The Annals of Statistics
1986, Vol. 14, No. 2, 691–707

INTERGROUP DIVERSITY AND CONCORDANCE FOR RANKING DATA: AN APPROACH VIA METRICS FOR PERMUTATIONS

BY PAUL D. FEIGIN[1] AND MAYER ALVO[2]

Technion and University of Ottawa

Motivated by the apportionment of diversity analysis due to C. R. Rao, a general approach to comparing populations of rankers is proposed. Each permutation metric corresponds to a particular population characteristic that forms the basis of the comparison. Tests of hypotheses concerning equality of characteristics are developed. Throughout, comparison is made with earlier work, most of which is based on the use of only the Spearman metric. Extension to tied rankings is discussed. Examples for two groups are presented which illustrate the computational feasibility as well as the value of the proposed procedures.

1. Introduction. There has recently been a revival of interest in the analysis of rank data (or rankings) and even some controversy concerning the appropriateness of various proposed statistics [see Hollander and Sethuraman (1978) plus comments by Schucany, and the paper by Kraemer (1981)]. Our aim here is to propose a general framework within which the various analyses can be investigated as well as to suggest some alternative analyses. We will concentrate on the situation of two or more populations of rankers or judges [for the single population case some related material may be found in Alvo, Cabilio, and Feigin (1982)], and our interest lies in determining whether and how the populations differ in the way they tend to rank a fixed set of r objects.

Although the rankers may apply some absolute scoring system to each object, with the resultant scores determining the ranking, these scores are not observed —the only datum each ranker provides is his ordering of the r objects with respect to some criterion. Various models have been proposed for such data, for example: ones that are based on an absolute scoring process [see the recent paper by Pettitt (1982) which also considers groups of judges]; or ones that try to reflect the paired comparisons approach to determining a ranking [see Mallows (1957) and Feigin and Cohen (1978)]. Here, we will not consider parametric models per se, but will be concerned with describing those characteristics of models with respect to which one may wish to compare different groups of judges.

The model or population characteristics were arrived at initially by applying methods of analysing diversity which have recently been developed by Rao

Received August 1984; revised May 1985.
[1]Research supported in part by VPR Fund—Lawrence Deutsch Research Fund.
[2]Research supported in part by NSERC Grant No. 9068.
AMS 1980 *subject classifications*. Primary 62G10, 62G05; secondary 62J10, 20B99.
Key words and phrases. Intergroup concordance, Kendall's tau, Spearman's rho, Spearman's footrule, tied rankings.

(1982a, 1982b). Although, in retrospect, one may also begin by defining the various characteristics directly, we feel that it is nevertheless instructive and useful to present the analysis of diversity as the point of departure (viz. Section 2).

Before proceeding with a general comparison of various approaches to analysing rankings, we pause to define a few quantities. A ranking of r objects can be described by a permutation, ω, which is an element of Ω, the set of all $k = r!$ permutations (or the symmetric group on r letters). The possibility of rankings with ties is discussed briefly later; meanwhile it is denied. A stochastic *model* for the selection of a ranking by a judge is a probability vector π, which is of dimension k and whose ith component $\pi(i)$, is the probability that $\omega_i (\in \Omega)$ is selected by the judge. The permutations $\omega_1, \ldots, \omega_k$ are numbered in an arbitrary but fixed way. Each judge in a particular group (or population) selects a ranking according to the same π and independently of the other judges. Equivalently, we can regard the population distribution of judges (as far as ranking the r objects is concerned) as given by π; and the particular group of judges as a random sample from this infinite population. Comparing g populations of judges then amounts to trying to compare the g probability vectors (models) $\pi_1, \pi_2, \ldots, \pi_g$. If a sample of n_i judges from population i is taken we denote by $\{X_{ij},\ j = 1, \ldots, n_i\}$ the set of rankings chosen by these judges. Summarizing the above, we have, for each $j = 1, \ldots, n_i$

$$P(X_{ij} = \omega_l) = \pi_i(l), \qquad l = 1, \ldots, k.$$

When referring to the comparison of groups of judges the terms *agreement* and *intergroup concordance* have been used in the literature. The question of comparing populations of judges then involves two stages: firstly, defining agreement or measuring degrees of agreement; and secondly, estimating or testing for levels of agreements based on the samples (or groups) of judges available.

The recent literature on agreement has focussed on analysis based on the average rank statistic for each group, whether or not agreement itself is defined in terms of the expected rank vector. One of the innovations proposed here is the possibility of considering other statistics corresponding to other characteristics of the population models. For example, the average rank vector is directly related to the Spearman rank correlation, ρ, whereas one may wish to consider that characteristic vector associated with Kendall's rank correlation, τ.

The characteristics of a population or model (determined by π) may be thought of as a mapping $T\pi$ of π into a lower dimensional Euclidean space $\mathbb{R}^s$ ($s < k = r!$)—we only consider linear maps T here. We note that Pettitt's (1982) approach, for example, may be regarded as involving a particular nonlinear map of π vectors which lie in a $(r-1)$-dimensional subspace of $\mathbb{R}^k$. Given T, agreement itself may be defined in terms of the vectors $T\pi_1, \ldots, T\pi_g$ [see, e.g., Kraemer (1981)] and/or the means of testing for agreement, however defined, may be based on the statistics $Tf_1, \ldots, Tf_g$; where f_i is the relative frequency vector of the permutations (rankings) in group i of judges.

One of the main reasons for considering the statistic Tf instead of f itself is the dimensionality of the latter. For r as small as 5, the analysis based directly on f would need to take place on the unit simplex in $\mathbb{R}^{120}$—a task that would

require more data than is usually available in order to apply standard asymptotics. Another important motivation for considering characteristics $T\pi$ is that permutations or rankings that differ do so to different degrees. In other words, one would like to incorporate the notion of distance between permutations when comparing groups of judges. This last point leads us to our approach via measures of diversity and forms the subject of the next section.

