Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 353–362

Measurement of Distance between Populations over Attributes under Multiple Samples

Valluri S. Rao and J.S. Murty[1]
Harvard Medical School, Boston, USA

SUMMARY

Some measures for the distance between two populations over attribute data have been defined when there are multiple samples from the same population. These distance measures are analogous to Mahalanobis distance. The relationships between the distance measures along with certain asymptotic properties have been discussed.

Key words : Distance measure, Attribute data, Mahalanobis distance, Asymptotic distribution, Asymptotic variance.

1. *Introduction*

Several measures of distances based on proportions of attributes have been proposed to measure the diversity between two populations, some of which deal in particular with genetic diversity based on gene frequencies (Bhattacharya [4], Sanghvi [20], Cavalli-Sforza and Edwards [6], Balakrishnan and Sanghvi [2], Morton *et al.* [12], Nei [13]). Only few of these measures use specifically the properties of the multinomial distribution of the frequencies of the attributes. Essentially, these distance measures are based on a single random sample of proportions from each of the two populations. When there are more samples (or subsets) from the same population, some studies attempt to test the diversity between and within populations through heterogeneity chi-squares (Balakrishnan [3]) while other studies attempt to understand the genetic diversity or relatedness either in terms of Wright's F-statistics (Nei [14], Long [10]) or in terms of intraclass correlations (Reynolds *et al.* [18],Weir and Cockerham [21]). However, none of these procedures uniquely define measures of distance between populaions. In this paper, some distance measures are defined based on the within sample covariance matrices estimated from several samples from each population. These distance measures are analogous to the Mahalanobis distance between two populations A and B given by

$$d^2(\underset{\sim}{\mu}_A, \underset{\sim}{\mu}_B ; \Sigma) = (\underset{\sim}{\mu}_A - \underset{\sim}{\mu}_B)' \Sigma^{-1} (\underset{\sim}{\mu}_A - \underset{\sim}{\mu}_B) \qquad (1.1)$$

[1] Department of Genetics, Osmania University, Hyderabad - 500 007, India

where $\underset{\sim}{\mu}_A$ and $\underset{\sim}{\mu}_B$ are p-dimensional mean vectors of the p-characters in the two populations and Σ is the common population variance covariance matrix the two populations possess.

2. *Some Distance Measures for Multiple Samples*

Consider sampling for p-attributes from two populations A and B. Let $\mathbf{X} = \{ x_{ij} \}$ be the set of n_1 samples from population A, x_{ij} being the proportion of the j-th attribute in the i-th sample. Similarly, let $Y = \{ y_{ij} \}$ be the set of n_2 samples from population B. Since x_{ij} and y_{ij} for all i and j are proportions of the p-attributes, they satisfy the conditions

$$\mathbf{X\,1} = \mathbf{1} \qquad \text{and} \qquad \mathbf{Y\,1} = \mathbf{1} \tag{2.1}$$

where $\mathbf{1}$ is a column vector of unity elements. Assume that this is the only linear constraint satisfied by the $\mathbf{X}$ and $\mathbf{Y}$ data matrices so that when $p < n_i$ for $i = 1, 2$, the rank of $\mathbf{X}$ and $\mathbf{Y}$ will be (p-1). The sample mean vectors from the populations A and B are given as:

$$\underset{\sim}{\overline{X}} = \frac{1}{n_1} \mathbf{X'1} \qquad \text{and} \qquad \underset{\sim}{\overline{Y}} = \frac{1}{n_2} \mathbf{Y'1} \tag{2.2}$$

One may regard the i-th set of proportions in A and j-th set of proportions in B as sample estimates from the multinomial populations $(m_{1i}, \pi_{i1}, \pi_{i2}, \ldots, \pi_{ip})$ and $(m_{2j}, \pi_{j1}, \pi_{j2}, \ldots, \pi_{jp})$ respectively. Corresponding to any set of sample proportions, say $(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})$, define a set of Bernoullian variables (U_{ikl}) with $P(U_{ikl} = 1) = \pi_k$ and $P(U_{ikl} = 0) = 1 - \pi_k$ for $k = 1, 2, \ldots, p$ and $l = 1, 2, \ldots, m_{1i}$ so that $x_{ik} = E(U_{ikl})$, is the mean of U_{ikl} over $l = 1, 2, \ldots, m_{1i}$. We then have the variance-covariance matrices between the attributes within the i-th sample in populations A and B respectively as

$$\begin{aligned} S_{iA} &= x_{ik}(1 - x_{ik}) && \text{for} \quad k = k' \\ &= -x_{ik}\,x_{ik'} && \phantom{\text{for}} \quad k \neq k' \end{aligned} \tag{2.3}$$

$$\begin{aligned} S_{jB} &= y_{jk}(1 - y_{jk}) && \text{for} \quad k = k' \\ &= -y_{jk}\,y_{jk'} && \phantom{\text{for}} \quad k \neq k' \end{aligned} \tag{2.4}$$

