Ann. Inst. Statist. Math.
Vol. 45, No. 2, 383–399 (1993)

RELATIVE DIFFERENCE IN DIVERSITY BETWEEN POPULATIONS

KHURSHEED ALAM AND CALVIN L. WILLIAMS

Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-1907, U.S.A.

(Received June 15, 1991; revised May 26, 1992)

Abstract. An entropy is conceived as a functional on the space of probability distributions. It is used as a measure of diversity (variability) of a population. Cross entropy leads to a measure of dissimilarity between populations. In this paper, we provide a new approach to the construction of a measure of dissimilarity between two populations, not depending on the choice of an entropy function, measuring diversity. The approach is based on the principle of majorization which provides an intrinsic method of comparing the diversities of two populations. We obtain a general class of measures of dissimilarity and show some interesting properties of the proposed index. In particular, it is shown that the measure provides a metric on a probability space. The proposed measure of dissimilarity is essentially a measure of relative difference in diversity between two populations. It satisfies an invariance property which is not shared by other measures of dissimilarity which are used in ecological studies. A statistical application of the new method is given.

Key words and phrases: Diversity, dissimilarity, cross entropy, majorization, Schur-convexity, ranking and selection.

1. Introduction

Diversity is a generic term used for variation in the data set (population). There is an extensive literature on the measurement and analysis of diversity. A comprehensive bibliography of papers on this subject has been compiled by Dennis *et al.* (1979). Patil and Taillie (1982) have given an interesting exposition of the concept of diversity. Rao (1982*a*, 1982*b*) has developed a unified approach to the measurement and analysis of diversity based on entropy functions. He has introduced a new measure of diversity, called the quadratic entropy, which is well-suited for studying diversity. Some natural conditions which a diversity measure should satisfy imply that it must have certain convexity properties (Rao (1984)). The convexity property leads to a meaningful measure of dissimilarity between populations.

The Gini-Simpson index (GS), due to Gini (1912) and Simpson (1949), and Shannon's entropy (SE) due to Shannon (1948) are two well-known measures of

diversity of a multinomial population π, given by

$$\text{GS} = 1 - \sum_{i=1}^{k} p_i^2 \quad \text{and} \quad \text{SE} = -\sum_{i=1}^{k} p_i \log p_i$$

where $p_1, \ldots, p_k$ denote the cell probabilities associated with π and $\sum_{i=1}^{k} p_i = 1$. More generally, we conceive of diversity as a functional on the space of probability distributions. Let $\mathcal{P}$ be a convex set of probability measures defined on a measure space $(\mathcal{X}, \mathcal{B})$, and let $P \in \mathcal{P}$. The quadratic entropy due to Rao, is defined as

$$Q(P) = \int K(x,y) dP(x) dP(y) \tag{1.1}$$

where $K(x,y)$ is a measurable kernel defined on $\mathcal{X} \times \mathcal{X}$, satisfying the condition (conditionally negative definite) that

$$\sum_i \sum_j K(x_i, x_j) a_i a_j \leq 0 \tag{1.2}$$

for all $x_1, \ldots, x_n \in \mathcal{X}$ and all real numbers $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$. If, for example, $\mathcal{X} = \mathcal{R}^1$, $K(x_1, x_2) = (x_1 - x_2)^2$ and $P = N(\mu, \sigma^2)$ then $Q(P) = 2\sigma^2$ is the variance functional of P. For another example, let $\mathcal{X}$ be a space of k points, P be the probability measure associated with a multinomial population π and $K(x_1, x_2) = 0(1)$ if $x_1 = (\neq) x_2$. Then (1.1) leads to the Gini-Simpson index GS. Letting $K(x_1, x_2)$ be a measure of difference between x_1 and x_2, a motivation for the expression (1.1) is that it represents the average difference between two individuals drawn at random from the population specified by the probability measure P.

Let H be a diversity measure defined on $\mathcal{P}$. It is a natural requirement that H be concave on $\mathcal{P}$, since the diversity of a mixed population should not be smaller than the average of the diversities within the individual populations. Let $P_1, \ldots, P_k \in \mathcal{P}$ with prior probabilities $\lambda_1, \ldots, \lambda_k$, where $\sum_{i=1}^{k} \lambda_i = 1$. Consider the difference between the diversity of the mixed population and the average diversity within populations, given by

$$D_H(P_1, \ldots, P_k) = H(\bar{P}) - \sum_{i=1}^{k} \lambda_i H(P_i) \tag{1.3}$$

where $\bar{P} = \sum_{i=1}^{k} \lambda_i P_i$. The difference is non-negative if H is concave on $\mathcal{P}$. It is a measure of overall differences among the probability measures P_i. Rao (1982a) has called it the *Jensen difference*. For $k = 2$ and $\lambda_1 = \lambda_2 = 1/2$, we have

$$D_H(P_1, P_2) = H\left(\frac{P_1 + P_2}{2}\right) - \frac{1}{2}H(P_1) - \frac{1}{2}H(P_2). \tag{1.4}$$

We shall refer to $D_H(P_1, P_2)$ as the dissimilarity between P_1 and P_2, induced by the diversity measure H. The dissimilarity measure is non-negative and symmetric in its arguments and vanishes when $P_1 = P_2$.

For example, consider the Gini-Simpson index GS. The dissimilarity between two multinomial populations π and π' is given by $D_{\mathrm{GS}}(\pi, \pi') = (1/4)\sum_{i=1}^{k}(p_i - p_i')^2$ where p_i (p_i') denotes the cell probabilities associated with π (π'). For another example, consider the quadratic entropy given by (1.1) with $K(x_1, x_2) = (x_1 - x_2)^2$, and let $P_1 = N(\mu_1, \sigma^2)$ and $P_2 = N(\mu_2, \sigma^2)$. The dissimilarity between P_1 and P_2 is given by $D_Q(P_1, P_2) = (\mu_1 - \mu_2)^2/2$.

