Journal of Classification 2:193-218 (1985)

Journal of
Classification

Comparing Partitions

Lawrence Hubert
The University of California, Santa Barbara

Phipps Arabie
University of Illinois at Champaign

Abstract: The problem of comparing two different partitions of a finite set of objects reappears continually in the clustering literature. We begin by reviewing a well-known measure of partition correspondence often attributed to Rand (1971), discuss the issue of correcting this index for chance, and note that a recent normalization strategy developed by Morey and Agresti (1984) and adopted by others (e.g., Milligan and Cooper 1985) is based on an incorrect assumption. Then, the general problem of comparing partitions is approached indirectly by assessing the congruence of two proximity matrices using a simple cross-product measure. They are generated from corresponding partitions using various scoring rules. Special cases derivable include traditionally familiar statistics and/or ones tailored to weight certain object pairs differentially. Finally, we propose a measure based on the comparison of object triples having the advantage of a probabilistic interpretation in addition to being corrected for chance (i.e., assuming a constant value under a reasonable null hypothesis) and bounded between ± 1.

Keywords: Measures of agreement; Measures of association; Consensus indices.

1. Introduction

The problem of measuring the correspondence between two partitions of an object set has attracted substantial interest in the literature of classification (e.g., see Arabie and Boorman 1973; Lerman 1973; Rohlf 1974, 1982; Fowlkes and Mallows 1983 for reviews). We will not try to review this literature comprehensively since that task would require the length of a monograph. Instead, we will concentrate on what has been called the Rand index, along with some of its possible variations and extensions. This measure appears to be one of the most popular alternatives for

William H.E. Day was Acting Editor for the reviewing of this paper. We are grateful to him, Ove Frank, Charles Lewis, Glenn W. Milligan, Ivo Molenaar, Stanley S. Wasserman, and anonymous referees for helpful suggestions. Lynn Bilger and Tom Sharpe provided competent technical assistance. Partial support of Phipps Arabie's participation in this research was provided by NSF Grant SES 8310866 and ONR Contract N00014-83-K-0733.

Authors' Addresses: Lawrence Hubert, Graduate School of Education, The University of California, Santa Barbara, California 93106, USA. Phipps Arabie, Department of Psychology, University of Illinois, 603 E Daniel Street, Champaign, Illinois 61820, USA.

comparing partitions and has a rather interesting history of being rediscovered and/or modified by different authors (e.g., Johnson 1968; Green and Rao 1969; Mirkin 1970; Rand 1971; Brennan and Light 1974; Hartigan 1975; Frank 1976; Hubert 1977; Fowlkes and Mallows 1983; Brook and Stirling 1984; Klastorin 1985; among others).

We begin by briefly reviewing the definition of the Rand index and how it might be normalized to provide a more appropriate descriptive measure. The latter part of the paper develops a much broader class of possible comparison measures under one common framework. This structure will include the Rand proposal as a special case along with several traditional indices from the statistical literature concerned with the assessment of association in contingency tables. Finally, we propose a measure of partition correspondence based on the comparison of object triples, rather than object pairs used by the Rand index and related alternatives. Our new measure has a clear probabilistic interpretation in the sense of Goodman and Kruskal (1954), is corrected for chance with respect to the null hypothesis, and is bounded between ± 1.

2. The Rand Index

Given an n object set $S = \{O_1, \ldots, O_n\}$, suppose $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ represent two different partitions of S, i.e., the entries in U and V are subsets of S; $\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$; $u_i \cap u_{i'} = \varnothing = v_j \cap v_{j'}$ for $1 \leqslant i \neq i' \leqslant R$ and $1 \leqslant j \neq j' \leqslant C$. Letting n_{ij} denote the number of objects that are common to classes u_i and v_j, the information on class overlap between the two partitions U and V can be written in the form of a contingency table (using standard "dot" notation for row and column sums) with $n_{i\cdot}$ and $n_{\cdot j}$ referring respectively to the number of objects in classes u_i (row i) and v_j (column j), as in Table 1.

Rand (1971), as well as others cited in the introduction, bases measures of correspondence between U and V on how object pairs are classified in the $R \times C$ contingency table. Specifically, there are four different types among the $\binom{n}{2}$ distinct pairs that could be found:

type (i): objects in the pair are placed in the same class in U and in the same class in V;

type (ii): objects in the pair are placed in different classes in U and in different classes in V;

type (iii): objects in the pair are placed in different classes in U and in the same class in V;

type (iv): objects in the pair are placed in the same class in U and in different classes in V.

TABLE 1

Notation for Comparing Two Partitions

		Partition V				
	Class	v_1	v_2	...	v_C	Sums
	u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}		n_{2C}	$n_{2\cdot}$
	.	.	.		.	.
Partition U	.	.	.		.	.
	.	.	.		.	.
	u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R\cdot}$
	Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

Types (i) and (ii) are typically interpreted as agreements in the classification of the objects from a pair; types (iii) and (iv) represent disagreements. Obviously, if A represents the total number of agreements and D the total number of disagreements, then $A+D = \binom{n}{2}$. Moreover, we can show (cf. Brennan and Light 1974) that

$$A = \binom{n}{2} + \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right]$$

$$= \binom{n}{2} + 2\sum_{i=1}^{R} \sum_{j=1}^{C} \binom{n_{ij}}{2} - \left[\sum_{i=1}^{R} \binom{n_{i\cdot}}{2} + \sum_{j=1}^{C} \binom{n_{\cdot j}}{2}\right] \quad , \tag{1}$$

where a binomial coefficient $\binom{m}{2}$ is defined as 0 when $m = 0$ or 1. In fact, as given in Table 2, explicit formulae can be obtained to express the number of object pairs of each type as a function of n, $n_{i\cdot}$, $n_{\cdot j}$, and n_{ij}. If we assume that the marginal sums are fixed in the $R \times C$ contingency table, then all of the formulae in Table 2, including those given for the sums A and D, are constant linear transformations of $\sum_{i,j} n_{ij}^2$, and thus, of each other.

Intuitively, two partitions that are similar produce relatively large values of A and small values for D. Thus, depending on how A and D are normalized, different raw measures of agreement are possible, e.g., Rand (1971) uses $A/\binom{n}{2}$; Johnson (1968), Mirkin (1970), and Arabie and Boorman (1973) adopt $D/\binom{n}{2}$; and Hubert (1977) suggests $(A - D)/\binom{n}{2}$. In all

TABLE 2

Formulae for the Number of (Unordered) Object Pairs of the Four Types

Type	Formula
(i)	$\frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(n_{ij}-1)$
(ii)	$\frac{1}{2}\,(n^2 + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - (\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2))$
(iii)	$\frac{1}{2}\,(\sum_{j=1}^{C} n_{\cdot j}^2 - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2)$
(iv)	$\frac{1}{2}\,(\sum_{i=1}^{R} n_{i\cdot}^2 - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2)$
(i) + (ii) = A =	$\binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\,(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2)$
(iii) + (iv) = D =	$\frac{1}{2}\,(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2) - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2$

three of these cases, the raw measures have straightforward probabilistic interpretations with respect to picking a pair of objects at random, i.e., $A/\binom{n}{2}$ is the probability of agreement; $D/\binom{n}{2}$ is the (complementary) probability of a disagreement, and $(A - D)/\binom{n}{2}$ is the difference between the probability of an agreement and a disagreement.

