MULTIVARIATE RANKING AND SELECTION PROCEDURES WITH AN APPLICATION TO OVERSEAS TRADE

A thesis presented in partial fulfillment

of the requirements for the degree

of

MASTER OF ARTS

in

STATISTICS

at

MASSEY UNIVERSITY

Name: MARY SHARMILA MENDIS
Year: 1983

To my husband Rohan, our son Ruvan

and

the memory of our son Gerard

ABSTRACT

An overview of some recent work in the field of Ranking and Selection with emphasis on aspects important to experimenters confronted with Multivariate Ranking and Selection problems is presented. Ranking and Selection procedures fall into two basic categories. They are:

1) Indifference Zone Approach
2) Subset Selection Approach.

In these approaches, the multivariate parameters are converted to univariate parameters. Various procedures using these real valued functions are given for both the Indifference Zone Approach and the Subset Selection Approach.

A new formulation that has recently been developed which selects the best multivariate population without reducing populations to univariate quantities is also described. This method is a Multivariate Solution to the Multivariate Ranking and Selection problem.

Finally a real life problem pertaining to New Zealand's overseas trade is discussed in the context of Multivariate Ranking.

ACKNOWLEDGEMENTS

I wish to thank my supervisor Dr. Howard Edwards for his interest and guidance in the preparation of this thesis. Thanks also to Mrs. S. Carlyle for her excellent typing of the manuscript. Finally, I wish to thank my husband Rohan for his help, encouragement and support.

TABLE OF CONTENTS

Chapter 6

LIST OF TABLES

LIST OF COMPUTER PRINTOUTS

CHAPTER 1

1. INTRODUCTION

In the mid nineteen fifties attention began to be drawn to a new type of problem which does not fit into the framework of testing hypotheses and for which no proper statistical approach has been developed. In this type of problem it is not necessary to refute a null hypothesis which is clearly false but rather answer a different type of question which deals with selecting the best or with the ranking of alternatives. This field of study is called RANKING AND SELECTION THEORY.

A statistical Selection procedure uses a random sample from each population to select the best population. The same sample of data is used to order the populations in statistical Ranking procedures. In these procedures it can be asserted with a specified level of confidence that the Selection or Ranking made is correct.

Procedures for Selection and Ranking problems were pioneered by R. E. Bechhofer in 1954 using normality and equal known variance. In the ensuing years such procedures have been developed for more complex problems and in more realistic settings.

This thesis presents an overview of some recent work in this field with emphasis on aspects important to experimenters confronted with Multivariate Ranking and Selection problems. An example pertaining to overseas trade using Multivariate Ranking is also discussed.

CHAPTER 2

2. RANKING AND SELECTION

2.1 POSSIBLE GOALS FOR RANKING AND SELECTION PROCEDURES

Ranking and Selection procedures include techniques appropriate for many different goals, although each different goal requires a careful formulation of the corresponding problem. For any given set of k populations some of the goals that can be accomplished by these methods are given below.

a) Selecting the one best population.

b) Selecting a random number of populations such that all populations better than the control population or the standard are included in the selected subset.

c) Selecting at least two, say t (≥ 2), best populations in an ordered or unordered manner.

d) Selecting a random number of populations, say r, that includes the t best populations.

e) Selecting a fixed number of populations, say f, that includes the t best populations.

f) Ordering or ranking all the k populations from best to worst or vice versa.

g) Ranking a fixed size subset of the k populations from best to worst or vice versa.

2.2 APPROACHES TO RANKING AND SELECTION PROCEDURES

Ranking and Selection procedures fall into two basic categories. They are:

a) the Indifference Zone Approach pioneered by R. E. Bechhofer (1954);

b) the Subset Selection Approach pioneered by S. S. Gupta (1956).

In this chapter the theory related to the two methods are explained in detail.

2.3 THE PHILOSOPHY OF THE INDIFFERENCE ZONE APPROACH

The essential problem formulation of the indifference zone approach pioneered by Bechhofer (1954) is as follows.

Let $\Pi_1, \Pi_2, \ldots., \Pi_k$ be k independent populations with underlying distribution functions $F(x, \theta_i)$, $i = 1, 2, \ldots., k$. The θ_i are unknown values of a quality characteristic which is used as the parameter for selecting the populations. Except for the value of θ_i, the distribution is assumed not to differ from population to population. Also, it is defined that if $\theta_i \geqslant \theta_j$ then Π_i is better than Π_j, although in some cases the inequality is reversed. Let the ordered θ_i be denoted by $\theta_{[1]} \leqslant \theta_{[2]} \leqslant \ldots. \leqslant \theta_{[k]}$. The experimenter is assumed to have no prior knowledge regarding the positions of the ordered and unordered θ_i.

The goal of the experimenter is to choose one of the populations and claim that is the best, the one associated with $\theta_{[k]}$.

$$\text{Max}(\theta_1, \theta_2, \ldots., \theta_k) = \theta_{[k]}$$

The selection is performed in such a way that the associated probability of a correct selection for a given selection rule R, P(CS/R) is at least as large as a predetermined P^* $(1/k < P^* < 1)$ whenever the distance (suitably defined) between the best and the second best populations denoted by $\delta = (\theta_{[k]} - \theta_{[k-1]})$, is at least as large as a specified constant δ^* (> 0).

$$P(CS/R) \geqslant P^* \qquad (1/k < P^* < 1)$$

if $\delta = (\theta_{[k]} - \theta_{[k-1]}) \geqslant \delta^*$ where $\delta^* > 0$.

The experimenter has the privilege of specifying P^* and δ^* satisfactory to himself. We will assume without loss of generality that the distance function is the usual difference $\delta(\theta_{[k]}, \theta_{[k-1]}) = \theta_{[k]} - \theta_{[k-1]}$.

This method does not explicitly seek to control the probability of a correct selection, P(CS/R) at the parameter points $\theta_{[k]}, \theta_{[k-1]}$. If the difference between the best and the second best are not sufficiently apart or if it is in the "ZONE" $\delta = (\theta_{[k]} - \theta_{[k-1]}) < \delta^*$, the experimenter is "INDIFFERENT" to which population is selected. Hence the name INDIFFERENCE ZONE.

In this approach the experimenter finds the smallest sample size (n) required from each population corresponding to the defined values of P^* and δ^*. Then the experimenter selects the best out of the k populations using the appropriate statistics $\hat{\theta}_i$, i = 1, 2, ..., k, where $\hat{\theta}_i$ is an estimate of θ_i.

The total space of the θ_i values is the union of the Indifference Zone (IZ) defined by $\delta = (\theta_{[k]} - \theta_{[k-1]}) < \delta^*$, and the Preference Zone (PZ) defined by $\delta = (\theta_{[k]} - \theta_{[k-1]}) \geqslant \delta^*$ for $\delta^* > 0$. δ^* (> 0) defines the threshold value to separate the Indifference Zone from the Preference Zone.

The experimenter is indifferent to which population is selected when $\delta = (\theta_{[k]} - \theta_{[k-1]}) < \delta^*$. δ^* or an indifference zone is still specified in recognition of the fact as $\delta^* \to 0$, $n \to \infty$. This indicates that a large sample per population may be necessary for assurance of trivial gains.

The indifference zone is feasible, in terms of the sample size n only if two conditions are met.

a) The number of populations, k, is not extremely large (eg $k \leqslant 50$).

b) The experimenter has some design control, via the choice of n.

There are many other variations of the preference zone as the situation warrants. The preference zone generally has an infinite number of points. In many cases there is some special configuration for which the probability of a correct selection is a minimum over all configurations in the preference zone. This configuration is called the least favourable configuration and denoted by $\underset{\sim}{\theta}_{LF}$.

$$P(CS/\underset{\sim}{\theta}) \geqslant P(CS/\underset{\sim}{\theta}_{LF}) \quad \text{for all } \underset{\sim}{\theta} \in PZ$$

where $\underset{\sim}{\theta}$ is the vector $(\theta_1, \theta_2, \ldots., \theta_k)$

and $\underset{\sim}{\theta}_{LF}$ is the vector $(\theta_{1,LF}, \theta_{2,LF}, \ldots., \theta_{k,LF})$.

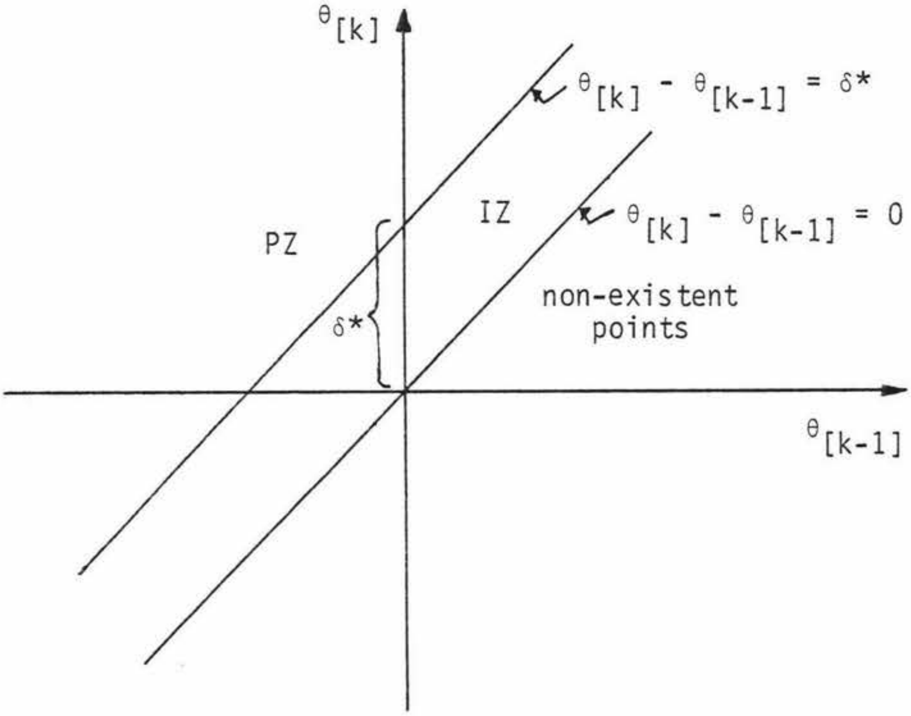
2.3.1 GRAPHICAL REPRESENTATION

For an arbitrary number of k populations suppose the distance measure $\delta = \theta_{[k]} - \theta_{[k-1]}$ and the parameter space is unlimited so that the values of $\underset{\sim}{\theta}$ vary on the entire real line. The preference zone is defined by,

$$PZ = \{\underset{\sim}{\theta} : \delta = (\theta_{[k]} - \theta_{[k-1]}) \geq \delta^*\}.$$

The indifference zone is defined by,

$$IZ = \{\underset{\sim}{\theta} : \delta = (\theta_{[k]} - \theta_{[k-1]}) < \delta^*\} \text{ for } \delta^* > 0.$$


$\theta_{[k]}$
$\theta_{[k]} - \theta_{[k-1]} = \delta^*$
IZ
$\theta_{[k]} - \theta_{[k-1]} = 0$
PZ
non-existent points
δ^*
$\theta_{[k-1]}$

2.4 THE PHILOSOPHY OF THE SUBSET SELECTION APPROACH

The goal of this approach is to select a non empty subset from the given populations so that it includes the best population.

The given set of k populations are divided into two identifiable subsets of random sizes in such a way that there is a high probability, P* (pre specified), that the selected subset contains the best population and the eliminated subset does not. There is no assertion made about which population is the best within the selected subset. Now a correct selection occurs if the selected subset contains the population with the parameter value $\theta_{[k]}$.

In this subset selection approach a random sample from each of the k populations is taken and an estimate $\hat{\theta}_i$ of the parameter θ_i is computed from the corresponding sample data of the i th population. Then for each population (i = 1, 2, ..., k) the selection rule is to place the i th population in the selected subset if and only if $\hat{\theta}_i$ is included in a certain region I. This region I is usually a closed interval of the form, $I = [\hat{\theta}_{[k]} - C, \hat{\theta}_{[k]}]$ where C (> 0) is to be determined.

The value of C should be as small as possible subject to the condition that the infimum of a correct selection for any rule R, over the whole parameter space of θ_i is at least P* or $P(CS/R) \geqslant P^*$ for whatever be the true configuration of the unknown θ_i. Here the subset selected is of random size and since $\theta_{[k]}$ is always contained in the region I, the selected subset cannot be empty.

The experimenter can select a rule R such that for the specified probability P* the expected value of the selected subset size is as small as possible for all rules R.

2.5 COMPARISON OF THE INDIFFERENCE ZONE AND THE SUBSET SELECTION APPROACHES

The main difference between the Indifference Zone Approach and the Subset Selection Approach is that in the latter there is no Indifference Zone. Also, in the Subset Selection Approach, the least favourable configuration is the one with all the θ_i equal. Hence it is almost impossible to compare the two approaches analytically.

In any given situation the preference of one over the other of the two approaches is mainly dictated by the objectives of the experimenter.

The Indifference Zone Approach is useful at the experimental design stage, where a common sample size is to be determined, whereas the Subset Selection Approach, in the main formulation, assumes that the sample size may be fixed arbitrarily or by other considerations.

When a subset is selected no single population within that subset is asserted to be the best one, except by implication if it happens that the subset selected is of size one. Hence the Subset Selection Approach gives less precise information but it provides more flexibility.

The infimum of the Probability of a Correct Selection in the Indifference Zone Approach is evaluated over the Preference Zone, whereas in the Subset Selection Approach it is over the entire parameter space.

The Subset Selection Approach is particularly useful in screening problems, for example, drug screening. It is also appropriate when k is very large and it is required to select a smaller number of populations to test further or to compare for secondary properties.

CHAPTER 3

3. MULTIVARIATE DISTRIBUTIONS

3.1 DEFINITION OF A MULTIVARIATE DISTRIBUTION

A multivariate distribution is the joint probability distribution of p ($\geqslant 2$) variables. A random sample of size n from a population having a multivariate distribution consists of n observations of p tuples (vectors) of measurements.

The most important multivariate distribution is the Multivariate Normal Distribution. The multivariate normal distribution has as parameters, not only the means and variances of each of the p variables, but also covariances or correlations between pairs of these components.