The concavity property of the diversity measure H enables us to apportion the total diversity in a population as due to differences between and within populations. More generally, we are given a number of populations grouped in hierarchical classifications. Given the distributions within populations and their *a priori* probabilities, we are required to compute the average diversity within groups at different levels of classification. In this case, we require that $D_H(P_1, P_2)$ be convex on $\mathcal{P}^2$. A higher order convexity property is required for higher order classifications. Rao (1984) has shown that the dissimilarity measure $D_H(P_1, P_2)$ is (completely) convex if H is a quadratic entropy, given by (1.1) and (1.2). A brief review of different measures of similarity and dissimilarity between populations is given by Gower (1985).

In this paper, we introduce a new measure of dissimilarity between two distributions. We shall deal with the analysis of categorical data. Therefore, we shall consider only multinomial populations. Our approach is based on the principle of majorization which provides an intrinsic criterion for comparing the diversities of two multinomial populations, π and π', say, with the associated probability vectors p and p'. We shall denote the dissimilarity between π and π' by $\tilde{\rho}(p, p')$. It will be seen that $\tilde{\rho}(p, p')$ is invariant under separate rearrangements of the components of p and p'. This is generally not true for the dissimilarity measure given by (1.4). The definition of $\tilde{\rho}(p, p')$ is given in Section 2, along with a preliminary discussion of the theory of majorization. Certain properties of the dissimilarity measure are given in Section 3. The definition of $\tilde{\rho}(p, p')$ is based on a choice of a vector norm or a matrix norm induced by a vector norm. Specifically, we consider the L_1 and L_2 norms. We denote the corresponding dissimilarity measures by $\tilde{\rho}(L_1)$ and $\tilde{\rho}(L_2)$ for the vector norm and by $\tilde{\rho}(IL_1)$ and $\tilde{\rho}(IL_2)$ for the induced vector norm. In Section 4, we consider an application of the new methodology to a problem of ranking and selection.

In order to motivate the reader at this stage, we present below the following data on the Danish educational aspirations of adolescent boys and girls classified in five categories. The data in Table 1 are reproduced from Table 1 of Light and Margolin (1971). We have computed from the data the dissimilarity between boys and girls (with respect to their educational aspirations), using several measures of dissimilarity. The summary statistics are shown in Table 2. We have included in the summary statistics the figures for the maximum dissimilarity between the uniform distribution (equi-probable categories) and the degenerate distribution (single category).

The standard chi-square analysis for studying the association between sex and educational aspiration yields a value of $\chi_4^2 = 47.0$ which is highly significant. However, we are studying here the dissimilarity between boys and girls with respect to the diversity in their educational aspirations. The summary statistics show

Table 1. Danish educational aspirations.

	Boys		Girls	
Educational Aspirations	Number	Proportion	Number	Proportion
Secondary school	62	0.1902	61	0.2096
Vocational training	121	0.3712	149	0.5120
Teacher college	26	0.0798	41	0.1410
Gymnasium	33	0.1012	20	0.0687
University	84	0.2576	20	0.0687
Total	326		291	

Table 2. Summary statistics.

	Gini-Simpson (GS)			Shannon's entropy (SE)		
	Boys	Girls	Uniform	Boys	Girls	Uniform
Diversity	0.7431	0.6646	0.8000	1.4665	1.3144	1.6094
	$\tilde{\rho}(L_1)$	$\tilde{\rho}(IL_1)$	$\tilde{\rho}(L_2)$	$\tilde{\rho}(IL_2)$	D_{GS}	D_{SE}
Dissimilarity between						
A: Boys-girls	0.2817	0.2882	0.1578	0.1746	0.0152	0.0409
B: Uniform-degenerate	1.6000	2.0000	0.8944	1.0955	0.2000	0.4228
Ratio A/B	0.1761	0.1441	0.1764	0.1594	0.0759	0.0967

that the diversity in educational aspirations is quite large both for boys and girls. But the diversity is slightly larger for boys, compared to girls. The dissimilarity in diversity between boys and girls is also substantial, equal to about 18% of the *maximal dissimilarity* between the uniform and the degenerate distributions, based on the measures $\tilde{\rho}(L_1)$ and $\tilde{\rho}(L_2)$. The ratio is between 14% and 16%, based on $\tilde{\rho}(IL_1)$ and $\tilde{\rho}(IL_2)$. However, the ratio is less than 10%, based on the measures D_{GS} and D_{SE}.

2. Dissimilarity measure $\tilde{\rho}$

A scalar index of diversity, such as an entropy function, imposes a linear ordering of populations with respect to their diversities. Two different indices may give different orderings. Therefore, we look for an intrinsic method of ordering diversity. One such method is based on the principle of majorization. An excellent presentation of the theory of majorization is given in the textbook by Marshall and Olkin (1979). In the following we shall use the short notation MAO for a reference to the book. Consider two multinomial populations π and π' with the associated probability vectors $p = (p_1, \ldots, p_k)$ and $p' = (p'_1, \ldots, p'_k)$ respectively, where $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} p'_i$. Let $p_{(1)} \leq \cdots \leq p_{(k)}$ and $p'_{(1)} \leq \cdots \leq p'_{(k)}$, denote the

ordered values of the components of p and p'. We say that p is *majorized* by p' ($p \prec p'$) if

$$\sum_{i=1}^{m} p_{(i)} \geq \sum_{i=1}^{m} p'_{(i)}, \qquad m = 1, \ldots, k-1. \tag{2.1}$$

The term *domination* is sometimes used for majorization. The criterion of majorization is used for comparing the diversities of two populations. We say that π is more diverse than π' if $p \prec p'$. Since $(1/k, \ldots, 1/k) \prec p \prec (1, 0, \ldots, 0)$ the uniform population is most diverse and a degenerate population is least diverse. An interesting application of majorization principle is given by Lorenz (1905). Consider a population of n individuals. Let $x = (x_1, \ldots, x_n)$ represent the wealth of individuals for the distribution of a total wealth W, and let $y = (y_1, \ldots, y_n)$ represent another distribution of the same total wealth. Let $x_{(i)}$ and $y_{(i)}$ denote the i-th smallest values among the components of x and y, respectively. According to Lorenz, x represents a more even distribution of wealth than y if and only if $\sum_{i=1}^{j} x_i \geq \sum_{i=1}^{j} y_i$, $j = 1, \ldots, n-1$. Of course, $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$. That is, the population associated with the distribution x is more even (diverse) than the population associated with the distribution y (equivalently, the distribution y is more concentrated than the distribution x) if and only if $x \prec y$.

