SELECTION AND SCREENING PROCEDURES IN MULTIVARIATE ANALYSIS

by

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399

and

S. Panchapakesan
Department of Mathematics
Southern Illinois University
Carbondale, IL 62901-4408

Technical Report # 92-15C

Department of Statistics
Purdue University

April, 1992

△

SELECTION AND SCREENING PROCEDURES IN MULTIVARIATE ANALYSIS*

by

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399

and

S. Panchapakesan
Department of Mathematics
Southern Illinois University
Carbondale, IL 62901-4408

ABSTRACT

This paper briefly reviews some of the past and recent developments in ranking and selection methodology for multivariate populations. This is done with a view to indicate and discuss directions for future research. The coverage of the subject matter includes: selection from a single multivariate normal population (Section 2), selection from several multivariate normal populations (Section 3), selection from a multinomial population (Section 4), selection from several multinomial populations (Section 5), selection of the best set of predictor variables in a linear regression model (Section 6), and classification procedures and selection in principal component analysis (Section 7). The final section provides additional comments on future directions.

AMS Classification: primary 62F07, secondary 62H99.

Key Words and Phrases: selection and ranking, indifference-zone, subset selection, multivariate normal, Mahalanobis distance, generalized variance, multiple correlation, multinomial, diversity measures, Shannon's entropy, Gini-Simpson index, linear regression, predictor variables, classification, principal components, future directions.

*This research was supported in part by NSF Grants DMS–8923071 and DMS–8717799.

1. INTRODUCTION

Statistical inference problems have been studied in the now familiar "ranking and selection" framework over the last four decades. These problems have been examined within this framework employing different goals and formulations under various model assumptions. As such, substantial progress has been achieved in studying selection from multivariate populations. Our objective in the present paper is to emphasize the need, potential, and some directions for future investigations. We review briefly some of the past and recent accomplishments in order to stress the need for further investigations of these aspects. We will also discuss aspects of multivariate selection problem that have been barely considered thus far. We divide the subject matter into: selection from a single multivariate normal population (Section 2), selection from several multivariate normal populations (Section 3), selection from a single multinomial distribution (Section 4), selection from several multinomial distributions (Section 5), selection from the predictor variables in a linear regression model (Section 6), and classification procedures and selection in principal component analysis (Section 7). The final section provides additional general remarks on future directions.

In most of the investigations, multivariate populations have been ranked in terms of a *scalar* function of the unknown parameters. This entails a complete (unknown) ordering of the populations according to the values of the chosen scalar function. The selection procedure in this situation depends on a suitably chosen (one-dimensional) statistic which has a univariate distribution. We then start with a brief introduction to the basic methodology of ranking and selection for univariate distributions.

Consider k independent populations $\pi_1, \ldots, \pi_k$ where π_i has the underlying distribution $F_{\theta_i}, i = 1, \ldots, k$. The θ_i are unknown real-valued parameters which represent the values of a quality characteristic of interest for these k populations. We define π_i to be *better than* π_j if $\theta_i > \theta_j$. The ordered θ_i are denoted by $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$. It is assumed that there is no prior knowledge regarding the correct pairing of the ordered and the unordered θ_i. Selection problems have been generally studied under one of two formulations, namely, (1) the *indifference-zone* and (2) the (random size) *subset* formulations.

For the basic problem of *selecting the best population* (i.e., the population associated with $\theta_{[k]}$), the indifference-zone formulation of Bechhofer (1954) stipulates that one of the

k competing populations be selected as the best. Selection of any population associated with $\theta_{[k]}$ results in a *correct selection* (CS). Any procedure R, to be *valid*, should guarantee a specified minimum probability of a correct selection (PCS), say $P^*(1/k < P^* < 1)$, whenever the best (assumed to be unique) and the second best populations are apart at least by a specified amount. Let $\delta(\theta_i, \theta_j)$ denote an appropriately defined nonnegative measure of the amount of separation between the population associated with θ_i and θ_j. For any specified $\delta^* > 0$, let Ω_{δ^*} be the subset of the parameter space $\Omega = \{\underset{\sim}{\theta} | \underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)\}$ defined by

$$\Omega_{\delta^*} = \{\underset{\sim}{\theta} | \delta(\theta_{[k]}, \theta_{[k-1]}) \geq \delta^*\}. \tag{1.1}$$

The subset Ω_{δ^*} is called the *preference-zone*. Letting $P(CS|R)$ denote the PCS of a rule R, in order to be valid, it should satisfy

$$P(CS|R) \geq P^* \text{ for all } \underset{\sim}{\theta} \in \Omega_{\delta^*}. \tag{1.2}$$

Both δ^* and P^* are specified by the experimenter in advance. Suppose R is based on samples of fixed size n from each population. The problem then is to determine the smallest sample size n for which the requirement (1.2) holds. The complement of Ω_{δ^*} w.r.t. Ω is the so-called *indifference-zone* where no requirement is made on the PCS.

In the subset selection formulation of Gupta (1956, 1965), the basic problem is to select a nonempty subset of the k given populations so that the best population is included in the selected subset with a guaranteed minimum probability P^*. In case of a tie for the best population, we assume that one of the contenders is tagged as the best. Selection of any subset that includes the best results in a correct selection. Any valid rule R should satisfy

$$P(CS|R) \geq P^* \text{ for all } \underset{\sim}{\theta} \in \Omega. \tag{1.3}$$

Note that the size S of the selected subset is not decided in advance but is determined by the data.

The requirements (1.2) and (1.3) are known as the *basic probability requirements* or the *P^*-conditions* of the two formulations. Any parametric configuration $\underset{\sim}{\theta}$ which yields the infimum of the PCS over Ω_{δ^*} or Ω, depending on the formulation, is called a *least favorable configuration* (LFC). For a valid subset selection rule R, the expected subset size, $E(S|R)$, has been generally used as a measure of its performance for a comparison with

another valid rule. Some alternative measures that have been used are $E(S)/P(CS|R)$ and $E(S) - P(CS|R)$, the latter being the expected number of non-best populations included in the selected subset.

Many variations and generalizations of the basic formulation using either of the two approaches have been extensively studied. There are also related problems such as selecting populations better than a standard or a control. A comprehensive survey of the developments involving all these aspects with an extensive bibliography is given by Gupta and Panchapakesan (1979). A critical review of developments in the subset selection theory with historical perspectives has been provided by Gupta and Panchapakesan (1985), who have reviewed in another paper (1987) subset selection procedures in multivariate models. Reference can also be made to Gibbons, Olkin and Sobel (1977, Chapter 15) who have given examples to illustrate the need for selection procedures for multivariate normal populations in terms of different criteria.

2. SELECTION FROM A SINGLE MULTIVARIATE NORMAL POPULATION

Consider a p-variate normal population $N_p(\underset{\sim}{\mu}, \Sigma)$ with mean vector $\underset{\sim}{\mu}' = (\mu_1, \ldots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{ij})$, which is assumed to be positive definite. We consider ranking the p components according to their means μ_i, and according to their variances σ_{ii}. The case of Σ with $\sigma_{ij} = 0, i \neq j$, is the case of independent components which has been extensively investigated in the literature and will not be discussed here.

2.1 Selection in Terms of the Means.

Our goal here is to select the component π_i associated with the largest mean, $\mu_{[p]}$. Let $\underset{\sim}{X}' = (X_1, \ldots, X_p)$ be the sample mean based on n independent (vector) observations.

Known Σ Case: Assume without loss of generality that $\sigma_{ii} = 1, i = 1, \ldots, p$. Gnanadesikan (1966) considered the subset selection procedure

$$R_1\text{: Select the component } \pi_i \text{ if and only if } X_i \geq X_{[p]} - \frac{d_1}{\sqrt{n}} \tag{2.1}$$

where $X_{[1]} \leq \ldots \leq X_{[p]}$ denote the ordered X_i, and $d_1 = d_1(n, p, \Sigma) > 0$ is the smallest number such that the P^*-condition is met. Letting $Y_i = \sqrt{n}(X_{(i)} - \mu_{[i]})$, where $X_{(i)}$ is the component of $\underset{\sim}{X}$ associated with $\mu_{[i]}$, it is easily seen that

$$\inf_{\Omega} P(CS|R_1) = Pr\{Y_p \geq Y_j - d_1, j = 1, \ldots, p-1\}. \tag{2.2}$$

In order to evaluate d_1 for which the right-hand side of (2.2) equals P^*, one needs to know $K = (\kappa_{ij})$, where $\kappa_{ij} = \text{cov}\,(Y_i, Y_j)$. Although Σ is known, the correspondence between the κ_{ij} and the k_{ij} is *not* known except when $p = 2$. For $p = 2$, we get

$$d_1 = d_1(n, 2, \Sigma) = \sqrt{2(1 - \sigma_{12})}\Phi^{-1}(P^*) \tag{2.3}$$

where $\Phi(\cdot)$ is the standard normal cdf. For $p > 2$, Gnanadesikan (1966) obtained two different lower bounds for the infimum of PCS, one using a well-known inequality of Slepian (1962) and the other using a Bonferroni inequality. These bounds yield conservative values for d_1. In the special case of equal positive correlation ρ, the value of d_1 is given by

$$\int_{-\infty}^{\infty} \Phi^{p-1}\left(x + \frac{d_1}{\sqrt{1-\rho}}\right) d\Phi(x) = P^* \tag{2.4}$$

and $H = d_1/\sqrt{2(1-\rho)}$ is tabulated for selected values of P^* and ρ by Gupta (1963a), and Gupta, Nagel and Panchapakesan (1973) who have also considered the selection problem in this special case.

For selecting the component associated with $\mu_{[p]}$ using the indifference-zone formulation with the preference-zone $\Omega_{\delta^*} = \{\underset{\sim}{\mu} | \mu_{[p]} - \mu_{[p-1]} \geq \delta^* > 0\}$, one can use the natural rule

R_1': Select the component that yields the largest X_i. (2.5)

In this case, we have

$$\inf_{\Omega_{\delta^*}} P(CS|R_1') = Pr\{Y_p \geq Y_j - \sqrt{n}\delta^*, j = 1, \ldots, p-1\}. \tag{2.6}$$

Comparing (2.6) and (2.2), we see that the minimum sample size needed is the smallest integer $n \geq (d_1/\delta^*)^2$, where d_1 is the constant to be used in rule R_1.