In terms of the notations of this paper, the controversy alluded to earlier may be paraphrased as follows. Hollander and Sethuraman (1978) define agreement between two populations of rankers as $\pi_1 = \pi_2$, and suggest testing agreement by comparing Tf_1 with Tf_2 where T is the mapping which gives the average (centred) ranks vector. Schucany and Frawley (1973), for example, regard complete absence of agreement as nonpositive correlation between the vectors $T\pi_1$ and $T\pi_2$ (same T as before). They are also concerned with the idea that if π_1 or, actually $T\pi_1$, is a constant vector then there is no consensus in population 1 and so no possibility for agreement between populations 1 and 2. Kraemer (1981) develops this idea further by defining a relative measure of intergroup concordance—relative to the average *intragroup* concordance. We will make some more explicit references to these approaches in the sequel, although our results, to some extent, develop along the lines proposed by Hollander and Sethuraman's (1978) analysis, but deal with a larger variety of statistics. Of particular interest is the characteristic corresponding to Kendall's τ, which forms the basis of the numerical examples presented in Section 5.

As the referee has pointed out to us, in his forthcoming monograph, Diaconis (1985) also presents one way to use metrics on the permutation group in order to test for agreement among populations of judges. His approach is based on constructing a minimal spanning tree of the data and comparing the number of edges joining rankings of different populations with the number of edges which join rankings from the same population. We hope to compare this rather different approach with ours in the future.

2. Measuring and apportioning diversity for populations of rankers. The measurement of diversity of a population has a long history, particularly in the biological sciences such as genetics. Coincidentally with the renewed interest in rank data, statisticians have recently returned to the problem of measuring diversity [see, for example, the paper (with comments) by Patil and Taillie (1982)] and one approach, espoused by Rao (1982a, 1982b), forms the basis of our analysis.

In contrast to the entropy-type measures of diversity which are simply functions of $\{\pi(1), \pi(2), \ldots, \pi(k)\}$ without any regard to the ordering of the categories (in our case—permutations), Rao suggests using measures which incorporate distances between categories. In light of the comments made earlier, this suggestion seems eminently appropriate to the analysis of rankings.

Consider a set Ω of k points $\omega_1, \omega_2, \ldots, \omega_k$ and let $\{\delta_{ij}: i, j = 1, \ldots, k\}$ denote the set of "distances" between pairs of points, i.e., δ_{ij} is the distance between ω_i and ω_j. In Rao's (1982a) formulation, δ need only be nonnegative and need not satisfy the properties of a metric—in other words, δ measures *some* concept of distance between points. The diversity coefficient of a population can now be

defined in terms of the expected distance between two members selected independently from the population according to π:

DEFINITION 2.1. The diversity coefficient (based on δ) of the population with probability vector $\pi' = (\pi(1), \ldots, \pi(k))$ on Ω is

$$H(\pi) = \pi' \Delta \pi \tag{2.1}$$

where $\Delta = (\delta_{ij})$ is a $(k \times k)$ matrix.

Applying this definition to the rankings situation simply involves choosing a measure of distance between permutations ($\mu, \eta \in \Omega$ say). Three such measures used in statistical applications are given below and are related to Spearman's $\rho(S)$, Kendall's $\tau(K)$ and Spearman's footrule (F), respectively:

$$d_S(\mu, \eta) = \frac{1}{2} \sum_{s=1}^{r} [\mu(s) - \eta(s)]^2 = [r(r+1)(2r+1)/6 - \mu'\eta]; \tag{2.2}$$

$$d_K(\mu, \eta) = \sum_{s<t}^{r} \{1 - \operatorname{sgn}[\mu(s) - \mu(t)] \operatorname{sgn}[\eta(s) - \eta(t)]\}; \tag{2.3}$$

$$d_F(\mu, \eta) = \frac{1}{2} \sum_{s=1}^{r} |\mu(s) - \eta(s)|. \tag{2.4}$$

Of course, many other metrics have been defined on the set of permutations and the particular one that the statistician chooses to use may involve consideration of the actual processes which determine the choice of a ranking by a judge. Alternatively, more robust conclusions may be reached by taking into account several different metrics when analysing the same set of data. It is this flexibility which we wish to pursue here, and not the determination of a particular metric as universally superior.

Thinking of diversity as a generalization of the notion of variance, and since we are interested in comparing several populations, the next step is defining the diversity between populations versus that within populations—or, as Rao (1982a) calls it: the apportionment of diversity. We note here that d_S is the *square* of a Euclidean metric and so leads directly to a standard analysis of variance. This fact gives further insight into the popularity of analyses based on Spearman's ρ. Although d_S is itself *not* a metric, we nevertheless will refer to it as the Spearman metric in the sequel.

Suppose g populations with probability vectors $\pi_1, \ldots, \pi_g$ are mixed together according to the proportions $\lambda_1, \ldots, \lambda_g$ such that $\lambda_1 + \lambda_2 + \cdots + \lambda_g = 1$, thereby forming a new population with probability vector $\pi = \sum_{i=1}^{g} \lambda_i \pi_i$. Following Rao (1982a) we now turn to:

DEFINITION 2.2. Suppose the diversity $H(\cdot)$ of (2.1) is a concave function on $\mathbb{S}_k$, the unit simplex in $\mathbb{R}^k$. Then the *total* diversity $H(\pi)$ can be apportioned into the *within* populations diversity

$$\sum_{i=1}^{g} \lambda_i H(\pi_i) \tag{2.5}$$

and the *between* populations diversity

$$-\sum_{i<j}^{g}\lambda_i\lambda_j(\pi_i-\pi_j)'\Delta(\pi_i-\pi_j). \tag{2.6}$$

The concavity requirement is to ensure that the between populations diversity (or *discrimination* coefficient) is nonnegative. In terms of the distance δ, this condition is equivalent to the requirement that

$$a'\Delta a \le 0 \quad \text{whenever} \sum_{s=1}^{k} a(s) = 0 \tag{2.7}$$

or equivalently, that

$$\Delta^* = \left(\delta_{i1} + \delta_{j1} - \delta_{ij}\right) \text{ be nonnegative definite.} \tag{2.8}$$

It is interesting to note that requiring that δ be a metric on Ω is not sufficient to ensure (2.7) or (2.8) hold, so that for the application to rankings one has to verify (2.7), say, for each potential distance measure.