A function ϕ which is order-preserving with respect to the majorization relation is called *Schur-convex* in honor of Schur (1923) who was the first to study such functions. That is, ϕ is Schur-convex if $\phi(x) \leq \phi(y)$ whenever $x \prec y$. If the reverse inequality holds, ϕ is called *Schur-concave*. Schur-concave functions have been used as indices of diversity. The Gini-Simpson index GS and Shannon's entropy SE, which have been mentioned in the preceding section are both Schur-concave functions. A Schur-concave function is a suitable choice for an index of diversity because of its isotonic property with respect to the majorization relation. However, the majorization relation is a partial ordering. We may have p and p' such that neither $p \prec p'$ nor $p' \prec p$ is true. In this case, π and π' are not comparable with respect to their diversities via majorization. We can find two Schur-concave functions ϕ and ψ for which $\phi(p) > \phi(p')$ and $\psi(p) < \psi(p')$. That is, π would be considered more diverse than π' if ϕ were used for an index of diversity. On the other hand, π' would be considered more diverse than π if ψ were used for the index of diversity. How should we compare the diversities of π and π' in this situation? We do not know a satisfactory answer to this question.

Now, we introduce the proposed measure of dissimilarity $\tilde{\rho}$ between the two populations π and π'. It is known that $p \prec p'$ if and only if there exists a $k \times k$ doubly stochastic matrix Q such that $p = Qp'$ (MAO, Theorem 2 B. 2.). Let $\mathcal{D}$ denote the set of all $k \times k$ doubly stochastic matrices. It is known that $\mathcal{D}$ is the convex hull of permutation matrices (MAO, Theorem 2 A. 2.). Conceptually, the diversity associated with p is invariant under a rearrangement of the components of p. We shall write $p \approx p'$ if $p = Pp'$, where P is a permutation matrix. Let $\|\cdot\|$ be a vector norm on $\mathcal{R}^k$. We shall assume that the vector norm is symmetric in the components of the vector. Let

$$\rho(p, p') = \inf_{Q \in \mathcal{D}} \|p - Qp'\| \tag{2.2}$$

and

$$\tilde{\rho}(p,p') = \rho(p,p') + \rho(p',p) \tag{2.3}$$

Note that $\tilde{\rho}(\cdot,\cdot)$ is symmetric in its arguments. Clearly, $\rho(p,p') = 0$ if $p \prec p'$. Moreover, $\rho(p,p') = 0 \Rightarrow p \prec p'$, since the norm is a continuous function and the set $\mathcal{D}$, denoting the convex hull of permutations matrices in the definition (2.2) is compact. Hence $\tilde{\rho}(p,p') = 0$ if and only if $p \prec p'$ and $p' \prec p$. Now, $p \prec p'$ and $p' \prec p$ if and only if $p = Pp'$, where P is a permutation matrix. That is $p \approx p'$. Therefore, $\tilde{\rho}(p,p') = 0$ if and only if $p \approx p'$. We consider $\tilde{\rho}$ as a measure of dissimilarity in diversity between the populations π and π'. Observe that $\tilde{\rho}(p,p') = \|p - p'\|$ if either p or p' represents the uniform distribution. Let $\xi = (1,0,\ldots,0)$ and let $\eta = (1/k,\ldots,1/k)$, then $\tilde{\rho}(\xi,\eta) = \|\xi - \eta\|$ represents the maximal dissimilarity between two multinomial populations.

We have observed that $\rho(p,p') = 0$ if and only if $p \prec p'$. Therefore, $\rho(p,p')$ is a measure of deficiency in the majorization relation $p \prec p'$. Let π, π' and π'', be three multinomial populations with the associated probability vectors p, p' and p'', respectively. It can be shown that $\rho(p,p'') \leq \rho(p',p'')$ if $p \prec p'$. That is, p is *closer* to being majorized by p'' than p' is to being majorized by p''. Similarly, we can show that $\rho(p,p'') \leq \rho(p,p')$ if $p' \prec p''$. That is p is closer to being majorized by p'' than p is to being majorized by p'. Since $\rho(p,p')$ is a symmetric function of the components of p and p', it follows that $\rho(p,p')$ is a Schur-convex (Schur-concave) function of $p(p')$. But $\rho(p,p')$ is not symmetric in its arguments. The proposed measure of dissimilarity $\tilde{\rho}(p,p')$ is obtained by symmetrizing $\rho(p,p')$.

In ecological studies, species composition is typically analyzed using various measures of similarity, which have been proposed in the literature. We have mentioned above a class of diversity measures, known as the generalized quadratic entropy, due to Rao (1982*a*, 1982*b*, 1984). The quadratic entropy provides a decomposition of diversity in the same way as variance is decomposed in the analysis of variance (ANOVA). The quadratic entropy is essentially based on the concept of a distance or dissimilarity. Smith (1989) has shown that a class of ecological similarity measures, called the expected species shared (ESS), leads to a partition of similarity, as the quadratic entropy leads to a decomposition of diversity. See also Smith *et al.* (1979) and Grassle and Smith (1976). A salient feature of the dissimilarity measure $\tilde{\rho}(p,p')$, we have proposed, is that it is invariant under separate (not necessarily the same) rearrangement of the components of p and p'. This property is not shared by any of the similarity measures which are used in ecological studies, such as the ESS. We elaborate this point below.