The density function of the multivariate normal distribution is

$$f(\underset{\sim}{X}_i) = \frac{1}{(2\pi)^{\frac{1}{2}p} \ |\Sigma_i|^{\frac{1}{2}}} \ \exp[-\frac{1}{2}(\underset{\sim}{X}_i - \underset{\sim}{\mu}_i)' \textstyle\sum_i^{-1} (\underset{\sim}{X}_i - \underset{\sim}{\mu}_i)]$$

where $\underset{\sim}{X}_i$ is a p variate random vector from the population Π_i. $\underset{\sim}{\mu}_i$ and Σ_i are the corresponding vector of means and the p×p variance-covariance matrix respectively of the population Π_i.

A p variate normal distribution with mean $\underset{\sim}{\mu}_i$ and variance-covariance matrix Σ_i will be denoted by $N_p(\underset{\sim}{\mu}_i , \Sigma_i)$. The inverse of Σ_i will be denoted by $\textstyle\sum_i^{-1}$.

3.2 NOTATIONS USED IN MULTIVARIATE RANKING AND SELECTION

Let Π_1, Π_2, ..., Π_k be k independent p variate normal populations with mean vectors $\underset{\sim}{\mu}_i$ and covariance matrices Σ_i, i = 1, 2, ..., k and denoted by $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$. All the vectors are column vectors and the Σ_i are assumed to be positive definite.

The sample mean vector of the i th population Π_i, based on a sample of size n is defined by

$$\underset{\sim}{\overline{X}}_i = \begin{bmatrix} \overline{X}_1^{(i)} \\ \vdots \\ \\ \cdot \\ \overline{X}_p^{(i)} \end{bmatrix}$$

where $\overline{X}_c^{(i)} = \dfrac{\sum\limits_{j=1}^{n} X_{cj}^{(i)}}{n}$ for c = 1, 2, ..., p and $X_{cj}^{(i)}$ denotes the c th component of the j th random vector observed from Π_i.

The sample variance-covariance matrix S_i of the i th population Π_i based on a sample of size n is defined by

$$S_i = \begin{bmatrix} s_{11}^{(i)} & s_{12}^{(i)} & \cdots & . \, s_{1p}^{(i)} \\ \vdots & & & \\ & & & \\ \cdot & & & \\ s_{p1}^{(i)} & & & s_{pp}^{(i)} \end{bmatrix}$$

$$\text{where} \quad s_{cd}^{(i)} = \frac{\sum_{j=1}^{n} (X_{cj}^{(i)} - \overline{X}_c^{(i)})(X_{dj}^{(i)} - \overline{X}_d^{(i)})}{(n - 1)}$$

for $c = 1, 2, \ldots., p$, $d = 1, 2, \ldots., p$, $i = 1, 2, \ldots., k$.

3.3 MULTIVARIATE RANKING AND SELECTION

The multivariate parameters $\underset{\sim}{\mu}_i$ and Σ_i, $i = 1, 2, \ldots., k$, are converted to the univariate parameters $\theta_1, \theta_2, \ldots., \theta_k$ by a scalar function $\phi(\underset{\sim}{\mu}_i, \Sigma_i)$, where

$$\begin{aligned} \theta_1 &= \phi(\underset{\sim}{\mu}_1, \Sigma_1) \\ \theta_2 &= \phi(\underset{\sim}{\mu}_2, \Sigma_2) \\ &\vdots \\ \theta_k &= \phi(\underset{\sim}{\mu}_k, \Sigma_k) \end{aligned}$$

Let $\theta_{[1]} \leqslant \theta_{[2]} \leqslant \ldots. \leqslant \theta_{[k]}$ denote the ordered parameters $\theta_1, \theta_2, \ldots., \theta_k$, where $\theta_{[k]}$ is the largest θ_i value. The best population is the one associated with $\theta_{[k]}$.

Using the θ_i, $i = 1, 2, \ldots., k$ values, it is necessary to select the best population out of the populations $\Pi_1, \Pi_2, \ldots., \Pi_k$. To achieve this it is necessary to develop a procedure R such that for a fixed P* the probability of a Correct Selection satisfies

$$\inf_{\Omega_p} P(CS/R) = P^*$$

where Ω_p is a subset of Ω the total parameter space of all admissible values of $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots., \theta_k)$ and Inf or Infimum denotes the greatest lower bound.

The best population could be selected by two different approaches.

a) Indifference Zone Approach.

Here the "CS" means the selection of the population associated with $\theta_{[k]}$. Ω_p is a proper subset of the total parameter space Ω and also it is the Preference Zone.

b) Subset Selection Approach.

Here the "CS" means the selection of a subset S from the populations $\Pi_1, \Pi_2, \ldots., \Pi_k$ such that S contains the populations associated with $\theta_{[k]}$ and $\Omega_p = \Omega$.

Various choices have been made of the real valued functions in papers presented in recent years. An overview of these papers is given in Chapter 4 and Chapter 5. The following list gives the real valued functions used in the overview.

$$\theta_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$$

$$\theta_i = \left(\underset{\sim}{\mu}_i' \underset{\sim}{\mu}_i\right)^{\frac{1}{2}}$$

θ_i = Generalized variance

θ_i = Multiple correlation coefficient

θ_i = Product moment correlation for the bivariate ($p = 2$) case

θ_i = Coefficient of alienation.

Each of the above functions reduces Multivariate parameters to a Univariate parameter. However, an attempt was made by Dudewicz and Taneja (1981) to give a multivariate solution to the multivariate ranking and selection problem. Chapter 6 is devoted to discussing this new technique.

CHAPTER 4

4. THE INDIFFERENCE ZONE APPROACH TO RANKING AND SELECTION OF SEVERAL MULTIVARIATE NORMAL POPULATIONS

As stated in the previous chapter, various real valued functions θ_i have been defined in recent work to select the populations.

Several real valued functions θ_i of the mean vector $\underset{\sim}{\mu}_i$ and the covariance matrix Σ_i are considered in this chapter. They are:

1) The Mahalanobis Distance

2) The Euclidean Distance

3) Multiple Correlation Coefficients

4) Sum of Bivariate Product Moment Correlations

5) Coefficient of Alienation.

4.1 SELECTION IN TERMS OF THE MAHALANOBIS DISTANCE

The selection parameter here is $\theta_i = \underset{\sim}{\mu}_i' \sum_i^{-1} \underset{\sim}{\mu}_i$ which is the Mahalanobis distance function. The cases studied in recent work involve that of Σ_i known and Σ_i unknown. Here, an overview is presented of the work carried out by Alam and Rizvi (1966) and Srivastava and Taneja (1972) using this parameter.

4.1.1 PROCEDURES STUDIED BY K. ALAM AND M. H. RIZVI (1966)

The procedure R selects t populations such that the infimum of the probability of a correct selection over a sub-space of the parameter space is equal to P*. The main exercise is to determine the least favourable configuration of the parameter space for which the probability of a correct selection is a minimum. The expression for the minimum value determines the smallest sample size needed to satisfy the P* condition.

The selection of k multivariate normal populations with mean vector $\underset{\sim}{\mu}_i$ and covariance matrices Σ_i (i = 1, ..., k) using the Mahalanobis distance function $\theta_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$ reduces to selecting from k, non central chi squared populations (in the case of Σ_i known) and non central F populations (in the case of Σ_i unknown) with respect to the non-centrality parameters. The best population is the one associated with $\theta_{[k]}$.

Let $\underset{\sim}{\overline{X}}_i$ and S_i denote the sample mean vector and sample covariance matrix (defined in Chapter 3.2), based on a sample of size n from Π_i. The P* condition may be satisfied only on a subset of Ω which may be termed a "preference zone". One such subset of Ω described here is

$$\Omega_p = \Omega_1 \cap \Omega_2$$

where $\Omega_1 = \{\underset{\sim}{\theta} : \theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta_1\}$

and $\Omega_2 = \{\underset{\sim}{\theta} : \theta_{[k-t+1]} / \theta_{[k-t]} \geq \delta_2\}$

for some $\delta_1 > 0$, $\delta_2 > 1$.

For selecting the t best populations such that the Probability of Correct Selection $\geq$ P* whenever $\underset{\sim}{\theta} \in \Omega_p$, Alam and Rizvi (1966) have proposed rules for the two cases Σ_i known and Σ_i unknown.

4.1.1.1 CASE 1 Σ_i KNOWN

Let $\quad U_i = \overline{\underset{\sim}{X}}_i' \, \Sigma_i^{-1} \, \overline{\underset{\sim}{X}}_i, \quad i = 1, 2, \ldots., k.$

Then nU_i has a non central χ^2 distribution with p degrees of freedom and non centrality parameter $\lambda_i' = n\,\theta_i = n\,\underset{\sim}{\mu}_i' \, \Sigma_i^{-1} \, \underset{\sim}{\mu}_i$. The goal is to select the t populations associated with $U_{[k-t+1]}, \ldots., U_{[k]}$ where $U_{[k]} = \max(U_1, \ldots., U_k)$. The least favourable configuration (LFC) is given by

$$\theta_{[1]} = \ldots. = \theta_{[k-t]} = \delta_1/(\delta_2 - 1)$$

$$\theta_{[k-t+1]} = \ldots. = \theta_{[k]} = \delta_1\delta_2/(\delta_2 - 1).$$

The smallest value of n required to satisfy the P* condition is obtained from,

$$t \int_0^\infty F_p^{k-t}\left(x, \frac{n\,\delta_1}{\delta_2 - 1}\right) \left\{1 - F_p\left(x, \frac{n\,\delta_1\delta_2}{\delta_2 - 1}\right)\right\}^{t-1} f_p\left(x, \frac{n\,\delta_1\delta_2}{\delta_2 - 1}\right) dx = P*$$

where $f_p(x, \lambda')$ and $F_p(x, \lambda')$ denote the probability density function and the cumulative distribution function respectively of a non central χ^2 random variable with p degrees of freedom and non centrality parameter λ'.

These functions are given by

$$f_p(x, \lambda') = \frac{e^{-(x+\lambda')/2}}{2^{\frac{1}{2}p}} \sum_{r=0}^{\infty} \frac{(\lambda')^r \, x^{\frac{1}{2}p+r-1}}{r! \, 2^{2r} \, \Gamma(\frac{1}{2}p+r)} \qquad x > 0, \quad \lambda' \geq 0$$

$$F_p(x, \lambda') = \int_0^x f_p(t, \lambda') dt$$

Srivastava and Taneja (1972) have stated that the tables for n are not yet available to carry out this procedure.

4.1.1.2 CASE 2 Σ_i UNKNOWN

Let $\quad V_i = \overline{\underset{\sim}{X}}{}_i' \, S_i^{-1} \, \overline{\underset{\sim}{X}}_i (n-p)/np, \quad i = 1, 2, \ldots., k.$

Then nV_i has a non central F distribution with p and (n-p) degrees of freedom and non centrality parameter $\lambda_i' = n \, \theta_i = n \, \underset{\sim}{\mu}{}_i' \, \Sigma_i^{-1} \, \underset{\sim}{\mu}{}_i$. The goal is to select the t populations associated with $V_{[k-t+1]}, \ldots., V_{[k]}$ where $V_{[k]} = \max(V_1, V_2, \ldots., V_k)$. The least favourable configuration (LFC) is given by

$$\theta_{[1]} = \ldots. = \theta_{[k-t]} = \delta_1/(\delta_2 - 1)$$

$$\theta_{[k-t+1]} = \ldots. = \theta_{[k]} = \delta_1\delta_2/(\delta_2 - 1),$$

for some $\delta_1 > 0$ and $\delta_2 > 1$. The smallest value of n required to satisfy the P* condition is obtained from

$$t \int_0^\infty G_{p,n-p}^{k-t} \left(x, \frac{n \, \delta_1}{\delta_2 - 1}\right) \left\{1 - G_{p,n-p} \left(x, \frac{n \, \delta_1\delta_2}{\delta_2 - 1}\right)\right\}^{t-1} g_{p,n-p}\left(x, \frac{n \, \delta_1\delta_2}{\delta_2 - 1}\right) dx = P*$$

where $g_{p,q}(x, \lambda')$ and $G_{p,q}(x, \lambda')$ denote the probability density function and the cumulative distribution function respectively of a non central F random variable with p and q = (n-p) degrees of freedom and non centrality parameter λ'.

These functions are given by

$$g_{p,q}(x,\ \lambda') = \frac{e^{-\frac{1}{2}\lambda'}}{\Gamma(\frac{1}{2}q)} \sum_{r=0}^{\infty} \frac{x^{\frac{1}{2}p+r-1}}{(1+x)^{\frac{1}{2}p+\frac{1}{2}q+r}} \frac{\Gamma(\frac{1}{2}p + \frac{1}{2}q + r)}{\Gamma(\frac{1}{2}p + r)} \cdot \frac{\lambda'^{r}}{2^{r} r!}, \quad x > 0$$

$$G_{p,q}(x,\ \lambda') = \int_0^x g_{p,q}(t,\ \lambda')dt$$

4.1.2 PROCEDURES STUDIED BY M. S. SRIVASTAVA AND V. S. TANEJA (1972)

The problem of sequential selection of the best of k, p variate normal populations with means $\underset{\sim}{\mu}_i$, i = 1, 2, ..., k and common covariance matrix Σ (for known and unknown Σ) are considered. Here we discuss the selection done with respect to the Mahalanobis distance function $\theta_i = \underset{\sim}{\mu}_i' \sum^{-1} \underset{\sim}{\mu}_i$.

Here Paulson's (1964) sequential procedure for selecting the normal population with the largest mean is extended to the multivariate case. Truncated and non truncated sequential procedures similar to those of Paulson (1964) and Hoel and Mazumdar (1968) are investigated. Hoel and Mazumdar (1968) have proposed a sequential method of selecting a member of an exponential family with the largest parameter.