An appropriate weighted estimate of the population variance-covariance matrix Σ from a sample i of A and j of B is

$$S_{ij} = \frac{m_{1i}\,S_{iA} + m_{2j}\,S_{jB}}{m_{1i} + m_{2j}} \tag{2.5}$$

In view of the constraint (2.1), matrices S_{iA}, S_{jB}, and S_{ij} are singular and of rank (p - 1), and require deletion of a row and a column from each of them for unique inverses. Analogous to (1.1), the distance between the i-th sample of population A and j-th sample of population B is given by

$$d^2 (x_i, y_j, S) = \left[\underset{\sim}{X}^* - \underset{\sim}{Y}^* \right]' [S_{ij}]^{-1} \left[\underset{\sim}{X}^* - \underset{\sim}{Y}^* \right] \qquad (2.6)$$

where $\underset{\sim}{X}^*$ and Y^* are the (p-1) dimensional vectors obtained by deleting the last character with the corresponding $(p-1) \times (p-1)$ ordered matrix S. (2.6) is also analogous to the distance defined by Balakrishnan and Sanghvi [2]. Evidently, this is an estimate of distance based on one sample from A and a sample from B. However, in practice, we need a distance which can be regarded as an 'oveall' distance between A and B. There are a number of ways of defining such a measure of distance and we shall consider some of them hereunder :

(i) By considering the mean vectors $\underset{\sim}{\overline{X}}$ and $\underset{\sim}{\overline{Y}}$, one may estimate the distance, as defined by Bhattacharya [4], Sanghvi [20], Cavalli-Sforza and Edwards [6], Nei [13] or Morton *et al.* [12]. These distance functions do not make use of the variance-covariance matrix (2.3) or (2.4) of sample proportions. Denote such a measure of distance by $d^2 (\underset{\sim}{\overline{X}} , \underset{\sim}{\overline{Y}})$. For example, the distance based on Sanghvi's measure is

$$d^2 (\underset{\sim}{\overline{X}} , \underset{\sim}{\overline{Y}}) = 2 \Sigma \frac{(\overline{X}_j - \overline{Y}_j)^2}{(\overline{X}_j + \overline{Y}_j)} \qquad (2.7)$$

(ii) Consider $\underset{\sim}{\overline{X}}$ as a sample of proportions from population A of size $m_{1}. = \Sigma m_{1i}$ and $\underset{\sim}{\overline{Y}}$ as a sample of proportions from population B of size $m_{2}. = \Sigma m_{2j}$. The within sample covariance matrices are then obtained from (2.3) and (2.4) replacing x_{ik}, y_{jk} by the means $\overline{X}_j, \overline{Y}_j$ so that the corresponding estimate of the population variance-covariance matrix over the two populations is

$$S_{\overline{xy}} = \frac{m_{1.} S_{\bar{x}} + m_{2.} S_{\bar{y}}}{m_{1.} + m_{2.}} \qquad (2.8)$$

The corresponding Balakrishnan-Sanghvi distance is then given by

$$d^2 (\underset{\sim}{\overline{X}} , \underset{\sim}{\overline{Y}}, S_{\overline{xy}}) = [\underset{\sim}{\overline{X}}^* - \underset{\sim}{\overline{Y}}^*]' [S^*_{\overline{xy}}]^{-1} [\underset{\sim}{\overline{X}}^* - \underset{\sim}{\overline{Y}}^*] \qquad (2.9)$$

where $\underset{\sim}{\overline{X}}^*$, $\underset{\sim}{\overline{Y}}^*$, and S^*_{xy} are truncated as defined in (2.6).

If $S_{\bar{x}}$ and $S_{\bar{y}}$ are based on maximum likelihood estimates of proportions (as for example the ABO blood groups gene frequencies), then the dispersion matrix of the estimates is given by

$$S_{\bar{x}\bar{y}} = \frac{m_{1.}^2 \, S_{\bar{x}} + m_{2.}^2 \, S_{\bar{y}}}{m_{1.} + m_{2.}} \qquad (2.10)$$

and the distance between the two populations is then obtained using (2.9).