In the example, given above, relating to the Danish educational aspirations of adolescent boys and girls, there are dissimilarities between the educational aspirations of boys and girls as well as between the diversity of their aspirations. However, the index $\tilde{\rho}$ is a measure of relative difference in diversities of two populations. We may construct an example where the multinomial probabilities are $p = (0.4, 0.3, 0.2, 0.1)$ for boys and $p' = (0.1, 0.2, 0.3, 0.4)$ for girls. In this example, boys and girls have the same diversity. But, whereas $\tilde{\rho}(p,p') = 0$ shows no dissimilarity between them, the Jensen difference, given by (1.4), assigns a positive measure of dissimilarity. Naturally, boys and girls have different preferences.

The Jensen difference takes into account the difference in preference for individual categories (aspirations), whereas $\tilde{\rho}$ takes into account only the difference in the apportionment of the preferences, disregarding the nature of the categories.

Let $p^*(p'^*)$ be obtained from $p(p')$ by adjoining a number of null multinomial classes. Clearly, $p \prec p' \Leftrightarrow p^* \prec p'^*$. Therefore, the definition of $\tilde{\rho}(p,p')$ may be generalized trivially to include the cases where p and p' are associated with unequal number of multinomial categories.

3. Properties of $\tilde{\rho}$

Let Q be a doubly stochastic matrix, and let $P_1,\dots,P_N$ denote the N permutation matrices, where $N = k!$. Since Q lies inside the convex hull of $P_1,\dots,P_N$, we have that $Q = \lambda_1 P_1 + \cdots + \lambda_N P_N$ where $\lambda_i \geq 0$, $i = 1,\dots,N$ and $\sum_{i=1}^{N} \lambda_i = 1$. Hence

$$\begin{aligned} \|Qp\| &\leq \lambda_1\|P_1 p\| + \cdots + \lambda_N\|P_N p\| \\ &= (\lambda_1 + \cdots + \lambda_N)\|p\| = \|p\|. \end{aligned} \tag{3.1}$$

Consider three multinomial probability vectors p, p' and p''. Let

$$\begin{aligned} \rho(p,p') &= \|p - Q_1 p'\| \\ \rho(p',p'') &= \|p' - Q_2 p''\| \end{aligned}$$

where Q_1 and Q_2 are doubly stochastic matrices. Since $Q_1 Q_2$ is a doubly stochastic matrix, we have

$$\begin{aligned} \rho(p,p'') &= \inf_{Q\in\mathcal{D}} \|p - Qp''\| \\ &\leq \|p - Q_1 Q_2 p''\| \\ &\leq \|p - Q_1 p'\| + \|Q_1(p' - Q_2 p'')\| \quad \text{by the triangle inequality} \\ &\leq \|p - Q_1 p'\| + \|p' - Q_2 p''\| \quad \text{by (3.1)} \\ &= \rho(p,p') + \rho(p',p''). \end{aligned} \tag{3.2}$$

Similarly, $\rho(p'',p) \leq \rho(p',p) + \rho(p'',p')$. Hence, $\tilde{\rho}(p,p'') \leq \tilde{\rho}(p,p') + \tilde{\rho}(p',p'')$. That is, $\tilde{\rho}$ satisfies the triangle inequality. The results given above show that

THEOREM 3.1. *$\tilde{\rho}(\cdot,\cdot)$ is a metric on the simplex $\{p : p_i \geq 0, i = 1,\dots,k, \sum_{i=1}^{k} p_i = 1\}$.*

3.1 *Isotonic property*

The dissimilarity measure $\tilde{\rho}$ is consistent with the majorization criterion in the following way.

THEOREM 3.2. *If $p \prec p' \prec p''$ then $\tilde{\rho}(p',p'') \leq \tilde{\rho}(p,p'')$ and $\tilde{\rho}(p,p') \leq \tilde{\rho}(p,p'')$.*

PROOF. Suppose that $p \prec p' \prec p''$. Let $p = Rp'$ and $p' = Sp''$, where R and S are doubly stochastic matrices. We have

$$\begin{aligned}
\tilde{\rho}(p', p'') &= \rho(p'', p'), \quad \text{since} \quad p' \prec p'' \\
&= \inf_{Q \in \mathcal{D}} \|p'' - Qp'\| \\
&\leq \inf_{Q \in \mathcal{D}} \|p'' - QRp'\|, \quad \text{since } QR \text{ is doubly stochastic} \\
&= \inf_{Q \in \mathcal{D}} \|p'' - Qp\| \\
&= \rho(p'', p) \\
&= \tilde{\rho}(p, p''), \quad \text{since} \quad p \prec p''.
\end{aligned}$$

Next, we have

$$\begin{aligned}
\tilde{\rho}(p, p'') &= \rho(p'', p) \\
&= \inf_{Q \in \mathcal{D}} \|p'' - Qp\| \\
&\geq \inf_{Q \in \mathcal{D}} \|Sp'' - SQp\|, \quad \text{by (3.1)} \\
&= \inf_{Q \in \mathcal{D}} \|p' - SQp\| \\
&\geq \inf_{Q \in \mathcal{D}} \|p' - Qp\| \\
&= \rho(p', p) \\
&= \tilde{\rho}(p, p'), \quad \text{since} \quad p \prec p'.
\end{aligned}$$

3.2 *Mixture of populations*

Let π' be a mixture of n multinomial populations with the associated probability vectors $q_1, \ldots, q_n$, say. That is,

$$p' = \sum_{i}^{n} \lambda_i q_i \tag{3.3}$$

where $\lambda_i \geq 0$, $i = 1, \ldots, n$ and $\sum_i^n \lambda_i = 1$. Assume that the components of p' and each q_i have been arranged in the same order of magnitude. It is natural to require that the dissimilarity between π and π' should not be greater than the average of the dissimilarities between π and the mixing components of π'. That this condition is satisfied by the measure $\tilde{\rho}$, is shown by the following theorem.

THEOREM 3.3. *Let p' be given by* (3.3). *Then* $\tilde{\rho}(p, p') \leq \sum_i^n \lambda_i \tilde{\rho}(p, q_i)$.