More recently, Wallace (1983) in commenting on Fowlkes and Mallows' (1983) paper has suggested two other raw measures of partition correspondence. Wallace bases his two measures

$$\sum_{i,j}\binom{n_{ij}}{2} \Big/ \sum_{i}\binom{n_{i\cdot}}{2} \quad \text{and} \quad \sum_{i,j}\binom{n_{ij}}{2} \Big/ \sum_{j}\binom{n_{\cdot j}}{2}$$

on type (i) object pairs alone and argues that those of type (ii) might be considered "neutral," and thus, should not be included in a raw index of correspondence. Considering the U partition as the fixed standard against which to compare V, the first index can be interpreted as the probability of a randomly selected object pair being within the same class in V *given* that it is within the same class in U; the second index is interpreted in a similar manner with V considered as the conditionalizing standard. The symmetric measure adopted by Fowlkes and Mallows (1983) is the simple geometric mean of the two nonsymmetric Wallace indices:

$$\sum_{i,j} \binom{n_{ij}}{2} \Big/ \sqrt{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}} \quad .$$

None of the indices mentioned thus far is "corrected for chance" in the sense that the index would take on some constant value (e.g., zero) under an appropriate null model of how the partitions have been chosen. Consequently, the relative sizes for each of these raw indices are difficult to evaluate and compare since they are neither measures of departure from a common baseline nor are they normalized to lie within certain fixed bounds (e.g., 0 and 1 or ± 1). As might be expected, a variety of suggestions have been made to eliminate these latter scaling difficulties.

2.1 Corrections for Chance

Probably the most obvious (null) model for randomness assumes that the $R \times C$ contingency table is constructed from the generalized hypergeometric distribution, i.e., the U and V partitions are picked at random, subject to having the original number of classes and objects in each. Under the hypergeometric assumption, we can show (see Hubert 1977 and follow-up by Klastorin 1985; Fowlkes and Mallows 1983):

$$E\left(\sum_{i,j} \binom{n_{ij}}{2}\right) = \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \Big/ \binom{n}{2} \quad . \qquad (2)$$

That is, (2) gives the expected number of object pairs of type (i), i.e., pairs in which the objects are placed in the same class in U and in the same class in V. This latter value is the number of distinct pairs that can be constructed within rows, $\sum_i \binom{n_{i\cdot}}{2}$, times the number of distinct pairs that can be formed from columns, $\sum_j \binom{n_{\cdot j}}{2}$, divided by the total number of pairs, $\binom{n}{2}$. Obviously, an analogue exists to the expected value for a cell entry in the traditional approach to contingency tables, defined by the number of objects in its row times the number of objects in its column divided by n. Now, however, the expected number of object pairs of type (i) attributable to a particular cell is the number of pairs in its row times the number of pairs in its column divided by $\binom{n}{2}$:

$$E(\binom{n_{ij}}{2}) = \binom{n_{i\cdot}}{2} \binom{n_{\cdot j}}{2} / \binom{n}{2} \quad .$$

All the formulae in Table 2 are constant linear transformations of $\sum_{i,j} \binom{n_{ij}}{2}$; consequently, the expectations for any of these expressions can be obtained directly from the expectation in (2). For example, using that result and the latter part of (1), we find that the Rand measure has expectation:

$$E(A/\binom{n}{2}) = 1 + 2 \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}^2$$

$$- \left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} \right] / \binom{n}{2} \quad . \tag{3}$$

Thus, using the general form of an index corrected for chance:

$$\frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}} \ , \tag{4}$$

which is bounded above by 1 and takes on the value 0 when the index equals its expected value, the corrected Rand index would have the form (assuming a maximum Rand index of 1):

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}} \ . \tag{5}$$

It might be appropriate to have a well-defined lower bound as well for the index in (4), but since negative values of the index have no substantive use, the required normalization would offer no practical benefits. Moreover, it might require an unusual bound on the maximum index.

In general, the normalized expression of (3) is invariant under positive linear transformations of the original raw index. Thus, because of the fixed marginal assumption on the $R \times C$ contingency table, the selection of which of the four types of pairs listed in Table 2 is used as a base in the general form of equation (4) is more or less irrelevant since all of the options in Table 2 are linear transformations of the number of unordered object pairs, $\sum_{i,j} \binom{n_{ij}}{2}$. (The normalization given in (4) assumes that larger positive values of the raw index imply greater degrees of partition correspondence. If the reverse is true, i.e., one has raw indices based on types (iii) and (iv) object pairs, then the directionality of the normalized index must be reversed by replacing the maximum in the denominator by the minimum. In this case, however, the use of type (iii) or (iv) object pairs and a minimum index of 0 would still lead directly to (5) and a normalized index based on $\sum_{i,j} \binom{n_{ij}}{2}$.)

A more important question concerns the choice of the "maximum index" in (4) for determining the upperbound on the raw index. The one currently given in (5) is, in effect, one possible bound on the number of type (i) object pairs, but others could be used instead. Thus, the two nonsymmetric Wallace indices mentioned earlier could be corrected for chance by replacing the term $\frac{1}{2} [\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}]$ in the denominator of (5) by either $\sum_i \binom{n_{i\cdot}}{2}$ or $\sum_j \binom{n_{\cdot j}}{2}$. Or, if an alternative symmetric index were desired as a competitor to (5), the expression $\frac{1}{2} [\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}]$ could be replaced by the sharper bound of minimum $[\sum_i \binom{n_{i\cdot}}{2}, \sum_j \binom{n_{\cdot j}}{2}]$. A corrected version of the Fowlkes and Mallows measure is a little more uncertain. One version would take the geometric mean of the two corrected Wallace measures, and would thus be equivalent to using a very complicated bound on $\sum_{i,j} \binom{n_{ij}}{2}$ in (4). Alternatively, one could rely on the more direct approach of bounding $\sum_{i,j} \binom{n_{ij}}{2}$ by $\sqrt{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}$, resulting in a much simpler expression. Obviously, there are many variations that could be pursued, depending on the bound we eventually choose. Moreover, the strategies discussed thus far do not begin to exhaust all possibilities. For instance, in a later section, two other measures will be introduced using a regression argument that indirectly generates two alternative bounds on $\sum_{i,j} \binom{n_{ij}}{2}$.

It is rather unfortunate all these variations exist and that a "best" bound cannot be derived with analytic finality. Constructing an exact bound, conditional on the fixed row and column totals of the given contingency table, is a very difficult problem of combinatorial optimization.

Consequently, we will be forced to rely on the less satisfactory options entailed by basing our measures exclusively on pairs of objects.

2.2 Commentary

Some additional observations should be made before we supply a more general framework for such measures as the Rand statistic. First, there are alternatives to the null model of randomness based on the generalized hypergeometric distribution. For example, Dubien and Warde (1981) develop two alternatives based on choosing partitions at random from either the set of all partitions or only those with a certain number of object classes. Given the availability of a Monte Carlo generation of partitions (e.g., Stam 1983) and/or the (second) moments of the Rand index (even under the Dubien and Warde alternatives), significance testing can be carried out routinely. In fact, as will be seen later, the Rand measure can be considered a special case in a more comprehensive framework, and thus, all the significance testing alternatives developed in that context apply immediately (for a review, see Hubert 1983).