For metrics on the set of permutations, we have that the desirable property of *right invariance* [see Diaconis and Graham (1977) or Alvo et al. (1982)] ensures that the k vector $e = (1, 1, \ldots, 1)'$ is an eigenvector of Δ. We therefore have:

LEMMA 2.1. *If δ is a right invariant metric on the set of permutations then there exists $c > 0$ such that*

$$\Delta e = ce$$

and $H(\cdot) = H_\Delta(\cdot)$ is concave if and only if

$$Q \equiv (c/k)J - \Delta \text{ is positive semidefinite,} \tag{2.9}$$

where $J = ee'$. Moreover, in this case $H(\pi)$ has the maximum value $\beta \equiv c/k$ at $\pi = u \equiv (1/k)e$.

PROOF. The existence of the eigenvalue c follows from the right invariance property as referred to above.

If (2.7) holds, then since $Qe = 0$, $x'Qx \ge 0$ for any $x = b + \alpha e$ with $b'e = 0$; but this includes all $x \in \mathbb{R}^k$. The converse is immediate.

Writing $\pi = u + (\pi - u)$ and since u is an eigenvector of Δ orthogonal to $(\pi - u)$ the result follows from

$$H(\pi) = u'\Delta u + (\pi - u)'\Delta(\pi - u) \le u'\Delta u = c/k. \qquad \square$$

This last result says that the uniform distribution over Ω is most diverse for diversity measures based on a right invariant metric.

The fact that (2.9) is valid for Δ based on d_S and d_K [see (2.2) and (2.3)] follows from Alvo et al. (1982) or from the form of (2.2) and (2.3) (see Lemmas 3.1 and 3.3). That (2.9) is also true for the footrule metric (2.4) is less obvious and is proved in Lemma 3.4. Note that the restriction to right invariant metrics is a natural one in the context of rankings.

If the rank of Q is less than or equal to s, then standard matrix theory implies the existence of an $(s \times k)$ matrix T such that

$$Q = T'T = (c/k)J - \Delta. \tag{2.10}$$

We can now gain further insight into the nature of the between groups diversity. For X a random vector we let var(X) denote its variance covariance matrix.

LEMMA 2.2. *If δ is a right invariant metric on the set of permutations Ω, then for H defined by* (2.1), *the between populations diversity is given by*

$$\sum_{i<j}^{g} \lambda_i \lambda_j \|T\pi_i - T\pi_j\|_s^2 = \operatorname{tr}\{\operatorname{var}(T\pi_I)\} \tag{2.11}$$

where $\|\cdot\|_m$ is the Euclidean norm in $\mathbb{R}^m$ and I has the distribution $P(I = i) = \lambda_i$, $i = 1, \ldots, g$.

PROOF. The result follows immediately from (2.6) by substituting $\Delta = (c/k)J - T'T$ and expanding the $\|\cdot\|^2$ term. □

The expression (2.11) allows us to interpret the apportionment of diversity based on δ in terms of the variability of $T\pi_i$, $i = 1, \ldots, g$. Moreover, it is the characteristic $T\pi$ of the model π which forms the basis for comparing populations if one does so using a diversity measure based on a right invariant metric δ. This interpretation of the apportionment of diversity is the topic of the next section.

3. Defining and interpreting model characteristics. We pursue the implications of Rao's apportionment of diversity for the analysis of rankings based on right invariant metrics. In so doing we arrive at a way of interpreting T matrices for given metrics as well as showing how the characteristics so defined have been or may be used to compare populations of judges. We concentrate on describing the T matrices corresponding to the Spearman, Kendall, and footrule metrics.

For the case $\delta_{ij} = d_S(\omega_i, \omega_j)$, we may identify T as follows. Define the centered rank vector

$$t_S: \Omega \to \mathbb{R}^r$$

by

$$t_S(\omega) = \left(\omega(1) - \frac{r+1}{2}, \ldots, \omega(r) - \frac{r+1}{2}\right)' \tag{3.1}$$

and let the $(r \times k)$ matrix T_S be defined by

$$T_S = (t_S(\omega_1), \ldots, t_S(\omega_k)). \tag{3.2}$$

LEMMA 3.1. *For δ corresponding to d_S, the matrix T in the decomposition* (2.10) *is given by T_S of* (3.2), *and the characteristic $T\pi$ corresponds to the expected centred rank vector.*

PROOF. Let the rank vector v: $\Omega \to \mathbb{R}^r$ be defined as $v(\omega) = t_S(\omega) + (r+1)/2e$ and

$$V = (v(\omega_1), \ldots, v(\omega_k)). \tag{3.3}$$

Then

$$T = T_S = \left(I - \frac{1}{r}J\right)V \tag{3.4}$$

and

$$\begin{aligned} T'T &= \left(V'V - \tfrac{1}{4}r(r+1)^2 J\right) \\ &= (1/12)r(r+1)(r-1)J - \Delta, \end{aligned} \tag{3.5}$$

the last inequality following from the definition of d_S (2.2). Since, by construction, $Te = 0$, we conclude that we have discovered the decomposition (2.10) with $c = kr(r+1)(r-1)/12$.

