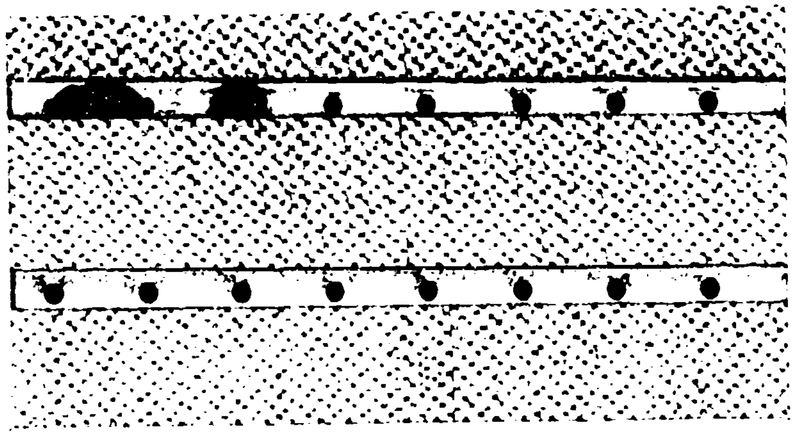
(2)

AFOSR-TR- 85-0749



Center for Multivariate Analysis

University of Pittsburgh

85 10 11 147

NEW MEASURES OF DIVERSITY*

Manzoor Ahmad
Universite du Quebéc á Montréal
and
Center for Multivariate Analysis
University of Pittsburgh

June 1985

Technical Report No. 85-23

Center for Multivariate Analysis
515 Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

*Part of the work was supported by the Air Force Office of Scientific Research under Contract F49620-85-C-0008. The U.S. Government is authorized to reporduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.
Author is thankful to NSERC of Canada for research support.

-A-

NEW MEASURES OF DIVERSITY

Manzoor Ahmad

ABSTRACT

The problem of measuring diversity within populations and dissimilarity or similarity between populations has been extensively treated in the literature. In this context a general procedure called Analysis of Diversity has been outlined and examined by C.R. Rao in a series of papers.

In this paper we proposes three new measuers of diversity and study related inference problems. Denote by S^k the simplex $S^k = \{\pi: \pi = (\pi_1, \ldots, \pi_k)', \pi_j \geq 0, \Sigma\pi_j = 1\}$. Then the proposed measures are of the form: $H_m(\pi) = 1 - a_m \Sigma_j \pi_j \phi_m(\pi_j)$, $m = 1,2,3$ where $\phi_1(x) = (1+k^{-1}-x)^{-\gamma}$, $\gamma \geq 0$, $\phi_2(x) = (2-k^{-\gamma}-x^{\gamma})^{-1}$, $\gamma \geq 0$, $\phi_3(x) = (a_3+(1-x)^{\gamma})^{-1}$, $0 < \gamma \leq 1$, and the a's are suitable normalizing constants. Estimation of $H_m(\pi)$, derivation of the penalty function and cross entropy and the problem of testing independence have been treated. Asymptotic distributions of relevant test statistics are indicated.

Additional keywords: Concavity; convexity; Computations.

1. INTRODUCTION

The problem of measuring diversity within populations and dissimilarity or similarity between populations has been extensively treated in the literature. This problem arises in a wide variety of domains; linguistics (Horvath, 1963; Zinger, 1982; Greenberg, 1956; Guirand, 1959; Herdan, 1964, 1966; Yule, 1944; Savchuc, 1964), sociology (Agresti and Agresti, 1978), biology (Sokal and Sneath, 1963; Pielou, 1975; Patil and Taillie, 1979), anthropology (Rao, 1971a, 1977b), to mention a few. An extensive bibliography of papers on measures of diversity and their ~~applications can be found in Dennis~~ et al (1979) and Patil and Taillie (1982).

Diversity within populations and dissimilarity between populations have been measured and interpreted differently. The choice of a diversity measure essentially depends on the context of a problem, however any diversity measure satisfying certain basic conditions can be used for partioning the total variability into a number of additive components, each of which can be used to test a certain null hypothesis or estimate a component of the variability. Rao ~~(1982,a,b)~~ outlined a general procedure called Analysis of Diversity (ANODIV) which is similar to the Analysis of Variance (ANOVA) for quantitative data. In this direction Light and Margolin (1971, 1974), Anderson and Landis (1980) have studied the Gini-Simpson index of diversity while Nayak (1984) has extended their results for Quadratic Entropy introduced by Rao (1982,b,c). ——→ See p-A-

Following the general procedure of Rao (1982,a,b) any function H defined on the simplex $S^k = \{\underset{\sim}{\pi}: \underset{\sim}{\pi} = (\pi_1,\ldots,\pi_k)'; \pi_j \geq 0, \Sigma\pi_j = 1\}$ of the Euclidean space R^k, is said to be a diversity measure if it satisfies the following conditions

(a) $H(\underset{\sim}{\pi}) \geq 0$, '$=0$', if and only if $\pi_j = 1$ for some j and $\pi_{j'} = 0$, $\forall j' \neq j$

(b) $H(\underset{\sim}{\pi}) \leq 1$ and '$=1$' if and only if $\underset{\sim}{\pi}$ is a uniform distribution i.e. $\pi_1 = \pi_2 = \ldots = \pi_k = \frac{1}{k}$

(c) $H(\underset{\sim}{\pi})$ is concave in $\underset{\sim}{\pi}$ on S^k.

While condition (a) is natural and (b) is standard normalization, the condition (c) fulfills the requirement that the diversity in a weighted mixture of populations should not be smaller than the weighted sum of diversitites within the individual populations. Gini-Simpson index of diversity, quadratic entropy of Rao, Shannon's entropy, α-degree entropy of Renyi (1961), α-degree entropy of Havrda and Charvat (1956), among others, satisfy conditions (a), (b) and (c). (See Nayak (1985a)).