A second and more important point to be made concerns the correction for chance suggested by Morey and Agresti (1984) and which has been used recently by Milligan and his coworkers (e.g., see Milligan, Soon, and Sokol 1983; Milligan and Cooper 1985) in their extensive Monte Carlo comparisons of the performance of clustering strategies and associated evaluation techniques. As defined, the Morey-Agresti correction inappropriately assumes that the expectation of a squared random variable is the square of the expectation. Specifically, Morey and Agresti assert that

$$E(\sum_{i,j} n_{ij}^2) = \sum_{i,j} n_{i.}^2 n_{.j}^2 / n^2 \ , \tag{6}$$

whereas our equation (2) could be rewritten to show

$$E(\sum_{i,j} n_{ij}^2) = \sum_{i,j} n_{i.}^2 n_{.j}^2 / (n(n-1)) \ + \ n^2/(n-1)$$
$$- \ (\sum_i n_{i.}^2 + \sum_j n_{.j}^2)/(n-1) \ . \tag{7}$$

In general, (7) is larger than (6) and the positive difference of

$$\frac{1}{n^2\,(n-1)} \ (n^2 - \sum_i n_{i.}^2)\ (n^2 - \sum_j n_{.j}^2)$$

is not necessarily small, depending on the sizes of the object sets and associated partitions being compared. As an example, we use the Morey-Agresti application taken from Rand (1971), shown in Table 3.

TABLE 3

Data Used by Morey and Agresti (1984) from Rand (1971)

		Partition V			
		B_1	B_2	B_3	
Partition U	A_1	2	1	0	3
	A_2	0	2	1	3
		2	3	1	

The Rand index, $A/\binom{n}{2}$, has a value of 9/15, and thus, the adjusted measure according to Morey and Agresti (1984, p. 36) should be

$$\frac{\frac{9}{15} - \frac{6}{15}}{1 - \frac{6}{15}} = \frac{1}{3} ,$$

using 6/15 as the expected value of $A/\binom{n}{2}$.

Our correction would give

$$\frac{\frac{9}{15} - \frac{8\frac{1}{5}}{15}}{1 - \frac{8\frac{1}{5}}{15}} = \frac{2}{17} ,$$

based on the expected value of $\frac{8\ 1/5}{15}$ calculated from (3). At present, it is unclear what effect the underadjustment for chance implied by the use of (6) might have had on the extensive Monte Carlo studies conducted by Milligan and his coworkers.

3. A More General Context

Although the Rand index and some of its simple transformations might be the most popular measures for assessing the correspondence between two partitions, we can also embed these indices in more all-inclusive schemes. These generalizations, in turn, lead directly to a number of alternative measures that could supplant the Rand statistic, including some based on traditional assessment of association in contingency tables. This framework facilitates for many special cases a comprehensive strategy for significance tests as well as a method of normalizing the indices.

3.0.1. An Alternative Development of the Rand Index

The easiest way to introduce this type of all-inclusive scheme is to provide an alternative derivation of the Rand index. Although this approach ostensibly requires more complicated notation, generalizations are readily forthcoming once this special case is available.

As we commented earlier, the Rand index is a constant linear transformation of $\sum_{i,j}\binom{n_{ij}}{2}$. In turn,

$$2\sum_{i,j}\binom{n_{ij}}{2} = \sum_{i,j} n_{ij}(n_{ij}-1) \quad ,$$

where the latter expression is the number of *ordered* object pairs of type (i). For convenience, it is this term that we will construct from an alternative notational scheme.

First, define two $n \times n$ (partitioned) binary matrices, **P** and **Q**, based on the categorization of the n objects in $S = \{O_1, \ldots, O_n\}$ according to U and V respectively as in Table 4. Both **P** and **Q** are assumed to have zeroes along their main diagonals with the indicated ones and zeroes defining all the entries in the corresponding submatrices. The rows and columns of **P** (**Q**) are partitioned according to the row (column) sums of the original contingency table; thus, there are $n_{i.}$ ($n_{.j}$) rows bracketed by the label A_i (B_j) in **P** (**Q**) for $1 \leqslant i \leqslant R$ ($1 \leqslant j \leqslant C$).

For notational simplicity, we will assume that the n objects in S are indexed according to the partition defined by U. Thus, $O_1, \ldots, O_{n_{1.}}$ belong to A_1; $O_{n_{1.}+1}, \ldots, O_{n_2}$ belong to A_2 ;...; $O_{n_{(R-1).}+1}, \ldots, O_{n_{R.}}$ belong to A_R. Obviously, the n row and column objects of **Q** are the same as the row and column objects of **P** but reordered to be consistent with the partition represented by V. In particular, we let $\psi_0(\cdot)$ denote the permutation on the first n integers that matches identical objects between the rows (and columns) of **P** and **Q**, i.e., if $\psi_0(r) = t$, then the *r-th* row (and column) in **P**, which corresponds to object O_r, is actually the *t-th* row (and column) in **Q**.

TABLE 4

Partitioned Matrices **P** and **Q**

$$
\mathbf{P} = \begin{array}{c|c|c|c|c|}
 & u_1 & u_2 & \cdots & u_R \\
\hline
u_1 & \begin{smallmatrix}0 & & \mathbf{1}\\ & \ddots & \\ \mathbf{1} & & 0\end{smallmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\
\hline
u_2 & \mathbf{0} & \begin{smallmatrix}0 & & \mathbf{1}\\ & \ddots & \\ \mathbf{1} & & 0\end{smallmatrix} & \cdots & \mathbf{0} \\
\hline
\vdots & \vdots & \vdots & & \vdots \\
\hline
u_R & \mathbf{0} & \mathbf{0} & \cdots & \begin{smallmatrix}0 & & \mathbf{1}\\ & \ddots & \\ \mathbf{1} & & 0\end{smallmatrix} \\
\end{array}
$$

$$
\mathbf{Q} = \begin{array}{c|c|c|c|c|}
 & v_1 & v_2 & \cdots & v_C \\
\hline
v_1 & \begin{smallmatrix}0 & & \mathbf{1}\\ & \ddots & \\ \mathbf{1} & & 0\end{smallmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\
\hline
v_2 & \mathbf{0} & \begin{smallmatrix}0 & & \mathbf{1}\\ & \ddots & \\ \mathbf{1} & & 0\end{smallmatrix} & \cdots & \mathbf{0} \\
\hline
\vdots & \vdots & \vdots & & \vdots \\
\hline
v_C & \mathbf{0} & \mathbf{0} & \cdots & \begin{smallmatrix}0 & & \mathbf{1}\\ & \ddots & \\ \mathbf{1} & & 0\end{smallmatrix} \\
\end{array}
$$

Given this framework, consider a measure of correspondence between the two matrices **P** and **Q** of the cross-product form

$$\Gamma(\psi_0) = \sum_{r,s} p_{rs}\, q_{\psi_0(r)\psi_0(s)} \quad , \tag{8}$$

which numerically equals

$$\sum_{i,j} n_{ij}(n_{ij}-1) \quad . \tag{9}$$

Intuitively, $\psi_0(\cdot)$ defines the observed contingency table (i.e., the *i-th* row and *j-th* column entry of the contingency table); n_{ij} specifies how many row (column) objects from the *i-th* class, A_i, of U are mapped to the *j-th* class, B_j, of V. Because all entries in **P** and **Q** are 0-1, the cross-product index in (8) merely counts the number of ordered object pairs placed in the same class in U *and* in the same class in V.