Furthermore,

$$T\pi = \sum_{i=1}^{k} \pi(i) t_S(\omega_i) = E_\pi t_S(\omega),$$

the expected centred rank vector. □

Much of the literature on analysing rankings is based upon the expected rank vectors. For the problem of comparing two populations of rankers, Schucany and Frawley (1973) and Li and Schucany (1975) consider a statistic which may be regarded as an estimator of $(T\pi_1)'(T\pi_2)$: a measure of "covariance" between the two vectors $T\pi_1$ and $T\pi_2$. Kraemer (1981) defines a measure of intergroup concordance ρ which, in our notation, is given by

$$\rho = \|V(\pi - u)\|_r^2 \Big/ \sum_{i=1}^{g} \lambda_i \|V(\pi_i - u)\|_r^2, \tag{3.6}$$

where $\pi = \Sigma \lambda_i \pi_i$ and $u = (1/k)e$ as before. Kraemer considers the case $\lambda_i = 1/g$; $i = 1, 2, \ldots, g$. We can rewrite ρ in terms of the within and total diversity.

LEMMA 3.2. *In terms of the apportionment of diversity (Definition 2.2) based on the Spearman metric, the intergroup concordance ρ is given by*

$$\rho = \|T\pi\|_r^2 \Big/ \left\{ \sum \lambda_i \|T\pi_i\|_r^2 \right\} \tag{3.7}$$

$$= (\beta - \text{total})/(\beta - \text{within}), \qquad \beta \equiv c/k = r(r+1)(r-1)/12. \tag{3.8}$$

PROOF. From (3.3) and (3.4) we have

$$V(\pi - u) = T(\pi - u) = T\pi$$

so that, via (2.9) and (2.10),

$$\|T\pi\|_r^2 = \pi' Q \pi = (c/k) - \pi' \Delta \pi = \beta - H(\pi)$$

and (3.7) and (3.8) follow. □

Note that if $\pi_i = u$ all i then ρ is undefined—otherwise it represents the proportion of concordance attainable between groups given the level within (Kraemer, 1981). In fact, we may regard concordance (C) as the complement of diversity (D) from the relationship

$$C = \beta - D. \tag{3.9}$$

Thus, Kraemer's coefficient measures the concordance ratio

$$\rho = C(\text{total})/C(\text{within})$$

whereas the apportionment of diversity would lead to the dispersion ratio

$$\alpha = D(\text{within})/D(\text{total}).$$

It then becomes a more philosophical issue whether dispersion (distance) or concordance (similarity) is the appropriate criterion. It is true that whenever C (within) = 0, ρ is undefined whereas $\alpha = 1$. This case corresponds to that of similar but completely internally discordant groups. Is it the similarity or is it the complete discordance that one wants to measure?

The form (3.7) could, of course, also serve to define a measure of intergroup concordance based on another right invariant metric, for example, that based on Kendall's tau [viz. (2.3)]. In order to interpret the latter we quote the following result.

LEMMA 3.3. *Let* $t_K: \Omega \to \{-1, +1\}^{\binom{r}{2}}$ *be defined by*

$$(t_K(\omega))(s) = (\operatorname{sgn}\{\omega(j) - \omega(i)\}), \qquad s = 1, 2, \ldots, \binom{r}{2}, \tag{3.10}$$

where

$$s = (i-1)(r - i/2) + (j - i), \qquad 1 \le i < j \le r.$$

Then the $\binom{r}{2} \times k$ *matrix* T

$$T = T_K \equiv (t_K(\omega_1), \ldots, t_K(\omega_k))$$

satisfies (2.10) *for* Δ *based on the Kendall tau metric* (2.3).

PROOF. Straightforward, since for the definition (2.3)

$$d_K(\mu, \eta) = \frac{r(r-1)}{2} - t_K'(\mu) t_K(\eta),$$

so that

$$\Delta = [r(r-1)/2]J - T'T. \qquad \square$$

The characteristic $T_K\pi$ may therefore be regarded as the expected pairwise concordance vector, where concordance is measured with respect to the identity permutation $(1, 2, \ldots, r)$. Thus the measure ρ of intergroup concordance with $T = T_K$ in (3.6) would look at relative agreement based on average pairwise decisions—a more sensitive criterion than that based on average ranks. Whether or not this extra sensitivity reflects relevant aspects of concordance or agreement

between judges must, of course, be ascertained from the particular context. At the very least, however, it provides a further tool for comparisons and contrasts.

In order to interpret the characteristic $T_F\pi$ corresponding to the diversity measure based on the Spearman footrule metric (2.4), we obtain the following result after some algebra.

LEMMA 3.4. *Let* t_F: $\Omega \to \mathbb{R}^{r^2}$ *be defined by*

$$(t_F(\omega))((i-1)r+j) = I[\omega(i) \le j] - j/r, \qquad 1 \le i, j \le r,$$

where $I[\omega(i) \le j]$ *equals one when* $\omega(i) \le j$ *and is zero otherwise. Then the* $(r^2 \times k)$ *matrix* T *given by*

$$T = T_F \equiv (t_F(\omega_1), \ldots, t_F(\omega_k))$$

satisfies (2.10) *for* Δ *based on Spearman's footrule.*

PROOF. The proof amounts to showing that

$$[t_F(\mu)]' t_F(\eta) = (r+1)(r-1)/6 - d_F(\mu, \eta)$$

using the fact that

$$\max(\mu(i), \eta(i)) = \tfrac{1}{2}[\eta(i) + \mu(i)] + \tfrac{1}{2}|\mu(i) - \eta(i)|. \qquad \square$$

The characteristic $T_F\pi$ may therefore be regarded as the set of (centred) distribution functions for the ranks of each item. If we write

$$\mathscr{F}_i(j) = P_\pi(\omega(i) \le j) - j/r, \qquad 1 \le j \le r,$$

then $T_F\pi$ is equivalent to the set $\{\mathscr{F}_1, \ldots, \mathscr{F}_r\}$. This characterization is only based on the marginal distributions of the ranks for each item, and so takes no account of possibly important dependence relationships between these ranks. The same is of course true for $T_S\pi$—which merely considers averages for each item—whereas $T_K\pi$ is sensitive to certain patterns of dependence in the allocation of ranks to each of the objects.

