Measuring Affinity of Distributions

by

Gerald van Belle and Ibrahim Ahmad

FSU Statistics Report M 255
AFOSR Technical Report No. 19

December 1973
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

This paper was presented at a Conference on Reliability and Biometry held at Florida State University, July 9 - July 27, 1973. This paper and others presented at that Conference will appear in the book Reliability and Biometry: Statistical Analysis of Life- length, edited by F. Proschan and R. J. Serfling, to be published by SIAM in the Spring of 1974.

Research for the first author supported by Training Grant Number 5T01 GM913 from the National Institute of General Medical Sciences.

Research for the second author sponsored by the Air Force Office of Scientific Research, AFSC, USAF, under Grant No. AFSOR 74-2581. The United States Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.
Measuring Affinity of Distributions

by

Gerald van Belle and Ibrahim Ahmad

The Florida State University

ABSTRACT

Morisita (Measuring of interspecific association and similarity between communities, Mem. Fac. Sci. Kyushu Univ. Ser. E. (1959), pp. 65-80) has discussed interspecific association (also called faunal affinity) and its measurement. One of his measures has become known as the Morisita-Ono index of faunal affinity and is routinely used in ecological studies; see, for example, Barnard (Benthic ecology of Bahia de San Quintin Baja California, Smithsonian Contributions to Zoology, 44 (1970), pp. 1-60) and references therein. However, very little is known about the statistical properties of this measure or of certain competitors. This paper considers two other measures as well as the Morisita-Ono index, deriving some of their statistical properties. The approach taken is that of measuring the closeness, or affinity, of two distributions as discussed by Matusita (Decision rules based on the distance, for problems of fit, two samples and estimation, Ann. Math. Statist., 26 (1955), pp. 631-640). Only the case of discrete distributions is considered. The relative merits of the three measures are discussed, and an application is made to the comparison of species distributions at two sites.
Measuring Affinity of Distributions

1. Introduction. Many biologists have studied the concept of interspecific association (also called community similarity or faunal affinity); a good review can be found in Pielou [13] and further discussion in a recent paper by Whittaker [16]. The latter, somewhat confusingly, discusses affinity of communities under the heading of β-diversity. Morisita [11] and Ono [12] have discussed the measurement of affinity and one of their measures has become routinely used and known as the Morisita-Ono index of faunal affinity; see, for example, Horn [7] and Barnard [1]. However, very little is known about the statistical properties of this measure or of certain competitors. This paper considers two other measures as well as the Morisita-Ono index, deriving a large sample theory and other statistical properties. The approach taken is that of measuring the closeness, or affinity, of two distributions, as discussed by Matusita [9]. Only the case of discrete distributions is considered.

Distance measures have played an important role in testing hypotheses about the closeness of two (or more) distributions, as, for example, the Kolmogorov distance $D = \sup_x |F_1(x) - F_2(x)|$ to test the hypothesis $H_0: F_1 = F_2 = G$, say, with corresponding test statistic

$$D_{n_1, n_2} = \sup_x |F_{1n_1}(x) - F_{2n_2}(x)|,$$

where $F_{1n_1}$ is the empirical cumu-
lative distribution function (c.d.f.), \( i = 1,2 \). Another formulation of this problem is to define a measure of the closeness of two c.d.f.'s and then to introduce a test statistic based on it. Since the goodness-of-fit problem can be reformulated as \( H_0: D = 0 \) (i.e., the distributions are the same), we shall require a measure of affinity, or closeness, to be such that \( H_0: \text{Affinity} = 1 \). The Morisita-Ono index and the measure defined by Matusita [9] have this property.

In Section 2 affinity measures are defined and interrelations among them are considered. In Section 3 the affinity measures are used to provide alternative test statistics for the hypothesis \( H_0: F_1 = F_2 \) in the case of \( F_1 \) and \( F_2 \) discrete. The limiting distributions of the proposed statistics are derived using results of Cramér [2] and Hsu [8]. An example and discussion are provided in Section 4, while the last section contains some extensions and remarks. Related work appears in Matusita [10] and references therein.

2. Definitions of affinity measures. Suppose we have two independent populations governed by c.d.f.'s \( F_1 \) and \( F_2 \) having the same support. Let \( \mu \) be a measure with respect to which \( F_1 \) and \( F_2 \) admit densities \( f_1 \) and \( f_2 \), respectively. The following measure of affinity was implicitly used by Morisita [11].