3.0.2. A Regression-Based Measure of Correspondence Between U and V

Although the use of matrices **P** and **Q** has been directed toward obtaining the number of ordered object pairs of type (i) in (9), it is of some interest to note that this redefinition by itself suggests another strategy for evaluating the correspondence between the two partitions U and V, based on predicting one matrix from the other. For example, considering U as the standard, suppose we predict the entries in **Q** from those in **P** using least-squares (i.e., $q_{\psi_0(r)\psi_0(s)}$ is predicted from p_{rs}). The regression coefficient obtained, say b_{QP}, can be written as

$$b_{\mathrm{QP}} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2} \Big/ \binom{n}{2}}{\sum_i\binom{n_{i\cdot}}{2}\left(1 - \dfrac{\sum_i\binom{n_{i\cdot}}{2}}{\binom{n}{2}}\right)} \quad .$$

Alternatively, treating V as the standard and predicting **P** from **Q**, b_{PQ} could be given in the same form with $\sum_j\binom{n_{\cdot j}}{2}$ replacing $\sum_i\binom{n_{i\cdot}}{2}$ in the denominator of b_{QP}. In both cases, the resulting indices are automatically corrected for chance, bounded above by 1, and can be rewritten to conform to the general structure of our index corrected for chance given in (4):

$$b_{QP} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2} / \binom{n}{2}}{\sum_{i}\binom{n_{i\cdot}}{2}\left(1 + \frac{\sum_{j}\binom{n_{\cdot j}}{2} - \sum_{i}\binom{n_{i\cdot}}{2}}{\binom{n}{2}}\right) - \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2} / \binom{n}{2}} ;$$

$$b_{PQ} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2} / \binom{n}{2}}{\sum_{j}\binom{n_{\cdot j}}{2}\left(1 + \frac{\sum_{i}\binom{n_{i\cdot}}{2} - \sum_{j}\binom{n_{\cdot j}}{2}}{\binom{n}{2}}\right) - \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2} / \binom{n}{2}} .$$

The two bounds on $\sum_{i,j}\binom{n_{ij}}{2}$ used in the latter representations for b_{QP} and b_{PQ}, respectively, are

$$\sum_{i}\binom{n_{i\cdot}}{2}\left(1 + \frac{\sum_{j}\binom{n_{\cdot j}}{2} - \sum_{i}\binom{n_{i\cdot}}{2}}{\binom{n}{2}}\right) ,$$

and

$$\sum_{j}\binom{n_{\cdot j}}{2}\left(1 + \frac{\sum_{i}\binom{n_{i\cdot}}{2} - \sum_{j}\binom{n_{\cdot j}}{2}}{\binom{n}{2}}\right) .$$

To obtain a symmetric measure, one option is to take the geometric mean of the two regression coefficients, thus yielding the conventional correlation between the off-diagonal entries of **P** and **Q**. Notice that this latter measure would also be the geometric mean of the two nonsymmetric Wallace measures that have been corrected for chance even though the two constituent components are slightly different.

TABLE 5

Weighting Schemes for Object Pairs in Matrices **P** and **Q**

$$
\mathbf{P} = \begin{array}{c|c|c|c|c|}
 & u_1 & u_2 & \cdots & u_R \\
\hline
u_1 & \begin{smallmatrix} 0 & \bar{\mathbf{p}}_{11} \\ \bar{\mathbf{p}}_{11} & 0 \end{smallmatrix} & \bar{\mathbf{p}}_{12} & \cdots & \bar{\mathbf{p}}_{1R} \\
\hline
u_2 & \bar{\mathbf{p}}_{21} & \begin{smallmatrix} 0 & \bar{\mathbf{p}}_{22} \\ \bar{\mathbf{p}}_{22} & 0 \end{smallmatrix} & \cdots & \bar{\mathbf{p}}_{2R} \\
\hline
\vdots & \vdots & \vdots & & \vdots \\
\hline
u_R & \bar{\mathbf{p}}_{R1} & \bar{\mathbf{p}}_{R2} & \cdots & \begin{smallmatrix} 0 & \bar{\mathbf{p}}_{RR} \\ \bar{\mathbf{p}}_{RR} & 0 \end{smallmatrix}
\end{array}
$$

$$
\mathbf{Q} = \begin{array}{c|c|c|c|c|}
 & v_1 & v_2 & \cdots & v_C \\
\hline
v_1 & \begin{smallmatrix} 0 & \bar{\mathbf{q}}_{11} \\ \bar{\mathbf{q}}_{11} & 0 \end{smallmatrix} & \bar{\mathbf{q}}_{12} & \cdots & \bar{\mathbf{q}}_{1C} \\
\hline
v_2 & \bar{\mathbf{q}}_{21} & \begin{smallmatrix} 0 & \bar{\mathbf{q}}_{22} \\ \bar{\mathbf{q}}_{22} & 0 \end{smallmatrix} & \cdots & \bar{\mathbf{q}}_{2C} \\
\hline
\vdots & \vdots & \vdots & & \vdots \\
\hline
v_C & \bar{\mathbf{q}}_{C1} & \bar{\mathbf{q}}_{C2} & \cdots & \begin{smallmatrix} 0 & \bar{\mathbf{q}}_{CC} \\ \bar{\mathbf{q}}_{CC} & 0 \end{smallmatrix}
\end{array}
$$

3.1 Generalizing P and Q

Although the use of the matrices **P** and **Q** and the index in (8) is a notationally cumbersome way of finding the number of ordered pairs of type (i), there is considerable generality inherent in this alternative development. In particular, the 0-1 nature of **P** and **Q** that we began with may be extended to allow for differential weighting functions in the classification of all object pairs by defining **P** and **Q** to have the generic forms given in Table 5.

Initially, the weights $\bar{p}_{ii'}$ for $1 \leqslant i, i' \leqslant R$ and $\bar{q}_{jj'}$ for $1 \leqslant j, j' \leqslant C$ are arbitrary but are assumed to be constant within the indicated submatrices (the on-diagonal submatrices of **P** and **Q** indicated in Table 5 are obtained when $i = i'$ and $j = j'$; the off-diagonal submatrices correspond to $i \neq i'$ and $j \neq j'$). The index in (8),

$$\Gamma(\psi_0) = \sum_{r,s} p_{rs}\, q_{\psi_0(r)\psi_0(s)} \quad ,$$

now assumes the reduced form:

$$\Gamma(\psi_0) = \sum_{i,i'} \sum_{j,j'} n_{ij} n_{i'j'} \bar{p}_{ii'} \bar{q}_{jj'} - \sum_{i,j} n_{ij} \bar{p}_{ii} \bar{q}_{jj} \quad . \tag{10}$$

Alternatively, we can express (10) as four additive components corresponding to weighted contributions from the four types of object pairs listed earlier, i.e.,

$$\begin{aligned}
\Gamma(\psi_0) = {} & \sum_{i,j} n_{ij}(n_{ij}-1)\bar{p}_{ii}\bar{q}_{jj} && \text{(type (i))} \\
& + \sum_{i \neq i'} \sum_{j \neq j'} n_{ij} n_{i'j'} \bar{p}_{ii'} \bar{q}_{jj'} && \text{(type (ii))} \\
& + \sum_{i \neq i'} \sum_{j} n_{ij} n_{i'j} \bar{p}_{ii'} \bar{q}_{jj} && \text{(type (iii))} \\
& + \sum_{j \neq j'} \sum_{i} n_{ij} n_{ij'} \bar{p}_{ii} \bar{q}_{jj'} \,. && \text{(type (iv))}
\end{aligned}$$

Obviously, the raw statistic in (9) defining the number of ordered object pairs of type (i) is obtained from (10) merely by letting $\bar{p}_{ii'} = 1$ for $i = i'$ and 0 otherwise; and $\bar{q}_{jj'} = 1$ for $j = j'$ and 0 otherwise.