We consider three measures which are of the form; $\forall\ \underset{\sim}{\pi} \in S^k$

(1.1) $$H_m(\underset{\sim}{\pi}) = 1 - a_m \sum_{j=1}^{k} \pi_j \phi_m(\pi_j), \quad m = 1,2,3$$

where

(1.2) $$a_1 = k^{-\gamma},\ \phi_1(x) = (1 + k^{-1} - x)^{-\gamma}, \quad \gamma \geq 0$$

(1.3) $$a_2 = 1 - k^{-\gamma},\ \phi_2(x) = (2 - k^{-\gamma} - x^{\gamma})^{-1}, \quad \gamma \geq 0$$

(1.4) $$a_3 = (1 - k^{-1})^{\gamma},\ \phi_3(x) = (a_3 + (1-x)^{\gamma})^{-1}, \quad 0 < \gamma \leq 1.$$

These functions vanish only at the vertices $\underset{\sim}{e}_j$, $j = 1,2,\ldots,k$ of S^k; where the probability vector $\underset{\sim}{e}_j$ represents a multinomial distribution whose j^{th} cell has cell frequency one and others zero. In section 2, we have shown that $H_1(\underset{\sim}{\pi})$ is concave for $\gamma \geq 0$, $H_2(\underset{\sim}{\pi})$ for $\gamma \geq 1$ and $H_3(\underset{\sim}{\pi})$ for $0 < \gamma \leq 1$. Further, $\forall\ \underset{\sim}{\pi} \in S^k$

(1.5) $$H_m(\underset{\sim}{\pi}) \geq 0, \quad m = 1,2,3$$

follows from the concavity of H_m since $\underset{\sim}{\pi} = \sum_{j=1}^{k} \pi_j \underset{\sim}{e}_j$, and $\Sigma\pi_j = 1$.
We have also shown in section 2 that these measures take their maximum value at $\underset{\sim}{\pi} = (1/k, 1/k, \ldots, 1/k)$, the most spread multinomial population. Define

(1.6) $$H_m^0 = \max_{\underset{\sim}{\pi} \in S^k} H_m(\underset{\sim}{\pi}) = 1 - a_m \phi_m(\tfrac{1}{k}).$$

Then we have

(1.7) $$H_1^0 = (k^\gamma - 1)/k^\gamma, \quad \gamma > 0$$
$$H_2^0 = H_3^0 = 1/2.$$

Further, since these functions are symmetric in $(\pi_1, \ldots, \pi_k)$, they turn out to be Schur-concave which is indeed a desirable property for measuring variability in a multinomial population. With such measures, the more spread-out the population the more diverse it turns out to be.

In section 3 we have treated the problem of estimating the diversity of a multinomail population, based on the measure H_m; $m = 1,2,3$.

Derivation of the penalty function (Haberman (1982)) and cross entropy (Rao (1982b)) for each of the proposed measures and the problem of testing independence has been treated in section 4.

2. CONCAVITY OF THE MEASURES H_m; $m = 1,2,3$

From (1.1) it is obvious that the concavity of $H_m(\underset{\sim}{\pi})$ would follow from the convexity of

(2.1) $$I_m(\underset{\sim}{\pi}) = \sum_{j=1}^{k} \pi_j \phi_m(\pi_j); \quad m = 1,2,3.$$

While proving their convexity, the functions I_m can be treated, without loss of generality, as functions of $\pi_1, \pi_2, \ldots, \pi_{k-1}$ and

$$\pi_k = 1-(\pi_1+\ldots+\pi_{k-1}); \quad \underset{\sim}{\pi} \in S^k.$$

2a. Convexity of $I_1(\underset{\sim}{\pi})$

Since

$$I_1(\underset{\sim}{\pi}) = \sum_{j=1}^{k} \pi_j (1+k^{-1}-\pi_j)^{-\gamma} = \sum_{j=1}^{k} \pi_j \eta_j^{-\gamma} \quad \text{(say)}$$

it follows that

(2.2)
$$\frac{\partial}{\partial \pi_j} I_1 = \eta_j^{-\gamma}+\gamma\pi_j \eta_j^{-\gamma-1} - \{\eta_k^{-\gamma}+\gamma\pi_k \eta_k^{-\gamma-1}\},$$

$$\frac{\partial^2}{\partial \pi_j^2} I_1 = \tau_j + \tau_k, \quad 1 \leq j \leq k-1$$

$$\frac{\partial^2}{\partial \pi_{j'} \partial \pi_j} I_1 = \tau_k, \quad 1 \leq j' \leq k-1, \quad j' \neq j$$

where

$$\tau_t = \gamma \eta_t^{-\gamma-2}[2(1+k^{-1}) + (\gamma-1)\pi_t]$$

and

$$\eta_t = 1+k^{-1}-\pi_t.$$

For a given vector $\underline{d} = (\underline{d}_1, \ldots, \underline{d}_p)'$, let $D_{\underline{d}}$ denote a diagonal matrix with elements $d_1, d_2, \ldots, d_p$. Now, the matrix $\nabla^2 I_1$ of second order derivatives of I_1 takes form

(2.3)
$$\nabla^2 I_1 = D_{\underline{\tau}} + \tau_k \underset{\sim}{1}\underset{\sim}{1}'$$

with

$$\underset{\sim}{\tau} = (\tau_1, \tau_2, \ldots, \tau_{k-1}),$$

which is certainly positive definite for $\forall\ \gamma \geq 0$.

2b. Convexity of $I_2(\underset{\sim}{\pi})$

In this case

$$I_2(\underset{\sim}{\pi}) = \sum_{j=1}^{k} \pi_j (\beta - \pi_j^\gamma)^{-1} \tag{2.3}$$

where $\beta = 2-k^{-\gamma}$, and straightforward computations yield

$$\frac{\partial}{\partial \pi_i} I_2 = (\beta-\pi_j^\gamma)^{-1} + \gamma\pi_j^\gamma(\beta-\pi_j^\gamma)^{-2} - (\beta-\pi_k^\gamma)^{-1} - \gamma\pi_k^\gamma(\beta-\pi_k^\gamma)^{-2}, \tag{2.4}$$

$$\frac{\partial^2}{\partial \pi_j^2} I_2 = \theta_j + \theta_k, \quad 1 \leq j \leq k-1$$

and

$$\frac{\partial^2}{\partial \pi_{j'} \partial \pi_j} I_2 = \theta_k, \quad 1 \leq j' \leq k-1;\ j' \neq j,$$

where

$$\theta_t = (\beta-\pi_t^\gamma)^{-3}\{\gamma(1+\gamma)\pi_t^{\gamma-1}(\beta-\pi_t^\gamma) + 2\gamma^2\pi_t^{2\gamma-1}\}.$$

Hence,

$$\nabla^2 I_2 = D_{\underset{\sim}{\theta}} + \theta_k \underset{\sim}{1}\underset{\sim}{1}' \tag{2.5}$$

with

$$\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_{k-1}),$$

is positive definite for $\gamma > 0$.