PROOF. Let $Q_1, \ldots, Q_n$ be doubly stochastic matrices, given by

$$\rho(q_i, p) = \inf_{Q_i \in \mathcal{D}} \|q_i - Q_i p\|, \quad i = 1, \ldots, n.$$

Since $\sum_i^n \lambda_i Q_i$ is a doubly stochastic matrix, we have

$$\begin{aligned}\rho(p',p) &= \inf_{Q\in\mathcal{D}} \|p' - Qp\| \\ &\le \left\| \sum_i^n \lambda_i q_i - \sum_i^n \lambda_i Q_i p \right\| \\ &\le \sum_i^n \lambda_i \|q_i - Q_i p\| \\ &= \sum_i^n \lambda_i \rho(q_i, p).\end{aligned} \tag{3.4}$$

Let $R_1, \ldots, R_n$ be doubly stochastic matrices, given by

$$\rho(p, q_i) = \inf_{R_i\in\mathcal{D}} \|p - R_i q_i\|, \qquad i = 1, \ldots, n.$$

Since $R_i q_i \prec q_i$, it follows from (2.1) that $\sum_i^n \lambda_i R_i q_i \prec \sum_i^n \lambda_i q_i$. Hence

$$\begin{aligned}\rho(p,p') &= \inf_{Q\in\mathcal{D}} \left\| p - Q\left(\sum_i^n \lambda_i q_i\right) \right\| \\ &\le \left\| p - \sum_i^n \lambda_i R_i q_i \right\| \\ &\le \sum_i^n \lambda_i \|p - R_i q_i\| \\ &= \sum_i^n \lambda_i \rho(p, q_i).\end{aligned} \tag{3.5}$$

From (3.4) and (3.5) we have that

$$\begin{aligned}\tilde{\rho}(p,p') &= \rho(p,p') + \rho(p',p) \\ &\le \sum_i^n \lambda_i \rho(q_i, p) + \sum_i^n \lambda_i \rho(p, q_i) \\ &= \sum_i^n \lambda_i \tilde{\rho}(p, q_i).\end{aligned}$$

3.3 *Convexity property*

Consider two pairs of multinomial populations (π_1, π_1') and (π_2, π_2') with the associated probability vectors (p, p') and (q, q'), respectively. It is a natural requirement that the dissimilarity between the two mixed populations $\lambda p + (1-\lambda)q$ and $\lambda p' + (1-\lambda)q'$ should not be greater than the average of the dissimilarities between π_1 and π_1' and between π_2 and π_2', where $0 \le \lambda \le 1$. That is,

$$J_1 = \lambda\tilde{\rho}(p,p') + (1-\lambda)\tilde{\rho}(q,q') - \tilde{\rho}(\lambda p + (1-\lambda)q, \lambda p' + (1-\lambda)q') \ge 0. \tag{3.6}$$

The above condition implies that the dissimilarity measure $\tilde{\rho}$ is a convex function on $\pi \times \pi$, the product space of two multinomial populations. With λ denoting the prior probability of the pair (p, p'), J may be interpreted as the difference in dissimilarity between the two pairs. The norm $\|\cdot\|$ used in the definition of $\tilde{\rho}$ needs to be chosen properly so that $\tilde{\rho}$ is convex on $\pi \times \pi$.

3.4 *Measure of diversity derived from* $\tilde{\rho}$

Rao and Nayak (1985) have considered a derivation of the entropy function from cross entropy, a measure of dissimilarity. We can similarly derive a diversity measure H from the dissimilarity measure $\tilde{\rho}$, as follows. Put

$$\tilde{\rho}(p, p') = \lim_{\lambda \to 0} \frac{1}{\lambda}[H(\lambda p + (1-\lambda)p') - \lambda H(p) - (1-\lambda)H(p')] \tag{3.7}$$

where $0 < \lambda \leq 1$. Following Rao and Nayak (1985) we assume that H has the smooth differentiability property

$$H(\lambda p + (1-\lambda)p') = H(p') + \lambda f(p', p - p') + o(\lambda)$$

where $f(p', p - p')$ is linear in $p - p'$ and $f(p', 0) = 0$. Then

$$\begin{aligned} \tilde{\rho}(p, p') &= f(p', p - p') + H(p') - H(p) \\ &= f(p, p' - p) + H(p) - H(p'). \end{aligned} \tag{3.8}$$

Let e_i denote the i-th co-ordinate vector so that $p = \sum_{i=1}^{k} p_i e_i$. Since e_i represents a degenerate distribution we substitute e_i for p' in (3.8), multiply both sides by p_i and add for $i = 1, \ldots, k$. As $f(p, \cdot)$ is linear in the second argument, we have

$$\begin{aligned} \sum_{i=1}^{k} p_i \tilde{\rho}(p, e_i) &= f\left(p, \sum_{i=1}^{k} p_i (e_i - p)\right) + H(p) \\ &= f(p, 0) + H(p) \\ &= H(p). \end{aligned} \tag{3.9}$$

The solution (3.9) for the diversity measure H resembles the representation given by Haberman (1982).

Now $\tilde{\rho}(p, e_i) = \rho(e_i, p)$. Moreover, $\rho(\cdot, \cdot)$ is invariant under separate permutations of the components of its arguments. Therefore, the sum on the left side of (3.9) is equal to $\rho(e_i, p)$. Thus,

$$H(p) = \rho(e_i, p). \tag{3.10}$$

We note that $H(p)$, given here, is a Schur-concave function of p, as it should be. Using the Euclidean norm, we get

$$H(p) = 1 + p_1^2 + p_2^2 - 2\max(p_1, p_2)$$

for $k = 2$. For $k > 2$ the computation is rather involved. It would be interesting to find a norm for which (3.10) gives a specific solution, such as the Gini-Simpson measure of diversity, for example. At this point, we have not been able to find such a norm. This is the subject of an ongoing investigation.