Unknown Σ *Case*: Assume that $\sigma_{ii} = \sigma^2, i = 1, \ldots, p$, that s_ν^2 is an estimator of σ^2 on ν degrees of freedom independent of the sample mean vector $\underset{\sim}{X} = (X_1, \ldots, X_p)$, and that $\nu s_\nu^2/\sigma^2$ has a chi-square distribution with ν degrees of freedom. In this case, for selecting the component associated with the largest mean, Gnanadesikan (1966) proposed a subset selection rule

R_2: Select the component π_i if and only if $X_i \geq X_{[p]} - \dfrac{d_2 s_\nu}{\sqrt{n}}$ (2.7)

where $d_2 = d_2(\nu, p, P^*) > 0$ is the smallest number for which the P^*-condition is met. Gnanadesikan (1966) has shown that $P(CS|R_2)$ is minimized when the μ_i are all equal. Letting Ω_0 denote the subset of Ω where the μ_i are equal, Gnanadesikan (1966) has shown that

$$P(CS|R_2) = Pr\{t_j \leq d_2/\sqrt{2(1-a_{pj})}, j = 1, \ldots, p-1\} \text{ for } \underset{\sim}{\mu} \in \Omega_0, \tag{2.8}$$

where $t_j = Z_j/s_\nu$, $\underset{\sim}{Z}' = (Z_1, \ldots, Z_{p-1})$ has $N_{p-1}(\underset{\sim}{0}, B)$, B has a known structure involving $a_{pj}, j = 1, \ldots, p-1$, and a_{pj} is the correlation between the components associated with $\mu_{[p]}$ and $\mu_{[j]}$. Letting $d_{03} = \min\{d_2/\sqrt{2(1-a_{pj})}, j = 1, \ldots, p-1\}$,

$$\begin{aligned} \inf_{\Omega} P(CS|R_2) &\geq Pr\{t_j \geq d_{02}, j = 1, \ldots, p-1\} \\ &\geq 1 - \sum_{j=1}^{p-1} Pr\{t_j \geq d_{02}\}. \\ &= (2-p) + (p-1)F_\nu(d_{02}), \end{aligned} \tag{2.9}$$

where $F_\nu(\cdot)$ is the cdf of a Student's t variable with ν degrees of freedom. Thus a conservative value for d_2 of the rule R_2 is given by d_{02} which is the solution of

$$(2-p) + (p-1)F_\nu(d_{02}) = P^*. \tag{2.10}$$

If we assume that $\sigma_{ij} = \rho\sigma^2$, where σ^2 is unknown and $\rho > 0$ is known, then d_{03}, a conservative value for d_3, can be evaluated as an equicoordinate percentage point of a multivariate t distribution. The d_{03} values in this case are tabulated by Gupta and Sobel (1957), Krishnaiah and Armitage (1966), and Gupta, Panchapakesan and Sohn (1985).

Frischtak (1973) considered the same problem using the indifference-zone approach with preference zone $\Omega_{\delta^*} = \{\underset{\sim}{\mu}, \Sigma | \mu_{[p]} - \mu_{[p-1]} \geq \delta^*\}$. However, he assumed that $\Sigma = \sigma^2\Gamma$, where σ^2 is *known* and $\Gamma = (\rho_{ij})$, the associated correlation matrix is unknown. Without loss of generality, we take $\sigma = 1$. Based on the mean $\underset{\sim}{X} = (X_1, \ldots, X_p)$ of n independent observations, Frischtak (1973) proposed the natural rule

$$R_2'\text{: Select the component that yields the largest } X_i, \tag{2.11}$$

where n is the minimum sample size needed to satisfy the P^*-condition. Letting, as before, $Y_i = \sqrt{n}(X_{(i)} - \mu_{[i]})$, it is easily shown that

$$P(CS|R_2') \geq Pr\{Z_j \leq a(n)(1-k_{jk})^{-1/2}, j = 1, \ldots, p-1\} \tag{2.12}$$

where $a(n) = \delta^*\sqrt{n/2}$, $\kappa_{jk} = \text{cov}\,(Y_j, Y_k)$, and $Z_j = (Y_j - Y_k)/\{2(1-\kappa_{jk})\}^{1/2}, j = 1, \ldots, k-1$. The Z_j are correlated standard normal random variables with

$$\text{cov}\,(Z_i, Z_j) = (1 - \kappa_{ik} - \kappa_{jk} + \kappa_{ij})/[2(1-\kappa_{ik})^{1/2}(1-\kappa_{jk})^{1/2}]. \tag{2.13}$$

Frischtak (1973) has shown that the infimum of P_1, the right-hand side of (2.12), occurs when $|\Gamma| = 0$ for $p = 3$. But no analytical result is available on the configuration of the ρ_{ij} that gives the global minimum of P_1. Further, very little small sample results are available for $p > 3$. One can, of course, get a conservative approximation to the sample size needed by using a Bonferroni inequality to obtain a lower bound for P_1.

If the correlations ρ_{ij} are all equal to $\rho > 0$, then

$$P(CS|R_2') \geq Pr\{Z_j \leq a(n)(1-\rho)^{-1/2}, j = 1, \ldots, p-1\} \tag{2.14}$$

where the Z_j are now equally correlated with correlation $\frac{1}{2}$. Since $-(p-1)^{-1} \leq \rho \leq 1$, we have

$$\inf_{\Omega_{S^*}} P(CS|R_2') = Pr\{Z_j \leq a(n)[(p-1)/p]^{1/2}, j = 1, \ldots, p\}. \tag{2.15}$$

The smallest sample size needed for satisfying the P^*-condition can now be obtained using the tables of Gupta (1963a) or Gupta, Nagel and Panchapakesan (1973).

2.2 Selection in Terms of the Variances

We now define the best component as the one associated with the smallest σ_{ii}. Let $S = (s_{ij})$ be the sample covariance matrix based on n independent (vector) observations from the population. Let $s_i^2 = s_{ii}$ so that $s_{[1]}^2 \leq \ldots \leq s_{[k]}^2$ are the ordered s_i^2. A natural subset selection rule for our goal is

$$R_3\text{: Select the component } \pi_i \text{ if and only if } s_{[i]}^2 \leq \frac{1}{c_3} s_{[1]}^2 \tag{2.16}$$

where $0 < c_3 = c_3(p, n, P^*) < 1$ is the largest number for which the P^*-condition is satisfied. Frischtak (1973) has considered this procedure R_3 and shown that the infimum of PCS is attained when $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} = 0$ for $p = 2$. Thus the constant c_3 in this case can be obtained for selected values of n and P^* from the tables of Gupta and Sobel (1962b) who have considered the selection problem in the uncorrelated case in a companion paper (1962a). For $p \geq 3$, Frischtak (1973) obtained only an asymptotic ($n \to \infty$) solution using

the asymptotic normality after suitable normalization of $\log(s^2_{(1)}/s^2_{(j)}), j = 2, \ldots, p$, where $s^2_{(i)}$ is the sample variance associated with the ith smallest σ_{ii}. The asymptotic solution for c_3 is given by

$$Pr\{Z_j \leq \sqrt{\frac{n-1}{2}} \log c_3, j = 2, \ldots, p\} = P^* \tag{2.17}$$

where the Z_j are equally correlated standard normal variables with correlation $\frac{1}{2}$, and thus can be obtained from the tables of Gupta (1963a) or Gupta, Nagel and Panchapakesan (1973).

Frischtak (1973) has also investigated the natural rule R'_3 which selects the component associated with the smallest s^2_i under the indifference zone formulation with $\Omega_{\delta^*} = \{(\underline{\mu}, \Sigma): \sigma^2_{[2]} \geq \delta^* \sigma^2_{[1]}, \delta^* > 1\}$, where $\sigma^2_i = \sigma_{ii}$. He has shown that, for $p = 2$, the infimum of PCS occurs when $\sigma^2_{[2]} = \delta^* \sigma^2_{[1]}$ and $\rho_{12} = 0$. For $p > 2$, no exact solution is available. One can obtain a conservative approximation or a large sample solution.

Frischtak (1973) also considered a generalization of the problem of selecting the component associated with the smallest σ_{ii}. Consider a partitioning of the p variates into $m (2 \leq m \leq p)$ subclasses, denoting the covariance matrices of these subclasses by $\Sigma_{ii}, i = 1, \ldots, m$. Then $\lambda_i = |\Sigma_{ii}|$ is the generalized variance of the ith subclass. The goal is to select the subclass associated with $\lambda_{[1]}$, the smallest generalized variance. Frischtak (1973) considered the natural procedure using the indifference zone formulation. He obtained only an asymptotic solution using the asymptotic normality of $(n-1)^{1/2}(\log|S_1| - \log\lambda_1, \ldots, \log|S_m| - \log\lambda_m)$.

2.3 Remarks and Future Directions

Although selection of the best component (suitably defined) from a multivariate normal population is a meaningful problem, there is very little published work on it. Most of the procedures discussed in this regard are from dissertations. For selection in terms of means, generally only conservative solutions are available based on lower bounds to the infimum of the PCS. In the case of R_2, it is assumed that an estimator of σ^2 independent of the X_i is available having (properly scaled) a chi-square distribution. One can obtain such an estimator from any one of the marginal distributions; however, using only one of them is expected to be inefficient. If we use the average of the estimators of σ^2 from all the p marginal distributions, as one would, its distribution is unknown and the setup of R_2 does not hold. This case needs further investigation.

In the independent case, where $\Sigma = \sigma^2 I$ and σ^2 is unknown, an indifference-zone procedure based on single sample does not exist. Bechhofer, Dunnett and Sobel (1954) studied a two-stage procedure in this case. It is interesting to ask: What is the correlated case analogue when Σ is unknown?

For selection in terms of the variances, there are no exact solutions for $p \geq 3$. In the case of R_3', the LFC is known only for $p = 2$.

Further, there are other goals of interest. For example, one may want to select the pair of components that have the largest (smallest) correlation or the largest (smallest) absolute correlation. These have not been studied.

3. SELECTION FROM SEVERAL MULTIVARIATE NORMAL POPULATIONS

Let $\pi_1, \ldots, \pi_k$ be k p-variate normal populations, $N_p(\underset{\sim}{\mu}_i, \Sigma_i), i = 1, \ldots, k$, where the $\underset{\sim}{\mu}_i$ are the mean vectors and the Σ_i are positive definite covariance matrices. We consider several measures for ranking the populations such as the generalized variance, Mahalanobis distance, and the multiple correlation coefficient. We also consider comparison with a control using as criteria linear combinations of the elements of the mean vector and those of the covariance matrix.