2c. Convexity of $I_3(\underset{\sim}{\pi})$

Note that

(2.6) $$I_3(\underset{\sim}{\pi}) = \sum_{i=1}^{k} \pi_i (b + \bar{\pi}_i^{\gamma})^{-1}$$

where $b = (1-k^{-1})^{\gamma}$ and $\bar{\pi}_j = 1-\pi_j$, and

(2.7) $$\frac{\partial}{\partial \pi_j} I_3 = (b+\bar{\pi}_j^{\gamma})^{-1} + \gamma\pi_j \bar{\pi}_j^{\gamma-1}(b+\bar{\pi}_j^{\gamma})^{-2}$$

$$- (b+\bar{\pi}_k^{\gamma})^{-1} - \gamma\pi_k \bar{\pi}_k^{\gamma-1}(b+\bar{\pi}_k^{\gamma})^{-2}$$

$$\frac{\partial^2}{\partial \pi_j^2} I_3 = \delta_j + \delta_k, \quad 1 \le j \le k-1$$

$$\frac{\partial^2}{\partial \pi_{j'}, \partial \pi_j} I_3 = \delta_k \quad 1 \le j' \le k-1, \quad j' \ne j$$

with

$$\delta_t = \gamma(2-(1+\gamma)\pi_t)\bar{\pi}_t^{\gamma-2}(b+\bar{\pi}_t^{\gamma})^{-2}$$

$$+ 2\gamma^2 \pi_t \bar{\pi}_t^{2\gamma-2}(b+\bar{\pi}_t^{\gamma})^{-3}.$$

Hence

(2.8) $$\nabla^2 I_3 = D_{\underset{\sim}{\delta}} + \delta_k \underset{\sim}{1}\underset{\sim}{1}',$$

where elements of $D_{\underset{\sim}{\delta}}$ are $(\delta_1, \ldots, \delta_{k-1})$, is p.d iff $0 < \gamma \le 1$.

2d. Maxima's of the Functions $H_m(\underset{\sim}{\pi})$; $m = 1,2,3$

Critical points of $H_m(\underset{\sim}{\pi})$ for $m = 1,2,3$, are solutions of the system of equations

(2.9) $$\frac{\partial}{\partial \pi_j} H_m(\underset{\sim}{\pi}) = - \frac{\partial}{\partial \pi_j} I_m(\underset{\sim}{\pi}) = 0; \quad j = 1,2,\ldots,k-1.$$

These equations, when $m = 1$, are of the form, for $j = 1,2,\ldots,k-1$

$$\eta_j^{-\gamma}+\gamma(1+k^{-1}-\eta_j)\eta_j^{-\gamma-1} = k^*(\text{const.})$$

Hence one can easily argue that the solutions must satisfy the condition

$$\eta_1=\eta_2=\ldots=\eta_{k-1} \tag{2.10}$$

or equivalently

$$\pi_1=\pi_2=\ldots=\pi_{k-1}=\frac{1}{k}$$

since the constant k^* is the value of the l.h.s. evaluated at η_k. Through analogous arguments, so turns out to be the case with $H_2(\underset{\sim}{\pi})$ and $H_3(\underset{\sim}{\pi})$.

3. ESTIMATION OF $H_m(\underset{\sim}{\pi})$; $m=1,2,3$

For inference problem, it is essential to estimate a measure of diversity $H(\underset{\sim}{\pi})$. A popular estimate based on sample proportions $p_1,p_2,\ldots,p_k$ would be $H(\hat{\underset{\sim}{\pi}})$ where $\hat{\pi}_j = p_j, \forall_j$. This is also the maximum likelihood estimator of $H(\underset{\sim}{\pi})$(Zehna (1966)). A Taylor series expansion of $H(\hat{\underset{\sim}{\pi}})$ around $\underset{\sim}{\pi}$ allows us to compute the asymptotic variance of $H(\hat{\underset{\sim}{\pi}})$. To do so we first express $H_m(\hat{\underset{\sim}{\pi}})$ as

$$H_m(\hat{\underset{\sim}{\pi}}) = 1-\sum_{j=1}^{k}\psi_m(\hat{\pi}_j) \tag{3.1}$$

with

$$\psi_m(x) = a_m x\phi_m(x), \quad m=1,2,3$$

and treat H_m as a function of (k-1) free variables $\pi_1,\pi_2,\ldots,\pi_{k-1}$ with $\pi_k = 1-\sum_{j=1}^{k-1}\pi_j$, $\pi_j \geq 0$ for $j=1,2,\ldots,k$, i.e. $H_m(\pi_1,\pi_2,\ldots,\pi_{k-1}) = 1-\sum_{j=1}^{k-1}\psi_m(\pi_j)-\psi_m(\pi_k)$. Then

$$\sum_{j=1}^{k-1}(\hat{\pi}_j-\pi_j)\,\frac{\partial}{\partial\pi_j}\,H_m(\underset{\sim}{\pi}) = \sum_{j=1}^{k-1}(\hat{\pi}_j-\pi_j)\{-\frac{\partial}{\partial\pi_j}\psi_m(\pi_j)+\frac{\partial}{\partial\pi_k}\psi_m(\pi_k)\} \tag{3.2}$$

$$= -\sum_{j=1}^{k}(\hat{\pi}_j-\pi_j)\frac{\partial}{\partial\pi_j}\psi_m(\pi_j) = -\sum_{j=1}^{k}(\hat{\pi}_j-\pi_j)d_{mj} \quad \text{(say)}$$

since $\frac{\partial}{\partial \pi_j} \pi_k = -1$, $\forall j$, and $\Sigma \hat{\pi}_j = \Sigma \pi_j = 1$. Now we can see easily that the asymptotic variances of $H_m(\hat{\underset{\sim}{\pi}})$, denotes by $\tilde{\sigma}_m^2$, is the variance of the linear combination $\sum_{j=1}^{k} d_{mj} \hat{\pi}_i$ where

$$d_{mj} = \frac{\partial}{\partial \pi_j} \psi_m(\pi_j), \quad j = 1,2,\ldots,k \tag{3.3}$$

$$= a_m\{\phi_m(\pi_j) + \pi_j \frac{d}{d\pi_j} \phi_m(\pi_j)\}.$$

It follows that, for $m = 1,2$ and 3, the asymptotic variance of the estimator $H_m(\hat{\underset{\sim}{\pi}})$ of the diversity measure $H_m(\underset{\sim}{\pi})$ is

$$n\tilde{\sigma}_m^2 = \sum_{j=1}^{k} \pi_j d_{mj}^2 - \{\sum_{j=1}^{k} \pi_j d_{mj}\}^2. \tag{3.4}$$