It is interesting also to derive the diversity measure H from the Jensen difference $D_H(p, p')$. We solve for $H(p)$ from equation (1.4), as follows. Let p' be fixed and let D be a functional operator, giving

$$Df(p) = 2f\left(\frac{p+p'}{2}\right) - f(p')$$

where f is a function defined on the space of probability vectors. Let I denote the identity transform. The equation (1.4) is written as $2D_H(p, p') = (D - I)H(p)$. Hence (formally)

$$\begin{aligned} H(p) &= \frac{2}{D-I} D_H(p, p') \\ &= -2\sum_{r=0}^{\infty} D^r D_H(p, p') + C(p') \end{aligned}$$

provided that the sum exists. Here $C(p')$ denotes a function of p'. Since $D_H(p, p') = 0$, we have for $r = 0, 1, 2, \ldots$

$$\begin{aligned} D^0 D_H(p, p') &= D_H(p, p'), \\ D^1 D_H(p, p') &= 2D_H\left(\frac{p+p'}{2}, p'\right), \\ D^2 D_H(p, p') &= 4D_H\left(\frac{p}{4} + \left(\frac{1}{2} + \frac{1}{4}\right)p', p'\right), \\ D^r D_H(p, p') &= 2^r D_H\left(\frac{p}{2^r} + \left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^r}\right)p', p'\right) \\ &= 2^r D_H\left(\frac{p}{2^r} + \left(1 - \frac{1}{2^r}\right)p', p'\right). \end{aligned}$$

Hence

$$H(p) = C(p') - \sum_{r=0}^{\infty} 2^{r+1} D_H\left(\frac{p}{2^r} + \left(1 - \frac{1}{2^r}\right)p', p'\right). \tag{3.11}$$

Putting $p = e_i$ $(i = 1, \ldots, k)$ in (3.11) and averaging we get

$$C(p') = \frac{1}{k}\sum_{r=0}^{\infty} 2^{r+1} \sum_{i=1}^{k} D_H\left(\frac{e_i}{2^r} + \left(1 - \frac{1}{2^r}\right)p', p'\right) \tag{3.12}$$

since $H(e_i) = 0$. Thus

$$\begin{aligned} H(p) = \sum_{r=0}^{\infty} 2^{r+1} \frac{1}{k} \sum_{i=1}^{k} \Bigg\{ & D_H\left(\frac{e_i}{2^r} + \left(1 - \frac{1}{2^r}\right)p', p'\right) \\ & - D_H\left(\frac{p}{2^r} + \left(1 - \frac{1}{2^r}\right)p', p'\right)\Bigg\}. \end{aligned} \tag{3.13}$$

Consider, for example the Jensen difference, corresponding to the Gini-Simpson index

$$H(p) = 1 - |p|^2 \tag{3.14}$$

as given by

$$D_H(p, p') = \frac{1}{4}|p - p'|^2$$

where $|\cdot|$ denotes the Euclidean norm. From the formula (3.13) we get $H(p) = C(p') - |p - p'|^2$ where

$$\begin{aligned} C(p') &= \frac{1}{4k} \sum_{r=0}^{\infty} 2^{r+1} \sum_{i=1}^{k} \left| \frac{e_i - p'}{2^r} \right|^2 \\ &= \frac{k-2}{k} + |p'|^2. \end{aligned}$$

Put $p' = (1/k, \ldots, 1/k)$. Then $H(p) = 1 - |p|^2$ as given by (3.14).

We can derive similarly a diversity measure $H(p)$ from the dissimilarity measure $\tilde{\rho}(p, p')$ by substituting $\tilde{\rho}(p, p')$ for $D_H(p, p')$ in (3.13). Thus we have

$$\begin{aligned} H(p) = \sum_{r=0}^{\infty} 2^{r+1} \frac{1}{k} \sum_{i=1}^{k} \Bigg\{ &\tilde{\rho}\left(\frac{e_i}{2^r} + \left(1 - \frac{1}{2^r}\right) p', p' \right) \\ &- \tilde{\rho}\left(\frac{p}{2^r} + \left(1 - \frac{1}{2^r}\right) p', p' \right) \Bigg\}. \end{aligned} \tag{3.15}$$

3.5 *Vector norm associated with* $\tilde{\rho}$

The measure of dissimilarity $\tilde{\rho}$ has been defined above with respect to any vector norm $\|\cdot\|$. If $\|\cdot\|$ is the $L_1(L_2)$ norm then the value of $\rho(p, p')$ is computed by the linear (quadratic) programming technique. A number of efficient algorithms are known for this programming method (Kostreva (1989)). It is interesting to consider a special class of norms, given as follows. Let $A = (a_{ij})$ be a lower triangular matrix, given by

$$a_{ij} = \begin{cases} 1, & 1 \le j \le i \le k \\ 0, & j \ge i. \end{cases}$$

Let $|\cdot|$ be any vector norm on $\mathcal{R}^k$ and let $\|\cdot\|$ denote the induced vector norm, given by

$$\|x\| = |Ax|. \tag{3.16}$$

Let T denote the subset of the simplex $\{p : 0 \le p_1 \le p_2 \le \cdots \le p_k, \sum_{i=1}^{k} p_i = 1\}$. From the definition of majorization it follows that for $p, p' \in T$

$$p \prec p' \Leftrightarrow Ap \ge Ap'.$$

Here $\geq$ means componentwise inequality.

We shall assume that $|x|$ is nondecreasing in the absolute values of the components of x in $\mathcal{R}^k$. This is true for the L_p norm, for example. Let $p, p' \in T$, q_i and q_i' denote the i-th component of Ap and Ap', respectively, $m_i = \max(q_i, q_i')$ and $m = (m_1, \ldots, m_k)'$. It is easily seen that with the vector norm $\|\cdot\|$ defined as above

$$
\begin{aligned}
\rho(p, p') &= |m - Ap|, \\
\rho(p', p) &= |m - Ap'|, \\
\tilde{\rho}(p, p') &= |m - Ap| + |m - Ap'| \\
&= \|p - p'\|^+ + \|p' - p\|^+
\end{aligned} \tag{3.17}
$$

where $\|x\|^+ = |(Ax)^+|$. Here $(x)^+$ denotes the positive part of x, component-wise. Note that $\tilde{\rho}(p, p') = \|p - p'\|$ if $p \prec p'$ or $p' \prec p$. Let $|\cdot|$ denote the L_1 norm. In this case we have

$$
\begin{aligned}
\tilde{\rho}(p, p') &= \sum_{i=1}^{k} (m_i - q_i) + \sum_{i=1}^{k} (m_i - q_i') \\
&= \sum_{i=1}^{k} |q_i - q_i'| \\
&= |Ap - Ap'| \\
&= \|p - p'\|.
\end{aligned} \tag{3.18}
$$