Although we might wish to generalize this structure even further by relaxing the cross-product restriction on the measure of correspondence between the matrices **P** and **Q** (e.g., see Hubert 1979 for a discussion of four-argument functions between the entries in **P** and **Q**), the form of (10) by itself still allows extensive flexibility for defining a number of alternative indices to assess the correspondence between the partitions U and V. For example, suppose $\bar{p}_{ii'}$ and $\bar{q}_{jj'}$ are zero unless $i = i'$ and $j = j'$ in which case they can be variously defined as below. Then, $\Gamma(\psi_0)$ reduces to an additive composite of type (i) ordered object pairs from each cell of the $R \times C$ contingency table, with a weighting scheme that may differ from cell to cell, i.e.,

$$\Gamma(\psi_0) = \sum_{i,j} n_{ij}(n_{ij}-1)\bar{p}_{ii}\bar{q}_{jj} \tag{11}$$

As pointed out recently by Brook and Stirling (1984), but reinterpreted in the present context, the use of unit weights allows each cell in the contingency table to contribute in direct relation to the number of object pairs it contains; thus, those classes in U and V with the larger marginal frequencies tend to dominate the raw index. If we wish to attenuate this effect, a different weighting function could be developed using some reciprocal function of the marginal totals. For instance, since the expected number of ordered object pairs in the cell formed by the *i-th* row and *j-th* column is $n_{i\cdot}(n_{i\cdot}-1)\,n_{\cdot j}(n_{\cdot j}-1)/(n(n-1))$, one possible raw index might be

$$\sum_{i,j} n_{ij}(n_{ij}-1) \Big/ \frac{n_{i\cdot}(n_{i\cdot}-1)\,n_{\cdot j}(n_{\cdot j}-1)}{n(n-1)} \tag{12}$$

Here, the contribution from each cell is taken relative to the number of pairs expected rather than to the absolute number of ordered object pairs generated from the cell irrespective of the corresponding marginal frequencies. This latter index may be obtained from the general scheme in (11) merely by letting

$$\bar{p}_{ii} = \sqrt{n(n-1)}\,/(n_{i\cdot}(n_{i\cdot}-1)) \quad ,$$

and

$$\bar{q}_{jj} = \sqrt{n(n-1)}\,/(n_{\cdot j}(n_{\cdot j}-1)) \quad .$$

As can be derived from a formula given in section 3.3 of this paper, the expectation of (12) is RC. Thus, given an upper bound on the raw index, such as

$$\frac{1}{2}\left[\sum_i \frac{n(n-1)}{n_{i\cdot}(n_{i\cdot}-1)} + \sum_j \frac{n(n-1)}{n_{\cdot j}(n_{\cdot j}-1)}\right] ,$$

an adjusted measure could be constructed as an analogue of (5):

$$\frac{\sum_{i,j} n_{ij}(n_{ij}-1) \Big/ \dfrac{n_{i\cdot}(n_{i\cdot}-1)\,n_{\cdot j}(n_{\cdot j}-1)}{n(n-1)} - RC}{\dfrac{1}{2}\,[\sum_i \dfrac{n(n-1)}{n_{i\cdot}(n_{i\cdot}-1)} + \sum_j \dfrac{n(n-1)}{n_{\cdot j}(n_{\cdot j}-1)}] - RC} .$$

As a somewhat different line of generalization leading to indices based on conventional tests for association in contingency tables, suppose only weighted object pairs of type (iv) are considered. Specifically, let $\bar{p}_{ii'}$ be zero unless $i = i'$, and conversely, let $\bar{q}_{jj'}$ be zero unless $j \neq j'$. Then

$$\Gamma(\psi_0) = \sum_{j \neq j'} \sum_i n_{ij}\, n_{ij'} \bar{p}_{ii} \bar{q}_{jj'} \quad . \tag{13}$$

As one example, if we define

$$\bar{p}_{ii} = 1/n_{i\cdot} \quad \text{and} \quad \bar{q}_{jj'} = 1/n_{\cdot j} \quad \text{when} \quad j \neq j' , \tag{14}$$

then $\Gamma(\psi_0)$ reduces to

$$C - \sum_{i,j} \frac{n_{ij}^2}{n_{i\cdot}\, n_{\cdot j}} ,$$

which is a constant linear transformation of the traditional chi-square statistic for evaluating association in an $R \times C$ contingency table, i.e.,

$$\chi^2 = n\,(\sum_{i,j} \frac{n_{ij}^2}{n_{i\cdot}\, n_{\cdot j}} - 1) ,$$

and thus,

$$\chi^2 = -n\,\Gamma(\psi_0) + n\,(C-1). \tag{15}$$

(Starting with type (iii) pairs, a similar equivalence results with C replaced by R.) Or, if we let **P** be defined as above and let

$$\bar{q}_{jj'} = \begin{cases} 1 & \text{when } j \neq j' \ ; \\ 0 & \text{otherwise} \ , \end{cases} \tag{16}$$

then the pairs of type (iv) are weighted differently to produce

$$\Gamma(\psi_0) = n - \sum_{i,j} \frac{n_{ij}^2}{n_{i\cdot}} \ ,$$

which is a constant linear transformation of an alternative index for measuring association in a contingency table, called the Goodman-Kruskal (1954) τ_b statistic. Specifically,

$$\tau_b = \frac{n \sum_{i,j} \frac{n_{ij}^2}{n_{i\cdot}} - \sum_j n_{\cdot j}^2}{n^2 - \sum_j n_{\cdot j}^2} \ ,$$

so that

$$\tau_b = \left(\frac{-n}{n^2 - \sum_j n_{\cdot j}^2} \right) \Gamma(\psi_0) + 1 \ . \tag{17}$$

Using different notation, this last equivalence between τ_b and the comparison of two matrices, **P** and **Q**, has been pointed out recently by Berry and Mielke (1985).

As these few illustrations suggest, a very wide variety of special cases of the index in (10) are possible. However, before proceeding further with these interpretations of the matrix comparison framework, we should discuss the general problems of significance testing and normalization of indices, irrespective of the weighting scheme chosen.

3.2 Significance Testing

The contingency table defined by the two partitions U and V was characterized by a permutation (or matching), ψ_0, of the row/column objects of **P** and **Q**. If we consider an arbitrary permutation ψ to effect such a matching and select ψ at random from all $n!$ possibilities, we generate a sampling distribution for the index $\Gamma(\cdot)$ based on the generalized hypergeometric (e.g., see Fowlkes and Mallows 1983). Each permutation ψ induces a contingency table on our original n objects where each table has the same row and column totals as the original. Stated differently, we produce a distribution for the index $\Gamma(\cdot)$ by picking two partitions of S at random that have the same number of classes and class sizes as U and V.