The sequences $\{d_{1j}\}_{j=1}^{k}$, $\{d_{2j}\}_{j=1}^{k}$ and $\{d_{3j}\}_{j=1}^{k}$ corresponding to $\tilde{\sigma}_1^2$, $\tilde{\sigma}_2^2$ and $\tilde{\sigma}_3^2$ respectively are

$$d_{1j} = a_1(1+k^{-1}-\pi_j)^{-\gamma-1}\{1+k^{-1}+(\gamma-1)\pi_j\}, \tag{3.5}$$

$$d_{2j} = a_2(2-k^{-\gamma}-\pi_j^{\gamma})^{-2}\{2-k^{-\gamma}+(\gamma-1)\pi_j^{\gamma}\}$$

and

$$d_{3j} = a_3\{(1-k^{-1})^{\gamma}+(1-\pi_j)^{\gamma}\}^{-2}\{(1-k^{-1})^{\gamma}+(1+(\gamma-1)\pi_j)(1-\pi_j)^{\gamma-1}\},$$

$j = 1,2,\ldots,k.$

Remark. Note that $n\tilde{\sigma}_m^2$ is equal to the variance of a random variable D_m which takes the value d_{mj} with probability π_j, $j = 1,2,\ldots,k$.

Case $\gamma = 1$

Each of the diversity measures $H_m(\underset{\sim}{\pi})$, $m = 1,2$ or 3, can be seen as a family of measures since it depends on a parameter of our choice γ. Any choice of γ within the range of values for which $H_m(\underset{\sim}{\pi})$ remains concave would lead to a specific measure. In practice as we will see in the sequal the choice of γ would depend upon the nature of the problem. However the choice $\gamma = 1$ should be emphasized since in this case H_m's take a simpler form. Define the diversity measure G_m as

$$(3.6) \qquad G_m(\underset{\sim}{\pi}) = H_m(\underset{\sim}{\pi}) \text{ with } \gamma = 1.$$

Then

$$(3.7) \qquad G_1(\underset{\sim}{\pi}) = 1 - k^{-1} \sum_{j=1}^{k} \pi_j (1+k^{-1}-\pi_j)^{-1},$$

$$G_2(\underset{\sim}{\pi}) = G_3(\underset{\sim}{\pi}) = 1 - (1-k^{-1}) \sum_{j=1}^{k} \pi_j (2-k^{-1}-\pi_j)^{-1}$$

and the variances of the estimators $G_1(\hat{\underset{\sim}{\pi}})$ and $G_2(\hat{\underset{\sim}{\pi}})$ respectively are

$$(3.8) \qquad n\tilde{\sigma}^2[\hat{G}_1] = k^{-2}(1+k^{-1})^2 \sum_j \pi_j \{(1+k^{-1}-\pi_j)^{-2} - \sum \pi_j (1+k^{-1}-\pi_j)^{-2}\}$$

and

$$n\tilde{\sigma}^2[\hat{G}_2] = (1-k^{-1})(2-k^{-1}) \sum \pi_j \{(2-k^{-1}-\pi_j)^{-2} - \sum \pi_j (2-k^{-1}-\pi_j)^{-2}\}.$$

4. DECOMPOSITION AND TEST OF INDEPENDENCE

Consider a population P of a nominal random variable Y that assumes the integral values j, $1 \le j \le k$, which is being viewed as a mixture of r populations $P_1, P_2, \ldots, P_r$ of Y identified according to r discrete levels of a factor X of some interest. Let $\underset{\sim}{\pi}_\ell = (\pi_{\ell 1}, \ldots, \pi_{\ell j}, \ldots, \pi_{\ell k})'$ be the probability vector of Y for the population P_ℓ, and λ_ℓ be the mixing weight of $\underset{\sim}{\pi}_\ell$ for the overall population P, $\lambda_\ell \ge 0$, $\ell = 1,2,\ldots,r$, $\Sigma\lambda_\ell = 1$. Hence Y is assumed to follow a multinomial distribution whose probability vector $\underset{\sim}{\pi}$. is the mixture $\sum_{\ell=1}^{r} \lambda_\ell \underset{\sim}{\pi}_\ell$. Based on the data classified in the above fashion, we are usually intereseed in a problem of prediction or testing a hypothesis of independence or testing a hypothesis H_0: $\underset{\sim}{\pi}_1 = \underset{\sim}{\pi}_2 = \cdots = \underset{\sim}{\pi}_r$. Such inference problems are handled through the analysis of Diversity (ANODIV).

In this regard the following decomposition, due to Rao (1982,a,b), is most natural;

$$(4.1)^{*} \qquad H(\underset{\sim}{\pi}.) = \Sigma\lambda_\ell H(\underset{\sim}{\pi}_\ell) + J_H(\{\lambda_\ell\},\{\underset{\sim}{\pi}_\ell\}).$$

The component $\Sigma\lambda_\ell H(\underset{\sim}{\pi}_\ell)$ is the average diversity within the populations, and the second term designated as "Jensen difference", defined by subtraction, represents the diversity between populations. The concavity of H ensures that $J_H \geq 0$. An alternative but similar decomposition, which provides an interpretation of J_H, can be obtained through the concept of 'Penalty function' associated with a diversity measure. (Rao and Nayak (1985).)

Let $\Delta(j,\underset{\sim}{\pi}^*)$ be the penalty (or the loss) to be incurred in a probabilistic prediction if a probability vector $\underset{\sim}{\pi}^*$ is used for prediction and the true category is j. Then expected penalty for using $\underset{\sim}{\pi}^*$ is $\Sigma\pi_j\Delta(j,\underset{\sim}{\pi}^*)$. If a diversity measure H is strictly concave then there exists a non-negative and possibly infinite function $\Delta(j,\underset{\sim}{\pi})$ such that

$$(4.2) \qquad \text{(i)} \quad H(\pi) = \Sigma\pi_j\Delta_H(j,\underset{\sim}{\pi}) \quad \forall\ \underset{\sim}{\pi}\in S^k$$

and

$$(4.3) \qquad \text{(ii)} \quad H(\underset{\sim}{\pi}) = \Sigma\pi_j\Delta_H(j,\underset{\sim}{\pi}) \leq \Sigma\pi_j\Delta_H(j,\underset{\sim}{\pi}^*)$$

for all $\underset{\sim}{\pi}$, $\underset{\sim}{\pi}^*\in S^k$, with equality only if $\underset{\sim}{\pi} = \underset{\sim}{\pi}^*$. The existence of Δ_H for every strictly concave function H is due to Haberman (1982), and it can be obtained as follows.