From the triangle inequality property of a norm it follows that $\tilde{\rho}$ given by (3.18), satisfies the relation (3.6). Therefore, the dissimilarity measure, derived from the L_1 norm, is first order convex. Let $\hat{\rho}(p, p') = \min\{\rho(p, p'), \rho(p', p)\}$. We may consider $\hat{\rho}(p, p')$ as a measure of deficiency in the majorization relation between p and p'. With respect to the induced L_1 norm, it is given by

$$
\hat{\rho}(p, p') = \sum_{i=1}^{k} m_i - \max\left(\sum_{i=1}^{k} q_i, \sum_{i=1}^{k} q_i' \right).
$$

4. Application

We consider a problem of selecting the most diverse population from a given set of $m \geq 2$ multinomial populations. Alam *et al.* (1986) and Rizvi *et al.* (1987) have proposed selection procedures based on the Gini-Simpson index, Shannon's entropy and some other indices of diversity. Selection procedures for binomial populations have been considered by Gupta and Huang (1976) and Dudewicz and Van der Meulen (1981). Gupta and Wong (1975) have considered a procedure for selecting a subset of the given multinomial populations which includes the most diverse population in the sense of majorization, assuming that there is at least one population among the m populations, whose probability vector is majorized by

each of the other corresponding vectors. Clearly, this assumption is very restrictive for applications in practice. We propose a new approach to the problem of selecting the most diverse population, under less restrictive conditions. We shall denote the i-th population and the associated probability vector by π_i and q_i, respectively, for $i = 1, \ldots, m$. Let

$$\xi_i = \max_{j \neq i} \rho(q_i, q_j), \tag{4.1}$$

$$\eta_i = \min_{j \neq i} \rho(q_j, q_i), \tag{4.2}$$

$$\xi_0 = \min(\xi_1, \ldots, \xi_m) \quad \text{and} \tag{4.3}$$

$$A_{ij} = \max_{\ell \neq i,j} \rho(q_i, q_\ell).$$

We shall call π_i the most diverse population among the m populations if $\xi_i = \xi_0$. With regard to the given criterion for the most diverse population, we note the following result. Suppose that $q_i \prec q_j$. Then $A_{ij} \leq A_{ji}$, since $\rho(p, p')$ is a Schur-convex function of p. Hence

$$\begin{aligned} \xi_i &= \max(\rho(q_i, q_j), A_{ij}) \\ &= A_{ij} \\ &\leq A_{ji} \\ &\leq \xi_j. \end{aligned}$$

Given a sample of n observations from each of the m populations, a procedure (R) for selecting the most diverse population is given as follows. Let $\hat{q}_i$ denote the maximum likelihood estimate of q_i from the given sample from π_i, and let $\hat{\xi}_i$ and $\hat{\eta}_i$ be given by (4.1) and (4.2) with the substitution $\hat{q}_i$ for q_i and $\hat{q}_j$ for q_j. Let $\hat{\xi}_0 = \min(\hat{\xi}_1, \ldots, \hat{\xi}_m)$. Select π_i for the most diverse population if $\hat{\xi}_i = \hat{\xi}_0$, breaking ties if any by randomization. For simplicity we shall ignore the event of a tie. Then the probability of a correct selection (PCS) for the procedure R (assuming without loss of generality that π_i is the best population) is given by

$$\begin{aligned} \text{PCS} &= P\{\hat{\xi}_i < \hat{\xi}_j, j = 1, \ldots, i-1, i+1, \ldots, m\} \\ &\geq P\{\hat{\xi}_i < \rho(\hat{q}_j, \hat{q}_i), j = 1, \ldots, i-1, i+1, \ldots, m\} \\ &= P\{\hat{\xi}_i < \hat{\eta}_i\}. \end{aligned} \tag{4.4}$$

Let P^* be a given number such that $1/m < P^* < 1$. Consider the configuration (C_i) of the parameter space, given by

$$C_i : \eta_i - \xi_i \geq \delta \tag{4.5}$$

where δ is a fixed positive number. Since $\hat{q}_i$ is a consistent estimator of q_i, it follows from (4.4) that a minimum value of the sample size n can be determined for which

$$\text{PCS} \geq P^*$$

inside the configuration C_i. The configuration C_i is called a preference zone in the usual terminology of ranking and selection procedures.

Schmidt and Strauss (1975) modeled occupational attainment in the United States, using several explanatory variables. The predicted probabilities for occupation in five occupation categories, given average schooling and experience, for four populations (i) Black females (ii) Black males (iii) White females and (iv) White males are reproduced from Table 9.6 of Agresti (1990) in Table 3. The five occupation categories are listed as Menial, Blue-collar, Craft, White collar and Professionals. Measuring the diversities of the four populations from the data, we find that the white male population is the most diverse population among the four populations, with respect to the Gini-Simpson index and Shannon's entropy, as well as the new criterion given above.

We have carried out a Monte Carlo study of the performance of the proposed procedure R for selecting the most diverse population from the sample data. The values of PCS, using the norms L_1, L_2, IL_1, and IL_2 for the sample size $n = 30$, 60, 100, 200, and 400 are shown in Table 4. Ten thousand simulations were carried out for each of the five sample sizes. It is seen from the table that the four choices of norm give comparable values of the PCS. The table gives also the values of the PCS for a selection procedure based on the Gini-Simpson index and Shannon's entropy, due to Alam *et al.* (1986). These values are also comparable.

Table 3. Occupational probabilities, given average schooling and experience.

		Occupation				
Race	Gender	Menial	Blue collar	Craft	White collar	Professional
Black	female	0.396	0.188	0.011	0.219	0.187
	male	0.222	0.368	0.136	0.073	0.202
White	female	0.153	0.146	0.018	0.492	0.192
	male	0.089	0.296	0.232	0.169	0.214

Table 4. Values of PCS for the most diverse population.