In the light of the above interpretation of model or population characteristics, we may look again at those analyses of agreements suggested previously. Hollander and Sethuraman's (1978) statistic is sensitive to departures from $T_S\pi_1 = T_S\pi_2$. An extension to the case of g groups was considered by Katz and McSweeney (1981): It amounts to treating disagreement as occurring if $T_S\pi_i \ne T_S\pi_j$ for some $1 \le i, j \le g$. Using other characteristics $T\pi$ instead of $T_S\pi$, other types of disagreement among populations of judges may be investigated. We develop this approach in the sequel, particularly for the pairwise concordance characteristic $T_K\pi$.

For other approaches to defining agreement, we refer to Li and Schucany's (1975) discussion of various comparisons of populations of judges. The approach ascribed to Quade amounts to saying that the two populations agree if

$$\pi_1' \Delta \pi_1 = \pi_2' \Delta \pi_2$$

where $\Delta = \Delta_S$ or Δ_K. This approach seems to ignore the fact that the two

populations could be equally concordant about different rankings. Linhart's (1960) definition is essentially equivalent to Quade's in that $\pi'\Delta_S\pi$ is directly related to the coefficient of concordance (viz. Kendall's W) for the population [see Alvo et al. (1982)].

Finally, we reiterate the point made in the introduction—that the characteristic t: $\Omega \to \mathbb{R}^s$ which describes the information used in comparing rankings may itself be chosen directly by the researcher and need not be derived from a diversity argument based on a particular metric. In this case, however, there is a certain ambiguity in the definition of the corresponding Δ (distance matrix). We assume T is chosen so that $Te = 0$ and then set

$$\Delta_\beta(T) \equiv \beta J - T'T.$$

DEFINITION 3.1. Δ is called the minimal distance matrix corresponding to T if it is nonnegative with at least one zero element (on the diagonal) and equals $\Delta_\beta(T)$ for some $\beta = \beta(T)$.

From the definition it is clear that

$$\beta(T) = \max\{(T'T)_{ii},\ i = 1, \ldots, k\},$$

and that if T is defined in terms of a metric δ then the minimal distance matrix corresponds to the original matrix $\Delta = (\delta_{ij})$ (with zeroes on the diagonal).

In the sequel we will refer to the *minimal* distance matrix corresponding to T as simply the distance matrix corresponding to T.

4. Inference for measures of agreement. We will refer to three essentially different ways of making inferences about the level of agreement observed among a set of g groups of rankers. They involve procedures:

A. based on the randomization distribution of a statistic;

B. based on jackknifing or related methods;

C. based on the asymptotic distributions of a statistic.

For each approach the actual statistic to be used may be chosen either *a priori* or based on the corresponding analysis for the asymptotic (multivariate normal) experiment. Furthermore, we have the option of deriving versions of a given statistic type by choosing that model characteristic $T\pi$ in terms of which we wish to assess the level of agreement.

The data available, although probably not presented in this way (see below concerning computational aspects), may be summarized as follows. For $i = 1, \ldots, g$, the n_i judges of group i assign rankings and the resulting relative frequency vector f_i is defined by

$$f_i(\ell) = n_i^{-1} \times (\text{no. of times } \omega_\ell \text{ assigned in group } i), \qquad \ell = 1, 2, \ldots, k \equiv r!. \tag{4.1}$$

The analysis of agreement may now proceed by defining statistics related to the corresponding population quantities—that is, we replace π_i by f_i in the relevant formulae. Given T [and the corresponding distance matrix Δ and value $\beta = \beta(T)$

—see Definition 3.1], we begin by considering the following two statistics:

$$\hat{\alpha} = W/S, \tag{4.2}$$

$$\hat{\rho} = (\beta - S)/(\beta - W) = (\beta/S - 1)/(\beta/S - \hat{\alpha}) \tag{4.3}$$

where

$$W = \sum_{i=1}^{g} \lambda_i f_i' \Delta f_i = \beta - \sum_{i=1}^{g} \lambda_i \|Tf_i\|^2,$$

$$S = f'\Delta f = \beta - \|Tf\|^2,$$

$$f = \sum_{i=1}^{g} \lambda_i f_i,$$

and

$$\lambda_i = n_i/N, \qquad N = n_1 + n_2 + \cdots + n_g, \tag{4.4}$$

or, possibly, $\{\lambda_i\}$ is an *a priori* set of weights for the g populations, independent of the sample sizes. We assume in the sequel that the λ_i are determined by (4.4).

At a descriptive level, both $\hat{\alpha}$ and $\hat{\rho}$ measure the relative degree of intergroup concordance—values near one indicating stronger agreement between the groups. The statistic $\hat{\alpha}$ corresponds to the proportion of total diversity due to intragroup differences (viz. Definition 2.2); whereas $\hat{\rho}$ is a generalized version of Kraemer's (1981) measure (viz. Lemma 3.2). These statistics to some degree answer the question: What level of agreement do the populations reveal? While α refers to the relative diversity (within/total) ρ relates to the relative similarity [(overall similarity)/(within similarity)].

To make inferences concerning the corresponding population quantities, we may employ procedures of type A or B. For testing, using A, with the null hypothesis of complete agreement

$$H_0\colon \pi_i = \pi, \qquad i = 1, \ldots, g \text{ (implies } \alpha = \rho = 1), \tag{4.5}$$

the randomization (permutation) distribution gives equal probability to each of the $M = N!/\prod_{i=1}^{g}(n_i!)$ partitions of the N judges into g groups of sizes $n_1, \ldots, n_g$. The significance level of the data is then the relative frequency of values of the statistic ($\hat{\alpha}$ or $\hat{\rho}$) that fall below the value observed for the actual partition. Approximations for moderately large N are also available. We refer the reader to Mielke et al. (1981) for some details and references. Hollander and Sethuraman (1978) derive their test statistic from the asymptotic approximation of the randomization distribution of $T_S f$. However, we find it more natural to discuss their approach with respect to procedures of type C.