The following form is taken by the procedures discussed here. Denote the ranked θ_i's $(= \underset{\sim}{\mu}_i' \sum^{-1} \underset{\sim}{\mu}_i)$ by $\theta_{[1]} \leqslant .... \leqslant \theta_{[k]}$. The problem is to design a procedure R for selecting the best population corresponding to the largest θ_i value such that

$$P(CS/R) \geqslant P^* \quad \text{whenever} \quad \theta_{[k]} - \theta_{[k-1]} \geqslant \delta^*$$

where δ^* and P^* are specified by the experimenter and $\Pi_{[k]}$ is the population associated with $\theta_{[k]}$.

In the sequential procedures developed here the inferior populations are eliminated before the final stage of the experiment which tends to decrease the number of observations required to reach a decision. It is also proved that in the non truncated case the proper sequential procedure terminates with probability one.

The following lemma by Bechhofer, Kiefer and Sobel (1968) is needed in the theory to follow.

Let Z_1, Z_2, ... be a sequence of independently distributed random variables having the same distribution as $Z = X - Y$, where X and Y are independent non central chi square random variables with p degrees of freedom and non centrality parameters λ_1' and λ_2' respectively.

Let $\lambda_1' < \lambda_2'$, $0 \leqslant \lambda < \lambda_2' - \lambda_1'$ and $b > 0$. Then

$$P\left\{\underset{n}{\text{Sup}} \sum_{j=1}^{n} (Z_j + \lambda) > b\right\} \leqslant e^{-t_0 b}$$

where 'Sup' or Supremum is the least upper bound and $t_0 > 0$ is the solution of

$$\max\left\{t : (1-4t^2)^{-\frac{1}{2}p} \, e^{t[\lambda-4\lambda t^2+2t(\lambda_2' + \lambda_1') - (\lambda_2' - \lambda_1')](1-4t^2)^{-1}} \leqslant 1\right\}$$

$$0 < t < \tfrac{1}{2} \qquad \text{.... (A)}$$

4.1.2.1 NON TRUNCATED SEQUENTIAL PROCEDURE

Let $U_{ij} = \underset{\sim}{X}_{ij}' \, \Sigma^{-1} \, \underset{\sim}{X}_{ij}$, $i = 1, \ldots., k$, $j = 1, 2, \ldots.$

$$Z_{in} = \sum_{j=1}^{n} U_{ij} \; .$$

Now for $\delta^* > 0$ define

$$C_{\delta^*} = t_0^{-1} \log\{(k-1)(1-P^*)^{-1}\}$$

where t_0 is the solution of equation (A) with $\lambda = 0$.

Srivastava and Taneja (1972) do not explain how to choose $\lambda_2' - \lambda_1'$ and $\lambda_2' + \lambda_1'$ which are in fact lower and upper bounds for the ordered differences and sums respectively of the non centrality parameters. It is reasonable to choose $\lambda_2' - \lambda_1' = \delta^*$ but the choice of $\lambda_2' + \lambda_1'$ is unclear.

The values of t_0 have been tabulated by Srivastava and Teneja (1972) for selected values of $(\lambda_2' - \lambda_1')$ and $(\lambda_2' + \lambda_1')$, for the number of variates, p = 2, 3, 4. The table is given below.

TABLE T 4.1 : Values of t for $\lambda = 0$

$\lambda_2' - \lambda_1'$	$\lambda_2' + \lambda_1'$	t values		
		p = 2	3	4
0.5	4.5 6.5	0.0381 0.0287	0.0325 0.0259	0.0287 0.0231
1.0	5.0 7.0	0.0709 0.0550	0.0625 0.0493	0.0550 0.0447
2.5	6.5 8.5	0.1478 0.1197	0.1328 0.1094	0.1197 0.1000
5.0	9.0 11.0	0.2321 0.1947	0.2134 0.1806	0.1966 0.1684
8.0	12.0 14.0	0.2931 0.2537	0.2753 0.2406	0.2594 0.2275
12.0	16.0 18.0	0.3437 0.3062	0.3287 0.2931	0.3137 0.2819

The sequential procedure R_1 is as follows.

Start with one observation on each population Π_i and compute Z_{i1}, i = 1, 2, ..., k. Eliminate from further consideration any population Π_s for which

$$Z_{s1} \leqslant \max_r Z_{r1} - C_{\delta*} \qquad .... (B)$$

If all but one population is eliminated terminate the experiment and select the remaining population as the best one. Otherwise go on to the second stage of the experiment (m = 2, 3, ...) and take one measurement on each population not eliminated after the (m-1)th stage and eliminate any population Π_s for which

$$Z_{sm} \leqslant \max_r Z_{rm} - C_{\delta*} \qquad .... (C)$$

where the maximum is taken over all populations left after the (m-1)th stage. We terminate the procedure when there is only one population left out and select it as the best.

It has been shown in Paulson (1964) that this procedure terminates with probability one, and that

$$P(CS/R_1) \geqslant P* \quad \text{whenever} \quad \theta_{[k]} - \theta_{[k-1]} \geqslant \delta*.$$

4.1.2.2 TRUNCATED SEQUENTIAL PROCEDURE

In the case of the Non Truncated Sequential Procedure it has not been possible to obtain any upper bound for the expected number of observations. For this reason a class of truncated procedures similar to Paulson's (1964) procedure has been considered.

Let η be chosen such that $0 < \eta < \delta^*$. Define

$$C_\eta = t_\eta^{-1} \log\{(k-1)(1-P^*)^{-1}\}$$

where t_η is the solution of equation (A) with $\lambda = \eta$. Let W_η be the largest integer less than C_η/η. The sequential procedure R_2 is as follows:

Start sampling as in the case of the non-truncated case with equations (B) and (C) replaced by

$$Z_{s1} \leq \max_r Z_{r1} - C_\eta + \eta$$

and

$$Z_{sm} \leq \max_r Z_{rm} - C_\eta + m\eta$$

respectively.

If more than one population remains after the W_ηth stage the experiment is terminated at the next stage by selecting the population with the largest Z value. For this procedure R_2 with $\eta \in (0, \delta^*)$

$$P(CS/R_2) \geq P^* \quad \text{whenever} \quad \theta_{[k]} - \theta_{[k-1]} \geq \delta^*$$

As in the case of Paulson's (1964) procedure, the optimum value of η is not known. However as recommended by Paulson (1964) the value of $\eta = \delta^*/4$ may be used.

The same comments on the choice of $\lambda_2' - \lambda_1'$ and $\lambda_2' + \lambda_1'$ stated in Chapter 4.1.2.1 apply here too.

The values of t_η have been tabulated by Srivastava and Taneja (1972) for selected values of $(\lambda_2' - \lambda_1')$ and $(\lambda_2' + \lambda_1')$ for the number of variates, p = 2, 3, 4. The table is given below.

TABLE T 4.2 : VALUES OF t

$\lambda_2' - \lambda_1'$	λ	$\lambda_2' + \lambda_1'$	t values		
			p = 2	3	4
0.5	0.1	4.5	0.0308	0.0267	0.0235
	0.1	6.5	0.0235	0.0211	0.0190
	0.2	4.5	0.0231	0.0200	0.0177
	0.2	6.5	0.0177	0.0158	0.0143
	0.3	4.5	0.0154	0.0133	0.0118
	0.3	6.5	0.0118	0.0105	0.0095
	0.4	4.5	0.0077	0.0067	0.0059
	0.4	6.5	0.0059	0.0053	0.0048
1.0	0.2	5.0	0.0574	0.0502	0.0446
	0.2	7.0	0.0446	0.0401	0.0364
	0.5	5.0	0.0359	0.0314	0.0279
	0.5	7.0	0.0279	0.0251	0.0228
	0.7	5.0	0.0215	0.0188	0.0167
	0.7	7.0	0.0167	0.0150	0.0137
	0.9	5.0	0.0072	0.0063	0.0056
	0.9	7.0	0.0056	0.0050	0.0045
2.5	0.6	6.5	0.1143	0.1019	0.0920
	0.6	8.5	0.0917	0.0836	0.0769
	1.2	6.5	0.0784	0.0699	0.0630
	1.2	8.5	0.0629	0.0573	0.0526
	1.8	6.5	0.0419	0.0374	0.0337
	1.8	8.5	0.0337	0.0307	0.0282
	2.4	6.5	0.0059	0.0053	0.0048
	2.4	8.5	0.0048	0.0044	0.0040
5.0	1.0	9.0	0.1911	0.1745	0.1604
	1.0	11.0	0.1590	0.1473	0.1372
	2.2	9.0	0.1355	0.1231	0.1129
	2.2	11.0	0.1124	0.1038	0.0965
	3.5	9.0	0.0716	0.0651	0.0597
	3.5	11.0	0.0597	0.0552	0.0513
	4.8	9.0	0.0092	0.0084	0.0077
	4.8	11.0	0.0077	0.0072	0.0067

TABLE T 4.2 (continued)

$\lambda_2' - \lambda_1'$	λ	$\lambda_2' + \lambda_1'$	t values		
			p = 2	3	4
8.0	1.6	12.0	0.2472	0.2298	0.2144
	1.6	14.0	0.2114	0.1985	0.1870
	3.5	12.0	0.1783	0.1644	0.1526
	3.5	14.0	0.1515	0.1416	0.1329
	6.5	12.0	0.0566	0.0524	0.0489
	6.5	14.0	0.0488	0.0457	0.0430
	7.8	12.0	0.0072	0.0067	0.0063
	7.8	14.0	0.0063	0.0059	0.0056
12.0	3.0	16.0	0.2821	0.2657	0.2507
	3.0	18.0	0.2465	0.2338	0.2223
	6.0	16.0	0.1932	0.1803	0.1691
	6.0	18.0	0.1678	0.1583	0.1498
	9.0	16.0	0.0920	0.0862	0.0811
	9.0	18.0	0.0810	0.0766	0.0726
	11.5	16.0	0.0141	0.0134	0.0127
	11.5	18.0	0.0127	0.0121	0.0115

4.2 SELECTION IN TERMS OF THE EUCLIDEAN DISTANCE

The selection parameter here is $\theta_i = (\underset{\sim}{\mu}_i' \underset{\sim}{\mu}_i)^{\frac{1}{2}}$ which is the Euclidean distance function. The cases studied in recent work involve that of the common covariance matrix Σ known and Σ unknown. Here, an overview is presented of the sequential procedures investigated by Srivastava and Taneja (1972).

4.2.1 PROCEDURES STUDIED BY M. S. SRIVASTAVA AND V. S. TANEJA (1972)

The problem of sequential selection of the best of k, p variate normal populations with means $\underset{\sim}{\mu}_i$, i = 1, 2, ..., k and common covariance matrix Σ (for known and unknown Σ) are considered. Here we discuss the selection done with respect to the Euclidean distance function $\theta_i = (\underset{\sim}{\mu}_i' \; \underset{\sim}{\mu}_i)^{\frac{1}{2}}$.

Suppose that the ordered set of $\theta_i = (\underset{\sim}{\mu}_i' \; \underset{\sim}{\mu}_i)^{\frac{1}{2}}$ values of populations $\Pi_1, \Pi_2, \ldots., \Pi_k$ are denoted by

$$\theta_{[1]} \leqslant \theta_{[2]} \leqslant \ldots. \leqslant \theta_{[k]} \cdot$$

The θ_i values are assumed to be unknown and the best population is the one which corresponds to $\theta_{[k]}$.

Chow and Robbin's (1965) sequential theory has been applied to design a set of rules R such that

$$\lim_{\delta^* \to 0} \; P(CS/R) \geqslant P^* \quad \text{whenever} \quad \theta_{[k]} - \theta_{[k-1]} \geqslant \delta^*$$

where P* and δ^* are preassigned constants.

Two cases are considered.

<u>4.2.1.1</u> <u>CASE 1 Σ KNOWN</u>

Let $\quad \underset{\sim}{\overline{X}}_{in} = \sum_{j=1}^{n} \underset{\sim}{X}_{ij}/n \quad i = 1, 2, \ldots., k,$

where $\underset{\sim}{\overline{X}}_{in}$ is the sample mean vector based on n independent observations from Π_i, i = 1, 2, ..., k. The procedure R_1 is as follows:

Take a sample of size n from each population where n is the smallest integer satisfying

$$n \geqslant a^2 \; \lambda_1 \; \delta^{*-2}$$

where $\lambda_1 = \max\limits_{\underset{\sim}{c}:\underset{\sim}{c}'\underset{\sim}{c}=1} \; \underset{\sim}{c}' \; \Sigma \; \underset{\sim}{c}$

i.e., λ_1 is the maximum characteristic root of Σ and 'a' is given by

$$\Phi\left(\frac{-a}{\sqrt{2}}\right) = (1 - P^*)(k - 1)^{-1}$$

where $\Phi(x) = \int_{-\infty}^{x} [e^{-t^2/2} (2\Pi)^{-\frac{1}{2}}]dt$ is the normal cumulative distribution function. The standard normal probability density function is given by

$$f(x) = \frac{1}{(2\Pi)^{\frac{1}{2}}} e^{-x^2/2}.$$

Now select the population associated with the largest $\overline{\underset{\sim}{X}}{}'_{in} \overline{\underset{\sim}{X}}_{in}$.

<u>4.2.1.2</u> <u>CASE 2 Σ UNKNOWN</u>

$$\text{Let} \qquad \overline{\underset{\sim}{X}}_{in} = \sum_{j=1}^{n} \underset{\sim}{X}_{ij}/n \qquad i = 1, 2, \ldots., k$$

$$S_n = (nk)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_{in})'(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_{in})$$

$$\lambda_{1n} = \max_{\underset{\sim}{b}:\underset{\sim}{b}'\underset{\sim}{b}=1} \underset{\sim}{b}' S_n \underset{\sim}{b}$$

Note that $\lim_{n\to\infty} \lambda_{1n} = \lambda_1$ almost surely.

Let $\{a_n\}$ be a sequence of positive constants such that $\lim_{n\to\infty} a_n = a$ where 'a' is defined by

$$\Phi\left(\frac{-a}{\sqrt{2}}\right) = (1 - P^*)(k - 1)^{-1}$$

and $\Phi(x)$ was defined in Chapter 4.2.1.1. The procedure R_2 is as follows:

Start by taking $n_0 > p$ observations from each population and then one observation at a time from each population and stop according to the stopping rule defined by

$$N = \text{smallest } n \geqslant n_0 \text{ such that } \lambda_{1n} \leqslant n\delta^{*2} a_n^{-2}.$$

When the sampling is stopped at $N = n$ select the population with the largest $\overline{\underset{\sim}{X}}'_{in} \overline{\underset{\sim}{X}}_{in}$ as the best population.