Let H^* be an extension of H to R_+^k such that $\forall\ \alpha \geq 0$

$$(4.4) \qquad H^*(\alpha\underset{\sim}{\pi}) = \alpha H^*(\underset{\sim}{\pi}).$$

Then for $\underset{\sim}{\pi}\in S^k$ with $\pi_j > 0$, $1 \leq j \leq k$, the penalty function $\Delta_H(j,\underset{\sim}{\pi})$ is given by

---

*is also expressed as (in analogy to variance decomposition) SST = SSW + SSB.

$$(4.5) \qquad \Delta(j,\underset{\sim}{\pi}) = \frac{\partial}{\partial\pi_j} H^*\Big|_{\underset{\sim}{\pi}} .$$

In terms of the penalty function associated with a strictly concave function H the following decomposition of the total diversity $H(\underset{\sim}{\pi}.)$ (or SST), obtained by Rao and Nayak (1985), allows an interpretation of the diversity between populations J_H (or SSB)

$$(4.6) \qquad H(\underset{\sim}{\pi}.) = \Sigma\lambda_\ell H(\underset{\sim}{\pi}_\ell) + \Sigma\lambda_\ell C_H(\underset{\sim}{\pi}_\ell,\underset{\sim}{\pi}.)$$

where

$$(4.7) \qquad C_H(\underset{\sim}{\pi},\underset{\sim}{\pi}^{\ }) = \sum_{j=1}^{k} \pi_j[\Delta_H(j,\underset{\sim}{\pi}^*) - \Delta_H(j,\underset{\sim}{\pi})].$$

The function $C_H(\cdot,\cdot)$, called as the 'Cross-entropy' induced by H, is non-negative but not necessarily symmetric. A more general discussion can be found in Rao and Nayak (1985).

Since $\underset{\sim}{\pi}_1,\ldots,\underset{\sim}{\pi}_r$ are associated with r levels of a factor X, the ratio

$$(4.8) \qquad \rho_H^2 = \frac{SSB}{SST} = \frac{\Sigma\lambda_\ell C_H(\underset{\sim}{\pi}_\ell,\underset{\sim}{\pi}.)}{H(\underset{\sim}{\pi}.)}$$

can be used as a measure of association between X and the response variable Y.

Now we give the extension H^*, and penalty function $\Delta(j,H^*)$ essentially needed to compute the cross-entropy (i.e. measure the dissimilarity between $\underset{\sim}{\pi}$ and $\underset{\sim}{\pi}^*$) for the proposed diversity measures H_1, H_2 and H_3.

Extensions H_1^*, H_2^* and H_3^* satisfying the condition (20) are, $\forall\ \underset{\sim}{\pi} \in R_+^k$

$$(4.9) \qquad H_1^*(\underset{\sim}{\pi}) = \sum_{j=1}^{k} \pi_j[1 - a_1(\Sigma\pi_\ell)^\gamma\{b_1\Sigma\pi_\ell - \pi_j\}^{-\gamma}]$$

$$(4.10) \qquad H_2^*(\underset{\sim}{\pi}) = \sum_{j=1}^{k} \pi_j[1 - a_2(\Sigma\pi_\ell)^\gamma\{b_2(\Sigma\pi_\ell)^\gamma - \pi_j^\gamma\}^{-1}]$$

and

$$(4.11)\qquad H_3^*(\underset{\sim}{\pi}) = \sum_{j=1}^{k} \pi_j[1 - a_3(\Sigma\pi_\ell)^\gamma\{b_3(\Sigma\pi_\ell)^\gamma + (\Sigma\pi_\ell - \pi_j)^\gamma\}^{-1}]$$

where

$$b_1 = 1+k^{-1},\ b_2 = 2-k^{-\gamma} \text{ and } b_3 = (1-k^{-1})^\gamma = a_3$$

the 'Σ' represents $\sum_{\ell=1}^{k}$.

The penalty functions Δ_{H_m} induced by the functions H_m, according to (21), turn out to be, for $m = 1,2,3$,

$$(4.12)\qquad \Delta_{H_1}(j,\underset{\sim}{\pi}) = \frac{a_1\tilde{\gamma}\pi_j - a_1 b_1}{(b_1-\pi_j)^{\gamma+1}} + 1 - \sum_{\ell=1}^{k} \frac{\gamma a_1 \pi_\ell^2}{(b_1-\pi_\ell)^{\gamma+1}}$$

$$(4.13)\qquad \Delta_{H_2}(j,\underset{\sim}{\pi}) = \frac{a_2\tilde{\gamma}\pi_j^\gamma - a_2 b_2}{(b_2-\pi_j^\gamma)^2} + 1 + \sum_{\ell=1}^{k} \frac{\gamma a_2 \pi_\ell^{\gamma+1}}{(b_2-\pi_\ell^\gamma)^2}$$

$$(4.14)\qquad \Delta_{H_3}(j,\underset{\sim}{\pi}) = \frac{a_3(\tilde{\pi}_j)^{-\tilde{\gamma}}(1+\tilde{\gamma}\tilde{\pi}_j) - a_3 b_3}{(b_3-\tilde{\pi}_j^\gamma)^2} + 1 + \sum_{\ell=1}^{k} \frac{a_3\pi_\ell^2}{(b_3-\tilde{\pi}_\ell^\gamma)^2}$$

where $\tilde{\gamma} = 1-\gamma$ and $\tilde{\pi}_j = 1-\pi_j$.

<u>In the case $\gamma = 1$</u>,

$$(4.15)\qquad \Delta_{G_1}(j,\underset{\sim}{\pi}) = \frac{-k^{-1}(1+k^{-1})}{(1+k^{-1}-\pi_j)^2} + 1 - \sum_{\ell=1}^{k} \frac{k^{-1}\pi_\ell^2}{(1+k^{-1}-\pi_\ell)^2}$$

and

$$(4.16)\qquad \Delta_{G_2}(j,\underset{\sim}{\pi}) = \Delta_{G_3}(j,\underset{\sim}{\pi}) = \frac{-(1-k^{-1})(2-k^{-1})}{(2-k^{-1}-\pi_j)^2} + 1 + \sum_{\ell=1}^{k} \frac{(1-k^{-1})\pi_\ell^2}{(2-k^{-1}-\pi_\ell^2)}\ .$$

Let us now consider the problem of testing the hypothesis H_0: $\underset{\sim}{\pi}_1 = \underset{\sim}{\pi}_2 = \ldots = \underset{\sim}{\pi}_r$. Following Rao (1982,a,b), a test of this hypothesis can be based on $\hat{J}_H$ (i.e. $\hat{SSB}$)

since under H_0, $J_H = 0$ in the population and conversely $J_H = 0$ implies $\underset{\sim}{\pi}_1 = \ldots = \underset{\sim}{\pi}_r$ provided H is strictly concave.

