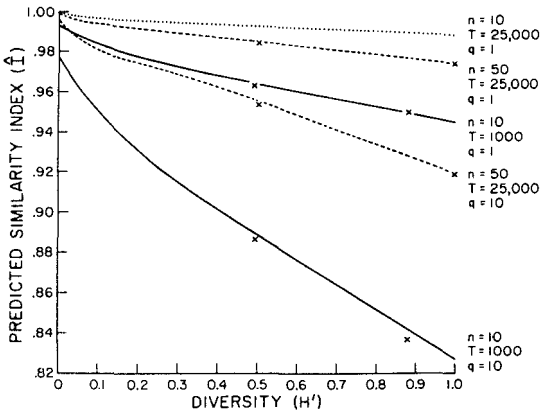

FIGURE 3.—Relationship between $\hat{I}$ and diversity (H') for associations with different numbers of species (n), total abundance (T), and heterogeneity (q). Curves are derived from Appendix Equation (7). X's indicate values of I observed in computer simulation and are included to indicate the accuracy of the equation.

and the underlying community structure is opposite in direction from that of $\hat{I}$. However, the behavior of $\hat{\sigma}^2(I)$ is mediated somewhat by the simultaneous dependence of the factor β on community structure (Figs. 4,5). Thus, although $\hat{I}$ shows a negative relationship with numbers of species, the relationship between $\hat{\sigma}^2(I)$ and n is also inverse, but much weaker (Kendall correlation, $0.05 < P < 0.10$). While $\hat{I}$ decreases continuously with increasing diversity, $\hat{\sigma}^2(I)$ increases with diversity, but stabilizes or decreases at high diversities. The dominant influence on the variance of I is the population heterogeneity, q. As evident from Appendix Equation (9), an order-of-magnitude increase in q produces an order-of-magnitude increase in $\hat{\sigma}^2(I)$.

CONSIDERATION OF ASSUMPTIONS

Two assumptions underlying this study are admittedly unrealistic and require further consideration:

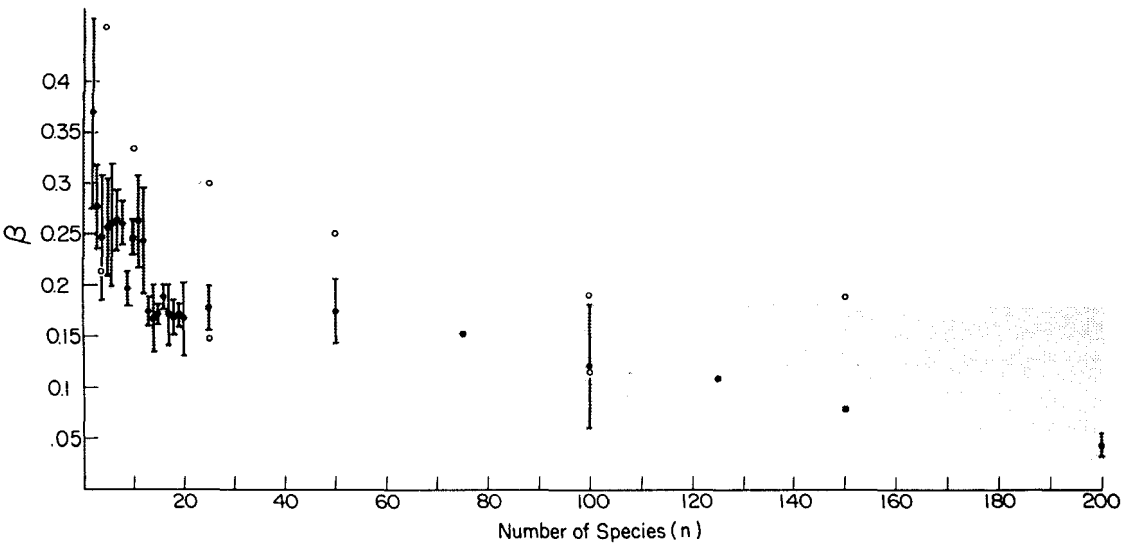

FIGURE 4.—Relationship between the value of β in Appendix Equation (9) and the number of species (n). Vertical bars are 95% confidence intervals from five estimates with diversity (H') = 1.0 and heterogeneity (q) = 0.1, 0.5, 1.0, 5.0, and 10. Dots are single estimates. Open circles are maximum values of β observed when H' varied from 0.0 to 1.0. Shaded area approximately delimits the range of observed values of β.