The first three moments of $\Gamma(\psi)$ have already been derived (e.g., see Mielke 1979), and thus, significance testing of the observed measure $\Gamma(\psi_0)$ can be approached by approximating the distribution of $\Gamma(\psi)$ by another, e.g., the Pearson Type III as suggested by Mielke, Berry, and Brier (1981). For contingency tables that are associated with large values of n, the direct use of Mielke's formulae, which operate on the $n \times n$ matrices $\mathbf{P}$ and $\mathbf{Q}$, may prove to be a computationally burdensome task. One option is to simplify expressions for the first three moments for the particular weighting options chosen, e.g., for weighted measures of type (i) or (iii). For the Rand and chi-square options, these formulae are available in Brook and Stirling (1984) and Mielke and Berry (1985), respectively. Alternatively, Monte Carlo significance testing can be used based on a random sample of all $n!$ permutations, or again, as a computationally more efficient strategy, the given Monte Carlo distribution could be generated directly from randomly selected $R \times C$ contingency tables (by, say, the method of Patefield 1981). For a comparison of the moment approximation and Monte Carlo methods, the reader is referred to Costanzo, Hubert, and Golledge (1983). Also, Brook and Stirling (1984), who rederived Mielke's (1979) third moment formula for the special case of Rand's measure, demonstrated through simulation the inadequacy of the normal approximation based on just the first two moments when the row or column categories have nearly an equal number of objects. Earlier, Mielke (1979) derived this result analytically and concluded that significance testing through a distributional approximation should be based on at least the third moment.

3.3 Normalization of Indices

For normalization of an index corrected for chance, it is necessary to have an explicit formula for the expectation of $\Gamma(\psi)$. Given the general partitioned structure of $\mathbf{P}$ and $\mathbf{Q}$ presented earlier, we can show that

$$E(\Gamma(\psi)) = \frac{1}{n(n-1)} [\sum_{i,i'} n_{i\cdot} n_{i'\cdot} \bar{p}_{ii'} - \sum_{i} n_{i\cdot} \bar{p}_{ii}] [\sum_{j,j'} n_{\cdot j} n_{\cdot j'} \bar{q}_{jj'} - \sum_{j} n_{\cdot j} \bar{q}_{jj}] \ .$$

Thus, assuming that large values of $\Gamma(\psi)$ should be keyed positively (e.g., indices based on weighted type (i) and (ii) object pairs), this expectation can be used in the general form of a normalized index:

$$\frac{\Gamma(\psi_0) - E(\Gamma(\psi))}{\max_{\psi} \Gamma(\psi) - E(\Gamma(\psi))} .$$

Typically, $\max_{\psi} \Gamma(\psi)$ is replaced by some upper bound since calculation of an exact solution is a very difficult computational task (e.g., see Garey and Johnson 1979). If small values of $\Gamma(\psi_0)$ are of interest (e.g., indices based on weighted type (iii) and (iv) object pairs) and are to be keyed positively, the $\max_{\psi} \Gamma(\psi)$ term is replaced by $\min_{\psi} \Gamma(\psi)$ (or, more usually, a lower bound):

$$\frac{E(\Gamma(\psi)) - \Gamma(\psi_0)}{E(\Gamma(\psi)) - \min_{\psi} \Gamma(\psi)} .$$

For example, in our two special cases of weighted type (iv) object pairs using the general index in (13), the specification in (14) leads to an expectation of $(n-R)\,(C-1)\,/\,(n-1)$ and the alternative in (16) leads to $(n-R)\,(n^2-\sum_j n_{.j}^2)\,/\,(n(n-1))$. If the bound on $\min_{\psi} \Gamma(\psi)$ is assumed zero, then normalized indices of the form given above are immediate. These latter measures are linear combinations of the usual χ^2 and τ_b statistics, respectively, which in turn have expectations of $n(R-1)(C-1)/(n-1)$ and $(R-1)/(n-1)$.

4. Comparison Measures Based on Object Triples

The measures discussed thus far for comparing two partitions have been based on the classification of all object pairs into one of four types. As a parallel development, it is possible to subdivide all $n(n-1)(n-2)$ ordered triples into separate categories depending on how the three distinct objects are classified by U and V. Although a wide variety of correspondence indices based on these counts could again be developed, our sole concern will be with concordant and discordant object triples. The latter can be defined using the 0-1 matrices **P** and **Q** representing U and V and the matching $\psi_0(\cdot)$ between identical row/column objects in **P** and **Q**:

Letting

$$\text{sign}\,(x) = \begin{cases} 1 & \text{for } x>0 \;; \\ 0 & \text{for } x=0 \;; \\ -1 & \text{for } x<0 \;, \end{cases}$$

a *concordant* triple of ordered objects $(0_r, 0_s, 0_t)$ is one for which $\text{sign}\,(p_{rs} - p_{rt})\,\text{sign}\,(q_{\psi_0(r)\psi_0(s)} - q_{\psi_0(r)\psi_0(t)}) = +1$; a *discordant* triple is one for which sign $(p_{rs} - p_{rt})$ sign $(q_{\psi_0(r)\psi_0(s)} - q_{\psi_0(r)\psi_0(t)}) = -1$. Consequently, a concordant triple has the property that either $p_{rs} = q_{\psi_0(r)\psi_0(s)} = 1$ and $p_{rt} = q_{\psi_0(r)\psi_0(t)} = 0$, or $p_{rs} = q_{\psi_0(r)\psi_0(s)} = 0$ and $p_{rt} = q_{\psi_0(r)\psi_0(t)} = 1$ (i.e., characterized as pairs, $(0_r, 0_s)$ is of type (i) *and* $(0_r, 0_t)$ is of type (ii), or conversely). A discordant triple is characterized by either $p_{rs} = q_{\psi_0(r)\psi_0(t)} = 1$ and $p_{rt} = q_{\psi_0(r)\psi_0(s)} = 0$, or $p_{rs} = q_{\psi_0(r)\psi_0(t)} = 0$ and $p_{rt} = q_{\psi_0(r)\psi_0(t)} = 1$ (i.e., $(0_r, 0_s)$ is of type (iv) *and* $(0_r, 0_t)$ is of type (iii), or conversely). If we consider the 0-1 matrices **P** and **Q** as providing order information in comparing object pairs, where +1 indicates a similar pair and 0 indicates a dissimilar pair, a concordant (or discordant) triple $(0_r, 0_s, 0_t)$ is one in which the order information for the pairs $(0_r, 0_s)$ and $(0_r, 0_t)$ is consistent (or inconsistent). For triples that are neither concordant or discordant, the values for p_{rs} and p_{rt} and/or $q_{\psi_0(r)\psi_0(s)}$ and $q_{\psi_0(r)\psi_0(t)}$ are tied at 0 or 1 and it is impossible to make an unambiguous determination of either consistency or inconsistency.

This development is directly analogous to assessing association between two 0-1 numerical sequences using, say, Kendall's τ or Goodman-Kruskal's γ. The numerators of both measures are based on the number of consistent pairs minus the number of inconsistent pairs where a consistency (or inconsistency) depends on whether there is the same (or different) order for a pair of objects in the two sequences. When the constituent objects in a pair are tied within either sequence, that occurrence is ignored since any clear determination of order consistency or inconsistency is impossible. In our definition for proximity matrices, triples are ignored when the two constituent pairs $(0_r, 0_s)$ and $(0_r, 0_t)$ are tied in **P** or in **Q** or both. Extending the analogy further, the Rand index is based on the number of exact matches between **P** and **Q**, which would correspond in the 0-1 numerical sequence framework to the use of the number of objects that are coded by a one in both sequences. This latter value, in effect, is the usual Fisher's exact test statistic in a 2×2 contingency table.