Based on a sample from a population P of the k-dimensional nominal r.v. Y divided into sub-samples according to r levels ($r \geq 2$) of the factor X, we are going to propose a criteria, based on $\hat{SSB}$, to test the null hypothesis that X and Y are independent, i.e. H_0: $\underset{\sim}{\pi}_1 = \underset{\sim}{\pi}_2 = \ldots = \underset{\sim}{\pi}_r$.

For the i^{th} level of the factor X, $i = 1,2,\ldots,r$, let n_{ij} be the observed frequency for the j^{th} category of Y, $j = 1,2,\ldots,k$, in a sample of size $n = \sum_{ij}\sum n_{ij}$. Further let

$$(4.17) \qquad n_{i.} = \sum_j n_{ij}, \quad n_{.j} = \sum_j n_{ij}, \quad v_i = (n_{i1},\ldots,n_{ik})'$$

$$p_{i.} = \frac{1}{n_{i.}} v_i \quad \text{and} \quad p_{..} = \frac{1}{n_{..}} \Sigma v_i$$

The total, within and between group diversities for the sample are

$$(4.18) \qquad \hat{SST} = H(p_{..})$$

$$\hat{SSW} = \Sigma \frac{n_{i.}}{n_{..}} H(p_{i.})$$

and

$$\hat{SSB} = \hat{SST} - \hat{SSW}.$$

Naik (1985) has shown that asymptotically, under H_0, (i) $\hat{SSB}$ is distributed as a linear combination of independent χ^2 variables and (ii) $\hat{SSB}$ and $\hat{SST}$ are independently distributed. For the sake of completeness and for determining the critical region for testing H_0, we give the basic assumptions and the main results of Naik (1985).

For a statistical analysis the sample vectors v_i, $i = 1,2,\ldots,r$ are assumed to be independently and multinomially distributed with parameters $n_{i.}$ and

$\underset{\sim}{\pi}_i = (\pi_{i_1}, \ldots, \pi_{i_k})'$.

Let H^* be a non-negative, strictly concave and twice differentiable function defined on R_+^k satisfying the condition (20). Let

$$(4.19) \qquad \nabla^2_{H^*}(x) = \left(\left(\frac{\partial^2 H^*(\underset{\sim}{\pi})}{\partial\pi_j, \partial\pi_j}\right)\right)_{\underset{\sim}{\pi}=x}$$

be the matrix of second order derivatives of H^*. Then, under H_0: $\underset{\sim}{\pi}_1 = \underset{\sim}{\pi}_2 = \ldots = \underset{\sim}{\pi}_r = \underline{\pi}$ as $n_{i.} \to \infty$ and $\frac{n_{i.}}{n_{..}} \to \lambda_i > 0$, asymptotically,

$$[\mathrm{I}] \qquad 2n_{..}\hat{SSB} \sim - \sum_{j=1}^{k-1} \beta_j \chi_j^2(r-1)$$

where $\chi_j^2(r-1)$, $i = 1, \ldots, k-1$ are i.i.d. χ^2 random variables with $(r-1)$ d.f. and β_j, $j = 1, 2, \ldots, (k-1)$ are possible non-zero eigenvalues of

$$(4.20) \qquad \nabla^2_{H^*}(\underset{\sim}{\pi}) \cdot D_{\underset{\sim}{\pi}} = \Delta^2_{H^*} \quad \text{(say)}$$

[II] $\hat{SST}$ and $\hat{SSB}$ are independently distributed.

For a proof of results (i) and (ii), see Naik (1985b).

For testing the null hypothesis that X and Y are independent, i.e. H_0: $\underset{\sim}{\pi}_1 = \ldots = \underset{\sim}{\pi}_r$ agianst the general alternative, a natural criteria, based on analysis of diversity using a diversity measure H, would be to reject H_0 at level α, if $\hat{SSB} \geq c$, choosing c such that

$$(4.21) \qquad P\{\hat{SSB} \geq c \mid H_0\} = \alpha .$$

The result [I] of Naik (1985), cited above, becomes useful for determining the critical value c. For each of the proposed, diversity measures H_m, $m = 1,2,3$

we have computed the matrix $\Delta^2_{H^*_m}$, $m = 1,2,3$ with the help of the extensions H^*_m given in (4.9), (4.10) and (4.11). Although in practice it is possible with the help of existing computer programme, to compute the eigenvalues of $\Delta^2_{\hat{H}^*_m}$ based on the estimate $\hat{\underset{\sim}{\pi}} = p_{..}$, the following approximation for the asymptotic distribution of $\hat{SSB}$ can be used. Approximating the eigenvalues α_i, $i = 1,2,\ldots,k-1$ by their average value

$$(4.22) \qquad \bar{\beta}_{H^*} = \frac{1}{k-1} \sum_{j=1}^{k-1} \beta_j$$

$$= \frac{1}{k-1} \operatorname{Tr}(\Delta^2_{H^*}) = \frac{1}{k-1} \Sigma \pi_j d_{jj}(\underset{\sim}{\pi})$$

where

$$d_{jj}(x) = \frac{\partial^2}{\partial \pi_j^2} H^*(\underset{\sim}{\pi}) \bigg|_{\underset{\sim}{\pi} = x} ,$$

the asymptotic distribution of $2n_{..}\hat{SSB}$ can be approximated by the distribution of $-\bar{\beta}_{H^*}\chi^2(r-1)(k-1)$. Further, if the estimator

$$(4.23) \qquad \hat{\bar{\beta}}_{H^*} = \frac{1}{k-1} \Sigma\, p_{..i} d_{ii}(p_{..})$$

is a consistent estimator of $\bar{\beta}_{H^*}$, then we shall have

$$(4.24) \qquad \frac{n_{..}\hat{SSB}}{-\bar{\beta}_{H^*}} \sim \chi^2(r-1)(k-1),$$

see Naik (1985). Light and Margolin (1971) using Gini-Simpson index of diversity and Nayak (1984) using quadratic entropy of Rao, have found the above approximation useful.