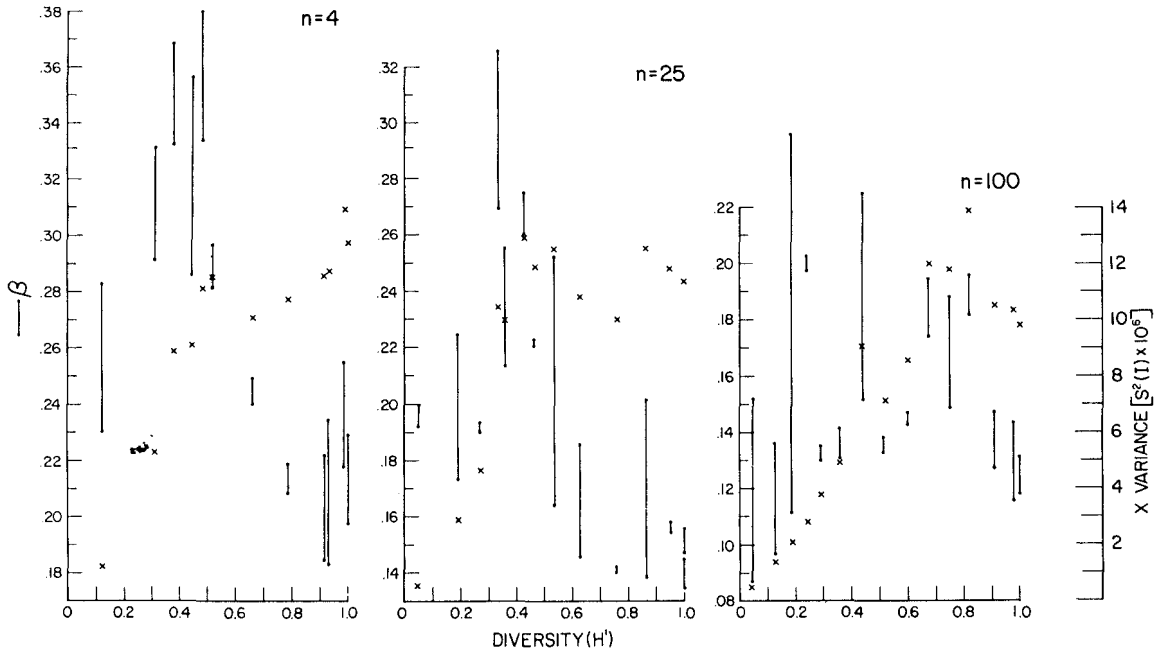

FIGURE 5.—Relationship between the value of β in Appendix Equation (9), the observed variance of I, and the diversity of the association (H'). X's represent the mean values of $s^2(I)$ observed in two or more sets of 100 replicate pairs of samples. Vertical bars represent the range of β values, each based upon single estimates of $\hat{\sigma}^2(I)$ from 100 samples, total abundance (T) = 12,500, heterogeneity (q) = 1.0.

1) The assumption of independence of species abundances may be justified in some situations, as when a sample is thoroughly mixed before subsamples are drawn, but it is probably unrealistic when applied to species in the field. However, this assumption is a convenience, not a necessity. If an independent measure of species covariance is available, the covariance between species i and the population total may be calculated and entered into Appendix Equation (5). Any positive covariance between component species increases the expected similarity index over that predicted by Appendix Equation (7) (decreasing bias). Perfect covariance between all species results in an index of 1.00. Thus, the effect of any positive covariance on the value of $\hat{I}$ is the greatest in those associations for which the expected bias is large, i.e., small samples from associations with many species, high diversity, and/or great heterogeneity.

The effect of negative covariance is less easily anticipated. For any two species, the value of $\sigma^2(x_i T)$ is decreased, lowering the value of $\hat{I}$. However, for associations of more than two species, perfect negative covariance does not exist. Large negative correlations between some species are likely to be accompanied by positive correlations between others, so that the overall effect on $\hat{I}$ may be minimal.

2) The assumption of normality of species distributions is necessitated by the use of the theoretical expected relationship between a range and a variance; however, this relationship has not been determined for other distributions. To examine the consequences of the use of the normal distribution, a final series of simulations was run to sample species distributed independently according to the negative binomial distribution which has given satisfactory fit to numerous field distributions (Bliss and Fisher 1953 and references therein; Holmes and Widrig 1956). The 39 simulations investigated values of q between 1.1 and 10. (The negative binomial is not defined at $q = 1$.) Corresponding values of k ranged between 0.44 and 900 depending upon the means and variances of the species.

In 38 of the 39 simulations, the value of I observed between replicate samples from negative binomial distributions was higher than the value predicted by Appendix Equation (7). Major deviations occur in those associations in which all species are heterogeneous and rare. In these cases, the normal distribution predicts large numbers of negative abundances, which are impossible in reality. For instance, in an association of 100 species, all with a mean abundance $\mu_i = 4$ and $q = 10$ ($k = 0.444$), the relative error of Appendix Equation (7) is 240%; in an association of 50

species, all with $\mu_i = 8$ and $q = 10.0$ ($k = 0.889$), the error drops to 25%. For the same two associations, when the heterogeneity is reduced so that $q = 1.1$ ($k = 40$ and 80, respectively), the error is reduced to 1.0 and 0.6%, respectively. This effect of rare, patchy species is less important in associations of lower diversity, dominated by a few abundant species. When such extreme associations were eliminated from consideration, the average relative error and bias were 1.6 and −1.6%, respectively, for 32 simulations. Thus, with the exception of the extreme case of small samples from a diverse, patchy association, the accuracy of Appendix Equation (7) appears to be independent of the underlying species frequency distributions. More important, the similarity index derived from negative binomial distributions shows the same relationships with the underlying community structure as does the index derived from normal distributions, decreasing either with increasing diversity, increasing numbers of species, or increasing heterogeneity (Fig. 6).

The variance between values of I from replicate samples of negative binomial distributions is satisfactorily predicted by Appendix Equation (9). In 38 of the 39 simulations, the observed variance fell within the predicted range (Fig. 7). Thus, it appears that use of the normal distribution in the present study does not restrict the applicability of the results and that the general conclusions of the paper are independent of the frequency distribution being sampled.