(iii) By finding the pooled mean proportions as

$$\bar{z}_j = \frac{m_{1.}\,\bar{x}_j + m_{2.}\,\bar{y}_j}{m_{1.} + m_{2.}} \qquad \text{for } j = 1, 2, \ldots, p \qquad (2.11)$$

and the within covariance matrix based on $\bar{z}_j$ using (2.3), one may define an index of distance based on Steinberg *et al.* (see Balakrishnan and Sanghvi [2]) as

$$d^2\,(\underset{\sim}{\bar{X}}\,,\ \underset{\sim}{\bar{Y}}\,,\ W_{\bar{z}}) = [\,\underset{\sim}{\bar{X}^*} - \underset{\sim}{\bar{Y}^*}\,]'\ [\,W_{\bar{z}}^*\,]^{-1}\ [\,\underset{\sim}{\bar{X}^*} - \underset{\sim}{\bar{Y}^*}\,] \qquad (2.12)$$

In this case, the inverse of within dispersion matrix can be found (see Kendall and Stuart [9]) as

$$W_{\bar{z}}^{*\,-1} = \begin{bmatrix} \frac{1}{\bar{z}_1} + \frac{1}{\bar{z}_p} & \frac{1}{\bar{z}_p} & \cdots & \frac{1}{\bar{z}_p} \\ \frac{1}{\bar{z}_p} & \frac{1}{\bar{z}_2} + \frac{1}{\bar{z}_p} & \cdots & \frac{1}{\bar{z}_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\bar{z}_p} & \frac{1}{\bar{z}_p} & \cdots & \frac{1}{\bar{z}_{p-1}} + \frac{1}{\bar{z}_p} \end{bmatrix} \qquad (2.13)$$

with $\underset{\sim}{\bar{X}^*}$., $\underset{\sim}{\bar{Y}^*}$ and $W^*_{\bar{z}}$ are truncated and as defined in (2.6), (2.12) on simplification yields

$$d^2\,(\underset{\sim}{\bar{X}}\,,\ \underset{\sim}{\bar{Y}}\,, W_{\bar{z}}) = \Sigma\ \frac{(\bar{X}_j - \bar{Y}_j)^2}{\bar{Z}_j} \qquad (2.14)$$

It may be noted that, if $m_{1.} = m_{2.}$ in (2.11), then (2.14) reduces to

$$d^2\,(\underset{\sim}{\bar{X}}\,,\ \underset{\sim}{\bar{Y}}\,,\ W_{\bar{z}}) = 2\,\Sigma\ \frac{(\bar{X}_j - \bar{Y}_j)^2}{(\bar{X}_j + \bar{Y}_j)} = d^2\ (\underset{\sim}{\bar{X}}, \underset{\sim}{\bar{Y}})\ \ (\text{of } (2.7)) \qquad (2.15)$$

We note that the distance in (2.12) is same as the one given by Chakraborty and Rao [7] and is equivalent to Sanghvi's distance (2.7) only if $m_{1.} = m_{2.}$.

We also note that (2.7) differs from that given by Chakraborty and Rao [7] by a multiple of 4. Further, deletion of frequency of one attribute is also necessary for finding (2.12) and (2.9) although it is not necessary for finding (2.7). Sanghvi's distance (2.7) is usually regarded as not taking into account the correlations between the multinomial proportions (Balakrishnan and Sanghvi [2]), but in view of the equivalence of (2.15) to (2.7), we note that it does not take into consideration the correlations in an indirect way, i.e. those between the pooled proportions.

[iv] Using the estimated covariance matrices S_{iA} and S_{jB} of (2.3) and (2.4) respectively, we estimate the Balakrishnan and Sanghvi's distance between the i-th sample of population A and j-th sample of population B. The average over the $n_1 \times n_2$ distances so generated may be used as an estimate of distance between the two populations:

$$d^{-2} (\underset{\sim}{X}_i , \underset{\sim}{Y}_j ; S_{xy}) = \frac{1}{n_1 n_2} \Sigma\, d_{ij}^2 (\underset{\sim}{\overline{X}}_{i}, , \underset{\sim}{\overline{Y}}_i ; S_{x_i y_j}) \qquad (2.16)$$