**Definition 2.1.** The \( C_\lambda \) measure of affinity is given by:

\[
C_\lambda(F_1, F_2) = \frac{2 \int f_1(x)f_2(x)d\mu(x)}{\int f_1^2(x)d\mu(x) + \int f_2^2(x)d\mu(x)},
\]

whenever the integrals are defined.
Remark. Definition 2.1 requires that $f_1$ and $f_2$ be square integrable with respect to $\mu$. If we take $\mu$ to be Lebesgue measure, and $f_1(x) = x^{-\frac{3}{2}}/2$ for $0 \leq x \leq 1$ or $f_2(x) = (3x)^{-\frac{1}{2}}/6$ for $0 \leq x \leq 1/3$, then $C_\lambda$ is not defined. Two measures of affinity which require less restrictive conditions are given below; the last measure appears to be new.

Definition 2.2. The $\rho$ measure of affinity, due to Matusita [9], is given by:

$$\rho(F_1, F_2) = \int (f_1(x) f_2(x))^{\frac{1}{2}} \, d\mu(x).$$

Definition 2.3. The $\tau$ measure of affinity is given by:

$$\tau(F_1, F_2) = 2 \int \left( \frac{1}{f_1(x)} + \frac{1}{f_2(x)} \right)^{-1} \, d\mu(x).$$

Some of the following properties of these measures follow immediately from the fact that $\rho$ is the integral of a geometric mean and $\tau$ the integral of a harmonic mean.

Properties of the affinity measures.

P(1) $0 \leq C_\lambda, \rho, \tau \leq 1$.

P(2) $C_\lambda = \rho = \tau = 1$ if and only if $f_1 = f_2$ a.e.($\mu$).

P(3) $\rho \geq \tau$.

P(4) $C_\lambda$ is not comparable with either $\tau$ or $\rho$ in that $C_\lambda$ may be $> \rho$ for one pair of distributions and $\tau > C_\lambda$ for another pair.
Remark. The quantity \( D_i = \int \hat{f}_i^2(x) d\mu(x) \), \( i = 1,2 \), which appears as a part of \( C_\lambda \), is well known to ecologists as a measure of clumping, and \( 1-D_i \), \( i = 1,2 \), is often used as a measure of the "diversity" of a population.

Now suppose that \( F_1 \) and \( F_2 \) are discrete distributions defined over the non-negative integers with associated probabilities \( p_{1j} \) and \( p_{2j} \), \( j = 0,1,2... \), respectively; thus \( 0 \leq p_{ij} \leq 1 \) and \( \sum_{j=0}^{\infty} p_{ij} = 1 \), \( i = 1,2 \). Definitions 2.1 - 2.3 reduce to:

\[
(2.1) \quad C_\lambda = 2 \sum p_{1j} p_{2j} / (\sum p_{1j}^2 + \sum p_{2j}^2),
\]

\[
(2.2) \quad \rho = \sum (p_{1j} p_{2j})^{1/2},
\]

and

\[
(2.3) \quad \tau = 2 \sum p_{1j} p_{2j} / (p_{1j} + p_{2j}),
\]

where \( \sum_{j=0}^{\infty} \).

Remark. For the discrete case the corresponding "diversity" measure is \( 1-D_i = \sum_{j=0}^{\infty} p_{ij}^2 \), \( i = 1,2 \) (see, for example, Pielou [13]). In this case \( D_i \) is the probability that two independent observations from \( F_i \) will be the same, \( i = 1,2 \). Similarly, the quantity \( \sum_{j=0}^{\infty} p_{ij} p_{2j} \) is the probability that two observations, one from each population, will take on the same value. Hence \( C_\lambda \) measures the similarity of two populations in units of the clumping of each of the two populations.
3. **Inference about affinity measures for discrete distributions.**

In many ecological applications the sample space is assumed to consist of a finite number of mutually exclusive and totally exhaustive events, labelled $1, 2, \ldots, k$, say, in which case $p_{ij} = P(X_i = j)$ for $j = 1, \ldots, k$, where $X_i$ is a random variable with c.d.f. $F_i$, $i = 1, 2$. To estimate the affinity measures, two sampling schemes are possible: sampling with replacement, and sampling without replacement. The first scheme leads to multinomial probabilities while the second gives multihypergeometric probabilities. Since many ecological applications deal with large samples from very large populations and since most of our results are asymptotic, it suffices to discuss the multinomial sampling scheme.