By analogy to rank correlation between two matched numerical sequences using concordant and discordant object pairs, one possible raw measure of correspondence between U and V could be given as Con − Dis, where Con and Dis represent the number of concordant and discordant triples, respectively:

$$\text{Con} - \text{Dis} = \sum_{\substack{r,s,t \\ \textit{distinct}}} \text{sign}\,(p_{rs} - p_{rt})\ \text{sign}\,(q_{\psi_0(r)\psi_0(s)} - q_{\psi_0(r)}\ q_{\psi_0(t)}) \quad . \tag{18}$$

Alternatively, Con − Dis can be obtained directly from the cell frequencies within the $R \times C$ contingency table:

$$\text{Con} - \text{Dis} = 2\left[(n-1)\sum_{i,j} n_{ij}(n_{ij}-1) - \sum_{i,j}(n_{i.}-1)(n_{.j}-1)n_{ij}\right]\ . \qquad (19)$$

We should emphasize that this formula is recommended for computational purposes, in contrast to the more transparent and general expressions given earlier. Recasting this last expression may clarify the relation to Rand's adjusted index. In particular, (19) can be given as

$$4(n-1)\left\{\sum_{i,j}\left[\binom{n_{ij}}{2} - \frac{\binom{n_{i.}}{2}\binom{n_{.j}}{2}}{\binom{n}{2}}\left[\frac{n_{ij}}{\frac{n_{i.}n_{.j}}{n}}\right]\right]\right\}, \qquad (20)$$

and the numerator of the adjusted Rand index in (5) as

$$\sum_{i,j}\left[\binom{n_{ij}}{2} - \frac{\binom{n_{i.}}{2}\binom{n_{.j}}{2}}{\binom{n}{2}}\right] \qquad (21)$$

Ignoring the multiplicative constant of $4(n-1)$ in (20), the difference between (20) and (21) resides in the multiplicative term $\left[\frac{n_{ij}}{\frac{n_{i.}n_{.j}}{n}}\right]$ applied to the expected number of object pairs in (20). This expression is greater or less than 1 depending on whether the observed cell frequency is greater or less than $n_{i.}n_{.j}/n$, the expected number of observations in the cell defined by A_i and B_j under the hypergeometric model. Thus, compared to the Rand measure in (21), the contribution of the number of object pairs $\binom{n_{ij}}{2}$ in relation to the expectation $\binom{n_{i.}}{2}\binom{n_{.j}}{2}/\binom{n}{2}$ in (20) is reduced when $n_{ij} > n_{i.}n_{.j}/n$ and increased when $n_{ij} < n_{i.}n_{.j}/n$. Given the form of Con $-$ Dis in (19), each cell of the contingency table will contribute positively or negatively depending on whether $n_{ij} - 1$ is greater or less than $(n_{i.}-1)(n_{.j}-1)/(n-1)$. Thus, for reasonable cell sizes that allow us to ignore the -1s in the above expression, each cell will contribute, as might be expected, positively or negatively depending on whether the number of observations in cell (i,j), n_{ij}, is greater than or equal to its expectation, $n_{i.}n_{.j}/n$.

4.1 Significance Testing

As can be seen from the form of (18), the agreement index, Con $-$ Dis, can be specified as a cross-product measure based on three-argument functions:

$$\Lambda(\psi_0) = \sum_{r,s,t} \tilde{p}_{rst} \; \tilde{q}_{\psi_0(r)\psi_0(s)\psi_0(t)} \quad ,$$

where

$$\tilde{p}_{rst} = \text{sign } (p_{rs} - p_{rt}) \quad ,$$

and

$$\tilde{q}_{\psi_0(r)\psi_0(s)\psi_0(t)} = \text{sign } (q_{\psi_0(r)\psi_0(s)} - q_{\psi_0(r)\psi_0(t)}) \quad .$$

Thus, as with the cross-product measure in (8) using two-argument functions, the distribution of Con $-$ Dis may be generated under the generalized hypergeometric model for the $R \times C$ contingency table by considering the distribution of $\Lambda(\psi)$ when ψ is picked at random from all $n!$ possible permutations. Again, using an efficient algorithm for randomly constructing $R \times C$ contingency tables with given row and column totals (see Patefield 1981), Monte Carlo significance testing of the observed index $\Lambda(\psi_0)$ may be routinely carried out through repeated evaluation of the raw index Con $-$ Dis given in reduced form in (19). We also note that the expectation of $\Lambda(\psi)$ is 0 (see Hubert, Golledge, Costanzo, and Gale 1985), and consequently, Con $-$ Dis as a raw measure of agreement is already corrected for chance.

4.2 Normalization of Index

Given the structure of an index corrected for chance in (3) and the expectation of 0 for Con $-$ Dis, normalized indices would have the form (Con $-$ Dis)/(Upper Bound). Generalizing from strategies of assessing rank correlation (see Reynolds 1977, Chapter 3, for a discussion), there are at least four different bounds that lead to clear probabilistic (or operational) meanings in the sense advocated by Goodman and Kruskal (1954) for the corresponding normalized measure:

(i) $n(n-1)(n-2)$;

(ii) $\sum_{\substack{r,s,t \\ \text{distinct}}} |\tilde{p}_{rst}\tilde{q}_{\psi_0(r)\psi_0(s)\psi_0(t)}| = \text{Con} + \text{Dis} =$

$$2\,[\sum_{i,j} n_{ij}(n_{ij}-1)(n-n_{i\cdot}-n_{\cdot j}+n_{ij}) + \sum_{i,j} n_{ij}(n_{i\cdot}-n_{ij})(n_{\cdot j}-n_{ij})];$$

(iii) $\sum_{\substack{r,s,t \\ \text{distinct}}} |\tilde{p}_{rst}| = 2\,[\sum_{i} n_{i\cdot}(n_{i\cdot}-1)(n-n_{i\cdot})];$

$$\text{(iv)} \qquad \sum_{\substack{r,s,t \\ distinct}} |\tilde{q}_{rst}| = 2\,[\sum_j n_{.j}(n_{.j}-1)(n-n_{.j})].$$

Using (i), for instance, and considering an object triple picked at random, (Con − Dis)/$(n(n-1)(n-2))$ can be interpreted as the difference between the probability of a concordant triple, Con/$(n(n-1)(n-2))$, and a discordant triple, Dis/$(n(n-1)(n-2))$, as an analogue of Kendall's (1970) τ_a.

Although the use of $n(n-1)(n-2))$ in (i) leads to a convenient probabilistic interpretation for the resulting measure, this bound does impose a penalty because of the existence of object triples other than those labeled concordant and discordant, e.g., the normalized index could never attain the value of 1.00 if U and V has more than two classes and/or if any class in U or V had more than two objects. To compensate for these "tied" object triples not included in either Con or Dis, we could consider the bound in (ii) (an analogue of Goodman-Kruskal's γ):

$$(\text{Con} - \text{Dis}) \;/\; (\text{Con} + \text{Dis}) \quad .$$

Here, the normalized measure would again be the difference between two probabilities, where the probabilities are now conditional on the randomly selected triple being either concordant or discordant, i.e., [Con/(Con + Dis)] − [Dis/(Con + Dis)]. Finally, if one particular partition is considered the standard and we wished to discount the difference, Con − Dis, when there are ties in the other partition, one could rely on the bounds in (iii) and (iv). For example, conditional on an untied triple in U, which is considered as the standard,

$$(\text{Con} - \text{Dis}) \;/\; \sum_{\substack{r,s,t \\ distinct}} |\tilde{p}_{rst}|$$

can be interpreted as the difference between the probability of a concordant and discordant triple. These two measures based on (iii) and (iv) are analogues of Somers' nonsymmetric rank correlation coefficients and may be generated alternatively as regression coefficients in predicting $\tilde{q}_{\psi_0(r)\psi_0(s)\psi_0(t)}$ from $\tilde{p}_{rst}$, or conversely. As in the earlier discussion of Wallace's measures, the geometric mean of the two indices based on (iii) and (iv) would lead to another possible bound,

$$\sqrt{\sum_{\substack{r,s,t \\ distinct}} |\tilde{p}_{rst}| \sum_{\substack{r,s,t \\ distinct}} |\tilde{q}_{rst}|} \quad ,$$

(and an analogue of Kendall's τ_b, which is based on Daniels' generalized correlation coefficient; see Kendall 1970).