3.1 Selection in Terms of Mahalanobis Distance

Let $\lambda_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$, the Mahalanobis distance from the origin. We discuss the selection of the population associated with the largest λ_i. The case of the smallest λ_i can be treated in an analogous manner.

3.1.1 *Subset Selection Procedures.*

Here we consider three cases: (i) known Σ_i, (ii) unknown Σ_i, not necessarily equal, and (iii) $\Sigma_1 = \ldots = \Sigma_k = \Sigma$, unknown. Let $\underset{\sim}{X}_{ij}, j = 1, \ldots, n$, denote n independent observations from $\pi_i, i = 1, \ldots, k$. Define $Y_{ij} = \underset{\sim}{X}_{ij}' \Sigma_i^{-1} \underset{\sim}{X}_{ij}, Y_i = \sum_{j=1}^{n} Y_{ij}, Z_i = \overline{\underset{\sim}{X}}_i' \Sigma_i^{-1} \overline{\underset{\sim}{X}}_i$, and $T_i = \overline{\underset{\sim}{X}}_i' S_i^{-1} \overline{\underset{\sim}{X}}_i, i = 1, \ldots, k$, where $\overline{\underset{\sim}{X}}_i$ and S_i are the sample mean vector and covariance matrices.

Case (i): The Σ_i are known. In this case, Gupta (1966) proposed the rule

$$R_4\text{: Select } \pi_i \text{ if and only if } Y_i \geq c_4 Y_{[k]} \tag{3.1}$$

where $0 < c_4 = c_4(k, p, n, P^*) < 1$ is the largest number for which the P^*-condition is satisfied. It has been shown [Gupta (1966) and Gupta and Studden (1970)] that the LFC

is $\underset{\sim}{\lambda} = (\lambda_1, \ldots, \lambda_k) = (0, \ldots, 0)$. Thus the constant c_4 is given by

$$\int_0^\infty G_\nu^{k-1}(x/c_4) dG_\nu(x) = P^* \tag{3.2}$$

where $G_\nu(x)$ is the cdf of a standardized (i.e. unit scale parameter) gamma variable with $\nu = np/2$ degrees of freedom. The values of c_4 are tabulated by Gupta (1963b), and Armitage and Krishnaiah (1964).

For the analogous procedure for selecting the population associated with the smallest λ_i, the appropriate constant can be obtained from the tables of Gupta and Sobel (1962b) and Krishnaiah and Armitage (1964).

It is natural to use the Z_i defined previously instead of the Y_i in rule R_4. In this situation however, the infimum of the PCS and hence the constant c_4 do not depend on n; this is an unsatisfactory feature. One can, of course, propose a different type of procedure; for example, R_4': Select π_i if and only if $Z_i \geq Z_{[k]} - d, d > 0$. Such a procedure has not been investigated.

Case (ii): The Σ_i are unknown and not necessarily equal. In this case, Gupta and Studden (1970) proposed the rule

$$R_5\text{: Select } \pi_i \text{ if and only if } T_i \geq c_5 T_{[k]} \tag{3.3}$$

where $0 < c_5 = c_5(k, p, n, P^*) < 1$ is to be chosen suitably to satisfy the P^*-condition. As in case (i), the LFC is $\underset{\sim}{\lambda} = \underset{\sim}{0}$ and the constant c_5 is given by

$$\int_0^\infty F_{p,n-p}^{k-1}(x/c_5) dF_{p,n-p}(x) = P^* \tag{3.4}$$

where $F_{p,n-p}(x)$ is the cdf of a central F-variable with p and $n-p$ degrees of freedom. Values of c_5 have been tabulated for selected values of k, n, p and P^* by Gupta and Panchapakesan (1969), who have also tables for the constant needed for the analogous rule for selecting the population associated with the smallest λ_i.

Case (iii): $\Sigma_1 = \ldots = \Sigma_k = \Sigma$ (unknown). In this case, one would use the rule R_6 which is R_5 with $T_i = \overline{\underset{\sim}{X}}_i' S^{-1} \overline{\underset{\sim}{X}}_i$, where S is the usual pooled estimator of Σ, and with c_6 in the palce of c_5. Chattopadhyay (1981) has investigated R_6 and its analogue for selecting the population associated with $\lambda_{[1]}$. In fact, his procedure is defined for unequal sample sizes. An exact evaluation of the infimum of $P(CS|R_6)$ is difficult. Chattopadhyay (1981)

obtained a lower bound for the infimum so that a conservative value for the constant c_6 is given by

$$\int_0^\infty G_\nu^{k-1}(x\theta_\epsilon/c_6)dG_\nu(x) = P^* + \epsilon \tag{3.5}$$

where $\nu = p/2, G_\nu$ is as defined in (3.2), and θ_ϵ is the ϵth percentile of the distribution of the ratio of the largest to the smallest of the characteristic roots of S. Values of c_6/θ_ϵ can be obtained from the tables of Gupta (1963b) for selected values of k, p, and $P^* + \epsilon$. Evaluation of θ_ϵ can be done rather easily for small values of p using the distribution result of Pillai, Al-Ani and Jouris (1969). Chattopadhyay (1981) has also mentioned the possibility of obtaining some improvement in the evaluation of c_6.

3.1.2. *Indifference-Zone Approach*

Alam and Rizvi (1966) have investigated procedures for selecting the populations associated with the t largest $\lambda_i, 1 \le t \le k-1$, by taking for the preference-zone $\Omega(\delta_1^*, \delta_2^*) = \Omega_{\delta_1^*} \cap \Omega_{\delta_2^*}$, where $\Omega_{\delta_1^*} = \{\underset{\sim}{\lambda}: \lambda_{[k-t+1]} - \lambda_{[k-t]} \ge \delta_1^*\}$ and $\Omega_{\delta_2^*} = \{\underset{\sim}{\lambda}: \lambda_{[k-t+1]}/\lambda_{[k-t]} \ge \delta_2^*\}$ for specified $\delta_1^* > 0$ and $\delta_2^* > 1$. Let Z_i and T_i be defined as before. Alam and Rizvi (1966) considered the natural selection rules, selecting the populations that yielded the t largest Z_i in the case of known Σ_i's and the t largest T_i when the Σ_i are unknown and not necessarily equal. In either case, the LFC is given by

$$\begin{cases} \lambda_{[1]} = \ldots = \lambda_{[k-t]} = \delta_1^*(\delta_2^* - 1)^{-1} \\ \lambda_{[k-t+1]} = \ldots = \lambda_{[k]} = \delta_1^*\delta_2^*(\delta_2^* - 1)^{-1}. \end{cases} \tag{3.6}$$

Alam and Rizvi (1966) did not consider the case where the Σ_i are all equal but unknown. An investigation parallel to Chattopadhyay (1981) is possible but has not been done.

3.2 Selection in Terms of the Generalized Variance

The generalized variance $\theta = |\Sigma|$ of a multivariate normal population with covariance matrix Σ serves as an over-all measure of the variability among the components. Gnanadesikan and Gupta (1970) considered selection of the population associated with the smallest θ_i. They proposed a subset selection rule

$$R_7\text{: Select } \pi_i \text{ if and only if } W_i \le \frac{1}{c_7} W_{[1]}$$

where $W_i = |S_i|$ is the sample generalized variance and $0 < c_7 = c_7(k, p, n, P^*) < 1$ is to be chosen to meet the P^*-condition. They have shown that

$$\inf_\Omega P(CS|R_7) = Pr\{Y_1 \le \frac{1}{c_7} Y_j, j = 2, \ldots, k\} \tag{3.7}$$

where the Y_i are i.i.d., each distributed as a product of p independent factors where the rth factor has a chi-square distribution with $(n-r)$ degrees of freedom. An *exact* solution for c_7 is obtained for $p=2$ and is tabulated by Gupta and Sobel (1962b) and Krishnaiah and Armitage (1964). For $p>2$, Gnanadesikan and Gupta (1970) have studied approximations using the normal approximation of $\log \chi^2$ and Hoel's approximation of the distribution of $Y_i^{1/p}$ by an appropriate gamma distribution.

Regier (1976) has considered two alternative procedures, namely, R_7': Select π_i if and only if $W_i \leq a\left(\prod_{j=1}^{k} W_j\right)^{1/k}$ and R_7'': Select π_i if and only if $W_i \leq b \sum_{j=1}^{k} W_j/k$. The evaluation of the constants a and b are again based on normal approximation to $\log \chi^2$ and the asymptotic distribution of the sample variance, respectively.

Eaton (1967) considered a decision-theoretic approach to ranking the k populations according to the values of the θ_i, assuming reasonable properties for the loss function which depends only on the θ_i. He showed that the natural rule which ranks the populations according to the values of W_i is minimax, admissible, and uniformly the best among rules that are invariant under permutations of $(W_1,\ldots,W_k)$.

3.3 Selection in Terms of Multiple Correlation Coefficient

We assume that the $\underline{\mu}_i$ and Σ_i are unknown. Let ρ_i denote the multiple correlation coefficient between the first component and the rest of $\pi_i, i=1,\ldots,k$. For selecting the population associated with $\rho_{[k]}$, Gupta and Panchapakesan (1969) studied the rule

$$R_8\text{: Select } \pi_i \text{ if and only if } R_i^{*2} \geq c_8 R_{[k]}^{*2} \tag{3.8}$$

where $R_i^{*2} = R_i^2/(1-R_i^2)$, R_i is the sample correlation coefficient, and $0 < c_8 = c_8(k,p,n-p,P^*) < 1$ is chosen suitably to meet the P^*-condition. They considered both the conditional (variables 2 to p are fixed) and unconditional (all variables are random) cases. It is shown that the LFC in both cases is given by $\rho_1 = \ldots = \rho_k = 0$ when R_i^{*2} has the same distribution. Values of c_8 have been tabulated for selected values of $k, p, n-p$, and P^* by Gupta and Panchapakesan (1969), who have also considered the analogous procedure for selecting the population associated with the smallest ρ_i and provided tables for the values of the constant.

3.4 Selection in Terms of Other Measures

For selecting the best multivariate normal populations, a few other measures have been considered in the literature. However, these investigations are either limited in scope or solved only asymptotically for large sample size. These measures are: the sum of bivariate product-moment correlations [Govindarajulu and Gore (1971)], the coefficient of Alienation between two partitioned sets of the components [Frischtak (1973)], and the conditional generalized variance of one set given another set in a two-set partition of the components [Gupta and Panchapakesan (1969)]. For more details, the reader is referred to Gupta and Panchapakesan (1979, Chapters 7 and 14).