The matrix $\Delta^2_{H^*_m}$ and the average eigenvalue $\bar{\beta}_{H^*_m}$ corresponding to each of the proposed diversity measures H_m, $m=1,2,3$, along with the elements of $\nabla^2_{H^*_m}$ are as follows.

For $m=1,2,3$ and $\underset{\sim}{\pi} \in S^k$, let

$$(4.25) \qquad d_{m;jj'} = \frac{\partial^2}{\partial\pi_j \partial\pi_{j'}} H^*_m(\underset{\sim}{\pi}); \quad 1 \le j, j' \le k,$$

then, for $m=1$,

$$(4.26) \qquad d_{1,jj} = a_1\gamma\{(2\pi_j-1)B_j(b_1-\pi_j)^{-\gamma-2}-S_1\}$$

$$(4.27) \qquad d_{1,jj'} = a_1\gamma\{\pi_j B_j(b_1-\pi_j)^{-\gamma-2}+\pi_{j'}B_{j'}(b_1-\pi_{j'})^{-\gamma-2}-S_1\}$$

where

$$a_1 = k^{-\gamma}, \quad b_1 = 1+k^{-1}$$

$$B_t = 2b_1-(1-\gamma)\pi_t, \text{ and } S_1 = \sum_{\ell=1}^{k} \frac{\pi_\ell^2 B_\ell}{(b_1-\pi_\ell)^{\gamma+2}}$$

For $m=2$

$$(4.28) \qquad d_{2,jj} = a_2\gamma\{(2\pi_j-1)\pi_j^{\gamma-1}C_j(b_2-\pi_j^\gamma)^{-3}-S_2\},$$

$$(4.29) \qquad d_{2,jj'} = a_2\gamma\{\pi_j^\gamma C_j(b_2-\pi_j^\gamma)^{-3}+\pi_{j'}^\gamma C_{j'}(b_2-\pi_{j'}^\gamma)^{-3}-S_2\}$$

where

$$a_2 = 1-k^{-\gamma}, \quad b_2 = 2-k^{-\gamma}$$

$$C_t = (1+\gamma)b_2-(1-\gamma)\pi_t^\gamma, \quad \text{and} \quad S_2 = \Sigma \frac{\pi_\ell^{1+\gamma}C_\ell}{(b_2-\pi_\ell^\gamma)^3}$$

Finally for $m = 3$,

(4.30) $d_{3,jj} = a_3\gamma\{(2\pi_j-1)\tilde{\pi}_j^{\gamma-2}D_j(b_3+\tilde{\pi}_j^{\gamma})^{-3}-S_3\}$

(4.31) $d_{3,jj'} = a_3\gamma\{\pi_j\tilde{\pi}_j^{\gamma-2}D_j(b_3+\tilde{\pi}_j^{\gamma})^{-3}+\pi_{j'}\tilde{\pi}_{j'}^{\gamma-2}D_{j'}(b_3+\tilde{\pi}_{j'})^{-3}-S_3\}$

where

$$a_3 = (1-k^{-1})^{\gamma},\ b_3 = (1-k^{-1})^{\gamma},\ \tilde{\pi}_t = 1-\pi_t,$$

$$D_t = b_3(1+\gamma)\tilde{\pi}_t+(1+\gamma)\tilde{\pi}_t^{\gamma}+(1-\gamma)\tilde{\pi}_t^{1+\gamma}+b_3(1-\gamma)$$

and

$$S_3 = \sum_{\ell=1}^{k}\frac{\pi_\ell^2\tilde{\pi}_\ell^{\gamma-2}C_\ell}{(b_3+\tilde{\pi}_\ell^{\gamma})^3}.$$

With these computations, the matrix $\Delta^2_{H_j^*}$, for $j = 1,2,3$ can be worked out as

(4.32) $$\Delta^2_{H_m^*}(\underset{\sim}{\pi}) = ((d_{m,jj'}))\cdot D_{\underset{\sim}{\pi}},$$

(4.33) $$\bar{\beta}_{H_m^*} = \frac{1}{k-1}\sum_{j=1}^{k}\pi_j d_{m,jj}.$$

Case $\gamma = 1$ For the diversity measures $G_1(\underset{\sim}{\pi})$ and $G_2(\underset{\sim}{\pi})$ ($= G_3(\underset{\sim}{\pi})$) defined in (15), we have

(4.34) (i): $$\Delta^2_{G_1^*}(\underset{\sim}{\pi}) = ((d_{jj'}))\cdot D_{\underset{\sim}{\pi}}$$

$$\bar{\beta}_{G_1^*} = \frac{1}{k-1}\sum_{j=1}^{k}d_{jj}$$

where

$$d_{jj} = 2a_1^* b_1^* \left[\frac{(2\pi_j - 1)}{(b_1^* - \pi_j)^3} - \Sigma \frac{\pi_\ell^2}{(b_1^* - \pi_\ell)^3} \right.$$

and

$$d_{jj'} = 2a_1^* b_1^* \left[\frac{\pi_j}{(b_1^* - \pi_j)^3} + \frac{\pi_{j'}}{(b_1^* - \pi_{j'})^3} - \Sigma \frac{\pi_\ell^2}{(b_1^* - \pi_\ell)^3} \right]$$

where $a_1^* = k^{-1}$, $b_1^* = 1+k^{-1}$

(4.35) (ii): $\Delta^2_{G_2^*}(\underset{\sim}{\pi}) = ((d'_{jj'}))\ .\ D_{\underset{\sim}{\pi}}$

where

$$d'_{jj} = 2a_2^* b_2^* \left[\frac{(2\pi_j - 1)}{(b_2^* - \pi_j)^3} - \Sigma \frac{\pi_\ell^2}{(b_2^* - \pi_\ell)^3} \right]$$

$$d'_{jj'} = 2a_2^* b_2^* \left[\frac{\pi_j}{(b_2^* - \pi_j)^3} + \frac{\pi_{j'}}{(b_2^* - \pi_{j'})^3} - \Sigma \frac{\pi_\ell^2}{(b_2^* - \pi_\ell)^3} \right.$$

$a_2^* = 1-k^{-1}$, and $b_2^* = 2-k^{-1}$

(4.36) (iii): $\bar{\beta}_{G_m^*} = \dfrac{-a_m^* b_m^*}{k-1} \displaystyle\sum_{\ell=1}^{k} \frac{\pi_\ell(1-\pi_\ell)}{(b_m^* - \pi_\ell)^3}$; $m = 1,2.$

ACKNOWLEDGEMENTS

The author wishes to thank Professor B.K. Sinha and Professor Tapan Nayak for their helpful comments.