Let $n_i$ be the size of a random sample from $F_i$ with $n_{ij}$ realized values equal to $j$, $j = 1, \ldots, k$; $i = 1, 2$. Then

$$P(n_{i1}, \ldots, n_{ik}) = \frac{n_i!}{k \prod_{j=1}^{k} p_{ij} n_{ij}!}, \quad i = 1, 2. \tag{3.1}$$

A. **Estimating Affinity.**

The maximum likelihood estimate (M.L.E.) of $p_{ij}$ is $\hat{p}_{ij} = n_{ij}/n_i$, $j = 1, \ldots, k$; $i = 1, 2$. Hence the M.L.E.'s of $C_\lambda$, $\rho$ and $\tau$ are, respectively:
\begin{align}
(3.2) \quad \hat{C}_\lambda &= 2 \sum \hat{p}_{1j} \hat{p}_{2j} / (\hat{\rho} \hat{p}_{1j}^2 + \hat{\rho} \hat{p}_{2j}^2), \\
(3.3) \quad \hat{\rho} &= \Sigma (\hat{p}_{1j} \hat{p}_{2j})^{1/2}, \\
(3.4) \quad \hat{\tau} &= 2 \Sigma \hat{p}_{1j} \hat{p}_{2j} / (\hat{p}_{1j} + \hat{p}_{2j}), \\
\text{where} \quad \Sigma &= \Sigma \\
\text{for} \quad j = 1, \ldots, k
\end{align}

It is easy to see that these estimates are strongly consistent.

A direct application of the dominated convergence theorem (Royden [15], p.229) proves that all the above estimates are consistent in the \( r \)-th mean for any \( r > 0 \).

**Remark.** The M.L.E.'s of \( D_i = \Sigma p_{ij}^2 \) are \( \hat{D}_i = \Sigma \hat{p}_{ij}^2 \), \( i = 1,2 \); these also are strongly consistent and consistent in the mean.

**B. Tests of homogeneity using affinity measures.**

Consider the hypothesis testing problem \( H_0: F_1 = F_2 \), where \( F_1 \) and \( F_2 \) are (unknown) discrete distributions, as defined in Section 2. The null hypothesis may be rewritten as \( H_0: C_\lambda = 1 \) (or \( \rho = 1 \), or \( \tau = 1 \)). The usual test statistic is given by

\[
\chi^2 = \sum \left( \frac{n_{1j} - n_1 \hat{p}_j}{n_1 \hat{p}_j} \right)^2 + \sum \left( \frac{n_{2j} - n_2 \hat{p}_j}{n_2 \hat{p}_j} \right)^2,
\]

where \( \hat{p}_j = (n_{1j} + n_{2j}) / (n_1 + n_2) \). Rao [14], p.333, proves that under the null hypothesis \( \chi^2 \) is asymptotically \( \chi^2(k - 1) \). Implicit in Rao's proof is that \( \min(n_1,n_2) \to \infty \). The statistic can be rewritten as
\[ x^2 = \frac{n}{2} \sum \left( \frac{\hat{p}_{i1} - \hat{p}_{i2}}{\hat{p}_j} \right)^2, \]

where \( \tilde{n} \) is the harmonic mean \( 2 \frac{n_1 n_2}{n_1 + n_2} \). This expression for the \( X^2 \) statistic will be useful in the proof of Theorem 3.1, which gives the limiting distributions of the three statistics \( \hat{C}_\lambda \), \( \hat{\rho} \) and \( \hat{\tau} \) both under \( H_0 \) and when \( H_0 \) is false. The null limiting distribution of \( \hat{C}_\lambda \) is a weighted sum of chi-squares whereas \( \hat{\tau} \) and \( \hat{\rho} \) have chi-square limiting distributions under the null hypothesis, suggesting that these latter two statistics are more practical.

**Theorem 3.1.** Let \( \min (n_1, n_2) \to \infty \).

(i) If \( H_0 \) is false then \( \sqrt{n} (\hat{C}_\lambda - C_\lambda) \), \( \sqrt{n} (\hat{\rho} - \rho) \) and \( \sqrt{n} (\hat{\tau} - \tau) \) are asymptotically normal (A.N.) with zero means and variances \( \sigma^2_1 \), \( \sigma^2_2 \), and \( \sigma^2_3 \), respectively, which are given by (3.5), (3.6) and (3.7).