3.5 Comparison with a Standard or Control

As we remarked in Introduction, related to selecting the best among a given set of populations is the goal of selecting those which are better than a standard or a control. Let $\pi_1, \ldots, \pi_k$ be the k given p-variate normal populations where π_i is $N_p(\underset{\sim}{\mu}_i, \Sigma_i), i = 1, \ldots, k$. The control population π_0 is $N_p(\underset{\sim}{\mu}_0, \Sigma_0)$. Krishnaiah and Rizvi (1966) have considered comparison with a control by defining positive (good) and negative (bad) populations using different criteria based on linear combinations of the elements of the mean vectors and also on distance functions. Krishnaiah (1967) based the comparison on linear combinations of the elements of the covariance matrices. Huang (1973) considered partitioning the set $\{\pi_1, \ldots, \pi_k\}$ into good and bad sets using comparison based on the generalized variance.

3.5.1. *Comparisons Based on Linear Combinations of Mean Vectors*

Let $\theta_{ic} = \underset{\sim}{a}_c' \underset{\sim}{\mu}_i (c = 1, \ldots, r; i = 0, 1, \ldots, k)$, where the $\underset{\sim}{a}_c'$ are specified vectors reflecting the economic weights assigned to the elements of the mean vectors. Krishnaiah and Rizvi (1966) considered three definitions of positive and negative populations as follows:

A. π_i is positive if $\theta_{ic} \geq \theta_{oc} + \Delta_c, c = 1, \ldots, r$, and negative if $\theta_{ic} \leq \theta_{oc}, c = 1, \ldots, r$, where the Δ_c are given positive constants;

B. π_i is positive if $(\theta_{ic} - \theta_{oc})^2 \geq \Delta_{1c}, c = 1, \ldots, r$, and negative if $(\theta_{ic} - \theta_{oc})^2 \leq \Delta_{2c}, c = 1, \ldots, r$, where $\Delta_{1c} \geq \Delta_{2c} > 0$ are known constants;

C. π_i is positive if $|\theta_{ic}| \geq |\theta_{oc}|, c = 1, \ldots, r$, and negative if $|\theta_{ic}| < |\theta_{oc}|, c = 1, \ldots, r$.

For any procedure δ defined to select the positive populations, let $P(\omega, \delta), S(\omega, \delta)$, and $R(\omega, \delta)$ denote, respectively, the probability of including all positive populations, the expected proportion of true positives, and the expected proportion of false positives; here

ω denotes a point in the appropriate parameter space. It should be noted that $P(\omega, \delta)$ and $S(\omega, \delta)$ are defined *only* for the set of parameters for which there is at least one positive population. We seek a procedure δ such that either

$$\inf_{\omega} P(\omega, \delta) \geq P^* \tag{3.9}$$

or

$$\inf_{\omega} S(\omega, \delta) \geq p^* \tag{3.10}$$

where P^* and p^* are appropriately specified constants. The rule δ proposed by Krishnaiah and Rizvi (1966) in each of the cases described previously is of the form:

$$R_9\text{: Select } \pi_i \text{ if and only if } T_{ic} \geq d, c = 1, \ldots, r, \tag{3.11}$$

where the choice of T_{ic} in each case is described as follows.

Let a sample of n_i observations be taken from $\pi_i, i = 0, 1, \ldots, k$, and let $\underset{\sim}{\overline{X}}_i$ denote the sample mean from π_i. Define

$$\begin{aligned} U_{ic} &= \frac{\underset{\sim}{a}_c'(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\overline{X}}_0)}{[\underset{\sim}{a}_c'(n_i^{-1}\Sigma_i + n_0^{-1}\Sigma_0)\underset{\sim}{a}_c]^{1/2}} \\ W_{ic} &= \frac{n^{1/2}[|\underset{\sim}{a}_c'\underset{\sim}{\overline{X}}_i| - |\underset{\sim}{a}_c'\underset{\sim}{\overline{X}}_0|]}{[\underset{\sim}{a}_c'\Sigma\underset{\sim}{a}_c]^{1/2}} \end{aligned} \tag{3.12}$$

where W_{ic} is defined when $n_i = n$ and $\Sigma_i = \Sigma, i = 0, 1, \ldots, k$. Krishnaiah and Rizvi (1966) used

$$T_{ic} = \begin{cases} U_{ic} & \text{in Case A,} \\ U_{ic}^2 & \text{in Case B,} \\ W_{ic} & \text{in Case C.} \end{cases} \tag{3.13}$$

In Case A, they obtained a lower bound for inf $P(\omega, \delta)$. However, solution for d such that this bound equals P^* is difficult in general and not obtained. In Cases B and C, Krishnaiah and Rizvi (1966) obtained lower bounds for inf $P(\omega, \delta)$ using a Bonferroni inequality and have no results regarding sup $R(\omega, \delta)$, which is a measure of the efficiency of δ.

The above discussion will indicate that the problem requires further investigations dealing with possible alternative procedures and efficiency comparisons. The same can be said about the procedures investigated using comparisons based on distance functions. For more details, see Gupta and Panchapakesan (1979, Chapter 20, Section 8).

3.5.2. *Partitioning Based on Comparing Generalized Variances*

Huang (1973), using a formulation of Tong (1969), considered a partition of $\mathcal{P} = \{\pi_1, \ldots, \pi_k\}$ into three subsets, $\mathcal{P}_G = \{\pi_i : |\Sigma_i| \leq \rho_1 |\Sigma_0|\}, \mathcal{P}_I = \{\pi_i : \rho_1 |\Sigma_0| < |\Sigma_i| < \rho_2 |\Sigma_0|\}$, and $\mathcal{P}_B = \{\pi_i : |\Sigma_i| \geq \rho_2 |\Sigma_0|\}$. He considered rules for partitioning $\mathcal{P}$ into two subsets S_G and S_B based on samples of size n from all the $(k+1)$ populations. A correct decision occurs when $\mathcal{P}_B \subset S_B$ and $\mathcal{P}_G \subset S_G$. For this problem, Huang (1973) considered a single-stage as well as a sequential procedure. We leave out details of these procedures which are given in Gupta and Panchapakesan (1979, Chapter 20, section 8). However, we emphasize on the need to revisit these problems.

3.6. Remarks and Future Directions

For the problem of selecting from multivariate normal populations in terms of the Mahalanobis distance function, the subset selection rules investigated were of the ratio type. For the procedures R_4 and R_5, the supremum of $E(S)$, the expected subset size turns out to be k. One has to investigate other types of rules and make efficiency comparisons. As pointed out in the case of R_4, we can use R_4', a difference type procedure. Panchapakesan and Santner (1977) have discussed selection of good populations defined using a class of functions. They considered also selecting a subset with a restricted size. Reference should be made to their application to selection in terms of Mahalanobis distance. One can use such a rule which selects the population that yields the largest value for the relevant statistic, say T_i, or uses a difference type rule, or a ratio type rule depending on the value of $T_{[k]}$ falling in one of three ranges. Further, the case of $\Sigma_1 = \ldots = \Sigma_k = \Sigma$ (unknown) has not been well investigated.

Regarding other goals, the problems described in Section 3.5 need to be examined further from the point of view of alternative procedures, efficiency comparisons, and optimality properties.

In the problems we have discussed so far, the multivariate normal populations have been ranked according to the values of a scalar function of the parameters. This reduces the problem to a univariate one. Instead, we can specify a partition $\{\Omega_1, \ldots, \Omega_k\}$ of the parameter space Ω so that if the true state is in Ω_i, then π_i is the best population. An attempt in this direction has been made by Dudewicz and Taneja (1981). In a recent paper, Bofinger (1992) has considered the problem of multiple comparisons with the best for

multivariate normal populations using a "multivariate" approach. Her results are mainly for bivariate normal populations. She finds that "for comparisons with the 'best' of each variate, repeated univariate comparisons appear to be almost as efficient as multivariate comparisons, at least for the bivariate case and, under certain circumstances, for higher dimensional cases." These aspects of the selection and related inference problems are worth exploring further.

4. SELECTION FROM A MULTINOMIAL POPULATION

Multinomial, as a prototype for many practical problems, is one of the most useful discrete multivariate distributions. When observations from a population are classified into a certain number of categories, it is natural to look for categories that occur very often or rarely. Consider a multinomial distribution on m cells with probabilities $p_1, \ldots, p_m$. Two goals of common interest are selecting the most and the least probable cells, that is, cells associated with $p_{[m]}$ and $p_{[1]}$. The early investigations of Bechhofer, Elmaghraby and Morse (1959), Gupta and Nagel (1967), and Cacoullos and Sobel (1966) generated substantial interest resulting in a considerable number of papers that followed. The investigations of multinomial selection problems reveal some interesting aspects. Analogous procedures for selecting the cells with probabilities $p_{[1]}$ and $p_{[m]}$ using either the indifference-zone or the subset selection approach do not have similar structure for the LFC; see Gupta and Nagel (1967) and Alam and Thompson (1972). Under the indifference-zone formulation, the preference-zone can be specified using either a ratio or a difference. For selecting the most probable cell, we can specify the preference-zone by $p_{[m]}/p_{[m-1]} \geq \delta^*$ or $p_{[m]} - p_{[m-1]} \geq \delta^*$. It is found that generally the ratio-type works well for $p_{[m]}$ and the difference for $p_{[1]}$ in dealing with the LFC. Chen and Hwang (1986) have surveyed the problem of LFC in multinomial problems.

Under the indifference-zone formulation, Bechhofer, Elmaghraby and Morse (1959) and Alam and Thompson (1972) studied fixed sample procedures for selecting the most probable and least probable cells, respectively. Cacoullos and Sobel (1966) investigated an inverse sampling procedure for the most probable cell. Alam (1971), Alam, Seo and Thompson (1971), Ramey and Alam (1979, 1980), and Bechhofer and Kulkarni (1984) considered sequential selection procedures for selecting the most probable cell.

Using the subset selection approach, Gupta and Nagel (1967) studied fixed sample

procedures for the most and the least probable cell. Panchapakesan (1971) and Chen (1985) considered inverse sampling procedures for the most probable and the least probable cells, respectively. Berger (1980, 1982) investigated minimax rules for selecting the most as well as the least probable cell.