For confidence intervals for the population quantities α or ρ, one may follow Kraemer's (1981) suggestion and compute the jackknife estimate of the parameter as well as its standard error viz. procedure B above. This procedure involves recalculating the statistic, leaving out each ranking one at a time, and then using the value from the t distribution on the appropriate number of degrees of freedom. Mosteller and Tukey (1977) give the general theory and Kraemer (1981) discusses the application to the particular case of ρ based on $T_S f$ (i.e., the average

ranks vector). The extension to α and to general T is straightforward in principle. Moreover, ascertaining the probability value of the hypothesis $\alpha = 1$ (or $\rho = 1$) can be determined approximately using the same t approximation.

The bootstrap is an approach related to the jackknife, which could also be applied to the analysis of $\hat{\alpha}$ or $\hat{\rho}$. The bootstrap samples would each contain N pairs (I, ω) randomly chosen with replacement from $\{(i, X_{ij})\colon\ j = 1, \ldots, n_i;\ i = 1, \ldots, g\}$ and then the standard Monte Carlo analysis would allow one to estimate biases and standard errors as well as to construct confidence intervals. (Here X_{ij} is the ranking of judge j in group i.) A succinct account of these ideas appears in Efron (1982).

The more classical statistical approach is to consider the asymptotic distributions of the statistics involved and this approach is readily applicable here also. We have, denoting the p-variate normal distribution by N_p:

THEOREM 4.1. *For* $n_i/N \to \lambda_i > 0$ *as* $N \to \infty$

$$\sqrt{n_i}\, T(f_i - \pi_i) \Rightarrow U_i, \qquad i = 1, \ldots, g,$$

where

$$U_i \sim N_s(\mathbf{0}, T\Sigma_i T'), \qquad U_1, \ldots, U_g \text{ are independent and}$$

$$\Sigma_i = \Pi_i - \pi_i \pi_i', \qquad \Pi_i = \operatorname{diag}(\pi_i(1), \ldots, \pi_i(k)).$$

PROOF. A straightforward application of the central limit theorem for multinomial vectors. □

Based on the asymptotic formulation, a test for agreement based on the characteristic $T\pi$, i.e., of

$$H_0(T)\colon T\pi_1 = T\pi_2 = \cdots = T\pi_g,$$

amounts to a multivariate analysis of variance with nonhomogeneous covariance matrices. If we regard the null hypothesis as

$$H_0\colon \pi_1 = \pi_2 = \cdots = \pi_g$$

then under H_0 the covariance matrices are homogeneous and classical MANOVA may be applied.

The approach proposed by Katz and McSweeney (1981) is designed for testing $H_0(T)$ and allows for the nonhomogeneity of variances. Although they only consider $T = T_S$, there is again no major difficulty in applying their approach to other T matrices.

We will concentrate on the situation of two groups of judges ($g = 2$).

THEOREM 4.2. *For* $g = 2$ *and under* $H_0(T)$, *the conditions of Theorem* 4.1 *imply*

$$\sqrt{N}\, T(f_1 - f_2) \Rightarrow N_S(0, T\Sigma T'),$$

where

$$\Sigma = \lambda_1^{-1}\Sigma_1 + \lambda_2^{-1}\Sigma_2. \tag{4.6}$$

PROOF. Straightforward. □

COROLLARY. *Suppose* $\hat{\Sigma}$ *is a consistent estimate of* Σ *and that* $\hat{D}$ *is the Moore–Penrose inverse of* $T\hat{\Sigma}T'$. *Then, under* $H_0(T)$,

$$\gamma_N \equiv N(f_1 - f_2)'T'\hat{D}T(f_1 - f_2) \Rightarrow \chi^2_v$$

where $v = \operatorname{rank}(T\Sigma T')$.

PROOF. Since $\hat{D}$ is consistent for D, where D is the Moore–Penrose (generalized) inverse of $T\Sigma T'$, the result follows from standard multivariate normal theory and the continuity theorem of weak convergence. □

One way to circumvent the need to use generalized inverses is by choosing T so that $T\Sigma T'$ is of full rank. Thus the $((r-1)\times k)$ matrix T_S^* only uses the ranks of the first $(r-1)$ objects and thus avoids the obvious singularity incurred by using T_S. Of course singularities may also derive from the particular form of Σ. For the case of $T = T_K$ [of dimension $\binom{r}{2} \times k$], there is no *a priori* singularity in the matrix $T\Sigma T'$.

It is now important to decide on how to estimate Σ of (4.6). We write

$$\hat{\Sigma}_i = n_i[F_i - f_i f_i'] \cdot (n_i - 1)^{-1}, \qquad i = 1, \ldots, g,$$

where

$$F_i = \operatorname{diag}(f_i(1), \ldots, f_i(k)).$$

There are basically three different estimates of Σ, depending to some extent on which null hypothesis is entertained:

$$\hat{\Sigma}_S = N(n_1^{-1}\hat{\Sigma}_1 + n_2^{-1}\hat{\Sigma}_2),$$

$$\hat{\Sigma}_P = [N^2/n_1n_2]((n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2)[1/(N-2)],$$

$$\hat{\Sigma}_C = \left[\frac{N-2}{N-1}\right]\hat{\Sigma}_P + \frac{N}{N-1}(f_1 - f_2)(f_1 - f_2)'.$$

The estimate $\hat{\Sigma}_S$ (separate) is appropriate when $H_0(T)$ is considered to be the null hypothesis since in this situation we may not assume that the variance matrices (dispersions) are equal. The estimate $\hat{\Sigma}_P$ (pooled) is obtained by pooling the estimates $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ and is appropriate under H_0. Hollander and Sethuraman (1978) actually use $\hat{\Sigma}_C$ (combined sample) which is the estimate of dispersion based on the combined sample of size N. In the MANOVA context, $(N-1)\hat{\Sigma}_C$ is the total dispersion whereas $(N-2)\hat{\Sigma}_P$ is the within groups dispersion—the latter is more commonly used to estimate the common dispersion.