APPLICATIONS

An earlier study of small-scale variability of oceanic diatoms (Venrick 1972) was based upon abundances in a series of 10 samples at each of three depths in each of two environments. The 10 samples from the 10 m depth in the subarctic Pacific were selected arbitrarily to examine the performance of Appendix Equations (7) and (9). The diatom flora consisted of nine species and was strongly dominated by one ($H' = 0.23$). Although the concordance between the four dominant species was marginally significant (Kendall concordance, $P \sim 0.10$), the species were assumed to be independently distributed. The necessary parameters for the formulae ($\bar{x}_i$, $\overline{T}$, and q) were calculated from the means of the 10 samples. Observed values of q were strongly correlated with mean abundance; a single, representative value was calculated from individual q values weighted by each species' mean proportion. (Individual q values could easily have been used.) Appendix Equation (7) predicts a similarity index between field samples of 0.9101. The actual observed values, calculated between five random independent pairs of samples,

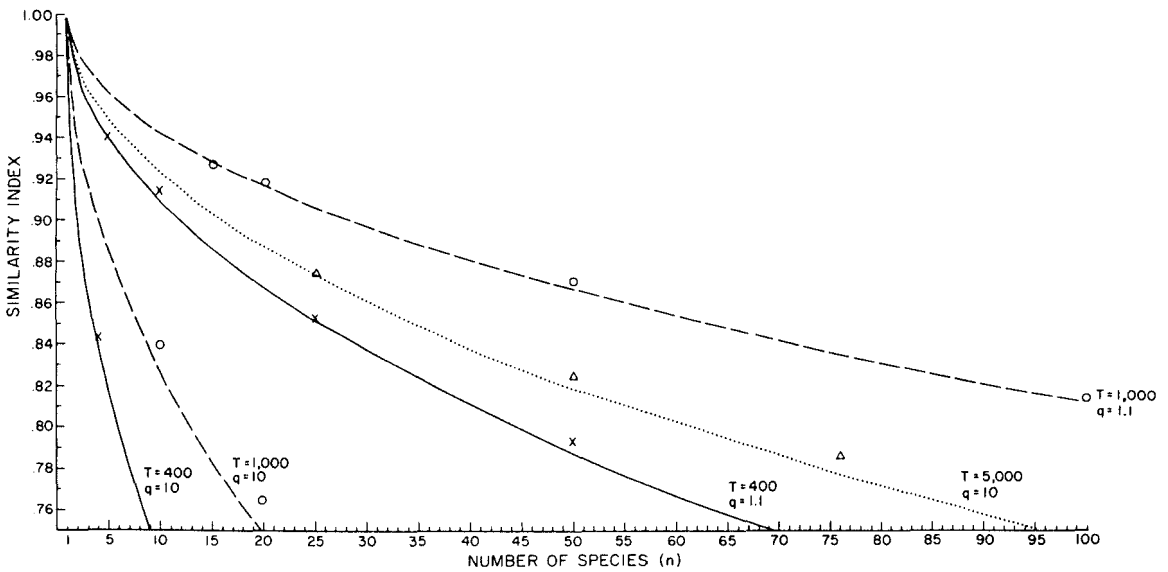

FIGURE 6.—Estimation of I from a negative binomial distribution. Curves are the value of $\hat{I}$ from appendix Equation (7) plotted against species number for five associations of different total abundance (T) and heterogeneity (q). For all associations diversity (H') = 1.0. Symbols indicate the value of I observed between replicate samples from corresponding associations of species distributed according to a negative binomial distribution. Each point is the mean of 100 replicate pairs.

range from 0.878 to 0.969 with a mean value of 0.9232. Appendix Equation (9) and Figure 4 predict a variance between replicate I values of between 1.00×10^{-3} and 2.54×10^{-3}. The observed variance is 1.48×10^{-3}.

This example is admittedly artificial; given replicate samples from the association of interest, the appropriate measure of the maximum expected similarity index is that observed between independent pairs of the replicate samples. Use of Appendix Equations (7) and (9) is unnecessary. Nevertheless, the example illustrates the accuracy of the equations when applied to field conditions, even when covariance between species is assumed to be negligible and the variances of species abundances are expressed as a simple function of the means.

McGowen and Walker (1979:211) present the percent similarity indices between samples of oceanic zooplankton. In order to estimate the bias of the index, they counted replicate aliquots of six samples and calculated the values of I between the replicates. They generously made their raw data available (five of the six samples), and the Appendix Equations (7) and (9) and Figure 4 were used to estimate the value of $\hat{I}$ expected from each single sample. A rough approximation of q between replicates was derived from a different set of 17 replicate counts of samples taken on the same cruise from the same location. Scanning the data suggested a relationship between q and the mean abundance, and the data were therefore arbitrarily divided into three categories according to abundance and separate values of q calculated for each category.

The results are presented in Table 1. The five values of I observed between the five replicate pairs of samples are compared with the 10 values and the probable ranges calculated from the equations, using the statistics observed in each sample. Only once does the observed value fall outside the estimated interval. This is good agreement (exact probability = 0.40). This is a situation in which, given some independent estimate of q, Appendix Equations (7) and (9) might have been used to estimate the magnitude of the bias in I introduced by laboratory procedures, thereby eliminating the necessity of counting replicate aliquots of single samples.