(ii) Under \( H_0 \), the limiting distribution of \( \tilde{n} (1 - \hat{C}_\lambda) \) is the distribution of a weighted sum of \( (k - 1) \) independent chi-squares each with one degree of freedom and with weights \( \theta_j^2 \), \( j = 2, \ldots, k \), as given in (3.8); \( 4 \tilde{n} (1 - \hat{\rho}) \) and \( 2 \tilde{n} (1 - \hat{\tau}) \) are each asymptotically \( \chi^2 (k - 1) \).

**Proof.** Some of the calculations are straightforward and will be omitted. Also note that \( \min (n_1, n_2) \to \infty \) implies \( \tilde{n} \to \infty \).
(1) The asymptotic variance of $\sqrt{n} (\hat{C}_\lambda - C_\lambda)$ is:

\[
\sigma_1^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \Sigma P_{1j} \left[ b_{1j} - (\Sigma P_{1j} b_{1j}) \right]^2,
\]

where $b_{1j} = 2 (P_{2j} - C_\lambda P_{1j})/(\Sigma P_{1j}^2 + \Sigma P_{2j}^2)$ and

$b_{2j} = 2 (P_{1j} - C_\lambda P_{2j})/(\Sigma P_{2j}^2 + \Sigma P_{1j}^2)$. It can be shown that $\sigma_1^2 = 0$

if and only if $H_o$ is true (i.e., $P_{1j} = P_{2j} = P_j$, $j = 1, \ldots, k$).

Using Theorem (ii) of Rao [14], p. 321, $\sqrt{n} (\hat{C}_\lambda - C_\lambda)$ is A.N. $(0, \sigma_1^2)$.

The asymptotic variance of $\sqrt{n} (\hat{\rho} - \rho)$ is:

\[
\sigma_2^2 = \frac{1}{2} (1 - \rho^2).
\]

Again, $\sigma_2^2 > 0$ if and only if $H_o$ is false, in which case $\sqrt{n} (\hat{\rho} - \rho)$

is A.N. $(0, \sigma_2^2)$.

Finally, the asymptotic variance of $\sqrt{n} (\hat{\tau} - \tau)$ is:

\[
\sigma_3^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} P_{1j} \left[ b_{1j}^* - (\Sigma P_{1j} b_{1j}^*) \right]^2
\]

where, $b_{1j}^* = 2[P_{2j}/(P_{1j} + P_{2j})]^2$ and $b_{2j}^* = 2[P_{1j}/(P_{1j} + P_{2j})]^2$.

Again, $\sigma_3^2 > 0$ if and only if $H_o$ is false, in which case $\sqrt{n} (\hat{\tau} - \tau)$

is A.N. $(0, \sigma_3^2)$.

(ii) We can write $1 - \hat{C}_\lambda = \Sigma (\hat{P}_{1j} - \hat{P}_{2j})^2/\Sigma (\hat{P}_{1j}^2 + \hat{P}_{2j}^2)$, which

implies that $\sqrt{n}(1 - \hat{C}_\lambda)$ has the same limiting distribution as

$$\sqrt{n} \Sigma (\hat{P}_{1j} - \hat{P}_{2j})^2/2 \Sigma P_j^2.$$ Let $Z_j = \sqrt{n} (\hat{P}_{1j} - \hat{P}_{2j})/\sqrt{2 \Sigma P_j^2}$,
\[ j = 1, \ldots, k. \text{ Then } Z = (Z_1, \ldots, Z_k) \text{ is } \text{A.N.}(Q, \Sigma), \text{ where} \]
\[ \sigma_{jj} = p_j (1 - p_j)/\Sigma p_j^2 \text{ and } \sigma_{jj'} = -p_j p_{j'}/\Sigma p_j^2 \text{ for } j \neq j'. \]

Also, \( \Sigma Z_j = 0 \). Let \( \Gamma = (\gamma_{ij}) \) be an orthogonal \((k \times k)\) matrix

with first row \((1/\sqrt{n}, \ldots, 1/\sqrt{n})\) and let \( \hat{Y} = \Gamma Z \).