Srivastava and Taneja (1972) state that as $\delta^* \to 0$, this procedure terminates with probability 1. Extensive work done by Starr (1966) on the univariate case suggests that the procedure and variations of it works for various values of δ^*.

It is not clear how n_0 is determined in practical applications. Presumably $n_0 > p$ is chosen so that it is large enough for S_{n_0} to be a reasonable approximation to Σ but not so large that the procedure terminates immediately. Srivastava and Taneja (1972) do not discuss this or give guidelines on how to choose n_0.

4.3 SELECTION IN TERMS OF MULTIPLE CORRELATION COEFFICIENTS

Here we consider the problem of selection of t largest from among k multiple correlation coefficients, each arising from one of k independent p variate normal populations with unknown mean vectors and unknown covariance matrices.

To arrive at a selection procedure a preassigned probability value $\binom{k}{t}^{-1} < P^* < 1$ is set and the requirement that the probability of a correct selection is not smaller than P^* whenever the square of the t largest multiple correlation coefficients

1) exceeds the square of each of the remaining multiple correlations by a magnitude δ_1, and simultaneously,

2) each is at least δ_2 times as large as each of the squares of the remaining multiple correlations

are met.

The separation thresholds $0 < \delta_1 < 1$ and $\delta_2 > 1$ are also preassigned. These two conditions specify a "preference zone" in the parameter space. The problem is formulated as follows:

Consider k ($\geqslant 2$) independent p variate ($p \geqslant 2$) normal populations $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$, i = 1, 2, ..., k. Here the mean vectors $\underset{\sim}{\mu}_i$ and the covariance matrices Σ_i are all unknown. For the i th population let θ_i denote the squared population multiple correlation coefficient between the first variate and the set of (p - 1) remaining variates. This is defined by

$$\theta_i = \rho^2_{i:1,\, 2,\, \ldots.,\, p} = 1 - \frac{|\Sigma_i|}{\sigma_{11}^{(i)} |\Sigma_{i(11)}|}$$

Here $\sigma_{11}^{(i)}$ is the leading element of Σ_i and $\Sigma_{i(11)}$ is the matrix obtained from Σ_i by deleting the first row and the first column.

Let the ordered values of the θ_i's be denoted by

$$0 \leqslant \theta_{[1]} \leqslant \theta_{[2]} \leqslant \ldots. \leqslant \theta_{[k]} < 1$$

The problem is the selection of the $t < k$ populations with the largest θ_i's on the basis of n sample observations from each of the k populations.

Let the parameter space Ω of the θ_i's be partitioned into a "preference zone" Ω_p and its complement the "indifference zone" Ω_I. For specified Ω_p and P^* $(\binom{k}{t}^{-1} < P^* < 1)$ a decision procedure R is required where

$$\underset{\Omega_p}{\text{Inf}} \; P(CS/R) \geq P^*.$$

Rizvi and Solomon (1973) have investigated a decision procedure R as an asymptotic (as $n \to \infty$) solution to this problem, with an explicit definition of Ω_p the preference zone.

Alam, Rizvi and Solomon (1975) subsequently investigated the procedure R in the exact sample case.

4.3.1 PROCEDURE INVESTIGATED BY M. H. RIZVI AND H. SOLOMON (1973)

The procedure R_1 is as follows:

Consider a random sample of size n where $n > p$ from each of the k populations. The sample squared multiple correlation coefficient

$$y_i = 1 - s_{11}^{(i)^{-1}} |S_i| |S_{i(11)}|^{-1} = r^2_{i:1,\ 2,\ \ldots.,\ p}$$

where $s_{11}^{(i)}$ is the leading element of the sample covariance matrix S_i and $|S_{i(11)}|$ is the cofactor of $s_{11}^{(i)}$; $(i = 1, \ldots., k)$, is then computed for each population. The y_i's are then ranked, breaking ties if any, with suitable randomization. The populations corresponding to the t largest y_i's are then selected.

For this problem the preference zone Ω_p is defined as $\Omega_1 \cap \Omega_2$ where

$$\Omega_1 = \{\underset{\sim}{\theta} \in \Omega : \theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta_1\}$$

$$\Omega_2 = \{\underset{\sim}{\theta} \in \Omega : \theta_{[k-t+1]} / \theta_{[k-t]} \geq \delta_2\}$$

and $0 < \delta_1 < 1$ and $\delta_2 > 1$ are specified constants.

Calculation of n

For the procedure R_1 and the preference zone Ω_p the probability requirement $\underset{\Omega_p}{\text{Inf}}\, P(CS/R) \geq P^*$ is employed and solved asymptotically for the common sample size n from each population. This value of n provides the sample size to incorporate in the selection procedure R_1 so that R_1 satisfies $\underset{\Omega_p}{\text{Inf}}\, P(CS/R_1) \geq P^*$ asymptotically.

For fixed p as $n \to \infty$, ny_i is asymptotically distributed as a non-central chi square random variable with $q = (p - 1)$ degrees of freedom, and non-centrality parameter $\lambda' = n\,\theta_i$. The non-central chi squared probability density function denoted here as $f_q(y, \lambda')$ and the cumulative distribution function denoted here as $F_q(y, \lambda')$ are given in Chapter 4.1.1.1.

The asymptotic probability of a correct selection $P_a(CS/R_1)$ can now be written as,

$$P_a(CS/R_1) = \sum_{i=k-t+1}^{k} \int_0^\infty \prod_{\beta=1}^{k-t} F_q(y, n\theta_{[\beta]}) \cdot \prod_{\substack{\alpha=k-t+1 \\ \alpha \neq i}}^{k} \left\{1 - F_q(y, n\theta_{[\alpha]})\right\}.$$

$$f_q(y, n\theta_{[i]})\,dy$$

Here, the least favourable configuration (LFC) is given by

$$\theta_{[1]} = \theta_{[2]} = \ldots\ldots = \theta_{[k-t]} = \delta_1/(\delta_2 - 1)$$

$$\theta_{[k-t+1]} = \ldots\ldots = \theta_{[k]} = \delta_1\delta_2/(\delta_2 - 1)$$

The smallest common sample size n is obtained as the solution of the integral equation

$$t\int_0^\infty F_q^{k-t}(y,n\delta_1/(\delta_2-1))[1-F_q(y,n\delta_1\delta_2/(\delta_2-1))]^{t-1}\ f_q(y,n\delta_1\delta_2/(\delta_2-1))dy = P^*$$

The tables used to solve the equation for $n\delta_1$ for $P^* = .90$ and $P^* = .95$ are given in Rizvi and Solomon (1973). Once $n\delta_1$ is known, n can be calculated.

4.3.2 PROCEDURE INVESTIGATED BY K. ALAM, M. H. RIZVI AND H. SOLOMON (1975)

The Selection Procedure R_2 explained here is the same as the procedure R_1 (given in Chapter 4.3.1) except for the specified preference zone and the fact that this procedure is studied in the exact sample size case. Here also, the sample squared multiple correlation coefficient y_i (defined in Chapter 4.3.1) for each population Π_i ($i = 1, 2, \ldots\ldots, k$) is used to select $t < k$ populations. The y_i's are ranked breaking ties if any with suitable randomization and the populations corresponding to the t largest y_i's selected.

For this problem, the preference zone Ω_p is defined as $\Omega_p = \Omega_1' \cap \Omega_2$ where

$$\Omega_1' = \{\underset{\sim}{\theta} \in \Omega : (1 - \theta_{[k-t]})/(1 - \theta_{[k-t+1]}) \geq \delta_1'\}$$

$$\Omega_2 = \{\underset{\sim}{\theta} \in \Omega : \theta_{[k-t+1]} / \theta_{[k-t]} \geq \delta_2\}$$

where $\delta_1' > 1$ and $\delta_2 > 1$ are specified constants.

Calculation of n

For the procedure R_2 and the preference zone Ω_p the probability requirement $\underset{\Omega_p}{\text{Inf}}\ P(CS/R_2) \geq P^*$ is employed and solved for the common sample size n for each population. This value of n provides the sample size to incorporate in the selection procedure R_2 so that R_2 satisfies $\underset{\Omega_p}{\text{Inf}}\ P(CS/R_2) \geq P^*$.

Some preliminaries concerning the distribution of a typical sample squared multiple correlation coefficient y_i based on a random sample of size n ($\geq p + 2$) and having population squared multiple correlation coefficient θ_i are given below. Let

$$J(a,b;c;x) = \sum_{r=0}^{\infty} \frac{(a)_r(b)_r}{(c)_r} \frac{x^r}{x!}$$

denote the hypergeometric function where $(a)_0 = 1$ and $(a)_r = a(a+1) \ldots. (a+r-1)$, $r = 1, 2, \ldots.$ The probability density function of y_i is given by

$$h_y(a,c,\theta_i) = (1 - \theta_i)^a B_y(c,a-c) J(a,a;c;\theta_i y), \quad 0 < y < 1$$

where

$$B_y(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\ \Gamma(b)} y^{a-1}(1-y)^{b-1},$$

i.e. a Beta probability density function. $a = (n-1)/2$, $c = (p-1)/2$ and $H_y(a,c,\theta_i)$ denotes the cumulative distribution function of y_i, where

$$H_y(a,c,\theta_i) = \int_0^y h_t(a,c,\theta_i)dt$$

The probability of a correct selection $P(CS/R_2)$ can be written as

$$P(CS/R_2) = \sum_{i=k-t+1}^{k} \int_0^1 \prod_{\beta=1}^{k-t} H_y(a,c,\theta_{[\beta]}) \; . \prod_{\substack{\alpha=k-t+1 \\ \alpha\neq i}}^{k} \{1 - H_y(a,c,\theta_{[\alpha]})\} \; .$$

$$h_y(a,c,\theta_{[i]})dy$$

Here, the least favourable configuration (LFC) is given by

$$\theta_{[1]} = \theta_{[2]} = \ldots . = \theta_{[k-t]} = (\delta_1' - 1)/(\delta_1'\delta_2 - 1)$$

$$\theta_{[k-t+1]} = \ldots . = \theta_{[k]} = \delta_2(\delta_1' - 1)/(\delta_1'\delta_2 - 1)$$

The sample size n is obtained as the solution of the integral equation

$$t \int_0^1 H_y^{k-1}(a,c,(\delta_1'-1)/\delta_1'\delta_2-1) \; . \; [1 - H_y(a,c,\delta_2(\delta_1'-1)/\delta_1'\delta_2-1)]^{t-1} \; .$$

$$h_y(a,c,\delta_2(\delta_1'-1)/\delta_1'\delta_2-1)dy = P*$$

where $a = (n-1)/2$ and $c = (p-1)/2$.

Tables to solve the equation are not given in Alam, Rizvi and Solomon (1975).

4.4 SELECTION IN TERMS OF THE SUM OF THE BIVARIATE PRODUCT-MOMENT CORRELATIONS

This is concerned with selecting the single largest population having the highest 'association' from among the set of k populations. Here, let θ_i be a measure associated with population Π_i, defined by

$$\theta_i = \sum_{\substack{c=1 \\ c \neq d}}^{p} \sum_{d=1}^{p} \frac{\rho_{cd}^{(i)}}{p(p-1)}, \quad i = 1, 2, \ldots., k$$

where $\rho_{cd}^{(i)}$ is the bivariate product-moment correlation coefficient between the c th and the d th coordinates of a vector $\underset{\sim}{X}$ from Π_i.

$$\rho_{cd}^{(i)} = \frac{\sigma_{cd}^{(i)}}{\left(\sigma_{cc}^{(i)} \; \sigma_{dd}^{(i)}\right)^{\frac{1}{2}}}$$

where

$$\Sigma_i = \begin{bmatrix} \sigma_{11}^{(i)} & \sigma_{12}^{(i)} & \ldots. & \sigma_{1p}^{(i)} \\ \sigma_{21}^{(i)} & & & \\ \vdots & & & \\ \sigma_{p1}^{(i)} & & & \sigma_{pp}^{(i)} \end{bmatrix}$$

Let $\theta_{[1]} \leqslant \theta_{[2]} \leqslant \ldots. \leqslant \theta_{[k]}$ be the ordered values of θ_i. Govindarajulu and Gore (1971) have studied a procedure for selecting the population associated with $\theta_{[k]}$ so that the P* condition $\inf_{\Omega_p} P(CS/R) \geqslant P^*$ is asymptotically $(n \to \infty)$ is satisfied. Here Ω_p is the preference zone.

4.4.1 PROCEDURE STUDIED BY Z. GOVINDARAJULU AND A. P. GORE (1971)

Define

$$V_i = \sum_{\substack{c=1 \\ c \neq d}}^{p} \sum_{d=1}^{p} \frac{r_{cd}^{(i)}}{p(p-1)}, \quad i = 1, 2, \ldots., k$$

where $r_{cd}^{(i)}$ is the sample correlation coefficient defined by

$$r_{cd}^{(i)} = \frac{s_{cd}^{(i)}}{\left(s_{cc}^{(i)}\ s_{dd}^{(i)}\right)^{\frac{1}{2}}}$$

and $s_{cd}^{(i)}$ is defined in Chapter 3.2. The procedure R selects the population associated with the largest V_i.

For the case $p \geqslant 3$ the authors have shown that for large n

$$P(CS/R) = P\left[U_i \leqslant \sqrt{n}\ \delta^* \left\{\frac{2(p+3)(p-3)}{p(p-1)}\right\}^{-\frac{1}{2}}, \quad i = 1, \ldots\ldots, k-1\right]$$

whenever $\theta_{[k]} - \theta_{[k-1]} \geqslant \delta^*$ and $U_i, \ldots\ldots, U_{k-1}$ are standard normal random variables with equal correlation ½.