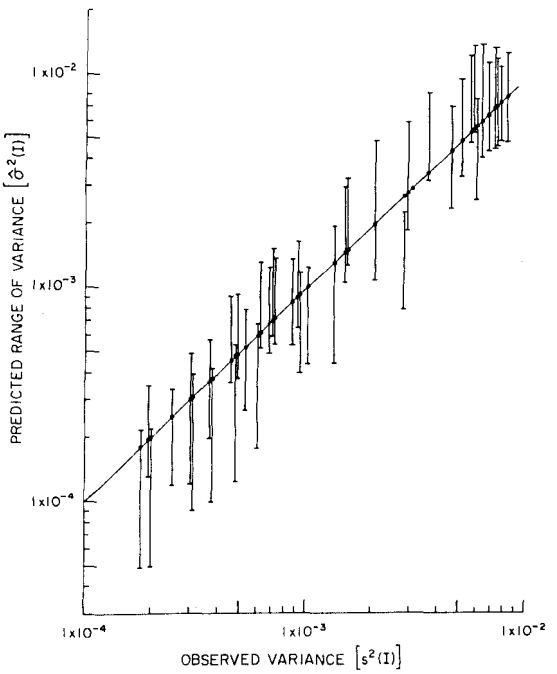

FIGURE 7.—Estimation of $\hat{\sigma}^2(I)$ from a negative binomial distribution. Vertical bars indicate the probable range of $\hat{\sigma}^2(I)$ derived from Appendix Equation (9) and Figure 4. Abscissa indicates the observed $s^2(I)$ between 100 values of I from replicate samples from associations of species distributed according to a negative binomial distribution. Values are from the simulations used for Figure 5. Diagonal line indicates values were $\hat{\sigma}^2(I) = s^2(I)$.

TABLE 1.—Similarity index (I) between counts from replicate aliquots of a single sample: observed (McGowan and Walker 1979) and expected ($\hat{I}$, Appendix Equation (7)). The probable range is based on the equation for the 95% confidence interval using a variance estimated from Appendix Equation (9) and Figure 4. Samples were collected in September 1968 near lat. 28°N, long. 155°W.

A. The Data Set

Sample no.	Depth interval (m)	Sample characteristics			
		T	H'	n	Maximum β
1a	100-225	671	0.757	58	0.24
1b		886	0.753	69	0.23
2a	225-350	847	0.686	54	0.25
2b		842	0.649	58	0.25
3a	0-25	576	0.779	35	0.28
3b		670	0.794	33	0.29
4a	25-50	844	0.643	52	0.25
4b		960	0.624	55	0.25
5a	50-75	835	0.740	59	0.24
5b		626	0.757	54	0.25

Values of q: $q = 5.6$ for $x_j \geq 100$
$q = 1.8$ for $100 > x_j \geq 10$
$q = 1.0$ for $10 > x_j$.

B. Results

Sample no.	Observed I	Predicted $\hat{I}$	Predicted $\hat{\sigma}^2(I)$ ($\times 10^{-3}$)	Probable range of I
1a	0.8472	0.8427	0.5438	0.7970-0.8884
1b		0.8367	0.6363	0.7873-0.8861
2a	0.8680	0.8628	0.6588	0.8125-0.9131
2b		0.8643	0.6742	0.8134-0.9152
3a	0.9168	0.8490	1.0592	0.7852-0.9128
3b		0.8589	0.9609	0.7981-0.9197
4a	0.8676	0.8648	0.8059	0.8092-0.9204
4b		0.8698	0.7478	0.8162-0.9234
5a	0.8701	0.8412	0.8464	0.7842-0.8982
5b		0.8304	0.9556	0.7679-0.8929

Venrick (1982) discussed data on the vertical distribution of phytoplankton samples from four stations at one location in the central Pacific. For the present study, counts from samples of 15 and 120 m depths (representing shallow and deep phytoplankton associations, respectively) were used to generate values of I between the field samples. Appendix Equations (7) and (9) were used to estimate the magnitude of $\hat{I}$ arising from laboratory subsampling error. A predetermined relationship between laboratory sampling error and mean abundance (Venrick 1982) is available from which to estimate the value of q. The parameters of each sample were used to calculate the value of $\hat{I}$ expected between replicate counts of that sample and the maximum probable range (Table 2). For the 15 m samples, one-half of the indices observed between field samples fall within the range expected from the equations. At least for these samples, it appears that differences between samples in the field may be largely attributed to handling and counting errors. For the 120 m samples, none of the observed indices fall within the expected range. At this depth there appear to be "real" differences between field samples.

The indices observed at 120 m are lower than those at 15 m. The extent to which this is due to heterogeneity of species abundances, as opposed to shifts in number of species, diversity, or total abundance, may be assessed by calculating the standardized I value:

$$\hat{I}' = I/\hat{I}$$

where I is the observed value and $\hat{I}$ is the maximum expected value calculated from Appendix Equation (7). For each observed value of I, two values of $\hat{I}$ are available, one from each sample. When two samples are similar in species content, a representative value of $\hat{I}$ may be obtained by calculating a new value of $\hat{I}$ from pooled data. This is time consuming and, when samples are dissimilar, the resultant value of $\hat{I}$ may not represent either of the original samples. In general, it seems preferable to use the mean of the individual $\hat{I}$ values.