Returning to testing for agreement, we recall that the relevant dispersion is $T\Sigma T'$ which is of dimension s, typically much less than $k = r!$. It is this lower dimensionality that makes both the inversion of $T\hat{\Sigma}T'$ feasible as well as the (asymptotic) normality a better approximation [see Remark 5(d) of Alvo et al. (1982)]. Based on the corollary, we therefore have an asymptotically χ^2 statistic for testing the hypothesis of complete agreement [$H_0(T)$ or H_0]. Alternatively, we may use the (approximate) F statistic appropriate to the (asymptotic) multivariate normal situation [see Rao (1965, Section 8d.3)],

$$\phi_N \equiv \{(N - v - 1)/[(N-2)v]\}\gamma_N \approx F_{v,\,N-v-1},$$

for γ_N based on the (generalized) inverse of $T\hat{\Sigma}_P T'$ (appropriate for the case of testing H_0). In this way we take into account the sampling error of estimating Σ.

The matrix T has columns $t_1, \ldots, t_k$ which in turn defines a mapping from Ω to $\mathbb{R}^s$ as follows: for $\omega \in \Omega$

$$t(\omega) = t_j \quad \text{if } \omega = \omega_j.$$

In Section 3 we have seen how to construct $t(\omega)$ for $T = T_S$, T_K or T_F. The matrix $T\hat{\Sigma}_i T'$ is simply the estimate of the $(s \times s)$ covariance matrix based on the vectors

$$\{t(X_{ij}): j = 1, \ldots, n_i\};$$

in other words,

$$T\hat{\Sigma}_i T' = (n_i - 1)^{-1} \sum_{j=1}^{n_i} \big(t(X_{ij}) - \bar{t}_{i\cdot}\big)\big(t(X_{ij}) - \bar{t}_{i\cdot}\big)' \tag{4.7}$$

where

$$\bar{t}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} t(X_{ij}). \tag{4.8}$$

These equations illustrate how the computations may be done in $\mathbb{R}^s$, with matrices of dimension $(s \times s)$, and there being no need to handle the $(k \times k)$ matrices $\hat{\Sigma}_i$. Even for $r = 10$ and $T = T_K$, we have $s = \binom{r}{2} = 45$—leading to matrix inversions well within the capabilities of today's mini-computers.

In terms of the mappings t we may also deal with the problem of ties. The idea is to extend the definition of t to the domain of possibly tied rankings in a linear way. For example, for the tied ranking

$$\zeta = (1\ 4\ 2.5\ 2.5\ 5)$$

we may consider it as the average of

$$\eta = (1\ 4\ 2\ 3\ 5)$$
$$\mu = (1\ 4\ 3\ 2\ 5)$$

and define

$$t(\zeta) = \tfrac{1}{2}(t(\eta) + t(\mu)). \tag{4.9}$$

For $t = t_S$ we simply have

$$t_S(\zeta) = (-2\ 1\ -0.5\ -0.5\ 2)',$$

whereas for $t = t_K$

$$t_K(\zeta) = (1\ 1\ 1\ 1\ -1\ -1\ 1\ 0\ 1\ 1)',$$

with the zero (0) deriving from (4.9) and coinciding with the natural extension of (3.10).

The fact that the sample space and hence the appropriate π vector is much enlarged by allowing tied rankings has no impact on the calculations outlined in (4.7) and (4.8) once the t mapping is appropriately extended. In this way we also induce an extension of the distance δ and the matrix Δ to the new sample space. It will, however, follow that the diagonal elements of the extended Δ matrix will

be positive corresponding to rankings with ties; that is, a ranking with ties is not at "distance" zero from itself! Although slightly disturbing at first, it is possible and thought-provoking to justify such a phenomenon in terms of comparing two judges who give identical rankings with or without ties. We leave the reader to ponder this aspect.

In the following section, we illustrate the application of the above ideas to various data sets. We focus attention on the analysis for $T = T_K$ and $T = T_S$.

5. Examples.

5.1 Sutton's data. Hollander and Sethuraman (1978) present data of C. Sutton on leisure time preferences for one group of 14 white females and one group of 13 black females. The analysis is summarized in Table 1.

The apportionment of diversity shows that 31% for the Spearman metric (and 27% for the Kendall metric) of the diversity is between groups. The percentages represent a sizeable proportion of the total diversity in the combined group of 27 women. The coefficient of intergroup concordance $\hat{\rho}$ indicates very high relative concordance of 0.97 between the groups for the Spearman case but only 0.64 for the Kendall case. In other words, intergroup concordance is not so high if one uses the more sensitive Kendall distance.

The significance tests (approximate) all lead to significant values, with the F statistics having "p values" less than 0.2%. We note that the analysis based on the combined sample estimate leads to more conservative values. This effect can be explained by the fact that the differences between the groups are blurred since the estimated covariance matrix is larger (in the ordering of positive definiteness). If one is testing H_0: $\pi_1 = \pi_2$ then the "pooled" estimate is preferable.

TABLE 1
Analysis of concordance: Sutton's data—black / white females

	Spearman		**Kendall**	
Apportionment				
Within	0.88	$(0.69 = \hat{\alpha})$	1.51	$(0.73 = \hat{\alpha})$
Between	0.41		0.54	
Total	1.29		2.05	
Kraemer's $\hat{\rho}$		0.97		0.64
Testing H_0, $H_0(T)$				
	$\chi^2(df)$	$F^1(df)$	$\chi^2(df)$	$F^1(df)$
Separate[2]	28.0(2)	12.8(2, 11)[3]	28.1(3)	7.8(3, 10)[3]
Pooled[2]	28.5(2)	13.7(2, 24)	28.5(3)	8.7(3, 23)
Combined[2]	13.8(2)	×	13.9(3)	×

2 groups; $n_1 = 14$, $n_2 = 13$; $N = 27$; $r = 3$.