In the bivariate case θ_i will be the product-moment correlation and V_i the sample correlation coefficient, and

$$P(CS/R) \geqslant P\left[U_i \leqslant \frac{\sqrt{n}\ \delta^*}{\sqrt{2}}, \quad i = 1, 2, \ldots\ldots, k-1\right].$$

4.5 SELECTION IN TERMS OF THE COEFFICIENT OF ALIENATION

Let $\underset{\sim}{X}_i = (\underset{\sim}{Y}_i, \underset{\sim}{Z}_i)'$ be a $(q_1 + q_2)$ dimensional random vector with covariance matrix

$$\sum_i = \begin{bmatrix} \sum_{yy}^{(i)} & \sum_{yz}^{(i)} \\ & \\ \sum_{zy}^{(i)} & \sum_{zz}^{(i)} \end{bmatrix}$$

where $\sum_{yy}^{(i)}$, $\sum_{yz}^{(i)}$, $\sum_{zy}^{(i)}$ and $\sum_{zz}^{(i)}$ are submatrices of dimension $q_1 \times q_1$, $q_1 \times q_2$, $q_2 \times q_1$ and $q_2 \times q_2$ respectively. Assume that $q_1 \leqslant q_2$. The coefficient of alienation between $\underset{\sim}{Y}_i$ and $\underset{\sim}{Z}_i$ is θ_i and defined by

$$\theta_i^2 = \frac{|\Sigma_i|}{|\Sigma_{yy}^{(i)}||\Sigma_{zz}^{(i)}|}$$

where the coefficient of alienation is a measure of association between the two sets of variables. For $q_1 = 1$ it is equal to $(1 - \rho^2)$ where ρ is the multiple correlation coefficient between y and $(z_1, z_2, \dots., z_{q_2})$. Let $\theta_{[1]} \leqslant \theta_{[2]} \leqslant \dots. \leqslant \theta_{[k]}$ be the ordered θ_i values.

The selection of the population associated with $\theta_{[1]}$ subject to the P* condition

$$\inf_{\Omega_p} P(CS/R) \geqslant P^* \quad \text{whenever} \quad \theta_{[2]}^2/\theta_{[1]}^2 \geqslant \delta^*$$

where $\delta^* > 1$, and has been considered by Frischtak (1973).

4.5.1 PROCEDURE PROPOSED BY R. M. FRISCHTAK (1973)

Let V_i be defined by

$$V_i^2 = \frac{|S_i|}{|S_{yy}^{(i)}||S_{zz}^{(i)}|}$$

where S_i is the sample covariance matrix based on n independent vector observations from Π_i. $S_{yy}^{(i)}$ and $S_{zz}^{(i)}$ are the appropriate submatrices of the partitioned S_i matrix

$$S_i = \begin{bmatrix} S_{yy}^{(i)} & S_{yz}^{(i)} \\ S_{zy}^{(i)} & S_{zz}^{(i)} \end{bmatrix}$$

where $S_{yy}^{(i)}$, $S_{yz}^{(i)}$, $S_{zy}^{(i)}$ and $S_{zz}^{(i)}$ are submatrices of dimension $q_1 \times q_1$,

$q_1 \times q_2$, $q_2 \times q_1$ and $q_2 \times q_2$ respectively.

The rule R is to select the population which gives the smallest V_i. An asymptotic ($n \to \infty$) lower bound on the probability of correct selection is given by

$$P\left\{U_i \leq \frac{n^{\frac{1}{2}} \log \delta^*}{2(2q_1)^{\frac{1}{2}}}, \quad i = 2, 3, \ldots., k\right\}$$

where $U_2, U_3, \ldots., U_k$ are standard normal variables with equal correlation 1/2.

CHAPTER 5

5. THE SUBSET SELECTION APPROACH TO RANKING AND SELECTION OF SEVERAL MULTIVARIATE NORMAL POPULATIONS

As in the preceding chapter, various real valued functions have been defined to select the populations.

These real valued functions are:

1) The Mahalanobis Distance
2) Generalized Variances
3) Multiple Correlation Coefficients
4) Measures of Association between Two Subclasses of Variates.

5.1 SELECTION IN TERMS OF THE MAHALANOBIS DISTANCE

The selection of a subset of k Multivariate Normal Populations that would include the population located farthest from the origin was considered by Gupta (1966).

This distance, known as the Mahalanobis distance, is defined as $\underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$ where $\underset{\sim}{\mu}_i$ is the mean vector and Σ_i is the covariance matrix of the i th population.

Let $Y_{ij} = \underset{\sim}{X}_{ij}' \Sigma^{-1} \underset{\sim}{X}_{ij}$. Here, all Σ_i are assumed to be equal to Σ, and $\underset{\sim}{X}_{ij}$, $i = 1, 2, \ldots., k$, $j = 1, 2, \ldots., n$ is a vector with p components of observations on the i th population. Then $Y_i = \sum_{j=1}^{n} Y_{ij}$ has a non-central chi squared distribution with np degrees of freedom and non centrality parameter $\lambda_i' = n\ \theta_i$ where $\theta_i = \underset{\sim}{\mu}_i' \sum^{-1} \underset{\sim}{\mu}_i$. The non-central chi squared probability density function and the cumulative distribution function are given in Chapter 4.1.1.1.

5.1.1 RULES PROPOSED BY S. S. GUPTA (1966)

5.1.1.1 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE LARGEST θ_i, $\Sigma_1 = \Sigma_2 = \ldots. = \Sigma_k = \Sigma$ KNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \sum^{-1} \underset{\sim}{\mu}_i ,$$
$$Y_{ij} = \underset{\sim}{X}_{ij}' \sum^{-1} \underset{\sim}{X}_{ij},$$
$$Y_i = \sum_{j=1}^{n} Y_{ij}.$$

The rule R_1 is as follows:

Select Π_i if and only if $Y_i \geqslant c \max(Y_1, Y_2, \ldots., Y_k)$ where $0 < c = c(k,n,p,P^*) \leqslant 1$ is determined to satisfy the P^* condition, $\inf_{\Omega} P(CS/R_1) \geqslant P^*$.

Gupta (1966) showed that

$$\inf_{\Omega} P(CS/R_1) = \inf_{\lambda' \geqslant 0} \int_0^\infty F_{\lambda'}^{k-1}(x/c)\ f_{\lambda'}(x)\ dx$$

where $f_{\lambda'}(x)$ and $F_{\lambda'}(x)$ are the probability density function and the cumulative distribution function respectively of a non central chi squared distribution with np degrees of freedom, and non centrality parameter λ'. These have been defined in Chapter 4.1.1.1.

The right hand side of the integral is non decreasing in λ'. This has been shown by Gupta (1966) for k = 2 populations and Gupta and

Studden (1970) for $k \geq 2$ populations. The integral is monotonically increasing in λ' so the infimum take place when $\lambda' = 0$. The property of monotonicity is that the probability of selecting a population with a larger value of λ' is at least as large as the probability of selecting a population with a smaller value of λ'. Therefore the problem reduces to selecting the gamma population with the largest scale parameter. Thus the constant c for this procedure is given by

$$\int_0^\infty G_\nu^{k-1} \left(\frac{x}{c}\right) g_\nu(x)\ dx = P*$$

where $g_\nu(x)$ and $G_\nu(x)$ are the probability density function and the cumulative distribution function respectively of a standardized gamma variable with $\nu = np/2$ degrees of freedom. These functions are given by

$$g_\nu(x) = e^{-x}\ x^{\nu-1}/\Gamma(\nu), \quad x > 0$$

$$G_\nu(x) = \int_0^x g_\nu(t)\ dt$$

The values for c are tabulated by Gupta (1963) and Armitage and Krishnaiah (1964).

5.1.1.2 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE SMALLEST θ_i, $\Sigma_1 = \Sigma_2 = \ldots\ldots = \Sigma_k = \Sigma$ KNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \Sigma^{-1} \underset{\sim}{\mu}_i$$

$$Y_{ij} = \underset{\sim}{X}_{ij}' \Sigma^{-1} \underset{\sim}{X}_{ij}$$

$$Y_i = \sum_{j=1}^{n} Y_{ij}$$

The rule R_2 is as follows:

Select π_i if and only if $Y_i \leq b \min(Y_1, Y_2, \ldots., Y_k)$ where $b = b(k,n,p,P^*) > 1$ is determined to satisfy the P* condition $\inf_{\Omega} P(CS/R_2) \geq P^*$.

Gupta (1966) showed that

$$\inf_{\Omega} P(CS/R_2) = \inf_{\lambda' \geq 0} \int_0^{\infty} [1 - F_{\lambda'}(x/b)]^{k-1} f_{\lambda'}(x)\, dx$$

where $f_{\lambda'}(x)$ and $F_{\lambda'}(x)$ are the probability density function and the cumulative distribution function respectively of a non central chi squared distribution with np degrees of freedom and non centrality parameter λ'. This has been defined in Chapter 4.1.1.1.

The right hand side of the integral is non decreasing in λ'. The integral is monotonically increasing in λ' so the infimum takes place when $\lambda' = 0$. Therefore the problem reduces to selecting the gamma population with the smallest scale parameter. Thus, the constant b for this procedure is given by

$$\int_0^{\infty} [1 - G_{\nu}(x/b)]^{k-1} g_{\nu}(x)\, dx = P^*$$

where $G_{\nu}(x)$ is the cumulative distribution function of a standardized gamma variable with np/2 degrees of freedom. This has been defined in Chapter 5.1.1.1. The values for b are tabulated by Gupta and Sobel (1962) and Armitage and Krishnaiah (1964).

5.1.2 PROCEDURES STUDIED BY S. S. GUPTA AND W. J. STUDDEN (1970)

5.1.2.1 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE LARGEST θ_i, Σ_i NOT NECESSARILY EQUAL BUT KNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$$
$$Z_{ij} = \underset{\sim}{X}_{ij}' \Sigma_i^{-1} \underset{\sim}{X}_{ij}$$
$$Z_i = \sum_{j=1}^{n} Z_{ij}$$

The procedure R_1 is as follows:

Select Π_i if and only if $c_1 Z_i \geq \max(Z_1, \ldots., Z_k)$, $c_1 > 1$ where c_1 is determined to satisfy the P* condition $\inf_{\Omega} P(CS/R_1) \geq P*$.

Gupta and Studden showed that

$$\inf_{\Omega} P(CS/R_1) = \int_0^\infty F_{np}^{k-1}(c_1 x)\, f_{np}(x)\, dx$$

where $f_{np}(x)$ and $F_{np}(x)$ are the probability density function and the cumulative distribution function respectively of a central chi squared distribution with np degrees of freedom. These functions are given by

$$f_{np}(x) = \frac{x^{\frac{1}{2}np-1}\, e^{-\frac{1}{2}x}}{2^{np/2}\, \Gamma(np/2)}$$

$$F_{np}(x) = \int_0^x f_{np}(t)\, dt$$

c_1 is chosen so that

$$\int_0^\infty F_{np}^{k-1}(c_1 x)\, f_{np}(x)\, dx = P*$$

5.1.2.2 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE SMALLEST θ_i, Σ_i NOT NECESSARILY EQUAL BUT KNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$$

$$Z_{ij} = \underset{\sim}{X}_{ij}' \Sigma_i^{-1} \underset{\sim}{X}_{ij}$$

$$Z_i = \sum_{j=1}^{n} Z_{ij}$$

The procedure R_2 is as follows:

Select Π_i if and only if $Z_i \leqslant b_1 \min(Z_1, \ldots, Z_k)$, $b_1 > 1$ where b_1 is determined to satisfy the P* condition $\underset{\Omega}{\text{Inf}}\, P(CS/R_2) \geqslant P^*$.

Gupta and Studden (1970) showed that

$$\underset{\Omega}{\text{Inf}}\, P(CS/R_2) = \int_0^\infty [1 - F_{np}(x/b_1)]^{k-1} f_{np}(x)\, dx$$

where $f_{np}(x)$ and $F_{np}(x)$ are the probability density function and the cumulative distribution function respectively of a central chi squared distribution with np degrees of freedom. These functions have been given in Chapter 5.1.2.1. b_1 is chosen so that

$$\int_0^\infty [1 - F_{np}(x/b_1)]^{k-1} f_{np}(x)\, dx = P^*$$

5.1.2.3 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE LARGEST θ_i, Σ_i ARE DIFFERENT AND UNKNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$$

$$\overline{\underset{\sim}{X}}_i = \sum_{j=1}^{n} \underset{\sim}{X}_{ij}/n$$

$$S_i = \sum_{j=1}^{n} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_i)(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_i)'/n-1$$

$$Z_i = \overline{\underset{\sim}{X}}_i' \, S_i^{-1} \, \overline{\underset{\sim}{X}}_i$$

The procedure R_3 is as follows:

Select Π_i if and only if $c_2 Z_i \geq \max(Z_1, \ldots., Z_k)$, $c_2 > 1$ where $c_2 = c_2(k,n,p,P^*)$ is determined to satisfy the P* condition $\inf_{\Omega} P(CS/R_3) \geq P^*$.

Gupta and Studden (1970) showed that

$$\inf_{\Omega} P(CS/R_3) = \int_0^{\infty} F_{p,n-p}^{k-1}(c_2 x) \; f_{p,n-p}(x) \; dx$$

where $f_{p,n-p}(x)$ and $F_{p,n-p}(x)$ are the probability density function and the cumulative distribution function respectively of a central F distribution with $\nu_1 = p$ and $\nu_2 = n-p$ degrees of freedom. These functions are given by

$$f_{\nu_1,\nu_2}(x) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} (\nu_1/\nu_2)^{\nu_1/2} \, x^{(\nu_1/2)-1} \, (1+(\nu_1/\nu_2)x)^{-\frac{1}{2}(\nu_1+\nu_2)},$$

$$\text{for } x > 0$$

$$F_{\nu_1,\nu_2}(x) = \int_0^{x} f_{\nu_1,\nu_2}(t) \; dt$$

c_2 is chosen so that

$$\int_0^{\infty} F_{p,n-p}^{k-1}(c_2 x) \; f_{p,n-p}(x) \; dx = P^*$$

The values of $1/c_2$ have been tabulated by Gupta and Panchapakesan (1969).