The comparison of standardized I' values for the phytoplankton data is presented in Table 3. In five of the six cases, the I' values at 120 m are lower than the corresponding value at 15 m. This shift in I' values with depth cannot be attributed only to changes in number of species or diversity. Assuming no depth-related change in the laboratory error, this indicates an increase in the spatial or temporal variability of abundances at greater depths. In the complete analysis (Venrick 1982), the source of this heterogeneity is postulated to be vertical displacement of vertically stratified populations.

TABLE 2.—Similarity index (I) observed between replicate field samples compared with maximum expected index calculated from Appendix Equation (7). The probable range is based on the equation for 95% confidence interval using a variance estimated from Appendix Equation (9) and the largest likely β from Figure 4. All samples were collected near lat. 28°N, long. 155°W.

A. Predicted Laboratory Bias

Sample depth and no.	Date	T	H'	n	Maximum β	$\hat{I}$	$\sigma^2(\hat{I})_2$ ($\times 10^{-2}$)	Probable range
15 m:								
1	6/05/77	1,574	0.438	37	0.28	0.8975	0.1976	0.8104-0.9846
2	6/13/77	1,051	0.669	40	0.27	0.8005	0.4321	0.6716-0.9293
3	6/20/77	664	0.538	36	0.28	0.7777	0.6956	0.6142-0.9411
4	8/19/78	1,897	0.328	43	0.27	0.9055	0.1347	0.8336-0.9774
120 m:								
1	6/05/77	1,475	0.672	57	0.24	0.8586	0.1540	0.7817-0.9355
2	6/13/77	1,051	0.669	51	0.25	0.8773	0.1009	0.8150-0.9396
3	6/20/77	1,196	0.639	59	0.24	0.7840	0.4372	0.6544-0.9136
4	8/19/78	597	0.768	55	0.25	0.7625	0.3170	0.6522-0.8729

q values: $q = 0.271$ species counted in entire 265 ml sample
$q = 2.13$ species counted in 44% of the sample
$q = 41.01$ species counted in 0.9% of the sample
(from Venrick 1982)

B. Observed I between field samples. Underlined values are those within the expected range if I values from replicate counts from same sample.

Sample depth and no.	1	2	3	Sample depth and no.	1	2	3
15 m:				120 m:			
2	0.579			2	0.573		
3	0.681	0.659		3	0.487	0.339	
4	0.732	0.502	0.649	4	0.403	0.299	0.462

TABLE 3.—Comparison of two phytoplankton associations using the standardized I: $I' = I/\hat{I}$. Original values of I are given in Table 2.

Sample depth and no.	Mean $\hat{I}$			I'		
	1	2	3	1	2	3
15 m:						
2	0.849			0.682		
3	0.838	0.789		0.813	0.835	
4	0.901	0.853	0.842	0.812	0.589	0.771
120 m:						
2	0.868			0.660		
3	0.821	0.831		0.593	0.408	
4	0.811	0.820	0.773	0.497	0.365	0.598

DISCUSSION AND CONCLUSIONS

In spite of the numerous approximations and assumptions which underlie the formulae for the percent similarity index and its variance, the formulae appear to be good predictors. This is true even when the equations are applied to actual species abundances which are unlikely to fulfill all the conditions met by computer simulation (i.e., normality and independence of species distributions and accurate knowledge of heterogeneity).

An important result of this study is the elucidation of the relationship between the bias of I and such community parameters as the number of species, their abundances, heterogeneity, and diversity. Decision about the importance of these dependencies is hampered by the vagueness of the concept "similar", i.e., that which is being measured by I. In my own mind, the concept is strongly linked to differences of relative abundances, and ultimately to q. In some situations the dependency of I on factors other than heterogeneity may be desirable, or at least irrelevant, as, for instance, when I values within one set of items are compared with I values between that set and a different set. Silver (1975) calculated values of I between the diatom associations in the stomachs of several salps and compared these with the indices between salps and nearby water samples. Finding no difference, she concluded that salps are nonselective feeders. In this comparison, any differences in any of the community parameters between the first set of indices (salp-salp) and the second (salp-water) are directly related to the concept of selective feeding and are validly confounded into a similarity index. A similar situation is presented by time series of I values (e.g., Miller 1970; McGowan and Walker 1979) where all comparisons are within the same general system and temporal changes in species number or diversity are important aspects of the evolution of the system, as measured by changes in I.

On the other hand, I values from within quite different systems are occasionally compared, leading to decisions about the relative similarity of items within the systems. In a study of plants ad homoptera in fields (Murdoch et al. 1972), several fields were surveyed for plant and insect abundances. Values of I between fields were lower for plants than for insects, leading to the conclusion that "the insect assemblages on different fields are more alike than are the plants." To the extent that the observed difference could reflect only different biases of the index in the two systems (caused, for instance, by different numbers of species of plants and insects), this conclusion seems unjustified. Such a comparison between plants and insects would be validated by the use of standardized I' values to remove the contribution of species number, abundances, and diversity so that the index accurately reflects the heterogeneity of the two systems.[4]

Numerous similarity indices have been proposed with different theoretical frameworks and different attributes. Intercomparisons have given different results depending upon the conditions of the comparison and the evaluation criteria (Morisita 1959; Grassle and Smith 1976; Pielou 1979; Bloom 1981; Wolda 1981). There is little evidence to suggest that other similarity indices are independent of the underlying community structure, nor is there reason to expect the relationships to be similar to those observed for the percent similarity index. The ultimate selection of a similarity index is less important than a thorough understanding of the behavior of that index under various conditions. Without such background information, interpretation of any similarity index is subject to serious error.