In short, each of the bounds in (i) through (iv) leads to an index that is corrected for chance, bounded between ± 1, and interpretable as the difference between (possibly conditional) probabilities of concordance and discordance. For each case, there is a direct analogue to well-known indices of rank correlation, and in fact, the discussion in this latter context extends directly to the task of comparing two partitions. When considering rank correlation, the ± 1 scoring function is constructed over object pairs through two given numerical sequences; for partition agreement, the ± 1 scoring function is on triples and is generated from dichotomous order information within two proximity matrices defined over the n given objects. Although we will not pursue the extensions to k-tuples when $k \geqslant 4$, similar probabilistic interpretations could be developed for selecting k-tuples at random.

References

ARABIE, P., and BOORMAN, S.A., (1973), "Multidimensional Scaling of Measures of Distance Between Partitions," *Journal of Mathematical Psychology, 10*, 148-203.

BERRY, K.J., and MIELKE, P.W., (1985), "Goodman and Kruskal's TAU-B Statistic," *Sociological Methods & Research, 13*, 543-550.

BRENNAN, R.L., and LIGHT, R.J., (1974), "Measuring Agreement When Two Observers Classify People into Categories not Defined in Advance," *British Journal of Mathematical and Statistical Psychology, 27*, 154-163.

BROOK, R.J., and STIRLING, W.D., (1984), "Agreement Between Observers When the Categories are not Specified in Advance," *British Journal of Mathematical and Statistical Psychology, 37*, 271-282.

COSTANZO, C.M., HUBERT, L.J., and GOLLEDGE, R.G., (1983), "A Higher Moment for Spatial Statistics," *Geographical Analysis, 15*, 347-351.

DUBIEN, J.L., and WARDE, W.D., (1981), *Some Distributional Results Concerning a Comparative Statistic Used in Cluster Analysis*, Unpublished manuscript, Department of Mathematics, Western Michigan University, Kalamazoo, Michigan.

FOWLKES, E.B., and MALLOWS, C.L., (1983), "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association, 78*, 553-569.

FRANK, O., (1976), "Comparing Classifications by the Use of the Symmetric Class Difference," in *Proceedings in Computational Statistics*, eds. J. Gordesch and P. Maeze, Würzburg: Physica Verlag, 84-96.

GAREY, M.R., and JOHNSON, D.S., (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, San Francisco: W.H. Freeman.

GOODMAN, L.A., and KRUSKAL, W.H., (1954), "Measures of Association for Cross-Classifications," *Journal of the American Statistical Association, 49*, 732-764.

GREEN, P.E., and RAO, V.R., (1969), "A Note on Proximity Measures and Cluster Analysis," *Journal of Marketing Research, 6*, 359-364.

HARTIGAN, J.A., (1975), *Clustering Algorithms*, New York: Wiley.

HUBERT, L.J., (1977), "Nominal Scale Response Agreement as a Generalized Correlation," *British Journal of Mathematical and Statistical Psychology, 30*, 98-103.

HUBERT, L.J., (1979), "Matching Models in the Analysis of Cross-Classifications," *Psychometrika, 44*, 21-41.

HUBERT, L.J., (1983), "Inference Procedures for the Evaluation and Comparison of Proximity Matrices," in *Numerical Taxonomy*, ed. J. Felsenstein, New York: Springer-Verlag, 209-228.

HUBERT, L.J., GOLLEDGE, R.G., COSTANZO, C.M., and GALE, N., (1985), "Order-Dependent Measures of Correspondence for Comparing Proximity Matrices and Related Structures," in *Measuring the Unmeasurable*, eds. P. Nijkamp and H. Leitner, The Hague: Martinus Nijhoff.

JOHNSON, S.C., (1968), "Metric Clustering," Unpublished manuscript, AT&T Bell Laboratories, Murray Hill, New Jersey.

KENDALL, M.G., (1970), *Rank Correlation Methods*, 4th Edition, London: Griffin.

KLASTORIN, T.D., (1985), "The *p*-Median Problem for Cluster Analysis: A Comparative Test Using the Mixture Model Approach," *Management Science, 31*, 84-95.

LERMAN, I.C., (1973), "Etude Distributionelle de Statistiques de Proximité entre Structures Finies de Même Type; Application à la Classification Automatique," *Cahier no. 19 du Bureau Universitaire de Recherche Opérationnelle*, Institut de Statistique des Universités de Paris.

MIELKE, P.W., (1979), "On Asymptotic Nonnormality of Null Distributions of MRPP Statistics," *Communications in Statistics - Theory and Methods, A8*, 1541-1550 (errata: *A10*, 1981, p. 1795 and *A11*, 1982, p. 847).

MIELKE, P.W., and BERRY, K.J., (1985), "Non-Asymptotic Inference Based on the Chi-square Statistic for r by c Contingency Tables," *Journal of Statistical Planning and Inference, 12*, 41-45.

MIELKE, P.W., BERRY, K.J., and BRIER, G.W., (1981), "Application of Multiresponse Permutation Procedures for Examining Seasonal Changes in Monthly Sea-Level Pressure Patterns," *Monthly Weather Review, 109*, 120-126.

MILLIGAN, G.W., and COOPER, M.C., (1985), "A Study of the Comparability of External Criteria across Hierarchy Levels," unpublished manuscript, Ohio State University, Columbus, Ohio.

MILLIGAN, G.W., SOON, S.C., and SOKOL, L.M., (1983), "The Effect of Cluster Size, Dimensionality, and the Number of Clusters on Recovery of True Cluster Structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5*, 40-47.

MIRKIN, B.G., and CHERNYI, L.B., (1970), "Measurement of the Distance Between Distinct Partitions of a Finite Set of Objects," *Automation and Remote Control, 31*, 786-792.

MOREY, L.C., and AGRESTI, A., (1984), "The Measurement of Classification Agreement: An Adjustment of the Rand Statistic for Chance Agreement," *Educational and Psychological Measurement, 44*, 33-37.

PATEFIELD, W.M., (1981), "Algorithm AS159. An Efficient Method of Generating Random $R \times C$ Tables with Given Row and Column Totals," *Applied Statistics, 30*, 91-97.

RAND, W.M., (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association, 66*, 846-850.

REYNOLDS, H.T., (1977), *The Analysis of Cross-Classifications*, New York: Free Press.

ROHLF, F.J., (1974), "Methods of Comparing Classifications," *Annual Review of Ecology and Systematics, 5*, 101-113.

ROHLF, F.J., (1982), "Consensus Indices for Comparing Classifications," *Mathematical Biosciences, 59*, 131-144.

STAM, A.J., (1983), "Generation of a Random Partition of a Finite Set by an Urn Model," *Journal of Combinatorial Theory A, 35*, 231-240.

WALLACE, D.L., (1983), "Comment" *Journal of the American Statistical Association, 78*, 569-579.