Ramey and Alam (1980) investigated a Bayes sequential procedure for selecting the most probable cell. Recently, Gupta and Liang (1989) have studied parametric empirical Bayes rules for selecting the most as well as the least probable cell. They assumed the loss $L(p,i) = p_{[m]} - p_i$ (or $p_i - p_{[1]}$), where i is the index of the selected cell and $p = (p_1,\ldots,p_m)$ which has a Dirichlet prior with hyperparameters $\underset{\sim}{\alpha} = (\alpha_1,\ldots,\alpha_m)$, where the α_i are positive but unknown. Gupta and Liang (1989) derived empirical Bayes rules when $\alpha_0 = \sum_{i=1}^{m} \alpha_i$ is known as well as unknown. Asymptotic optimality of these rules have also been established by them. This study has been later generalized by Gupta and Hande (1992) for more general loss functions.

In recent years, the problem of estimation after selection has received increasing attention. Recently Gupta and Miescke (1990a, b) studied this problem for the normal means selection problem and the binomial selection problem under the general decision-theoretic framework. Reference can be made to Gupta and Miescke (1990a) for a list of earlier papers dealing with this problem. Gupta and Hande (1992) have considered the problem of simultaneous selection and estimation for the most and the least probable cells using an empirical Bayes setup.

We have not referred to several papers dealing with selecting the best multinomial cell. We have confined ourselves to a few that will indicate the basic developments. Because of the difficulties with finding the LFC, several procedures have not been completely investigated. Further, there have been no systematic comparative studies of competing procedures.

Finally, it should be noted that the importance of multinomial selection rules is enhanced by the fact that they provide distribution-free procedures. Suppose that $\pi_1,\ldots,\pi_k$ have continuous distributions $F_{\theta_i}, i = 1,\ldots,k$, and $\{F_{\theta_k}\}$ is a stochastically increasing family in θ. Let p_i denote the probability that in a set of k observations, one from each population, the observation from π_i is the largest, $i = 1,\ldots,k$. The problem of selecting the stochastically largest (smallest) population can now be converted to selecting the most

(least) probable cell in a multinomial distribution.

5. SELECTION FROM SEVERAL MULTINOMIAL POPULATIONS

Although selecting the best cell from a single multinomial population has been investigated over a period of more than thirty years, selecting the best of several multinomial populations has not received attention until recently except for the paper by Gupta and Wong (1977). For ranking multinomial populations, we need a measure of diversity within a population. The need for such a measure arises in ecology, sociology, genetics, economics and other disciplines. Diversity in ecological contexts has been discussed by Pielou (1975), and Patil and Taillie (1982). Consider a multinomial population with m categories (cells) and probability vector $p = (p_1, \ldots, p_m)$ where p_i denotes the proportion of the population in category i. Three indices of diversity widely used in ecological studies are: (1) the species count, (2) Shannon's entropy, and (3) the Gini-Simpson (GS) index. Of these, the species count, which is defined as $m - 1$, is not of interest here because we are comparing populations having the same number of categories. The entropy function of Shannon (1949) is defined by

$$H(p_1, \ldots, p_m) = -\sum_{j=1}^{m} p_j \log p_j \tag{5.1}$$

and is a measure of uncertainty in an m-state system used in information theory. The GS index is defined by

$$\psi(p_1, \ldots, p_m) = 1 - \sum_{j=1}^{m} p_j^2 \tag{5.2}$$

was introduced by Gini (1912) and Simpson (1949). There are various other measures of diversity available in the literature. Rao (1982) introduced a unified approach to the measurement of population diversity from which the Mahalanobis distance function can be derived as a measure.

Now, let $\pi_1, \ldots, \pi_k$ be k independent multinomial populations with m cells and let the unknown cell probability vector of π_i be $p_i = (p_{i1}, \ldots, p_{im}), i = 1, \ldots, k$. We discuss selection from multinomial populations in terms of the entropy function, the GS index, and two other measures (defined later). As a preliminary to our discussions, let $\underset{\sim}{a} = (a_1, \ldots, a_m)$ and $\underset{\sim}{b} = (b_1, \ldots, b_m)$ such that $\sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i$. Then $\underset{\sim}{a}$ is to *majorize* $\underset{\sim}{b}$ (written $\underset{\sim}{a} > \underset{\sim}{b}$) if $\sum_{i=r}^{m} a_{[i]} \geq \sum_{i=r}^{m} b_{[i]}, r = 2, \ldots, m$, where the $a_{[i]}$ and $b_{[i]}$ are the ordered a_i and b_i as previously.

Further, a function f defined on the m-dimensional space is *Schur-concave* if $f(\underset{\sim}{x}) \leq f(\underset{\sim}{x}')$ whenever $\underset{\sim}{x} > \underset{\sim}{x}'$.

5.1 Selection in Terms of the Entropy Function

For $m = 2$ (binomial case), H_i and H_j are equal if and only if π_i and π_j have the same probability vectors. However for $m > 2$, while $\underset{\sim}{p}_i = \underset{\sim}{p}_j$ implies that H_i and H_j are equal, the converse is not true. We assume that there exists a population whose probability vector is majorized by that of any other population. This implies that there exists a population with the largest H_i because the entropy function is Schur-concave. Let $\varphi_i = \varphi\left(\frac{x_{i1}}{n}, \ldots, \frac{x_{im}}{n}\right)$, where φ is Schur-concave and $X_{i1}, \ldots, X_{im}$ are the cell counts based on n independent observations from $\pi_i, i = 1, \ldots, k$. Gupta and Wong (1977) proposed the subset selection rule

$$R_{10}\text{: Select } \pi_i \text{ if and only if } \varphi_i \geq \varphi_{[k]} - d_{10} \tag{5.3}$$

where $d_{10} = d_{10}(k, m, n, P^*)$ is the smallest positive constant for which the P^*-condition is met. They have shown that the PCS is minimized when all the $\underset{\sim}{p}_i$ are equal to $\underset{\sim}{p}_0$ (say). However, the value of $\underset{\sim}{p}_0$ for which the PCS will attain its infimum is not known. Gupta and Wong (1977) obtained a conservative solution for d, using the idea of conditioning as done by Gupta and Huang (1976) who have studied selection in terms of entropy in the binomial ($m = 2$) case.

Alam, Mitra, Rizvi and Saxena (1986) considered among other things the problem of selecting the population having the largest H_i using the indifference-zone approach. Continuing with our earlier notations, let $T_i^* = -n^{-1} \sum_{j=1}^{m} X_{ij} \log(X_{ij}/n), i = 1, \ldots, k$. The preference-zone Ω_{δ^*} is given by the set of configurations for which

$$H_{[k]} - H_{[k-1]} \geq \delta^* \tag{5.4}$$

where δ^* is a specified number such that $0 < \delta^* \leq \log m$. Alam et al. (1986) proposed the natural rule

$$R_{11}\text{: Select the population that gives the largest } T_i^*, \tag{5.5}$$

using randomization to break any tie. Alam et al. has no exact solution for the minimum sample size n needed to satisfy the P^*-condition. They obtained a lower bound for $P(CS|R_{11})$ when n is large. Using this lower bound one can obtain a large sample

(conservative) solution for n for given k, m, and δ^* but this involves computing certain asymptotic variance which has been tabulated by Alam et al. (1986) for $m = 2(1)10$.

5.2 Selection in Terms of the Gini-Simpson Index

We are interested in selecting the population associated with the largest (smallest) ψ_i. Let $\theta_i = \sum_{j=1}^{m} (p_{ij} - \frac{1}{m})^2$. Then $\theta_i = \sum_{j=1}^{m} p_{ij}^2 - \frac{1}{m} = \frac{m-1}{m} - \psi_i$. Some investigations have been done in terms of the θ_i.

Gupta and Leu (1990) considered selection of the population associated with the smallest θ_i and proposed the subset selection rule

$$R_{12}\text{: Select } \pi_i \text{ if and only if } Y_i \leq Y_{[1]} + d_{12} \tag{5.6}$$

where $Y_i = \sum_{j=1}^{m} (X_{ij} - \frac{1}{m})^2$ and d_{12} is the smallest positive number such that the P^*-condition is satisfied. Assuming that there exists a p_i which is majorized by all other p_j, Gupta and Leu (1990) showed that the infimum of PCS takes place when $p_1 = \ldots = p_k = p_0$ (say). However, the p_0 where the infimum takes place is not known. Gupta and Leu (1990) have obtained a conservative solution for d_{12} using the idea of conditioning as done by Gupta and Huang (1976). They have also obtained a large sample solution.

Recall that we assumed that there exists a p_i which is majorized by all other probability vectors. Instead of this, under some other restrictions, Gupta and Leu (1990) obtained some partial solutions. They have also discussed selecting the population with the largest θ_i under both sets of assumptions. Details of these are omitted here.

Liang and Panchapakesan (1992) have derived an empirical Bayes procedure for selecting the population associated with the largest ψ_i relative to Dirichlet product prior G (unknown) of the p_i, assuming the loss $L(\psi, i) = \psi_{[k]} - \psi_i$ corresponding to the true state of nature $\psi = (\psi_1, \ldots, \psi_k)$ and the decision to choose π_i. They have shown the asymptotic optimality of their procedure, showing that its Bayes risk converges to the minimum Bayes risk at an exponential rate.

Gupta and Leu (1990) also considered selecting all *good* populations which are defined to be those π_i's for which $\theta_i \leq \delta$ for a specified δ such that $0 < \delta < 1 - \frac{1}{m}$. A correct selection occurs if the selected subset contains all good populations. They proposed a subset selection rule

$$R_{13}\text{: Select } \pi_i \text{ if and only if } Y_i \leq c_{13} \tag{5.7}$$

where Y_i is the statistic used in R_{12}, and the constant $c_{13} > \delta$ is the smallest constant such that the P^*-condition is satisfied. Gupta and Liang (1991) have shown that the rule R_{13} is a Bayes rule relative to a symmetric Dirichlet prior and consequently is admissible for an additive loss function which is made up of a loss $(\theta_0 - \theta_i)$ for not including a good population π_i and a loss $(\theta_i - \theta_0)$ for including a bad (not good) population π_i. They have also derived an empirical Bayes rule relative to a symmetric Dirichlet prior G_α, where the hyperparameter α is unknown. They have established the asymptotic optimality of their procedure, showing that its Bayes risk converges exponentially to the minimum Bayes risk.