ACKNOWLEDGMENTS

I am grateful to John McGowan and Patricia Walker for making available their raw zooplankton data, to discussions with Tom Hayward and Patricio Bernal which demonstrated the vague nature of my own concept of similarity and gave direction to my concluding section, and to many fellow computer users who waited patiently through my endless simulation studies.

The work was supported in part by the Marine Life Research Group of Scripps Institution of Oceanography and in part by a grant from the Office of Naval Research.

[4]Elsewhere in the paper these authors use a value of I which has been corrected for internal heterogeneity of the associations.

LITERATURE CITED

ABRAMSKY, Z., M. I. DYER, AND P. D. HARRISON
1979. Competition among small mammals in experimentally perturbed areas of the shortgrass prairie. Ecology 60:530-536.

BLISS, C. I., AND R.A. FISHER.
1953. Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. Biometrics 9:176-200.

BLOOM, S. A.
1981. Similarity indices in community studies: potential pitfalls. Mar. Ecol. Prog. Ser. 5:125-128.

BOESCH, D. F.
1977. Application of numerical classification in ecological investigations of water pollution. Va. Inst. Mar. Sci. Spec. Sci. Rep. 77.

DIXON, W. J., AND F. J. MASSEY, JR.
1969. Introduction to statistical analysis. 3d ed. McGraw-Hill, N.Y., 638 p.

DONALDSON, H. A.
1975. Vertical distribution and feeding of sergestid shrimps (Decapoda: Natantia) collected near Bermuda. Mar. Biol. (Berl.) 31:37-50.

FAGER, E. W.
1972. Diversity: a sampling study. Am. Nat. 106:293-310.

GRASSLE, J. F., AND W. SMITH.
1976. A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. Oecologia (Berl.) 25:13-22.

HAEDRICH, R. L., AND G. KREFFT.
1978. Distribution of bottom fishes in Denmark Strait and Irminger Sea. Deep-Sea Res. 25:705-720.

HAEDRICH, R. L., G. T. ROW, AND P. T. POLLONI.
1975. Zonation and faunal composition of epibenthic populations on continental slope south of New England. J. Mar. Res. 33:191-212.

HICKS, K. L., AND J. O. TAHVANAINEN.
1974. Niche differentiation by crucifer-feeding flea beetles (Coleoptera: Chrysomelidae). Am. Midl. Nat. 91:406-423.

HOLMES, R. W., AND T. M. WIDRIG.
1956. The enumeration and collection of marine phytoplankton. J. Cons. Cons. Int. Explor. Mer 22:21-32.

HORN, H. S.
1966. Measurement of "overlap" in comparative ecological studies. Am. Nat. 100:419-424.

LANCE, G. N., AND W. T. WILLIAMS.
1966. A generalized sorting strategy for computer classification. Nature (Lond.) 212:218.

MCGOWAN, J. A., AND P. W. WALKER.
1979. Structure in the copepod community of the North Pacific central gyre. Ecol. Monogr. 49:195-226.

MILLER, C. B.
1970. Some environmental consequences of vertical migration in marine zooplankton. Limnol. Oceanogr. 15:727-741.

MORISITA, M.
1959. Measuring of interspecific association and similarity between communities. Mem. Fac. Sci. Kyushu Univ., Ser. E (Biol.) 3:65-80.

MURDOCK, W. W., F. C. EVANS, AND C. H. PETERSON.
1972. Diversity and pattern in plants and insects. Ecology 53:819-829.

PIELOU, E. C.
1979. Interpretation of paleoecological similarity matrices. Paleobiology 5:435-443.

REID, F. M. H., E. STEWART, R. W. EPPLEY, AND D. GOODMAN.
1978. Spatial distribution of phytoplankton species in chlorophyll maximum layers off southern California. Limnol. Oceanogr. 23:219-226.

SEBER, G. A. F.
1973. The estimation of animal abundance, and related parameters. Hafner Press, N.Y., 506 p.

SILVER, M. W.
1975. The habitat of *Salpa fusiformis* in the California Current as defined by indicator assemblages. Limnol. Oceanogr. 20:230-237.

SILVER, M. W., A. L. SHANKS, AND J. D. TRENT.
1978. Marine snow: microplankton habitat and source of small-scale patchiness in pelagic populations. Science (Wash., D.C.) 201:371-373.

VENRICK, E. L.
1972. Small-scale distributions of oceanic diatoms. Fish. Bull., U. S. 70:363-372.
1978. The implications of subsampling. *In* A. Sournia (editor), Phytoplankton manual, p. 75-87. Monogr. Oceanogr. Methodol. 6.
1982. Phytoplankton in an oligootrophic ocean; observations and questions. Ecol. Mongr. 52:129-154.

VENRICK, E. L., J. R. BEERS, AND J. F. HEINBOKEL.
1977. Possible consequences of containing microplankton for physiological rate measurements. J. Exp. Mar. Biol. Ecol. 26:55-76.

WHITTAKER, R. H.
1952. A study of summer foliage insect communities in the Great Smoky Mountains. Ecol. Monogr. 22:1-44.