[1]The F statistic is Hotelling's T^2 statistic (see Section 4).

[2]See the discussion of estimating $\hat{\Sigma}$ in Section 4.

[3]Approximate (conservative) F approximation using $\min(n_1, n_2) - 1$ for degrees of freedom of estimate $\hat{\Sigma}_S$.

TABLE 2
Analysis of concordance: Latrobe Valley data—male / female residents

	Spearman		Kendall	
Apportionment				
Within	28.41	$(0.98 = \hat{\alpha})$	20.81	$(0.99 = \hat{\alpha})$
Between	0.46		0.29	
Total	28.86		21.10	
Kraemer's $\hat{\rho}$		0.997		0.96
Testing H_0, $H_0(T)$				
	$\chi^2(df)$	$F^1(df)$	$\chi^2(df)$	$F^1(df)$
Separate[2]	7.5(7)	0.9(7, 40)[3]	40.7(28)	0.6(28, 19)[3]
Pooled[2]	7.5(7)	1.0(7, 87)	40.6(28)	1.0(28, 66)
Combined[2]	7.0(7)	×	28.6(28)	×

2 groups; $n_1 = 47$, $n_2 = 48$; $N = 95$; $r = 8$.
[1]As for Table 1.
[2]As for Table 1.
[3]As for Table 1.

In this example there is some contradiction between the diversity analysis and the intergroup concordance result (at least for the Spearman case). The former indicates group differences whereas the latter indicates high intergroup agreement. The same is not true for the Kendall metric. An explanation is that Kraemer's $\hat{\rho}$ depends on the maximum possible diversity (β) and if this is far from being attained within or in total then $\hat{\rho}$ will be large. That $\hat{\rho}$ seems to be so sensitive to the metric used is a disadvantage compared to $\hat{\alpha}$. Alternatively, one may claim that the intergroup concordance based on Kendall's metric is more meaningful than that originally proposed by Kraemer. This claim is borne out by other examples which were investigated and in which the between group differences are significant.

5.2 Latrobe Valley data. Residents in the Latrobe Valley of Victoria, Australia were asked to rank eight sectors in order of the degree that they would be affected by proposed industrial developments. Of the 95 respondents, 47 were male and 48 were female and the researcher wanted to know if there was a difference due to sex.

The results of the concordance analysis (Table 2) show quite unequivocally that there is no difference in the rankings of the eight sectors between males and females. The ability to cross-check such conclusions by using two (or more) metrics is one of the main contributions of the proposed methodology.

6. Conclusions. We have approached the comparison of groups of rankers from the point of view of analysing diversity based on various metrics for ranks. We have shown how this approach is related to others based on measuring intergroup concordance as well as interpreting the analysis as the comparison of population (or model) characteristics.

A class of descriptive and test statistics have been proposed to allow the researcher to quantify the proportion of diversity ascribed to differences between

groups as well as to test hypotheses of equality of populations or of their characteristics.

Experience with examples indicates that the new types of statistics proposed as well as the extension of intergroup concordance originally proposed by Kraemer (1981) provide useful extra information for comparing groups of rankers.

Acknowledgment. The data is available from the first author. We acknowledge the researchers of the Division of Building Research at CSIRO Victoria, Australia, who collected and made the data available to us.

REFERENCES

ALVO, M., CABILIO, P. and FEIGIN, P. D. (1982). Asymptotic theory for measures of concordance with special reference to average Kendall tau. *Ann. Statist.* **10** 1269–1276.

DIACONIS, P. (1985). *Group Theory in Statistics.* IMS Lecture Notes–Monograph Series. Forthcoming.

DIACONIS, P. and GRAHAM, R. L. (1977). Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B* **39** 262–268.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS–NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

FEIGIN, P. D. and COHEN, A. (1978). On a model for concordance between judges. *J. Roy. Statist. Soc. Ser. B* **40** 203–213.

HOLLANDER, M. and SETHURAMAN, J. (1978). Testing for agreement between two groups of judges. *Biometrika* **65** 403–411.

KATZ, B. M. and MCSWEENEY, M. (1981). Some tests for ranked data in repeated measures multi-group designs. *Amer. Statist. Assoc. Proc. Soc. Statist. Sec.* 476–481.

KRAEMER, H. C. (1981). Intergroup concordance: definition and estimation. *Biometrika* **68** 641–644.

LI, L. and SCHUCANY, W. R. (1975). Some properties of a test for concordance of two groups of rankings. *Biometrika* **62** 417–423.

LINHART, H. (1960). Approximate test for m rankings. *Biometrika* **47** 476–480.

MALLOWS, C. L. (1957). Non-null ranking models I. *Biometrika* **44** 114–130.

MIELKE, P. W., BERRY, K. J., BROCKWELL, P. J. and WILLIAMS, J. S. (1981). A class of nonparametric tests based on multiresponse permutation procedures. *Biometrika* **68** 720–724.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, Mass.

PATIL, G. P. and TAILLIE, C. (1982). Diversity as a concept and its measurement. *J. Amer. Statist. Assoc.* **77** 548–561.

PETTITT, A. N. (1982). Parametric tests for agreement amongst groups of judges. *Biometrika* **69** 365–375.

RAO, C. R. (1982a). Diversity and dissimilarity coefficients: A unified approach. *J. Theoret. Pop. Biol.* **21** 24–43.

RAO, C. R. (1982b). Gini–Simpson index of diversity: a characterization generalization and applications. *Utilitas Math.* **21B** 273–282.

SCHUCANY, W. R. and FRAWLEY, W. H. (1973). A rank test for two group concordance. *Psychometrika* **38** 249–258.

FACULTY OF INDUSTRIAL ENGINEERING
AND MANAGEMENT
TECHNION
HAIFA 32000
ISRAEL

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OTTAWA
OTTAWA, CANADA