5.3 Selection in Terms of Other Measures of Diversity

For the population $\pi_i, i = 1, \ldots, k$, define

$$q_{ij} = \sum_{t=m-j+1}^{m} p_{i[t]} \tag{5.8}$$

where $p_{i[1]} \leq \ldots \leq p_{i[m]}$ are the ordered $p_{ij}, j = 1, \ldots, m$. It follows that $1/m \leq q_{i1} \leq \ldots \leq q_{im} = 1$ and that $q_{ij} \geq j/m, j = 1, \ldots, m; i = 1, \ldots, k$. The population π_i degenerates when $q_{i1} = \ldots = q_{im} = 1$, and it is uniform when $q_{ij} = \ldots = q_{im} = 1/m$. Motivated by these considerations, Rizvi, Alam and Saxena (1987) proposed two diversity measures, namely,

$$D_i = \sum_{j=1}^{m}(1 - q_{ij}) = \sum_{j=1}^{m}(m - j)p_{i[j]} \tag{5.9}$$

and

$$D_i^* = m + \sum_{j=1}^{m} sgn(c - q_{ij}), \tag{5.10}$$

where c is a fixed number between 0 and 1, and

$$sgn(x) = \begin{cases} 1 & \text{for } x \geq 0, \\ -1 & \text{for } x < 0. \end{cases}$$

The two measures have the minimum value zero when π_i is degenerate and attain their different maximum values when π_i is uniform. The measure D_i is related to the area under the Lorenz curve used in economic inequalities studies. Finally, both D_i and D_i^* are Schur-concave functions of p_i.

Rizvi, Alam and Saxena (1987) considered procedures for selecting the population associated with the largest D_i and the one with the largest D_i^*. In each case, they have

proposed procedures both under the indifference-zone and subset selection formulations, considering two possibilities according as the ordering of magnitudes of the cell probabilities in each p_i are, a priori, known (Case A) or unknown (Case B). These procedures are based on estimators $\hat{T}_i$ and $\hat{T}_i^*$ of D_i and D_i^* in Case A, and on estimators $\tilde{T}_i$ and $\tilde{T}_i^*$ in Case B. These estimators are obtained by replacing q_{ij} in (5.9) and (5.10) by X_{ij} associated with the cell probability p_{ij} in Case A, and by $X_{i[j]}$ in Case B. The indifference-zone procedure proposed by Rizvi et al. (1987) selects in each case the population that yields the largest value of the appropriate statistic. Their subset selection procedure is R_{10} with ϕ_i being the appropriate estimator.

5.4 Remarks

As we mentioned earlier, the diversity measures are of practical importance. Shannon's entropy has been used by Lewontin (1972) in biology. Applications of the GS index have been discussed by Agresti and Agresti (1978), Greenberg (1956), and Lieberson (1969), in the areas of sociology and linguistics. For further investigations of selection problems, it will be interesting to consider other measures of diversity. Nayak (1985) has discussed several diversity measures based on entropy functions. Reference should be made also to Rao and Nayak (1985) who have discussed cross entropy and dissimilarity measures. Cross entropy could be used to define the best population by considering the amount of dissimilarity of each population from the rest. In discussing the procedures of Rizvi et al. (1987) in Section 5.3, we mentioned Case A, where the ordering among the cell probabilities in each p_i is a priori known. In this case, it is more appropriate to use isotonic estimators of the cell probabilities. One has to consider procedures based on such estimators and compare their performances with those of Rizvi et al. Several procedures have been investigated when selection is in terms of different diversity measures. It is important to explore a unified approach to the selection problems.

6. SELECTION OF VARIABLES IN LINEAR REGRESSION

In applying regression analysis in practical situations for prediction purposes such as economic forecasting and weather prediction, one is faced with a large number of predictor (independent) variables. While the prediction can be made more accurate by bringing in as many relevant predictor variables as possible, some of them may be highly correlated among themselves and some others may contribute only very marginally. In these situations,

an "adequate" prediction may well be possible by considering a smaller number of the predictor (independent) variables. Thus arises the problem of choosing a "good" subset of these variables. Hocking (1976) and Thompson (1978a, b) have reviewed several criteria and techniques that have been used in practice. However, these procedures are ad hoc in nature and are not designed to control the probability of selecting the important variables. McCabe and Arvesen (1974), and Arvesen and McCabe (1975) were first to formulate this problem in the framework of subset selection theory.

Consider the standard linear model

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \tag{6.1}$$

where $\underset{\sim}{Y}' = (Y_1, \ldots, Y_n)$ is a vector of random observations, $X = [\underset{\sim}{1}, \underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_{p-1}]$ is an $n \times p$ matrix of known constants, $\underset{\sim}{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is a vector of unknown parameters, and $\underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma_0^2 I_n)$. Here $\underset{\sim}{1}$ is a column vector of 1's and I_n is an $n \times n$ identity matrix. The model (6.1) with $p-1$ independent variables is considered as the "true" model. Any reduced model whose "X matrix" has r columns is obtained by retaining any $r-1$ of the $p-1$ independent variables, where $2 \leq r \leq p$. For each r, there are $k_r = \binom{p-1}{r-1}$ such models, which are indexed arbitrarily $i = 1, \ldots, k_r$. These models, referred to as models M_{ri}, are then described by

$$\underset{\sim}{Y} = \underset{\sim}{X}_{ri}\underset{\sim}{\beta}_{ri} + \underset{\sim}{\epsilon}_{ri}, i = 1, \ldots, k_r, \tag{6.2}$$

where $\underset{\sim}{X}_{ri}$ is an $n \times r$ matrix (of rank r), $\underset{\sim}{\beta}_{ri}$ is a $r \times 1$ parameter vector, and $\underset{\sim}{\epsilon}_{ri} \sim N(\underset{\sim}{0}, \sigma_{ri}^2)$.

Arvesen and McCabe (1975) considered all possible subsets of an arbitrary size $t(= r-1)$ of the independent variables. These reduced models are considered for prediction purposes and must be compared under the true model assumptions. The expectations of residual mean squares in the corresponding ANOVA evaluated under the true model assumptions are $\sigma_{ri}^2, i = 1, \ldots, k_r$. Arvesen and McCabe (1975) considered the goal of selecting the model associated with the smallest σ_{ri}^2. They proposed the rule

$$R_{14}\text{: Select the model } M_{ri} \text{ if and only if } SS_{ri} \leq \frac{1}{c_{14}} SS_{r[1]} \tag{6.3}$$

where SS_{ri} is the residual sum of squares in the ANOVA corresponding to model M_{ri}, and $0 < c_{14} = c_{14}(p, t, n, P^*) < 1$ is to be chosen to satisfy the P^*-condition. An exact

evaluation of c_{14} is difficult. Arvesen and McCabe showed that the PCS is asymptotically $(n \to \infty)$ minimized when $\underset{\sim}{\beta} = \underset{\sim}{0}$. The evaluation of c_{14} is not easy even under this asymptotic LFC. McCabe and Arvesen (1974) has given an algorithm for determining c_{14} under the asymptotic LFC using Monte Carlo methods.

Now, for any reduced model M_{ri} in (6.2),

$$SS_{ri}/\sigma_0^2 \sim \chi^2\{\nu_r, \lambda_{ri}\} \tag{6.4}$$

where $\nu_r = n - r$ is the degrees of freedom and λ_{ri} is the noncentrality parameter. This gives

$$E(SS_{ri}) = \nu_r\sigma_0^2 + 2\sigma_0^2\lambda_{ri}. \tag{6.5}$$

Since σ_0^2 is fixed, it is clear from (6.5) that λ_{ri} should not be large for a good model. This motivates the criterion employed by Gupta and Huang (1988), namely, any reduced model M_{ri} with the associated noncentrality parameter λ_{ri} is defined to be *inferior* if $\lambda_{ri} \geq \Delta$, where $\Delta > 0$ is a specified constant. The goal is to eliminate all inferior models from the set of $2^{p-1} - 1$ regression models including the true model. For this goal, Gupta and Huang (1988) proposed and studied a two-stage procedure. In the first stage, inferior models are eliminated. Then, in the second stage, one of the models from the retained set (if it has more than one) is selected. Their procedure is based on the estimate $\hat{\lambda}_{ri}$ of λ_{ri} given by

$$\hat{\lambda}_{ri} = \frac{n-p}{2}\frac{1-R_{ri}^2}{1-R^2} - \frac{\nu_r}{2} \tag{6.6}$$

where R and R_{ri} are the multiple correlation coefficients of the models (6.1) and (6.2), respectively. The two-stage procedure R_{15} of Gupta and Huang (1988) is as follows.

R_{15}: At Stage 1, eliminate all models M_{ri} for which

$$\hat{\lambda}_{ri} \geq d_r \tag{6.7}$$

and at Stage 2, select from all the models that are retained after Stage 1 that model which has the smallest

$$\hat{\Gamma}_{ri} = \frac{n-p-2}{n-p}[2\hat{\lambda}_{ri} + (p-r)] - (p-2r). \tag{6.8}$$

The constant d_r in (6.7) is chosen to satisfy

$$D_r = \left[(d_r + \frac{\nu_r}{2})\frac{2}{n-p} - 1\right]\frac{n-p}{p-r} \tag{6.9}$$

where D_r is the 100 $(1-P^*)$ percent point of the noncentral F distribution with $p-r$ and $n-p$ degrees of freedom and noncentrality parameter Δ. It has been shown that, for the rule R_{15},

$$Pr\{\text{all inferior models } M_{ri} \text{ are eliminated}\} \geq P^*. \tag{6.10}$$

Several authors have studied the influence on the fitted regression line when a part of the data is deleted. Recently Gupta and Huang (1992) have integrated the concept of influential data with their procedure R_{15}.

Now, consider all reduced models M_{ri}. Let $\theta_{ri} = E(1-R_{ri}^2)$ where R_{ri} is the multiple correlation coefficient associated with M_{ri}. Any reduced model M_{ri} is called *inferior* if $\theta_{p-1,1} \leq \delta^* \theta_{ri}$, where $0 < \delta^* < 1$ is specified. ($\theta_{p-1,1}$ is associated with the true model.) A correct decision is selection of any subset of all possible models which does not include any inferior model. For this goal, Huang and Panchapakesan (1982) proposed the rule

$$R_{16}\text{: Exclude model } M_{ri} \text{ if and only if } \hat{\theta}_{ri} \geq \frac{c_{16}}{\delta^*}\hat{\theta}_{p-1,1} \tag{6.11}$$

where $\hat{\theta}_{ri} = 1-R_{ri}^2$, and $c_{16} = c_{16}(n,p,P^*) > \delta^*$ is determined so that the P^*-condition is satisfied. The LFC is established only asymptotically ($n \to \infty$). For evaluating c_{16} under the asymptotic LFC ($\underset{\sim}{\beta} = \underset{\sim}{0}$), Huang and Panchapakesan (1982) used an algorithm similar to that of McCabe and Arvesen (1974).