REFERENCES

[1] Agresti, A. and Agresti, B.F. (1978). Some statistical analysis of qualitative variation. Social Methodology (K.F. Schussler, Ed.) 204-237.

[2] Anderson, R.J. and Landis, J.R. (1980). CATANOVA for multidimensional contingency tables: nomianl scale response. Comm. Stat. A9(11), 1191-1206.

[3] Dennis, B., Patil, G.P., Rossi, O., Stehman, S. and Taillie, C. (1979). A bibliography of literature on ecological diversity and related methodology, in Ecological Diversity in Theory and Practice, Vol. 1 CPH, 319-354.

[4] Greenberg, Joseph H. (1956). The measurement of linguistic diversity, Language, 32, 109-115.

[5] Guiraud, P. (1959). Problemes et Methodes de la Statistique Linguistique, Dordrecht: D Reidel.

[6] Haberman, S.J. (1982). Analysis of dispersion of multinomial responses. J. Amer. Stat. Asso. 77, 568-580.

[7] Havrda, M.E. and Charvat, F. (1967). Quantification method of classification processes: Concept of structural α- entropy. Kybernetika 3, 30-35.

[8] Herdan, G. (1964). Quantitative Linguistics, Washington, DC: Butterworth, Inc.

[9] Herdan, G. (1966). The Advanced Theory of Language as Choice and Chance, New York: Springer-Verlag.

[10] Horvath, W.J. (1963). A stochastic model for word association tests. Psychological Review, 70, 361-364.

[11] Light, R.J. and Margolin, B.H. (1971). An analysis of variance for categorical data. J. Amer. Stat. Asso. 66, 534-544.

[12] Light, R.J. and Margolin, B.H. (1974). An analysis of variance for categorical data, II: small sample comparisons with chi-square and other competitors. J. Amer. Stat. Asso. 69, 755-764.

[13] Nayak, T.K. (1984). An analysis of diversity using Rao's quadratic entropy. Tech. Rept. 84-15, Center for Multivariate Analysis, University of Pittsburgh.

[14] Nayak, T.K. (1985a). On diversity measures based on entorpy functions. Comm. in Statistics, 14, 203-215.

[15] Nayak, T.K. (1985b). On asymptotic sampling distributions useful in analysis of diversity [unpublished].

[16] Patil, G.P. and Taillie, C. (1979). An overview of diversity, in Ecological Diversity in Theory and Practice, 1, CPH, 3-28.

[17] Patil, G.P. and Taillie, C. (1982). Diversity as a concept and its measurement. J. Amer. Stat. Asso. 77, 548-567.

[18] Pielou, E.C. (1975). Ecological Diversity, Wiley, New York.

[19] Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. J. Roy. Statist. Soc. B 10, 159-193.

[20] Rao, C.R. (1982a). Gini-Simpson index of diversity: a characterization, generalization and applications. Utilitas Mathematics 21, 273-282.

[21] Rao, C.R. (1982b). Diversity: its measurement, decomposition, apportionment and analysis. Sankhya, A44, 1-21.

[22] Rao, C.R. (1982c). Diversity and dissimilarity coefficients: a unified approach. Theo. Popln. Bio., 21, 24-43.

[23] Rao, C.R. and Nayak, T.K. (1985). Cross entorpy, dissimilarity measures and characterizations of quadratic entropy. To appear in IEEE Trans. Inf. Th.

[24] Renyi, A. (1961). On measures of entropy and information, in Proceedings, Fourth Berkeley Symp. Vol. 1, 547-561.

[25] Savchuk,A.P. (1964). On estimates for the entropy of a language according to Shannon. Th. Prob. Appln. 9, No. 1, 138-141.

[26] Sokal, R.R. and Sneath, P.H.A. (1963). Principles of Numerical Taxonomy Freeman, San Francisco.

[27] Yule, G. Udny (1944) The Statistical Study of Literary Vocabulary, London, Cambridge University Press.

[28] Zehna, P.W. (1966). Invariance of maximum likelihood estimation. Ann. Math. Statist, 37, 744.

[29] Zinger, A. (1982). Word association tests: A statistical approach. Psychological Reports, 1982, 50, 283-287.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	AFOSR-TR- 85-0748
2. GOVT ACCESSION NO.	AD-A160 873
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle)	New Measures of Diversity
5. TYPE OF REPORT & PERIOD COVERED	Technical - June, 1985
6. PERFORMING ORG. REPORT NUMBER	85-23
7. AUTHOR(s)	Manzoor Ahmad
8. CONTRACT OR GRANT NUMBER(s)	F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS	Center for Multivariate Analysis 515 Thackeray Hall University of Pittsburgh, Pittsburgh, PA15260
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS	Air Force Office of Scientific Research Department of the Air Force Bolling Air Force Base, DC 20332
12. REPORT DATE	June, 1985
13. NUMBER OF PAGES	24
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	
15. SECURITY CLASS. (of this report)	Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

The problem of measuring diversity within populations and dissimilarity or similarity between populations has been extensively treated in the literature. In this context a general procedure called Analysis of Diversity has been outlined and examined by C.R. Rao in a series of papers.

In this paper we propose three new measures of diversity and study related inference problems.
Continued

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Denote by S^k the simplex $S^k = \{\underset{\sim}{\pi}: \underset{\sim}{\pi} = (\pi_1, \ldots, \pi_k)', \pi_j \geq 0, \Sigma\pi_j = 1\}$. Then the proposed measures are of the form: $H_m(\underset{\sim}{\pi}) = 1 - a_m \Sigma_j \pi_j \phi_m(\pi_j)$, $m = 1,2,3$ where $\phi_1(x) = (1+k^{-1}-x)^{-\gamma}$, $\gamma \geq 0$, $\phi_2(x) = (2-k^{-\gamma}-x^{\gamma})^{-1}$, $\gamma \geq 0$, $\phi_3(x) = (a_3+(1-x)^{\gamma})^{-1}$, $0 < \gamma \leq 1$, and the a's are suitable normalizing constants. Estimation of $H_m(\underset{\sim}{\pi})$, derivation of the penalty function and cross entropy and the problem of testing independence have been treated. Asymptotic distributions of relevent test statistics are indicated.