WHITTAKER, R. H., AND C. W. FAIRBANKS.
1958. A study of plankton copepod communities in the Columbia Basin, southeastern Washington. Ecology 39:46-65.

WOLDA, H.
1981. Similarity indices, sample size and diversity. Oecologia (Berl.) 50:296-302.

YATES, F.
1953. Sampling methods for censuses and surveys. 2d ed. Hafner Publ. Co., N.Y., 401 p.

APPENDIX

Derivation of Formulae for $\hat{I}$ and $\hat{\sigma}^2$ (I), the Percent Similarity Index and Its Variance

The Percent Similarity Index

General Case

In the equation defining the percentage similarity index,

$$I = 1 - 0.5 \sum_{i=1}^{n} |E(p_{i,1}) - E(p_{i,2})|, \qquad (1)$$

where n is the total number of species in samples 1 and 2, $p_{i,1}$ and $p_{i,2}$ are the proportions of species i in samples 1 and 2, and the expression $|p_{i,1} - p_{i,2}|$ is the range (w) of a sample of size two and its expected value can be related to the standard deviation of the underlying normal population by the equation $\sigma_i = 0.8862\, w_i$ (Dixon and Massey 1969, table A-8b (2)). Thus

$$E[|p_{i,1} - p_{i,2}|] = \sigma(p_i)/0.8862.$$

Substitution of this expression in Equation (1) gives

$$\hat{I} = 1 - 0.5642 \sum_{i=1}^{n} \sigma(p_i). \qquad (2)$$

The proportional abundance of species i, p_i, is the ratio of the abundance of that species, x_i (or μ_i) to the total number of individuals in the sample, T (or τ). The variability of p_i is a function of the variance of x_i and the variance of T. When variances are small relative to mean values, the variance of p_i may be approximated by

$$\hat{\sigma}^2(p_i) = [\tau^2\sigma^2(x_i) - 2\mu_i\tau\sigma^2(x_i, T) + \mu_i^2\sigma^2(T)]/\tau^4 \qquad (3)$$

and

$$\hat{\sigma}(p_i) = \{|\tau^2\sigma^2(x_i) - 2\mu_i\tau\sigma^2(x_i^2, T) + \mu_i^2\sigma^2(T)|/\tau^4\}^{1/2} \qquad (4)$$

(Yates 1953; the equation may also be derived using the differential theory of variances, or delta method, Seber 1973).

The substitution of Equation (4) into Equation (2) gives an equation for $\hat{I}$:

$$\hat{I} = 1 - \frac{0.5642}{\tau^2} \sum_{i=1}^{n} \{|\tau^2\sigma^2(x_i) - 2\mu_i\tau\sigma^2(x_i, T) + \mu_i^2\sigma^2(T)|\}^{1/2} \qquad (5)$$

Single Sample Case

In order to estimate $\hat{I}$ from a single sample, some independent method of estimating $\sigma^2(x_i)$, $\sigma^2(T)$, and $\sigma^2(x_i, T)$ must be available. In the following derivation, two assumptions are made: 1) The variance can be expressed as a function of the mean, e.g., $\sigma^2(x_i) \sim (q)(\mu_i)$; and 2) species are independently distributed so that $\sigma^2(x_i, T) = \sigma^2(x_i)$. Values of x_i and T from a single sample are unbiased estimates of μ_i and τ.

When the above approximations are introduced into Equation (4), the expression for $\hat{\sigma}(p_i)$ becomes

$$\hat{\sigma}(p_i) = [(q/T^3)(Tx_i - x_i^2)]^{1/2}$$

and

$$\sum_{i=1}^{n} \hat{\sigma}(p_i) = \sum_{i=1}^{n} [(q/T^3)(Tx_i - x_i^2)]^{1/2} \qquad (6)$$

The accuracy of Equation (6) was examined over a spectrum of values of q and T using computer simulation. Associations of 10 species with prescribed means and variances were sampled 10 times. The abundances of the species in each sample were converted to proportions and, for each species, the standard deviation of these proportions within the 10 samples was calculated. These observed standard deviations were then summed over all species to give one simulated value of $\sum_{i=1}^{n} \sigma(p_i)$. For comparison, the observed values of x_i and T from each sample and the prescribed value of q were entered into Equation (6) to give 10 predicted estimates of $\sum_{i=1}^{n} \hat{\sigma}(p_i)$. Each set of 10 samples was repeated 10 times in a run. Over 44 runs, sampling associations with a broad range of diversities and values of q from 0.1 to 50, the mean relative error and bias of the estimate of $\sum_{i=1}^{n} \hat{\sigma}(p_i)$ were 2.9 and -2.5%, respectively.