Then the limiting distribution of \( \tilde{n}(1 - \hat{C}_\lambda) \) is the same as the

distribution of \( \sum_{j=2}^{k} \gamma_j^2 \), where \( \gamma \) is \text{A.N.}(Q, \Sigma^*),
\[ \Sigma^* = (\sigma_{ij}^*) (k-1) \times (k-1), \quad \sigma_{ij}^* = 0 \text{ if } i \neq j \text{ and } \sigma_{jj}^* = 1/\theta_j^2, \]

and

\[ (3.8) \quad \theta_j^2 = \frac{\sum_{j=1}^{k} \gamma_{jj} p'_{j} (1 - p_j') - \frac{1}{2} \gamma_{jj'} p_j' p_{j'}'}{\sum_{j=1}^{k} \gamma_{jj} p_j' p_{j'}}. \]

\[ j = 2, \ldots, k. \text{ Then the limiting distribution of } \tilde{n}(1 - \hat{C}_\lambda) \text{ is the} \]

same as the distribution of \( \sum_{j=2}^{k} \theta_j^2 \chi_j^2 \), i.e., a weighted sum of \( \chi_j^2 \), \( j = 2, \ldots, k \) independent chi-squares each with one degree of freedom.

Next, write \( (1 - \hat{\rho}) = \frac{1}{2} \sum (\sqrt{\hat{p}_{1j}} - \sqrt{\hat{p}_{2j}})^2 = \frac{1}{2} \sum (\hat{p}_{1j} - \hat{p}_{2j})^2 / (\sqrt{\hat{p}_{1j}} + \sqrt{\hat{p}_{2j}})^2 \). Under \( H_0 \), \( \tilde{n}(1 - \hat{\rho}) \) has the same limiting distribution as \( \frac{\tilde{n}}{8} \sum (\hat{p}_{1j} - \hat{p}_{2j})^2 / p_j \) and, by the result stated at the beginning of this section, \( 4\tilde{n}(1 - \hat{\rho}) \) has a limiting distribution which is chi-square with \((k - 1)\) degrees of freedom.
Finally, write \( 1 - \hat{\tau} = \frac{1}{2} \sum (\hat{p}_{1j} - \hat{p}_{2j})^2 / (\hat{p}_{1j} + \hat{p}_{2j}) \). Under \( H_0 \),
\( \hat{\mu} (1 - \hat{\tau}) \) has the same limiting distribution as \( \frac{\hat{\mu}}{4} \sum (\hat{p}_{1j} - \hat{p}_{2j})^2 / p_j \),
so that \( 2 \hat{\mu} (1 - \hat{\tau}) \) has a limiting distribution which is chi-square with \((k - 1)\) degrees of freedom.

Remarks. (i) On the basis of the properties of the limiting distributions, the measures \( \hat{\rho} \) and \( \hat{\tau} \) are preferred to \( \hat{\zeta} \), since the latter under \( H_0 \) has a weighted chi-square distribution, which is not tabulated. A further advantage of the measure \( \hat{\rho} \) is that the limiting variance of \( \sqrt{\hat{\mu}} (\hat{\rho} - \rho) \) is \( \frac{1}{2} (1 - \rho^2) \), which is very simple and can be simplified even further by a variance-stabilizing transformation; see, for example, Rao [14], p. 357. For \( \hat{\rho} \), a variance stabilizing transformation is \( g(\hat{\rho}) = \sqrt{2} \sin^{-1}(\hat{\rho}) \), so that
\( \sqrt{\hat{\mu}} (g(\hat{\rho}) - g(\rho)) \) is A.N. \((0,1)\). This result will be useful for comparing the degree of affinity of different pairs of distributions. Alternatively, the maximum value of the variance of \( \sqrt{\hat{\mu}} (\hat{\rho} - \rho) \) is \( \frac{1}{2} \) (i.e., when \( \rho = 0 \)) and conservative test procedures can be developed.