Hsu and Huang (1982) considered the goal of selecting a subset of the models that contains all the *superior* models, namely, all models for which $\sigma_{ri}^2 \leq \Delta\sigma_0^2$, where $\Delta > 1$ is a specified constant. For this problem, they investigated a sequential procedure.

Gupta, Huang and Chang (1984) studied the problem of eliminating inferior models, using the expected mean squares as the criterion for comparing any model with the true model. Their approach differs from other papers in that they considered tests of a family of hypotheses in constructing their procedure.

Finally, Ramberg (1977) considered an indifference-zone approach for selecting the best predictor variate among $X_1,\ldots,X_k$ to predict X_0, assuming that $\underset{\sim}{X} = (X_0, X_1, \ldots, X_k)$ has a multivariate normal distribution with an unknown mean vector $\underset{\sim}{\mu}$ and unknown covariance matrix $\Sigma = (\sigma_{ij})$. Let $\sigma_{0.i}^2$ denote the conditional variance X_0 given X_i. The goal is to select the variate associated with the smallest $\sigma_{0.i}^2$. Since $\sigma_{0.i}^2 = \sigma_{00}(1-\rho_{0i}^2)$, the problem is equivalent to selecting the variate associated with the largest ρ_{0i}^2. The

preference-zone Ω_{δ^*} is defined by $\sigma^2_{0\cdot[1]} \leq \sigma^2_{0\cdot[2]}/\delta^*$, where $\delta^* > 1$ is a specified constant. Based on a sample of size n, Ramberg (1977) proposed the natural rule that selects the variate which yields the smallest sample conditional variance $s^2_{0\cdot i}$. For the minimum sample size needed to satisfy the P^*-condition, he obtained an asymptotic solution for $k = 2$ and discussed some approximations for $k \geq 3$.

7. SOME MISCELLANEOUS PROBLEMS

In this section, we discuss selection and ranking approach to classification problems and to the problem of determining the appropriate number of components in principal component analysis. Both these problems and their applications are well-known in multivariate analysis.

7.1 Classification Problem

Classification problems typically arise when an investigator makes measurements on characteristics of an individual with a view to classify the individual into one of several possible categories. Assuming that the individual actually belongs to one of the specified categories may not be realistic. What we are really looking for is the category to which the individual is closest. In doing so, we want to control the probability of a correct classification (CC). An approach based on the concept of ranking and selection was considered by Cacoullos (1973) and Gupta and Govindarajulu (1973, 1985). However, their results are too conservative and limited.

Let P(CC|R) denote the probability of a correct classification (PCC) using the rule R. We want the rule R to guarantee a minimum value of P^*, $1/k < P^* < 1$, for the PCC. Gupta and Leu (1989) have considered selection procedures, mostly using subset selection approach, based on Mahalanobis distance function.

Let $\pi_i, i = 0, 1, \ldots, k$, be $k+1$ populations, where π_0 is to be classified as one of the remaining π_i. We assume that $\pi_i \sim N_p(\underset{\sim}{\mu}_i, \Sigma_i), i = 0, 1, \ldots, k$. Gupta and Leu (1989) have considered various cases depending on whether the $\underset{\sim}{\mu}_i$ are known or not, and whether the Σ_i are known or unknown (with a possibility that they are known to be equal). The procedures are developed in a way analogous to the selection rules in terms of Mahalanobis distance discussed in Section 3.1. The details of these are omitted here.

7.2. Procedure for Principal Component Analysis

Let $\underset{\sim}{X}' = (X_1, \ldots, X_p)$ be a random (observable) vector with mean $\underset{\sim}{\mu}$ and covariance matrix Σ having the characteristic roots $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Of the principal components $Y_i, i = 1, \ldots, p$ (which are certain linear combinations of the X_i), we want to determine the important components. This problem has been considered by Kaiser (1958), Cattell (1966), Horn (1965), and Horn and Engstrom (1979), who have suggested heuristic methods. Recently, Huang and Tseng (1992) have formulated this problem using a selection approach. Let

$$g_m(\underset{\sim}{\lambda}) = (\lambda_1 + \ldots + \lambda_m)/(\lambda_1 + \ldots + \lambda_p) \tag{7.1}$$

and

$$\Omega_k(\delta) = \{\underset{\sim}{\lambda} | g_k(\underset{\sim}{\lambda}) \geq \delta\} \tag{7.2}$$

where δ is a fixed constant and k is an integer less than p. We want to choose an integer m from $\{1, 2, \ldots, p\}$ which corresponds to choosing the first m principal components. A correct decision (CD) occurs if $\underset{\sim}{\lambda} \in \Omega_k(\delta)$ and $m = k$. We want a decision rule R for which $P_{\underset{\sim}{\lambda}}(CD|R) \geq P^*$, where $0 < P^* < 1$ is specified in advance.

Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_N$ be N independent observations on $\underset{\sim}{X}$ and $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p \geq 0$ denote the p characteristic roots of the sample covariance matrix. Huang and Tseng (1992) proposed the rule

R_{17}: Select the number of components up to the smallest integer m for which $g_m(\underset{\sim}{\ell}) \geq c_{17}$ where c_{17} is a fixed constant. They have discussed the determination of (N, c_{17}) when $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$. They have also done a simulation study of the robustness of the rule R_{17} derived under normality assumption when $\underset{\sim}{X}$ follows a multivariate t distribution.

8. CONCLUDING REMARKS

In the preceding sections, we have discussed several selection procedures for multivariate populations and also have, in several places, made specific comments relating to future directions of research. Related to the basic problem of selecting the best population, there are problems such as estimating the true PCS, estimating the appropriate parameter of the selected population, and comparison of the selected population with the best. Such problems have been examined in the case of selection from univariate populations. These problems, for example, should be of interest in selection problems discussed in this paper.

Further, Bayes and empirical Bayes procedures have been studied generally only for multinomial populations. From the point of view of reliability studies multivariate exponential populations and multivariate analogues of increasing failure rate (IFR) distributions are important. Selection problems in these contexts have been studied for univariate populations. Reference should be made to Gupta and Panchapakesan (1985, 1988).

Essentially, a vast literature is available detailing techniques and formulations in the case of selection from univariate populations; see Gupta and Huang (1981) and Gupta and Panchapakesan (1979, 1985, 1988). These provide natural avenues for future investigations of selection from multivariate populations.

REFERENCES

Agresti, A. and Agresti, B. F. (1978). Statistical analysis of qualitative variation. *Statistical Methodology* (ed. K. F. Schuessler), 204–237.

Alam, K. (1971). On selecting the most probable category. *Technometrics*, **13**, 843–850.

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

Alam, K. and Rizvi, M. H. (1966). Selection from multivariate populations. *Ann. Inst. Statist. Math.*, **18**, 307–318.

Alam, K., Seo, K. and Thompson, J. R. (1971). A sequential sampling rule for selecting the most probable multinomial event. *Ann. Inst. Statist. Math.*, **23**, 365–374.

Alam, K. and Thompson, J. R. (1972). On selecting the least probable multinomial event. *Ann. Math. Statist.*, **43**, 1981–1990.

Armitage, J. V. and Krishnaiah, P. R. (1964). Tables for the studentized largest chi-square distribution and their applications. ARL 64–188, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Dayton, OH.

Arvesen, J. N. and McCabe, G. P. (1975). Subset selection problems of variances with applications to regression analysis. *J. Amer. Statist. Assoc.*, **70**, 166–170.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.*, **25**, 16–39.

Bechhofer, R. E., Dunnett, C. W. and Sobel, M. (1954). A two-sample multiple-decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika*, **41**, 170–176.

Bechhofer, R. E., Elmaghraby, S.A., and Morse, N. (1959). A single-sample multiple-decision procedure for selecting the multinomial event which has the largest probability. *Ann. Math. Statist.*, **30**, 102–119.

Bechhofer, R. E. and Kulkarni, R. V. (1984). Closed sequential procedure for selecting the multinomial events which have the largest probabilities. *Commun. Statist.-Theor. Meth.*, **13**, 2997–3031.

Berger, R. L. (1980). Minimax subset selection for the multinomial distribution. *J. Statistical Planning Inf.*, **4**, 391–402.

Berger, R. L. (1982). A minimax and admissible subset selection rule for the least probable multinomial cell. *Statistical Decision Theory and Related Topics–III*, Vol. 2 (eds. S. S. Gupta and J. O. Berger), Academic Press, New York, 143–156.

Bofinger, E. (1992). Multiple comparisons with 'best' for multivariate normal populations. *Commun. Statist.–Theor. Meth.*, **21**, 915–941.

Cacoullos, T. (1973). Distance, discrimination and error. *Discriminant Analysis and Applications* (ed. T. Cacoullos), Academic Press, New York, 61–75.

Cacoullos, T. and Sobel, M. (1966). An inverse sampling procedure for selecting the most probable event in a multinomial distribution. *Multivariate Analysis* (ed. P. R. Krishnaiah), Academic Press, New York, 423–455.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behav. Res.*, **1**, 140–161.

Chattopadhyay, A. K. (1981). Selecting the normal population with largest (smallest) value of Mahalanobis distance from the origin. *Commun. Statist. A–Theor. Meth.*, **10**, 31–37.

Chen, P. (1985). Subset selection for the least probable multinomial cell. *Ann. Inst. Statist. Math.*, **37**, 303–314.

Chen, R.W. and Hwang, F. K. (1986). Least favorable configurations in the multinomial selection problem: a survey. *Amer. J. Math. Management Sci.*, **6**, 13–25.

Dudewicz, E. J. and Taneja, V. S. (1981). A multivariate solution of the multivariate ranking and selection problem. *Commun. Statist. A–Theor. Meth.*, **10**, 1849–1868.

Eaton, M. L. (1967). The generalized variance: testing and ranking problem. *Ann. Math. Statist.*, **38**, 941–943.

Frischtak, R. M. (1973). Statistical Multiple-Decision Procedures for Some Multivariate Selection Problems. Ph.D. Thesis (Tech. Report No. 187), Dept. of Operations Res., Cornell Univ., Ithaca, NY.

Gibbons, J. D., Olkin, I. and Sobel, M. (1977). *Selecting and Ordering Populations.* John Wiley, New York.

Gini, C. (1912). Variabilita e Mutabilitá. Studi Economico-aguridici della facotta di Giurisprudegza dell, Universite di Cagliari, III, Part II.

Gnanadesikan, M. (1966). Some Selection and Ranking Procedures for Multivariate Normal Populations. Ph.D. Thesis, Dept. of Statistics, Purdue Univ., West Lafayette, IN.