(v) Estimate the weighted averages of the covariance matrices S_{iA} and S_{jB} of (2.3) and (2.4) for the populations A and B respectively as

$$\overline{W}_A = \frac{1}{m_{1.} - n_1} [(m_{11} - 1) S_{1A} + (m_{12} - 1) S_{2A} + \ldots + (m_{1n_1} - 1) S_{n_1 A}]$$

and

$$\overline{W}_B = \frac{1}{m_{2.} - n_2} [(m_{21} - 1) S_{1B} + (m_{22} - 1) S_{2B} + \ldots (m_{2n_2} - 1) S_{n_2 B}]$$

The pooled covariance matrix over the two populations, A and B will be

$$\overline{S} = \frac{1}{m_{1.} + m_{2.} - n_1 - n_2 - 2} [(m_{1.} - n_1 - 1) \overline{W}_A + (m_{2.} - n_2 - 1) \overline{W}_B]$$

which is used to obtain the distance as

$$d^2 (\underset{\sim}{\overline{X}}, \underset{\sim}{\overline{Y}}, \overline{S}) = [\underset{\sim}{\overline{X}}^* - \underset{\sim}{\overline{Y}}^*]' [\overline{S}]^{-1} [\underset{\sim}{\overline{X}}^* - \underset{\sim}{\overline{Y}}^*] \qquad (2.17)$$

(vi) Considering the sets of sample proportions as independent samples from the same multinomial population, we define an alternative Mahalanobis like distance measure as follows :

We define the 'within population covariance matrices' as :

$$S_A = \frac{1}{n_1 - 1} X' H_X X$$

and

$$S_B = \frac{1}{n_2 - 1} Y' H_Y Y \tag{2.18}$$

where

$$H_X = I - \frac{1}{n_1} 1 1' \text{ and } H_Y = I - \frac{1}{n_2} 1 1'$$

are the centering matrices of order $n_1 \times n_1$ and $n_2 \times n_2$ respectively. It may be noted that the effect of the constraint (2.1) is to make $|S_X| = 0$ and $|S_Y| = 0$. Under the usual assumptions made in the case of Mahalanobis distance, one can show that these pooled within variance-covariance matrices are unbiased estimates of true covariance matrices obtained by replacing the estimates by their corresponding parameter values in the respective populations. For example, in the population A,

$$E[S_A(j, j)] = E\left[\frac{1}{n_1} \Sigma (x_{ij} - \bar{x}_j)^2\right]$$

$$= \frac{1}{n_1}\left[\Sigma [E(x_{ij}^2) - [E(\bar{x}_j)]^2\right]$$

and this by assumption is equivalent to

$$= \frac{1}{n_1}\left[\Sigma\left[\frac{\Pi_j(1 - \Pi_j)}{n_1} + \Pi_j^2\right] - n_1 \Pi_j^2\right]$$

$$= \frac{1}{n_1} \Pi_j (1 - \Pi_j)$$

Note the divisor to be n_1 for the variance to be unbiased. Also, observe that

$$E(S_A(j, k)) = -\frac{1}{n_1} \Pi_j \Pi_k$$

Similarly, analogous expressions may be obtained even for population B using S_B and π'. These are the variances and covariances of multinomial proportions for the j-th and jk-th characters in the two populations. As it is, the variance covariance matrices in the two populations will be in terms of π and π', hence will be unequal. Under the assumption of $\pi = \pi'$, an estimate of the common covariance matrix may be obtained by pooling the sample covariance matrices as given by

$$S_n^* = \frac{1}{n_1 + n_2}\left[n_1 \left(\frac{n_1 - 1}{n_1}\right) S_A + n_2 \left(\frac{n_2 - 1}{n_2}\right) S_B \right] \qquad (2.19)$$

A distance measure based on Mahalanobis distance is then given by

$$d^2 (\underset{\sim}{\overline{X}} , \underset{\sim}{\overline{Y}} , S_n) = [\underset{\sim}{\overline{X}}^* - \underset{\sim}{\overline{Y}}^*]\ [S_n^*]^{-1}\ [\underset{\sim}{\overline{X}}^* - \underset{\sim}{\overline{Y}}^*] \qquad (2.20)$$

where $\underset{\sim}{\overline{X}}^*$, $\underset{\sim}{\overline{Y}}^*$, and S_n^* are truncated and as defined in (2.6).