(ii) For completeness we give the corresponding asymptotic theory for the measures of clumping, \( D_i \), \( i = 1,2 \). For simplicity consider \( \hat{D} \) associated with a discrete distribution \( F \). Consider \( H_0: p_j = 1/k, j = 1, \ldots, k \). The following result can be
derived. If $H_0$ is false, then $\sqrt{n} (\hat{D} - D)$ is A.N. $(0, \sigma^2)$, where

$$\sigma^2 = \sum 4 p_j^3 - (2D)^2.$$  Under $H_0$, $\chi^2 _n (k - 1)$ is asymptotically

4. An Example. Hopkins [6] studied the plankton community of the St. Andrew Bay System located on the North Gulf Coast of Florida. Copepods are the most important zooplankton group and the data in Table 1 consists of the species distribution at two selected sites in the Bay System. Omitted from this tabulation are Copepod nauplii, i.e., those copepods that were in a larval stage and could not be classified. The hypothesis testing problem is to determine on the basis of the observed data whether the two sites are identical with respect to the species distribution. Table 2 contains the values of the three statistics $\hat{C}_\lambda$, $\hat{\rho}$ and $\hat{r}$ as well as related material to be discussed.

The three statistics differ little and the corresponding chi-square values for $\hat{\rho}$ and $\hat{r}$ are similar and very large. These two chi-square values, in turn, are similar to the $\chi^2$ statistic value, leading to the rejection of the null hypothesis that the two sites have identical species distributions. One reason for the very large chi-square values is the unusually large sample size: almost 25,000 observations.
Table 1. Species distribution at two sites (Hopkins [6] data).

<table>
<thead>
<tr>
<th>Copepoda</th>
<th>Site 1 ($N_1$)</th>
<th>Site 2 ($E_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acartia tonsa</td>
<td>8525</td>
<td>3489</td>
</tr>
<tr>
<td>Calanopia americana</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Centropages furcatus</td>
<td>105</td>
<td>2</td>
</tr>
<tr>
<td>Centropages hamatus</td>
<td>69</td>
<td>3</td>
</tr>
<tr>
<td>Corycaeus americanus</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Corycaeus giesbrechti</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Eucalanus pileatus</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>Euterpina acutifrons</td>
<td>161</td>
<td>6</td>
</tr>
<tr>
<td>Harpaticoids (benthic)</td>
<td>31</td>
<td>7</td>
</tr>
<tr>
<td>Hemicyclops type immatures</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Kelleria sp.</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Labidocera aestiva</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>Oithona brevicornis</td>
<td>5097</td>
<td>450</td>
</tr>
<tr>
<td>Oithona nana</td>
<td>190</td>
<td>6</td>
</tr>
<tr>
<td>Oithona simplex</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Oncaea curta</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Paracalanus crassirostris</td>
<td>5319</td>
<td>479</td>
</tr>
<tr>
<td>Paracalanus parvus</td>
<td>110</td>
<td>17</td>
</tr>
<tr>
<td>Pseudodiaptomus coronatus</td>
<td>299</td>
<td>123</td>
</tr>
<tr>
<td>Temora turbinata</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>Tortanus setacaudatus</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20105</strong></td>
<td><strong>4593</strong></td>
</tr>
</tbody>
</table>

The usual $X^2$ statistic could have been used for testing the null hypothesis that the sites are identical, but given its rejection, there is the further statistical problem of characterizing the degree of similarity between the two sites. This can be done with one of
Table 2. Summary of analysis of data of Table 1.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Estimate</th>
<th>Chi-square value under H₀ (see Thm. 3.1)</th>
<th>d.f. = k-1</th>
<th>Z = \sqrt{2}X^2 - \sqrt{2}(d.f.)-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₀</td>
<td>C₀</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>\hat{ρ} = .931</td>
<td>2074.6</td>
<td>20</td>
<td>58.2</td>
</tr>
<tr>
<td>τ</td>
<td>\hat{τ} = .870</td>
<td>1949.0</td>
<td>20</td>
<td>56.2</td>
</tr>
<tr>
<td></td>
<td>X²</td>
<td>1828.0</td>
<td>20</td>
<td>54.2</td>
</tr>
</tbody>
</table>

* As the d.f. \( \rightarrow \), \( Z = \sqrt{2}X^2 - \sqrt{2}(d.f.)-1 \) has a standard normal limiting distribution.

the three measures of faunal affinity. Using \( \hat{ρ} \) we can construct confidence limits for the degree of affinity between the two sites.

On the basis of the first remark at the end of Theorem 3.1, we know that \( \sqrt{n} [\sqrt{2} \sin^{-1}(\hat{ρ}) - \sqrt{2} \sin^{-1}(ρ)] \) is A.N. (0,1), so that approximate 95% confidence limits for \( \sin^{-1}(ρ) \) are \( \sin^{-1}(\hat{ρ}) \pm 1.96/\sqrt{2n} \). For the Hopkins data \( n = 7477.7 \), so that \( 1.96/\sqrt{2n} = .0160 \). The value of \( \sin^{-1}(0.9306) \) is 1.1946 and approximate 95% confidence limits on \( \sin^{-1}(ρ) \) are thus \( 1.1946 \pm .0160 \). This corresponds to lower and upper limits, in units of \( ρ \), of .924 and .936.