	Procedure R					
	Norm					
Sample size	L_1	L_2	IL_1	IL_2	Gini-Simpson	Shannon's entropy
30	0.6412	0.6385	0.6350	0.6369	0.6244	0.6507
60	0.7658	0.7682	0.7566	0.7628	0.7675	0.7645
100	0.8473	0.8496	0.8446	0.8476	0.8493	0.8317
200	0.9377	0.9383	0.9362	0.9380	0.9399	0.9212
400	0.9874	0.9888	0.9864	0.9868	0.9869	0.9793

5. Concluding remarks

We are familiar with the analysis of variance and its applications. This technique is used to analyze differences between populations, using the variance as a measure of variability. However, the method is not applicable to discrete data. We have proposed a new measure of the difference between two multinomial populations which is based on the dissimilarity between the intrinsic diversities of the two populations. It may be noted that the proposed measure is invariant under permutations of the probability vector, whereas some other known measures of dissimilarity which have been cited in this paper, depending on species identification, do not share this property. Certain mathematical properties of the new measure are given with applications. In a subsequent paper, we shall examine the problem of decomposing the total dissimilarity between a collection of populations into dissimilarities between and within subgroups of populations, using the proposed measure of dissimilarity. This paper is mainly concerned with the exposition of certain mathematical properties of the proposed measure of dissimilarity. A comparative study of other approaches using dissimilarity measures and cross entropy ideas would be appropriate. This is the topic of a subsequent paper.

Acknowledgements

Professor Michael M. Kostreva is acknowledged for his assistance in the quadratic programming needed for the results given in Section 1. We appreciate the referee's comments on this paper. Particularly, the comments with regard to the properties of the proposed measures of dissimilarity were very useful.

References

Agresti, A. (1990). *Categorical Data Analysis*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Alam, K., Mitra, A., Rizvi, M. H. and Saxena, K. M. Lal (1986). Selection of the most diverse multinomial population, *Amer. J. Math. Management Sci.*, Special Volume **6**, 65–86.

Dennis, B., Patil, G. P., Rossi, O. and Taillie, C. (1979). A bibliography of literature on ecological diversity and related methodology, *Ecological Diversity in Theory and Practice*, 319–354, International Co-operative Publishing House, Jerusalem.

Dudewicz, E. J. and Van der Meulen, E. C. (1981). Selection procedures for the best binomial population with generalized entropy goodness, *Tamkang J. Math.*, **12**, 206–208.

Gini, C. (1912). Variabilitá e Mutabilitá, Studi Economico-Giuridici della facoltá di Giurisprodenza dell, Universitá di Cagliari, Anno 3, Part 2, p. 80.

Gower, J. C. (1985). Measures of similarity, dissimilarity, and distance, *Encyclopedia of Statistical Sciences* (eds. S. Kotz and N. L. Johnson), **5**, 397–405, Wiley-Interscience, New York.

Grassle, J. F. and Smith, W. K. (1976). A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities, *Oecologia*, **25**, 13–25.

Gupta, S. S. and Huang, D. Y. (1976). On subset selection procedures for the entropy function associated with the binomial populations, *Sankhyā Ser. A*, **38**, 153–173.

Gupta, S. S. and Wong, W. Y. (1975). Subset selection procedures for finite schemes in information theory, *Colloq. Math. Soc. János Bolyai*, **16**, 279–291.

Haberman, S. J. (1982). Analysis of dispersion of multinomial responses, *J. Amer. Statist. Assoc.*, **77**, 568–580.

Kostreva, M. M. (1989). Generalization of Murty's direct algorithm to linear and convex quadratic programming, *J. Optim. Theory Appl.*, **62**, 63–76.

Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data, *J. Amer. Statist. Assoc.*, **66**, 534–544.

Lorenz, M. O. (1905). Methods of measuring concentration of wealth, *J. Amer. Statist. Assoc.*, **9**, 209–212.

Marshall, A. V. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*, Academic Press, San Diego.

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

Rao, C. R. (1982*a*). Diversity and dissimilarity coefficients, a unified approach, *Theoret. Population Biol.*, **21**, 24–43.

Rao, C. R. (1982*b*). Diversity: its measurement, decomposition, appartionment and analysis, *Sankhyā Ser. A*, **44**, 1–22.

Rao, C. R. (1984). Convexity properties of entropy functions and analysis of diversity, *Inequalities in Statistics and Probability*, IMS Lecture Notes - Monograph Series, Vol. 5, 68–77, Hayward, California.

Rao, C. R. and Nayak, T. K. (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy, *IEEE Trans. Inform. Theory*, **31** (5), 589–593.

Rizvi, M. H., Alam, K. and Saxena, K. M. Lal (1987). Selection procedure for multinomial populations with respect to diversity indices, *Contribution to the Theory and Application of Statistics* (ed. A. E. Gelfard), Academic Press, New York.

Schmidt, P. and Strauss, R. P. (1975). The prediction of occupation using multiple logit models, *Internat. Econom. Rev.*, **16**, 471–486.

Schur, I. (1923). Uber eine Klasse von Mittelbildungen mit Anwendungen die Determinanten, *Theorie Sitzungsber. Berlin Math. Gesellschaft*, **22**, 9–20 (*Issai Collected Works* (eds. A. Brauer and H. Rohrbach), Vol. II, 416–427, Springer, Berlin, 1973).

Shannon, C. E. (1948). A mathematical theory of communication, *Bell Syst. Tech. J.*, **27**, 379–423 and 626–656.

Simpson, E. H. (1949). Measurement of diversity, *Nature*, **163**, 688.

Smith, W. (1989). ANOVA-like similarity analysis using expected species shared, *Biometrics*, **45**, 873–881.

Smith, W., Grassle, J. F. and Kravitz, D. (1979). Measures of diversity with unbiased estimators, *Ecological Diversity in Theory and Practice*, 177–191, International Co-operative Publishing House, Jerusalem.