Remark: While the distance measures (2.12) and (2.9) essentially use the variance covariance matrix of Bernoullian variables, the distance (2.20) uses the variance covariance matrix of sample proportions and hence each element in the covariance matrix will have the number of samples as the divisor. Observe that under the assumption $m_1. = m_2.$ and $n_1 = n_2$,

$$d^2 (\underset{\sim}{\overline{X}} , \underset{\sim}{\overline{Y}} , S_n) = (n_1 + n_2)\ d^2 (\underset{\sim}{\overline{X}} , \underset{\sim}{\overline{Y}} , S_{\overline{xy}}) \qquad (2.21)$$

3. *Distribution of Properties of Distance*

The problem of classification of populations involves construction of an index, for example the distance, by which one can measure the resemblance or divergence between two populations. The choice of an appropriate distance measure poses a problem since each measure has its own relative merits and demerits (see for example Chakraborty and Rao [7]). One useful criterion could be the sampling variance of the distance measure but it is not always easy to obtain the variance of the estimated distance measure. Asymptotic variances of some distance measures have been obtained before (e.g., Nei and Roychoudhury [15]). A direct estimation of the variances of the measures defined here is formidable. However, one can consider estimates of the asymptotic distribution of the distance measure defined, since all these are analogous to Mahalanobis D^2. Hotelling [8], Bose and Roy [5] have shown that D^2 follows an F-distribution under the assumptions of asymptotic normality of mean vectors. In the present measures, although the mean vectors are vectors of multinomial proportions, one can apply the result given in Anderson ([1], pp. 163) which states that for large samples, the distribution of T^2 (and hence that of D^2) is approximately valid even if the parent distribution is not normal; and it is in this sense that the T^2 test is a robust procedure.

Hence

$$\frac{k_1 k_2}{k_1 + k_2}\ \frac{D^2}{k_1 + k_2 - 2}\ \frac{k_1 + k_2 - p - 1}{p} \sim F_{p, k_1 + k_2 - p - 1} \qquad (3.1)$$

where k_1 and k_2 are the appropriate sample sizes for the two populations. Under such assumption of asymptotic normality, one can regard the proposed distance measures (2.9), (2.12), (2.16), (2.17), and (2.20) as following an F-distribution with appropriate degrees of freedom. Then the variance for d^2 in (2.9), (2.12) and (2.20) will be

$$V_{d^2} = \frac{2p\,(m_{1.} + m_{2.})^2\,(m_{1.} + m_{2.} - 2)^2\,(m_{1.} + m_{2.} - 3)}{m_{1.}^2\,m_{2.}^2\,(m_{1.} + m_{2.} - p - 3)^2\,(m_{1.} + m_{2.} - p - 5)}$$

Similarly, corresponding estimates for (2.16) and (2.17) can be obtained as

$$\sum_{i,j} \frac{2p\,(m_{1i} + m_{2j})^2\,(m_{1i} + m_{2j} - 2)^2\,(m_{1i} + m_{2j} - 3)}{m_{1i}^2\,m_{2j}^2((m_{1i} + m_{2j} - p - 3)^2\,(m_{1i} + m_{2j} - p - 5)}$$

and

$$V_{d^2} = \frac{2p\,(\eta_1 + \eta_2)^2\,(\eta_1 + \eta_2 - 2)^2\,(\eta_1 + \eta_2 - 3)}{\eta_1^2\,\eta_2^2\,(\eta_1 + \eta_2 - p - 3)^2\,(\eta_1 + \eta_2 - p - 5)}$$

respectively with $\eta_1 = (m_{1.} - n_1)$ and $\eta_2 = (m_{2.} - n_2)$ in the above equation. It may be noted however that the distribution of d^2 in (2.7) is a chi-square since the F distribution for (2.12) tends to a chi-square when $m_{1.} = m_{2.}$ and is large.

However for purposes of clustering, one uses 'd' rather than d^2. Correspondingly, we obtain the distribution of 'd' which can be shown to be a beta distribution with parameters (α, β) given by (Rao [16]).

$$f(D) = \frac{1}{\sqrt{C}}\,\frac{p}{k_1 + k_2 - p - 1}\,\frac{1}{B(\alpha, \beta)}\,\frac{(kD^2)^{\alpha - 1}}{(1 + kD^2)^{\alpha + \beta}} \quad \text{for} \quad 0 \leq D < \infty$$

with k_1 and k_2 being the sample sizes of the respective populations measuring 'p' characters, and

$$\alpha = \frac{p + 1}{2}, \quad \beta = \frac{k_1 + k_2 - p - 2}{2}$$

and constants

$$C^2 = 2\left[\frac{k_1\,k_2\,p}{(k_1 + k_2)\,(k_1 + k_2 - 2)\,(k_1 + k_2 - p - 1)}\right]$$

and $$k = \frac{k_1 k_2}{(k_1 + k_2)(k_1 + k_2 - 2)}$$

Hence for measures (2.9), (2.12), (2.16), (2.17), and (2.20), the asymptotic mean and variance of 'd' will be

$$\mu = \frac{\alpha}{\beta - 1} \quad ; \quad \sigma^2 = \frac{\alpha(\alpha + \beta - 1)}{(\beta - 1)^2 (\beta - 2)}$$