Summarizing, we conclude that the two sites are not identical - this could have been concluded using any of the three statistics \( 4n(1 - \hat{ρ}) \), \( 2n(1 - \hat{τ}) \) or \( X^2 \), each with limiting chi-square distribution under the null hypothesis. The three measures \( C₀, \hat{ρ} \) and \( \hat{τ} \) provide
estimates of the degree of similarity between the two sites, but only for $\rho$ can confidence limits be easily constructed. The observed value of $\hat{\rho} = .931$ and the corresponding approximate 95% confidence limits indicate that the two sites are very similar (but not identical) in the distribution of species.

5. Extension and concluding remarks. Matusita [10] extended the $\rho$ measure to the $c$-sample situation by defining

$$
(5.1) \quad \rho(F_1, \ldots, F_c) = \int \prod_{i=1}^{c} f_i(x)^{\frac{1}{c}} d\mu(x).
$$

In the same fashion the $C_\lambda$ and $\tau$ measures can be extended to the $c$-sample situation by defining

$$
(5.2) \quad C_\lambda(F_1, \ldots, F_c) = \frac{c \int \prod_{i=1}^{c} f_i(x) d\mu(x)}{c \sum_{i=1}^{c} \int f_i(x) d\mu(x)}
$$

and

$$
(5.3) \quad \tau(F_1, \ldots, F_c) = c \int \left[\prod_{i=1}^{c} 1/f_i(x)\right]^{-1} d\mu(x).
$$

The estimates of $C_\lambda$, $\rho$ and $\tau$ discussed in Sections 2 and 3 extend in a more or less straightforward manner.
Properties of the affinity-based goodness-of-fit tests such as unbiasedness, consistency and efficiency (relative, say, to Kolmogorov-Smirnov tests) need to be studied.

Matusita [10] studied some of the mathematical properties of $\rho(F_1, \ldots, F_c)$. For example, he showed that

$$0 \leq \rho^c(F_1, \ldots, F_c) \leq \rho^{c-1}(F_1, \ldots, F_{c-1}) \leq \ldots \leq \rho^2(F_1, F_2) \leq 1.$$ 

A similar study can be made of the extended measures $C_\lambda$ and $\tau$.

Goodman and Kruskal [3], [4], [5] have studied many measures of association for contingency tables. Some of their asymmetrical measures could also be used for measuring the affinity of discrete distributions.

Finally, the problem of extension to continuous distributions is worth studying. For example, biologists often consider species biomass (in addition to species frequency) and would be very interested in knowing whether there are comparable measures for this situation.

Acknowledgments. The authors are grateful to Mr. J. Marum, Department of Oceanography, Florida State University, for bringing this problem to their attention. One of the authors (GvB) received a grant from the Committee on Faculty Research, F.S.U., for this project. Comments of the editors led to many improvements.
REFERENCES


Measuring Affinity of Distributions

### Abstract

Morisita (Measuring of interspecific association and similarity between communities, Mem. Fac. Sci. Kyushu Univ. Ser. E. (1959), pp. 65-80) has discussed interspecific association (also called faunal affinity) and its measurement. One of his measures has become known as the Morisita-Ono index of faunal affinity and is routinely used in ecological studies; see, for example, Barnard (benthic ecology of Bahia de San Quintin Baja California, Smithsonian Contributions to Zoology, 44 (1970), pp. 1-60) and references therein. However, very little is known about the statistical properties of this measure or of certain competitors. This paper considers two other measures as well as the Morisita-Ono index, deriving some of their statistical properties. The approach taken is that of measuring the closeness, or affinity, of two distributions as discussed by Matusita (Decision rules based on the distance, for problems of fit, two samples and estimation, Ann. Math. Statist., 26 (1955), pp. 631-640). Only the case of discrete distributions is considered. The relative merits of the three measures are discussed, and an application is made to the comparison of species distributions at two sites.

### Key Words

- Faunal affinity
- Morisita-Ono index
- Closeness of distributions
- Measures of faunal affinity