5.1.2.4 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE SMALLEST θ_i, Σ_i ARE DIFFERENT AND UNKNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \, \Sigma_i^{-1} \, \underset{\sim}{\mu}_i$$

$$\underset{\sim}{\overline{X}}_i = \sum_{j=1}^{n} \underset{\sim}{X}_{ij}/n$$

$$S_i = \sum_{j=1}^{n} (\underset{\sim}{X}_{ij} - \underset{\sim}{\overline{X}}_i)(\underset{\sim}{X}_{ij} - \underset{\sim}{\overline{X}}_i)'/n-1$$

$$Z_i = \underset{\sim}{\overline{X}}_i' \, S_i^{-1} \, \underset{\sim}{\overline{X}}_i$$

The procedure R_4 is as follows:

Select Π_i if and only if $Z_i \leqslant b_2 \min(Z_1, \ldots., Z_k)$, $b_2 > 1$ where b_2 is determined to satisfy the P* condition $\inf_{\Omega} P(CS/R_4) \geqslant P^*$.

Gupta and Studden (1970) showed that

$$\inf_{\Omega} P(CS/R_4) = \int_0^{\infty} [1 - F_{p,n-p}(x/b_2)]^{k-1} f_{p,n-p}(x) \, dx$$

where $f_{p,n-p}(x)$ and $F_{p,n-p}(x)$ are the probability density function and the cumulative distribution function respectively of a central F distribution with $\nu_1 = p$ and $\nu_2 = n-p$ degrees of freedom. These functions have been given in Chapter 5.1.2.3. b_2 is chosen so that

$$\int_0^{\infty} [1 - F_{p,n-p}(x/b_2)]^{k-1} f_{p,n-p}(x) \, dx = P^*$$

5.1.3 PROCEDURE CONSIDERED BY K. ALAM AND M. H. RIZVI (1966)

5.1.3.1 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE LARGEST θ_i, Σ_i NOT NECESSARILY EQUAL BUT KNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$$

$$\overline{\underset{\sim}{X}}_i = \sum_{j=1}^{n} \underset{\sim}{X}_{ij}/n$$

$$T_i = \overline{\underset{\sim}{X}}_i' \Sigma_i^{-1} \overline{\underset{\sim}{X}}_i$$

The procedure R is as follows:

Select Π_i if and only if $T_i \geq c_3 \max(T_1, \ldots., T_k)$ where $0 < c_3 < 1$. The smallest value of c_3 required to satisfy the P* condition $\inf_{\Omega} P(CS/R) \geq P^*$ is determined by the equation

$$\int_0^\infty F_p^{k-1}(c_3 x)\ f_p(x)dx = P^*$$

where $f_p(x)$ and $F_p(x)$ are the probability density function and the cumulative distribution function respectively of a central chi squared distribution with p degrees of freedom. These functions have been defined in Chapter 5.1.2.1.

Gupta and Panchapakesan (1979) state that this procedure is unsatisfactory as the constant c_3 does not depend on n.

5.1.4 UNSOLVED PROCEDURES

5.1.4.1 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE LARGEST θ_i, $\Sigma_1 = \Sigma_2 = \ldots. = \Sigma_k = \Sigma$ KNOWN

Here
$$\theta_i = \underset{\sim}{\mu}_i' \textstyle\sum^{-1} \underset{\sim}{\mu}_i$$

$$U_i = \underset{\sim}{\overline{X}}_i' \textstyle\sum^{-1} \underset{\sim}{\overline{X}}_i$$

The procedure R is as follows:

Select Π_i if and only if $U_i \geqslant \max(U_1, \ldots., U_k) - d$ where d is determined to satisfy the P* condition $\underset{\Omega}{\text{Inf}}\ P(CS/R) \geqslant P^*$

$$\underset{\Omega}{\text{Inf}}\ P(CS/R) = \underset{\lambda' \geqslant 0}{\text{Inf}} \int_0^\infty F_{\lambda'}^{k-1}(x+d)\ f_{\lambda'}(x)\ dx$$

where $f_{\lambda'}(x)$ and $F_{\lambda'}(x)$ are the probability density function and the cumulative distribution function respectively of a non central chi squared distribution with np degrees of freedom and non centrality parameter λ'. These functions are defined in Chapter 4.1.1.1. d is chosen so that

$$\underset{\lambda' \geqslant 0}{\text{Inf}} \int_0^\infty F_{\lambda'}^{k-1}(x+d)\ f_{\lambda'}(x)\ dx = P^*$$

Since the monotone behaviour of the integral involving d is not known procedures of the above type when $\Sigma_i = \Sigma$ known or Σ_i not equal but known have not been determined explicitly.

Another unsolved problem is the case of $\Sigma_1 = \Sigma_2 = \ldots. = \Sigma_k = \Sigma$ unknown, and a pooled estimate is used for Σ.

5.2 SELECTION IN TERMS OF THE GENERALIZED VARIANCES

The Covariance Matrix is regarded as the natural measure of dispersion for a multivariate normal distribution. However, a univariate measure of dispersion need be defined for the purpose of selection. Various measures of dispersion have been considered in the statistical literature, but none of these is uniformly best in the sense of being a robust estimator of the dispersion. A frequently used measure of dispersion is the generalized variance.

In this section, selection in terms of the univariate measure $\theta_i = |\Sigma_i|$ the generalized variance associated with the population Π_i, is discussed.

5.2.1 RULE PROPOSED BY M. GNANADESIKAN AND S. S. GUPTA (1970)

5.2.1.1 SELECTION OF A SUBSET CONTAINING THE POPULATION WITH THE SMALLEST $|\Sigma_i|$ BASED ON THE SAMPLE COVARIANCE MATRICES S_i, $i = 1, 2, \ldots., k$

Assume Σ_i and $\underset{\sim}{\mu}_i$ are unknown. The rule R is as follows:

Select Π_i if and only if $|S_i| \leqslant \frac{1}{b} \min(|S_1|, |S_2|, \ldots., |S_k|)$ where $0 < b = b(k,p,n,P^*) \leqslant 1$ is the largest value to satisfy the P^* condition $\inf\limits_{\Omega} P(CS/R) \geqslant P^*$. The $|S_i|$ is distributed as $|\Sigma_i| / (n-1)^p$ times the product of p independent chi squared factors with (n-1), (n-2), ..., (n-p) degrees of freedom.

Using this fact

$$\inf_{\Omega} P(CS/R) = P(Y_1 \leqslant \frac{1}{b} Y_i), \quad i = 2, 3, \ldots., k,$$

where Y_1, Y_2, ..., Y_k are independent identically distributed random variables each being the product of p independent factors where the r th factor is distributed as a chi squared with (n-r) degrees of freedom.

The constant b is the 100(1-P*) percentage point of

$$Y' = \min_{2 \leq i \leq k} \left\{\frac{Y_i}{Y_1}\right\}$$

The exact distribution of Y_i is not known, except when p = 2. In this case

$$\operatorname*{Inf}_{\Omega} P(CS/R) = P\left(Z_1 \leq \frac{1}{\sqrt{b}} Z_i\right), \quad i = 2, 3, \ldots., k$$

where Z_1, Z_2, ..., $Z_k = 2(n-1)^{\frac{1}{2}p}(|S_i| \,/\, |\Sigma_i|)^{\frac{1}{2}}$ are independent identically distributed chi squared random variables with 2(n-2) degrees of freedom.

Here, $\sqrt{b}$ is the 100(1-P*) percentage point of

$$F_{\min} = \min_{2 \leq i \leq k} \left\{\chi^2_{\nu,i}/\chi^2_{\nu,1}, \quad \nu = 2n - 4\right\}$$

b can be obtained from the tables of Gupta and Sobel (1962) and Krishnaiah and Armitage (1964).

5.2.1.2 APPROXIMATIONS TO THE DISTRIBUTION OF Y_i WHEN $p > 2$

Gnanadesikan and Gupta (1970) have considered the relative merits of different approximations to $Y_i = U_{i1} \cdot U_{i2} \cdot \ldots. \cdot U_{ip}$ where U_{id} is independent and has a chi squared distribution with (n-d) degrees of freedom.

CASE 1

This was suggested by Hoel (1937).

Approximating the distribution of $Y_i^{1/p}$ by the gamma distribution with density function

$$g(x) = \frac{\lambda^{\frac{1}{2}p(n-p)} \, x^{[\frac{1}{2}p(n-p)-1]} \, e^{-\lambda x}}{\Gamma(\frac{1}{2}p(n-p))}$$

$$\lambda = \frac{p}{2}\left(1 - \frac{(p-1)(p-2)}{2n}\right)^{1/p}$$

The approximation of $Y_i^{1/p}$ decreases in accuracy as p increases.

CASE 2

(log Y_i)/p using the normal approximation of log χ^2 as suggested by Bartlett and Kendall (1946).

1) Here the approximation of the distribution of log χ^2 by the normal distribution improves with the degrees of freedom of the chi squared variable.

2) The normal approximation to the distribution of the log (generalized variance) improves with both p and n. Approximating the distribution of (1/p) log Y_i by the normal distribution gives

$$\inf_{\Omega} P(CS/R) \simeq \int_{-\infty}^{\infty} \Phi^{k-1} (x-b_0) \, d\Phi(x)$$

where $\Phi(x)$ is the standard normal cumulative distribution function and is defined in Chapter 4.2.1.1, and

$$b_o = \log b \Big/ \left\{ \sum_{d=1}^{p} \mathrm{Var}(\log \chi^2_{n-d}) \right\}^{\frac{1}{2}}$$

where, for large n, $\mathrm{Var}(\log \chi^2_n) \simeq 2/(n-1)$.
The values of b_o have been tabulated by Gupta (1963) and Gupta, Nagel and Panchapakesan (1973) for various values of k and P*. Therefore, the value of b can be easily calculated.

5.2.2 ALTERNATIVES TO PROCEDURE R PROPOSED BY M. H. REGIER (1976)

5.2.2.1 PROCEDURE BASED ON THE GEOMETRIC MEAN

The procedure R_1 is as follows:

Select π_i if and only if $|S_i| \leqslant a\left(\prod_{i=1}^{k} |S_i|\right)^{1/k}$, where 'a' is determined subject to the P* condition $\inf_{\Omega} P(CS/R_1) \geqslant P^*$.

The approximate value of 'a' is based on the normal approximation to $\log \chi^2$, given by Bartlett and Kendall (1946). The probability condition is approximately satisfied if

$$\log a = Z_{P*} \left(\frac{k-1}{k}\right)^{\frac{1}{2}} \cdot \left(\sum_{d=1}^{p} \mathrm{Var}(\log \chi^2_{n-d})\right)^{\frac{1}{2}} - H$$

Here $\Phi(Z_{P*}) = P^*$, where $\Phi(x)$ denotes the standard normal distribution function and is defined in Chapter 4.2.1.1, and $H \geqslant 0$ is a known lower bound for

$$\frac{1}{k} \log \prod_{i=1}^{k} \frac{|\Sigma_i|}{|\Sigma|_{[1]}}$$

where $|\Sigma|_{[1]} \leqslant \ldots\ldots \leqslant |\Sigma|_{[k]}$. If no information is available on

$\sum_{i=1}^{k} \log\left(|\Sigma_i| \;/\; |\Sigma|_{[1]}\right)$ then H assumes its lowest possible value, namely zero.

Values for $\left(\sum_{d=1}^{p} \mathrm{Var}(\log \chi^2_{n-d})\right)^{\frac{1}{2}}$ are specified in Regier (1976).

5.2.2.2 PROCEDURE BASED ON THE ARITHMETIC MEAN

The procedure R_2 is as follows:

Select Π_i if and only if $|S_i| \leqslant b \sum_{i=1}^{k} |S_i| \;/\; k$ where $b > 0$ is determined subject to the P* condition $\inf_{\Omega} P(CS/R_2) \geqslant P^*$.

The asymptotic distribution of the sample variance is used for determining b. Clearly, b must be less than k; otherwise the selected subset would include all k populations.

For n sufficiently large, this condition is approximately satisfied by $b = k/(1 + B)$, where B is a solution of the equation

$$((n-1)/2p)^{\frac{1}{2}} (M-B) \;/\; ([M-(k-2)]^2 + (k-2) + B^2)^{\frac{1}{2}} = Z_{P*}$$

Here, $\Phi(Z_{P*}) = P^*$, where $\Phi(x)$ is the standard normal cumulative distribution function and is defined in Chapter 4.2.1.1, and $M \geqslant k-1$ is a lower bound for

$$\sum_{\substack{i=1\\ i\neq[1]}}^{k} \left(|\Sigma_i| \;/\; |\Sigma|_{[1]}\right).$$

If no information is available on $\sum_{\substack{i=1\\ i\neq[1]}}^{k} \left(|\Sigma_i| / |\Sigma|_{[1]}\right)$, then M assumes the lowest possible value k-1. In this case

$$B = [(k-1) - ((k-1)(k-X)X)^{\frac{1}{2}}]/(1-X)$$

where $X = (2pZ^2_{p*})/(n-1)$.

5.2.3 A COMPARISON OF THE THREE PROCEDURES R, R_1 AND R_2

1) All three procedures share the monotone property, i.e. $P(\Pi_i$ included in the selected subset) decreases as $|\Sigma_i| / |\Sigma|_{[1]}$ increases.

2) For all three procedures

$$E(\text{size of subset}) \leqslant k\ P(\Pi_{[1]} \text{ included})$$

3) In all three procedures an exact evaluation of the P(CS) depends on the knowledge of the ratios $|\Sigma_i| / |\Sigma|_{[1]}$, $i = 1, 2, \ldots., k$.

   However, R depends on the values of $|\Sigma_i| / |\Sigma|_{[1]}$

   R_1 depends on the $\prod_{i=1}^{k} |\Sigma_i| / |\Sigma|_{[1]}$

   R_2 depends on the $\sum_{i=1}^{k} |\Sigma_i| / |\Sigma|_{[1]}$.

4) The three procedures differ in the extent to which they require special tables for the evaluation of the constants needed. R and R_1 require special tables, whereas for R_2 no special tables are necessary. However, if the approximation $\text{Var}(\log \chi^2_n) \simeq 2/(n-1)$ is used (for $n \geqslant 10$), the need for special tables is eliminated for R_1.

5) When comparing the performance of the three procedures when applied to the same data, it can be seen that all three

procedures behave in a similar manner, however, R_2 consistently appears to be more conservative than the other two. When known lower bounds on H and M are used, both R_1 and R_2 result in a smaller expected subset size.