The closeness of these asymptotic variances to the true variances depends upon the rate of convergence of the true distribution of D^2 or of D to the asymptotic form.

As against such asymptotic distributions of D^2 or of D, in practice these distributions will be based on finite sample sizes. Extensive simulations are also carried out to examine the nature of these distributions and are reported elsewhere (Rao and Murty [17]). Except in cases of small samples, the distribution of D generally agreed well with that of beta distribution (Rao [16]).

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to the Indian Council of Medical Research, New Delhi, for providing a fellowship to VSR.

REFERENCES

[1] Anderson, T.W., 1984. *An Introduction to Multivariate Statistical Analysis*. 2nd ed. John Wiley and Sons, New York.

[2] Balakrishnan, V. and Sanghvi, L.D., 1968. Distance between populations on the basis of attribute data. *Biometrics*, **24**, 859-865.

[3] Balakrishnan, V., 1976. A method to study allelic variability. In : *The Origin of the Australians*, (R.L. Kirk and A.G. Thorne, eds.), 411-414, Australian Institute of Aboriginal Studies, Canberra.

[4] Bhattacharya, A., 1946. On a measure of divergence between two multinomial populations. *Sankhya*, **7**, 401-406.

[5] Bose, R.C. and Roy, S.N., 1938. The distribution of the studentized D-statistic. *Sankhya*, **4**, 19-38.

[6] Cavalli-Sforza, L.L. and Edwards, A.W.F., 1967. Phylogenetic analysis : models and estimation procedures. *Am. J. Hum. Genet.*, **19** (3), 233.

[7] Chakraborty, R. and Rao, C.R., 1991. Measurement of genetic variation for evolutionary studies. In : *Handbook of Statistics*, **8**, (C.R. Rao and R. Chakraborty, eds.), 271-316, Elsevier Science Publishers.

[8] Hotelling, H., 1931. The generalization of Student's ratio. *Ann. Math. Stat.*, **2**, 360-378.

[9] Kendall, M. and Stuart, A., 1977. *The Advanced Theory of Statistics*. I. ed IV, Griffin, London.

[10] Long, C.J., 1986. The allelic correlation structure of Gainj- and Kalam-speaking people. 1. The Estimation and interpretation of Wright's F-statistics. *Genetics*, **112**, 629-647.

[11] Mahalanobis, P.C., 1936. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. (India)*, **12**, 49-55.

[12] Morton, N.E., Yee, S., Harris, D.E. and Lew, R., 1971. Bioassay of kinship. *Theor. Popul. Biol.*, **2**, 507-524.

[13] Nei, M., 1972. Genetic distance between populations. *Amer. Natrualist*, **106**, 283-292.

[14] Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci.*, USA, **70**, 3321-3323.

[15] Nei, M. and Roychoudhury, A.K., 1974. Sampling variance of heterozygosity and genetic distance. *Genetics*, **76**, 379-390.

[16] Rao, V.S., 1991. On multivariate distance measures. Ph.D. thesis (unpublished). Osmania University, India.

[17] Rao, V.S. and Murty, J.S., 1996. Distribution of distance between populations based on attribute data from multiple samples: Simulation studies on the role of nonlinear transformations. (In preparation).

[18] Reynolds, J., Weir, B.S. and Cockerham, C.C., 1983. Estimation of the coancestry coefficient : Basis for a short-term genetic distance. *Genetics*, **105**, 767-779.

[19] Nei, M. and Roychoudhury, A.K., 1974. Sampling variance of heterozygosity and genetic distance. *Genetics*, **76**, 379-390.

[20] Sanghvi, L.D., 1953. Comparison of genetical and morphological methods for a study of biological differences. *Amer. J. Phys. Anthrop.*, **11**, 385-404.

[21] Weir, B.S. and Cockerham, C.C., 1984. Estimating F-statistic for the analysis of population structure. *Evolution*, **38**, 1358-1370.