Gnanadesikan, M. and Gupta, S. S. (1970). Selection procedures for multivariate normal distributions in terms of measures of dispersion. *Technometrics*, **12**, 103–117.

Govindarajulu, Z. and Gore, A. P. (1971). Selection procedures with respect to measures of association. *Statistical Decision Theory and Related Topics* (ed. S. S. Gupta and J. Yackel), Academic Press, New York, 313–345.

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language*, **32**, 109–115.

Gupta, A. K. and Govindarajulu, Z. (1973). Some new classification rules for c univariate normal populations. *Canad. J. Statist.*, **1**, 139–157.

Gupta, A. K. and Govindarajulu, Z. (1985). On minimum distance classification rules for c multivariate populations. *Statistica*, **45**, 101–104.

Gupta, S. S. (1956). On a decision rule for a problem of ranking means. Mimeograph Series No. 150, Institute of Statistics, Univ. of North Carolina, Chapel Hill, NC.

Gupta, S. S. (1963a). Probability integrals of the multivariate normal and multivariate t. *Ann. Math. Statist.*, **34**, 792–828.

Gupta, S. S. (1963b). On a selection and ranking procedure for gamma populations. *Ann. Inst. Statist. Math.*, **14**, 199–216.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, **7**, 225–245.

Gupta, S. S. (1966). On some selection and ranking procedures for multivariate normal populations using distance functions. *Multivariate Analysis* (ed. P.R. Krishnaiah), Academic Press, New York, 457–475.

Gupta, S. S. and Hande, S. N. (1992). Single-stage Bayes and empirical Bayes rules for ranking multinomial events. *J. Statist. Planning Inf.*, to appear.

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

Gupta, S. S. and Huang, D.-Y. (1981). *Multiple Decision Theory: Recent Developments*. Lecture Notes in Statistics, Vol. 6, Springer-Verlag, New York.

Gupta, S. S. and Huang, D.-Y. (1988). Selecting important independent variables in linear regression models. *J. Statist. Planning Inf.*, **20**, 155–167.

Gupta, S. S. and Huang, D.-Y. (1992). On detecting influential data and selecting regression variables. Tech. Report No. 89–28C, Dept. of Statistics, Purdue Univ., West Lafayette, IN.

Gupta, S. S., Huang, D.-Y., and Chang, C.-L. (1984). Selection procedures for optimal subsets of regression variables. *Design of Experiments* (eds. T. J. Santner and A. C. Tamhane), Marcel Dekker, New York, 67–75.

Gupta, S. S. and Leu, L.-Y. (1989). On a classification problem: ranking and selection approach. Tech. Report No. 89–27C, Dept. of Statistics, Purdue Univ., West Lafayette, IN.

Gupta, S. S. and Leu, L.-Y. (1990). Selecting the fairest of $k(\geq 2)$ m-sided dice. *Commun. Statist. A–Theor. Meth.*, **19**, 2159–2177.

Gupta, S. S. and Liang, T. (1989). Parametric empirical Bayes rules for selecting the most probable multinomial event. *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin* (eds. L. Gleser et al.), Springer-Verlag, New York, 318–328.

Gupta, S. S. and Liang, T. (1991). Selecting multinomial populations. Tech. Report No. 90–01C, Dept. of Statistics, Purdue Univ., West Lafayette, IN. To appear in Bahadur Festschrift (eds. J. K. Ghosh et al.).

Gupta, S. S. and Miescke, K. J. (1990a). On finding the largest normal mean and estimating the selected mean. *Sankhyā Ser. B*, **52**, 144–157.

Gupta, S. S. and Miescke, K. J. (1990b). On combining selection and estimation in the search for the largest binomial parameter. Tech. Report No. 90–33, Dept. of Statistics, Purdue Univ., West Lafayette, IN.

Gupta, S. S. and Nagel, K. (1967). On selection and ranking procedures and order statistics

from multinomial distribution. *Sankhyā Ser. B*, **29**, 1–34.

Gupta, S. S., Nagel, K. and Panchapakesan, S. (1973). On the order statistics from equally correlated normal random variables. *Biometrika*, **60**, 403–413.

Gupta, S. S. and Panchapakesan, S. (1969). Some selection and ranking procedures for multivariate normal populations. *Multivariate Analysis–II* (ed. P. R. Krishnaiah), Academic Press, New York, 475–505.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. John Wiley, New York.

Gupta, S. S. and Panchapakesan, S. (1985). Subset selection procedures: review and assessment. *Amer. J. Math. Management Sci.*, **5**, 235–311.

Gupta, S. S. and Panchapakesan, S. (1987). Statistical selection procedures in multivariate models. *Advances in Multivariate Statistical Analysis* (ed. A. K. Gupta), D. Reidel Publishing Co., Dordrecht, Holland, 141–160.

Gupta, S. S. and Panchapakesan, S. (1988). Selection and ranking procedures in reliability models. *Handbook of Statistics 7: Quality Control and Reliability* (eds. P. R. Krishnaiah and C. R. Rao), North-Holland, Amsterdam, 131–156.

Gupta, S. S., Panchapakesan, S. and Sohn, J. K. (1985). On the distribution of the studentized maximum of equally correlated normal random variables. *Commun. Statist.–Simul. Comput.*, **14**, 103–135.

Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. *Ann. Math. Statist.*, **28**, 957–967.

Gupta, S. S. and Sobel, M. (1962a). On selecting a subset containing the population with the smallest variance. *Biometrika*, **49**, 495–507.

Gupta, S. S. and Sobel, M. (1962b). On the smallest of several correlated F-statistics. *Biometrika*, **49**, 509–523.

Gupta, S. S. and Studden, W. J. (1970). On some selection and ranking procedure with applications to multivariate populations. *Essays in Probability and Statistics* (ed. R. C. Bose et al.), Univ. of North Carolina Press, Chapel Hill, 327–338.

Gupta, S. S. and Wong, W.-Y. (1977). Subset selection procedures for finite schemes in information theory. *Colloquia Mathematica Societatis János Bolyai, Topics in Information Theory*, 279–291.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, **30**, 179.

Horn, J. L. and Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behav. Res.*, **14**, 283–300.

Hsu, T.-A. and Huang, D.-Y. (1982). On eliminating inferior regression models. *Commun. Statist. A–Theor. Meth.*, **11**, 751–759.

Huang, D.-Y. and Panchapakesan, S. (1982). On eliminating inferior regression models. *Commun. Statist. A–Theor. Meth.*, **11**, 751–759.

Huang, D.-Y. and Tseng, S.-T. (1992). A decision procedure for determining the number of components in principal component analysis. *J. Statist. Planning Inf.*, **30**, 63–71.

Huang, W.-T. (1973). On partitioning k multivariate normal populations with respect to a control. *Bull. Inst. Math. Acad. Sinica*, **1**, 191–206.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

Krishnaiah, P. R. (1967). Selection procedures based on covariance matrices of multivariate normal populations. *Blanch Anniversary Volume*, Aerospace Research Laboratories, U.S. Air Force, Dayton, OH, 147–160.

Krishnaiah, P. R. and Armitage, J. V. (1964). Distribution of the studentized smallest chi-square, with tables and applications. ARL 64–218, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Dayton, OH.

Krishnaiah, P. R. and Rizvi, M. H. (1966). Some procedures for selection of multivariate normal populations better than a control. *Multivariate Analysis* (ed. P. R. Krishnaiah), Academic Press, New York, 477–490.

Lewontin, R. C. (1972). The apportionment of human diversity. *Evolutionary Bio.*, **6**, 381–398.

Liang, T. and Panchapakesan, S. (1992). An empirical Bayes procedure for selecting the most homogeneous multinomial population according to the Gini-Simpson index. To appear in the *Proceedings of the 1990 International Statistical Symposium* held in

Taipei, Republic of China.

Lieberson, S. (1969). Measuring population diversity. *Amer. Soc. Rev.*, **34**, 850–862.

McCabe, G. P. and Arvesen, J. N. (1974). A subset selection procedure for regression variables. *J. Statist. Comput. Simul.*, **3**, 137–146.

Nayak, T. K. (1985). On diversity measures based on entropy functions. *Commun. Statist. A–Theory Methods*, **14**, 203–215.

Panchapakesan, S. (1971). On a subset selection procedure for the most probable event in a multinomial distribution. *Statistical Decision Theory and Related Topics* (eds. S. S. Gupta and J. Yackel), Academic Press, New York, 275–298.

Panchapakesan, S. and Santner, T. J. (1977). Subset selection procedures for Δ_p-superior populations. *Commun. Statist. A–Theor. Meth.*, 1081–1090.

Patil, G. P. and Taillie, C. (1982). Diversity as a concept and its measurement. *J. Amer. Statist. Assoc.*, **77**, 548–567.

Pielou, E. C. (1975). *Ecological Diversity*. John Wiley, New York.

Pillai, K. C. S., Al-Ani, S. and Jouris, G. M. (1969). On the distributions of the ratios of the roots of a covariance matrix and Wilk's criterion for tests of three hypotheses. *Ann. Math. Statist.*, **40**, 2033–2040.

Ramberg, J. S. (1977). Selecting the best predictor variate. *Commun. Statist. A–Theor. Meth.*, **6**, 1133–1147.

Ramey, J. and Alam, K. (1979). A sequential procedure for selecting the most probable multinomial event. *Biometrika*, **66**, 171–173.

Ramey, J. and Alam, K. (1980). A Bayes sequential procedure for selecting the most probable multinomial event. *Commun. Statist. A–Theor. Meth.*, **9**, 265–276.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theo. Popln. Bio.*, **21**, 24–43.

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

Regier, M. H. (1976). Simplified selection procedures for multivariate normal populations. *Technometrics*, **18**, 483–489.

Rizvi, M. H., Alam, K. and Saxena, K. M. L. (1987). Selection procedure for multinomial populations with respect to diversity indices. Contributions to the Theory and

Application of Statistics, a Volume in Honor of Herbert Solomon (ed. A. E. Gelfand), Academic Press, New York, 485–509.

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

Slepian, D. (1962). On the one-sided barrier problem for gaussian noise. *Bell System Tech. J.*, **41**, 463–501.

Thompson, M. L. (1978a). Selection of variables in multiple regression: Part I. A review and evaluation. *Int. Statist. Rev.*, **46**, 1–10.

Thompson, M. L. (1978b). Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples. *Int. Statist. Rev.*, **46**, 129–146.

Tong, Y. L. (1969). On partitioning a set of normal populations by their locations with respect to a control. *Ann. Math. Statist.*, **40**, 1300–1324.