5.3 SELECTION IN TERMS OF THE MULTIPLE CORRELATION COEFFICIENTS

In some situations, it may be interesting to compare populations in terms of the association between a particular component and the rest. A measure of this association in the population Π_i is

$$\theta_i = \rho_{i:1,\,2,\,\ldots.,\,p}\ .$$

The squared multiple correlation coefficient between X_{i1} and $\{X_{i2}, \ldots., X_{ip}\}$ is defined in Chapter 4.3 by

$$\rho^2_{i:1,\,2,\,\ldots.,\,p} = 1 - \frac{|\Sigma_i|}{\sigma^{(i)}_{11}\ |\Sigma_{i(11)}|}\ .$$

The random vector $\underset{\sim}{X}_i$ has a multivariate normal distribution $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ where $\underset{\sim}{\mu}_i$ and Σ_i are unknown. Here $\sigma^{(i)}_{11}$ is the leading element of Σ_i and $\Sigma_{i(11)}$ is the matrix obtained from Σ_i by deleting the first row and the first column. The multiple correlation coefficient $\rho_{i:1,\,2,\,\ldots.,\,p}$ is the positive square root of $\rho^2_{i:1,\,2,\,\ldots.,\,p}$ and is the maximum of the correlation between X_{i1} and a linear combination of $X_{i2}, \ldots., X_{ip}$ over all possible linear combinations and as such, is a measure of the dependence of X_{i1} on $X_{i2}, \ldots., X_{ip}$.

Gupta and Panchapakesan (1969) investigated procedures for selecting a subset containing the population associated with $\theta_{[k]} = \rho_{[k]}$ or $\theta_{[1]} = \rho_{[1]}$.

Let $R_i = R_{i:1, 2, \ldots., p}$. The sample squared multiple correlation coefficient between X_{i1} and $X_{i2}, \ldots., X_{ip}$, is defined as

$$R_i^2 = 1 - \frac{|S_i|}{s_{11}^{(i)} \; |S_{i(11)}|}$$

where S_i is the sample covariance matrix, $s_{11}^{(i)}$ the leading element of S_i, and $S_{i(11)}$ the matrix obtained from S_i by deleting the first row and the first column. Two cases arise,

1) The case in which $X_{i2}, \ldots., X_{ip}$ are fixed, called the conditional case.
2) The case in which $X_{i2}, \ldots., X_{ip}$ are random, called the unconditional case.

5.3.1 PROCEDURES INVESTIGATED BY S. S. GUPTA AND S. PANCHAPAKESAN (1969)

5.3.1.1 SELECTION OF $\theta_{[k]} = \rho_{[k]}$

The procedure D_1 is as follows:

Select Π_i if and only if $R_1^{*2} \geq c \max(R_1^{*2}, R_2^{*2}, \ldots., R_k^{*2})$ where $R_i^{*2} = R_i^2 \,/\, (1 - R_i^2)$ and $0 < c \leq 1$ and is determined to satisfy the P* condition $\inf\limits_{\Omega} P(CS/D_1) \geq P^*$. The density of R_i^{*2} can be written as

$$u_{\lambda_i}(x) = \sum_{j=0}^{\infty} \frac{\Gamma(q+m+j)\, \lambda_i^j}{\Gamma(q+m)\; j!} (1 - \lambda_i)^{q+m} \; f_{2(q+j),2m}(x)$$

- unconditional case

$$u_{\lambda_i}(x) = \sum_{j=0}^{\infty} \frac{e^{-m\lambda_i}(m\lambda_i)^j}{j!} \; f_{2(q+j),2m}(x)$$

- conditional case

where $\lambda_i = \rho_i^2$, $q = (p-1)/2$, $m = (n-p)/2$ and $f_{r,s}(x)$ is the probability density function of a central F distribution with r and s degrees of freedom. This is given in Chapter 5.1.2.3. The distribution of R_i^{*2} is stochastically increasing in λ. Hence

$$\underset{\Omega}{\text{Inf}}\ P(CS/D_1) = \underset{\lambda \geq 0}{\text{Inf}} \int_0^\infty U_\lambda^{k-1}\ (x/c)\ u_\lambda(x)\ dx$$

where $U_\lambda(x)$ is the cumulative distribution function corresponding to $u_\lambda(x)$.

Gupta and Panchapakesan (1969) have shown that the integral is non decreasing in λ in both the unconditional and conditional cases and therefore the constant c is obtained in both cases from

$$\int_0^\infty F_{2q,2m}^{k-1}\ (x/c)\ f_{2q,2m}(x)\ dx = P^*$$

where $f_{2q,2m}(x)$ and $F_{2q,2m}(x)$ are the probability density function and the cumulative distribution function of a central F variable with (2q,2m) degrees of freedom. These functions are given in Chapter 5.1.2.3.

The values of c which are the same in both cases are given in Gupta and Panchapakesan (1969).

5.3.1.2 SELECTION OF $\theta_{[1]} = \rho_{[1]}$

The procedure D_2 is as follows:

Select Π_i if and only if $R_i^{*2} \leqslant \frac{1}{b} \min(R_1^{*2}, R_2^{*2}, \ldots., R_k^{*2})$ where $0 < b = b(k,n,p,P^*) \leqslant 1$ and is determined to satisfy the P^* condition $\underset{\Omega}{\text{Inf}}\ P(CS/D_2) \geqslant P^*$.

In an analogous manner to procedure D_1 it follows for both the conditional and unconditional cases

$$\operatorname*{Inf}_{\Omega} P(CS/D_2) = \int_0^{\infty} [1 - F_{2q,2m}(bx)]^{k-1} f_{2q,2m}(x)dx = P*$$

Since $1 - F_{2q,2m}(bx) = F_{2m,2q}(1/bx)$ the constants b can be obtained from constants c by interchanging q and m.

5.4 SELECTION IN TERMS OF MEASURES OF ASSOCIATION BETWEEN TWO SUB-CLASSES OF VARIATES

When comparing k, p variate normal distributions the p variates can be considered to be made up of two subsets of q_1 and q_2 $(q_1 + q_2 = p)$ variates. The populations can be selected according to a suitable measure of association between the two sets of variates in these populations. There are various possible measures considered, but in this case, the two measures considered by Gupta and Panchapakesan (1969) and Frischtak (1973) are presented.

Let $\underset{\sim}{X}_i = (\underset{\sim}{Y}_i, \underset{\sim}{Z}_i)'$ be a $(q_1 + q_2)$ dimensional random vector with covariance matrix

$$\Sigma_i = \begin{bmatrix} \Sigma_{yy}^{(i)} & \Sigma_{yz}^{(i)} \\ \Sigma_{zy}^{(i)} & \Sigma_{zz}^{(i)} \end{bmatrix}$$

where $\Sigma_{yy}^{(i)}$, $\Sigma_{yz}^{(i)}$, $\Sigma_{zy}^{(i)}$ and $\Sigma_{zz}^{(i)}$ are submatrices of dimension $q_1 \times q_1$, $q_1 \times q_2$, $q_2 \times q_1$ and $q_2 \times q_2$ respectively. Assume $q_1 \leq q_2$.

Let Π_i ($i = 1, 2, \ldots, k$) be a p variate normal distribution with mean vector $\underset{\sim}{\mu}_i$ and covariance matrix Σ_i. The p variates are partitioned into two sets of q_1 and q_2 components. Let the corresponding sample covariance matrix S_i based on n independent vector observations from Π_i be denoted as

$$S_i = \begin{bmatrix} S_{yy}^{(i)} & S_{yz}^{(i)} \\ S_{zy}^{(i)} & S_{zz}^{(i)} \end{bmatrix}$$

where $S_{yy}^{(i)}$, $S_{yz}^{(i)}$, $S_{zy}^{(i)}$ and $S_{zz}^{(i)}$ are submatrices of dimension $q_1 \times q_1$, $q_1 \times q_2$, $q_2 \times q_1$ and $q_2 \times q_2$ respectively.

Selection in terms of the Conditional Generalized Variance of $\underset{\sim}{Z}_i$, given $\underset{\sim}{Y}_i$, is defined by

$$\theta_i = |\Sigma_{z.y}^{(i)}| = |\Sigma_i| \;/\; |\Sigma_{yy}^{(i)}| = |\Sigma_{zz}^{(i)} - \Sigma_{zy}^{(i)} \Sigma_{yy}^{(i)-1} \Sigma_{yz}^{(i)}|$$

and has been considered by Gupta and Panchapakesan (1969).

Selection in terms of the Coefficient of Alienation between $\underset{\sim}{Y}_i$ and $\underset{\sim}{Z}_i$ is defined by θ_i, where

$$\theta_i^2 = \frac{|\Sigma_{z.y}^{(i)}|}{|\Sigma_{yy}^{(i)}|} = \frac{|\Sigma_i|}{|\Sigma_{yy}^{(i)}|\,|\Sigma_{zz}^{(i)}|}$$

and has been considered by Frischtak (1973).

5.4.1 PROCEDURE CONSIDERED BY S. S. GUPTA AND S. PANCHAPAKESAN (1969)

Let $\theta_i = |\Sigma_i| \,/\, |\Sigma_{yy}^{(i)}|$, the conditional generalized variance of $\underset{\sim}{Z}_i$ given $\underset{\sim}{Y}_i$. The procedure here selects the subset containing the population with the smallest θ_i. Define

$$V_i = |S_i| \,/\, |S_{yy}^{(i)}|$$

The procedure R_1 is as follows:

Select Π_i if and only if $V_i \leqslant \frac{1}{b} \min(V_1, V_2, \ldots., V_k)$ where $0 < b = b(k,P^*,n,q_1,q_2) \leqslant 1$ is determined to satisfy the P^* condition $\inf\limits_{\Omega} P(CS/R_1) \geqslant P^*$.

Gupta and Panchapakesan (1969) showed that

$$\inf_{\Omega} P(CS/R_1) = \int_0^{\infty} [1 - G(bx)]^{k-1} \, g(x) \, dx$$

where $g(x)$ and $G(x)$ are the probability density function and the cumulative distribution function of a random variable that is distributed as the product of q_2 independent chi squared variables with degrees of freedom $(n-q_1-1)$, $(n-q_1-2)$, $\ldots.$, $(n-q_1-q_2)$ respectively. The problem of evaluating b is similar to that encountered in Chapter 5.2.1 when evaluating b.

5.4.2 PROCEDURE CONSIDERED BY R. M. FRISCHTAK (1973)

Let $$\theta_i^2 = \frac{|\Sigma_i|}{|\Sigma_{yy}^{(i)}|\,|\Sigma_{zz}^{(i)}|}$$

where θ_i is the coefficient of alienation, between $\underset{\sim}{Y}_i$ and $\underset{\sim}{Z}_i$. The procedure here selects the subset containing the population with the smallest θ_i. Define

$$V_i^2 = \frac{|S_i|}{|S_{yy}^{(i)}|\,|S_{zz}^{(i)}|}$$

The procedure R_2 is as follows:

Select Π_i if and only if $V_i^2 \leqslant \frac{1}{b}\min(V_1^2, V_2^2, \ldots., V_k^2)$ where $0 < b = b(k,P^*,n,q_1,q_2) \leqslant 1$ is determined to satisfy the P^* condition $\inf\limits_{\Omega} P(CS/R_2) \geqslant P^*$.

Frischtak has obtained an asymptotic ($n \to \infty$) solution and the value of b is given by

$$P\left\{U_i \leqslant \frac{-\sqrt{n-1}\ \log b}{2(2q_1)^{\frac{1}{2}}}\ ,\ i = 1, 2, \ldots., k-1\right\} = P^*$$

where the U_i are standard normal variables with equal correlation coefficient ½.

Also note that for $q_1 = 1$, θ_i^2 is equal to $(1 - \rho^2)$ where ρ is the multiple correlation coefficient between y and $(z_1, z_2, \ldots., z_{q_2})$

CHAPTER 6

6. THE MULTIVARIATE SOLUTION TO THE MULTIVARIATE RANKING AND SELECTION PROBLEM

In this chapter the new formulation by Dudewicz and Taneja (1981) that selects the best multivariate population without reducing populations to univariate quantities is described. The solution developed for both the known and the unknown variance-covariance matrices are considered.

This multivariate solution to the multivariate ranking and selection problem allows for such occurrences as $\Pi_1 \succ \Pi_2 \succ \Pi_3 \succ \Pi_1$, where $\succ$ means "is preferred to". This would be an anomaly in previous chapters, however, it is expected in truly multivariate problems. They are problems in which one cannot associate a univariate measure of goodness or number $\theta_i = \phi(\underset{\sim}{\mu}_i, \Sigma_i)$ with a given population but must rather compare different $(\underset{\sim}{\mu}_i, \Sigma_i)$ pairs themselves, in order to determine which is preferred.

This method is also applicable to situations where each population Π_i has associated with it a numerical measure of goodness $u_i = u(\underset{\sim}{\mu}_i)$, $1 \leqslant i \leqslant k$ such that Π_t is preferred to Π_m if and only if $u_t > u_m$ $(1 \leqslant t, m \leqslant k)$.

In this situation, not only linear functions but also quadratic, polynomial, exponential and power series or even Fourier series functions are allowed.

In the new theory $g(\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots., \underset{\sim}{\mu}_k)$ is an experimenter specified function with range space $\{1, 2, \ldots., k\}$ such that $g(\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots., \underset{\sim}{\mu}_k) = t$ if and only if given a choice of $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots., \underset{\sim}{\mu}_k$, $\underset{\sim}{\mu}_t$ is preferred.

6.1 THE MULTIVARIATE PREFERENCE SELECTION PROCEDURE - R_{mvp}

A random sample is taken from each of the populations $\Pi_1, \Pi_2, \ldots., \Pi_k$. $\underset{\sim}{\mu}_i$ is estimated by the sample mean vector $\underset{\sim}{\overline{X}}_i (i = 1, \ldots., k)$ and $\Pi_{g(\underset{\sim}{\overline{X}}_1, \underset{\sim}{\overline{X}}_2, \ldots., \underset{\sim}{\overline{X}}_k)}$ selected.