Substitution of Equation (6) into Equation (2) gives an expression for $\hat{I}$ in which all parameters may be estimated from one sample:

$$\hat{I} = 1 - 0.5642\,(q/T^3)^{1/2} \sum_{i=1}^{n} (Tx_i - x_i^2)^{1/2}.$$

The factor 0.5642 is expected to be increased somewhat by the demonstrated bias in the estimation of

$\sum_{i=1}^{n} \hat{\sigma}(p_i)$ and may also be affected by any biases resulting from approximating $|p_{i,1} - p_{i,2}|$ by $\alpha(p_i)$. Thus, the equation for $\hat{I}$ was expressed as

$$\hat{I} = 1 - \alpha(q/T^3)^{1/2} \sum_{i=1}^{n} (Tx_i - x_i^2)^{1/2},$$

and the magnitude and properties of α were investigated by computer simulation (described in Methods). In a total of 260 runs, the mean value of α was 0.5765 (95% confidence interval: 0.5751 - 0.5780). The magnitude of α appears to be independent of the number of species in the association (n varied from 5 to 200; Kendall correlation, $P > 0.20$) and their diversity (H' varied from 1.0 to 0.03; run test, $P > 0.20$). There is a relationship between the magnitude of α and the value of q (Friedman two-way ANOVA over 20 values of n and 5 values of q; $P < 0.01$). However, over the range of q values investigated, the change in the value of α is small (Appendix Table 1). For practical purposes, this correlation may be ignored and the overall mean value of α employed. Thus, the equation for estimating the percent similarity index between replicate samples becomes

$$\hat{I} = 1 - 0.5765(q/T^3)^{1/2} \sum_{i=1}^{n} (Tx_i - x_i^2)^{1/2}. \quad (7)$$

The relative error of this estimate, determined from computer simulation, is small and independent of the number of species, their abundances, and their diversity. There is a direct relationship with the square root of q, reflecting the dependence of α on q. For values of q of 0.1, 1.0, and 10, the mean relative error was 0.005, 0.022, and 0.53%, respectively.

APPENDIX TABLE 1.—The relationship between α and q (population heterogeneity). Each value α is the mean of 40 runs, with n varying between 3 and 200 and diversity varying between 0.50 and 1.00. Friedman 2-way ANOVA is significant and may indicate a linear trend. (Friedman 2-way ANOVA: $\omega = 0.0915$, $m = 40$, $n = 5$, $P \sim 0.01$.)

$q = 0.1$	0.5	1.0	5.0	10.0
$\alpha = 0.5749$	0.5763	0.5789	0.5761	0.5797

Variance of the Percent Similarity Index

A first approximation to the variance of the similarity index, like Equations (2), (5), and (7), is based upon the analogy between the absolute value of a difference and the range of a sample of size two. The expected relationship between the variance of a standard deviation estimated from a range and the variance of the population being sampled is known (Dixon and Massey 1969, table A-8b(1)):

$$\hat{\sigma}^2(0.886|p_{i,1} - p_{i,2}|) = 0.571\ \sigma^2(p_i)$$
$$\hat{\sigma}^2(|p_{i,1} - p_{i,2}|) = 0.7274\ \sigma^2(p_i).$$

An expression for the variance of the similarity index then becomes

$$\hat{\sigma}^2(I) = \sigma^2(1 - 0.5 \sum_{i=1}^{n} |p_{i,1} - p_{i,2}|)$$
$$= 0.25 \sum_{i=1}^{n} \sigma^2(|p_{i,1} - p_{i,2}|)$$
$$= 0.1818 \sum_{i=1}^{n} \sigma^2(p_i).$$

Using the delta approximation (Equation (3)) this becomes

$$\hat{\sigma}^2(I) = \frac{0.1818}{\tau^4} \sum_{i=1}^{n} [\tau^2\sigma^2(x_i) - 2\mu_i\tau\sigma^2(x_i, T) + \mu_i^2\sigma^2(T)].$$

Squaring Equation (6) and substituting gives an expression which may be used with single samples:

$$\hat{\sigma}^2(I) = 0.1818\ (q/T^3) \sum_{i=1}^{n} (Tx_i - x_i^2). \quad (8)$$

However, this equation, based on the addition of variances, assumes independence of the components which is not valid in the present case where the components are fractional parts of a sample and must sum to 1.0. The consequences of these interdependencies were investigated empirically by expressing Equation (8) as

$$\hat{\sigma}^2(I) = \frac{\beta q}{T^3} \sum_{i=1}^{n} (Tx_i - x_i^2) \quad (9)$$

and examining the effect on β of varying the underlying population parameters.

β is dependent upon the number of species (Kendall correlation, $P < 0.01$), decreasing as n increases (Fig. 4). The value is independent of T (Kendall correlation, $P > 0.20$) and, unlike α, appears independent of q (Friedman 2-way ANOVA, $P > 0.25$). The relationship of β to diversity is nonlinear and appears linked to the relationship between the variance of I and diversity (Fig. 5). At low diversities both $\hat{s}^2(I)$ and β increase as H' increases. At higher diversities, $\hat{s}^2(I)$ reaches a plateau or decreases while β decreases

more sharply. Thus, minimum values of β are associated with values of $H' < 0.25$ and $H' \sim 1.0$, while maximum values occur at some intermediate value of H', possibly influenced by the number of species.

The shaded area in text Figure 4 approximately encompasses the maximum and minimum values of β observed empirically over a broad range of H'. Much of the variability in the estimated β apparent in Figures 4 and 5 is due to errors in the empirical determination of the true variance of I; with 100 samples in each estimate, 95% confidence intervals are $0.47\,s^2 - 1.35\,s^2$. Although the value of β cannot be determined with sufficient accuracy to allow Equation (9) to be used to establish confidence intervals about predicted values of $\hat{I}$, nevertheless the figure can be used to provide conservative estimates of β so that Equation (9) may aid in decisionmaking.