Let $\underset{\approx}{\overline{X}} = (\underset{\sim}{\overline{X}}_1, \underset{\sim}{\overline{X}}_2, \ldots., \underset{\sim}{\overline{X}}_k)$. The following cases are considered in selecting Π_t, where $t = g(\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots., \underset{\sim}{\mu}_k)$.

a) $\Sigma_1 = \ldots. = \Sigma_k = \sigma^2 I$ with σ^2 known

b) $\Sigma_1 = \ldots. = \Sigma_k = \Sigma$ with Σ known

c) $\Sigma_1, \ldots., \Sigma_k$ are known but unequal

d) $\Sigma_1, \ldots., \Sigma_k$ are unknown and not necessarily equal

6.1.1 SELECTING THE BEST WHEN $\Sigma_1 = \ldots. = \Sigma_k = \sigma^2 I$ WITH σ^2 KNOWN

Let Π_i be $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ for $i = 1, 2, \ldots., k$. Assume $p \geqslant 1$ and $\Sigma_1 = \ldots. = \Sigma_k = \sigma^2 I$. $I = p \times p$ identity matrix and σ^2 is known. $g(\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k)$ is an experimenter specified function with range space $\{1, 2, \ldots., k\}$. $g(\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k) = t$ denotes that among $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots., \underset{\sim}{\mu}_k$, $\underset{\sim}{\mu}_t$ is preferred.

The Selection Procedure $R_{mvp}(\sigma^2 I)$ is as follows: Observe n independent observations from $\Pi_1, \ldots., \Pi_k$. Estimate $\underset{\sim}{\mu}_i$ by the sample mean vector $\underset{\sim}{\overline{X}}_i$ $(1 \leqslant i \leqslant k)$. Select $\Pi_{g(\underset{\sim}{\overline{X}}_1,\ \underset{\sim}{\overline{X}}_2,\ ....,\ \underset{\sim}{\overline{X}}_k)}$.

Choice of n

If the true means $\underset{\approx}{\mu} = (\underset{\sim}{\mu}_1, ..., \underset{\sim}{\mu}_k)$ are such that $g(\underset{\approx}{\mu}) = t$ while $g(\underset{\approx}{\mu} + \underset{\approx}{\varepsilon}) = m$ $(m \neq t)$ for a matrix $\underset{\approx}{\varepsilon}$ of small numbers then $P(CS/R_{mvp}(\sigma^2 I))$ will not be much larger than $1/k$.

Therefore a method is required on how to specify the sample size n per population so that for a reasonable preference zone Ω_p and for a fixed P^* $(1/k < P^* < 1)$ and $\delta^* > 0$ the procedure $R_{mvp}(\sigma^2 I)$ satisfies $\underset{\Omega_p}{\operatorname{Inf}} P(CS/R_{mvp}(\sigma^2 I)) \geqslant P^*$. Let

$$P_t = \{\underset{\approx}{\mu} : g(\underset{\approx}{\mu}) = t\}, \quad t = 1, 2, ..., k.$$

Note: $P_1, ..., P_k$ are disjoint preference sets whose union is R^{kp}.

The Euclidean distance

$$d(\underset{\approx}{a}, \underset{\approx}{b}) = \left(\sum_{h=1}^{kp} (a_h - b_h)^2\right)^{\frac{1}{2}}$$

defines the distance between any two points a and b of R^{kp}. The distance from $\underset{\approx}{\mu}$ to the boundary of $P_{g(\underset{\approx}{\mu})}$ is denoted by

$$d_B(\underset{\approx}{\mu}) = \operatorname{Inf}\{d(\underset{\approx}{\mu}, \underset{\approx}{b}) : \underset{\approx}{b} \in P_{g(\underset{\approx}{\mu})}\}.$$

The probability requirement for any procedure R is set as $P(CS/R) \geqslant P^*$, whenever $d_B(\underset{\approx}{\mu}) \geqslant \delta^*$.

Whenever $d_B(\underset{\approx}{\mu}) \geqslant \delta^*$ we have,

$$\begin{aligned}
P(CS/R_{mvp}(\sigma^2 I)) &= P(\underset{\approx}{\overline{X}} \in P_{g(\underset{\approx}{\mu})}) \\
&\geqslant P(d(\underset{\approx}{\mu}, \underset{\approx}{\overline{X}}) \leqslant d_B(\underset{\approx}{\mu})) \\
&\geqslant P(d(\underset{\approx}{\mu}, \underset{\approx}{\overline{X}}) \leqslant \delta^*) \\
&= P\left(\sum_{i=1}^{k} \sum_{c=1}^{p} (\overline{X}_{ic} - \mu_{ic})^2 \leqslant (\delta^*)^2\right) \\
&= P\left(\sum_{i=1}^{k} \sum_{c=1}^{p} \left(\frac{\overline{X}_{ic} - \mu_{ic}}{\sigma/\sqrt{n}}\right)^2 \leqslant n\left(\frac{\delta^*}{\sigma}\right)^2\right) \\
&= P(Y \leqslant n(\delta^*)^2/\sigma^2)
\end{aligned}$$

where Y has a central chi squared distribution with kp degrees of freedom. The probability density function and the cumulative distribution function of a central chi squared variable are defined in Chapter 5.1.2.1.

Therefore, the selection procedure $R_{mvp}(\sigma^2 I)$ satisfies the probability requirement if the sample size n per population satisfies

$$n \geqslant \chi^2_{kp}\ (P^*)\sigma^2/(\delta^*)^2,$$

where χ^2_{kp} (P*) is the value a central chi squared random variable with kp degrees of freedom fails to exceed with probability P*.

Choice of P*

The choice of P* is similar to the choice of power in tests of hypotheses. Normally, P* = 0.95 or a similar high value.

Choice of δ^*

For any two possible $\underset{\approx}{\mu}$'s, say $\underset{\approx}{a}$ and $\underset{\approx}{b}$, which satisfies

$$\text{Max}(\mu_{1c}, \mu_{2c}, \ldots., \mu_{kc}) - \text{Min}(\mu_{1c}, \mu_{2c}, \ldots., \mu_{kc}) \geq \Delta_c$$

$$d(\underset{\approx}{a}, \underset{\approx}{b}) \geq \Delta_c/2$$

where Δ_c is the minimum range between the largest of $\mu_{1c}, \mu_{2c}, \ldots., \mu_{kc}$ and the smallest of $\mu_{1c}, \mu_{2c}, \ldots., \mu_{kc}$ which the experimenter wishes to detect. Hence the choice of δ^* could be

$$\delta^* = \text{Min}(\Delta_1, \Delta_2, \ldots., \Delta_p)/2$$

6.1.2 SELECTING THE BEST WHEN $\Sigma_1 = \ldots. = \Sigma_k = \Sigma$, Σ KNOWN

Let Π_i be $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ for i = 1, 2, ..., k. Assume $p \geq 1$ and $\Sigma_1 = \Sigma_2 = \ldots. = \Sigma_k = \Sigma$. Σ is the common p×p variance-covariance matrix. It is assumed to be positive definite and known. Let $\lambda_1 \leq \lambda_2 \leq \ldots. \leq \lambda_p$ denote the characteristic roots of Σ. $g(\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k)$ is an experimenter specified function with range space {1, 2, ..., k}. $g(\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k) = t$ denotes that among $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots., \underset{\sim}{\mu}_k$, $\underset{\sim}{\mu}_t$ is preferred.

The selection procedure $R_{mvp}(\Sigma)$ is as follows: Observe n independent observations from $\Pi_1, \ldots., \Pi_k$. Estimate $\underset{\sim}{\mu}_i$ by the sample mean vector $\underset{\sim}{\overline{X}}_i$ $(1 \leq i \leq k)$. Select $\Pi_{g(\underset{\sim}{\overline{X}}_1, \underset{\sim}{\overline{X}}_2, \ldots., \underset{\sim}{\overline{X}}_k)}$.

Choice of n

To find the n the common sample size per population such that the probability requirement $P(CS/R) \geq P^*$ whenever $d_B(\underset{\approx}{\mu}) \geq \delta^*$, where $d_B(\underset{\approx}{\mu}) = \text{Inf}\{d(\underset{\approx}{\mu}, \underset{\approx}{b}) : \underset{\approx}{b} \in P_{g(\underset{\approx}{\mu})}\}$ the following procedure is used.

The result by Rao (1965) is used as explained below.

Let A be any symmetric p×p matrix. Let $\gamma_1 \leq \gamma_2 \leq \ldots. \leq \gamma_p$ be the characteristic roots of A. Let $\underset{\sim}{X}$ be any p×1 vector. Then

$$\gamma_1 \underset{\sim}{X}'\underset{\sim}{X} \leq \underset{\sim}{X}'A\underset{\sim}{X} \leq \gamma_p \underset{\sim}{X}'\underset{\sim}{X}$$

Choose $A = \sum^{-1}$ so that γ_1 is the smallest characteristic root of $\sum^{-1}$, i.e. $1/\lambda_p$. Then, whenever $d_B(\underset{\approx}{\mu}) \geq \delta^*$

$$\begin{aligned}
P(CS/R_{mvp}(\Sigma)) &= P(\underset{\approx}{\overline{X}} \in P_{g(\underset{\approx}{\mu})}) \\
&\geq P\left[\sum_{i=1}^{k} \sum_{c=1}^{p} (\overline{X}_{ic} - \mu_{ic})^2 \leq (\delta^*)^2\right] \\
&= P\left[\sum_{i=1}^{k} (\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i)'(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i) \leq (\delta^*)^2\right] \\
&= P\left[\sum_{i=1}^{k} n \frac{1}{\lambda_p} (\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i)'(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i) \leq \frac{n}{\lambda_p} (\delta^*)^2\right] \\
&\geq P\left[\sum_{i=1}^{k} n(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i)' \textstyle\sum^{-1} (\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i) \leq \frac{n}{\lambda_p} (\delta^*)^2\right] \\
&= P[Y \leq n(\delta^*)^2/\lambda_p]
\end{aligned}$$

where Y has a central chi-squared distribution with kp degrees of freedom. The probability density function and the cumulative distribution function of a central chi-squared variable are defined in Chapter 5.1.2.1.

Therefore the selection procedure $R_{mvp}(\Sigma)$ satisfies the probability requirement if the sample size n per population is such that,

$$n \geqslant \chi^2_{kp}\ (P^*)\lambda_p/(\delta^*)^2,$$

where $\chi^2_{kp}\ (P^*)$ is the value a central chi-squared random variable with kp degrees of freedom fails to exceed with probability P* and λ_p is the largest characteristic root of Σ.

6.1.3 SELECTING THE BEST WHEN $\Sigma_1, \ldots., \Sigma_k$ KNOWN

Let Π_i be $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ for i = 1, 2, ..., k. Assume $p \geqslant 1$ and $\Sigma_1, \ldots., \Sigma_k$ are known and they are p×p positive definite matrices. $g(\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k)$ is an experimenter specified function with range space {1, 2, ..., k}. $g(\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k) = t$ denotes among $\underset{\sim}{\mu}_1, \ldots., \underset{\sim}{\mu}_k$, $\underset{\sim}{\mu}_t$ is preferred.

The selection procedure $R_{mvp}(\Sigma_1, \ldots., \Sigma_k)$ is as follows: Observe n_i independent observations from Π_i $(1 \leqslant i \leqslant k)$. Estimate $\underset{\sim}{\mu}_i$ by the sample mean vector $\underset{\sim}{\overline{X}}_i$ $(1 \leqslant i \leqslant k)$. Select $\Pi_{g(\underset{\sim}{\overline{X}}_1,\, \underset{\sim}{\overline{X}}_2,\, \ldots.,\, \underset{\sim}{\overline{X}}_k)}$.

Choice of $n_1, \ldots., n_k$

To find the sample sizes $n_1, \ldots., n_k$ from the k populations such that the probability requirement $P(CS/R) \geqslant P^*$ whenever $d_B(\underset{\approx}{\mu}) \geqslant \delta^*$ is satisfied, where $d_B(\underset{\approx}{\mu}) = \text{Inf}\{d(\underset{\approx}{\mu}, \underset{\approx}{b}) : \underset{\approx}{b} \in P_{g(\underset{\approx}{\mu})}\}$ the following method is used.

Let λ_{ip} denote the largest characteristic root of Σ_i, i.e. $1/\lambda_{ip}$ is the smallest characteristic root of $\sum_i^{-1}$. Using the results developed previously, it then follows that whenever $d_B(\underset{\approx}{\mu}) \geqslant \delta^*$ we have (letting $n_{[1]} = \text{Min}(n_1, \ldots., n_k)$ and $\lambda_{[k]p} = \text{Max}(\lambda_{1p}, \ldots., \lambda_{kp})$)

$$P(CS/R_{mvp}(\Sigma_1, \ldots., \Sigma_k))$$

$$\geq P\left[\sum_{i=1}^{k} (\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i)'(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i) \leq (\delta^*)^2\right]$$

$$\geq P\left[\sum_{i=1}^{k} n_i \frac{1}{\lambda_{ip}} (\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i)'(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i) \leq \frac{n_{[1]}}{\lambda_{[k]p}} (\delta^*)^2\right]$$

$$\geq P\left[\sum_{i=1}^{k} n_i(\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i)' \Sigma_i^{-1} (\underset{\sim}{\overline{X}}_i - \underset{\sim}{\mu}_i) \leq \frac{n_{[1]}}{\lambda_{[k]p}} (\delta^*)^2\right]$$

$$= P\left[\sum_{i=1}^{k} Y_i \leq n_{[1]} (\delta^*)^2/\lambda_{[k]p}\right]$$

where $Y_1, \ldots., Y_k$ are independent central chi-squared random variables with p degrees of freedom. The probability density function and the cumulative distribution function of a central chi-squared variable are defined in Chapter 5.1.2.1.

Therefore, the selection procedure $R_{mvp}(\Sigma_1, \ldots., \Sigma_k)$ satisfies the probability requirement if the sample sizes $n_1, \ldots., n_k$ are such that

$$n_{[1]} \geq \chi^2_{kp} (P^*)\lambda_{[k]p}/(\delta^*)^2,$$

where χ^2_{kp} (P*) is the value a central chi-squared random variable with kp degrees of freedom fails to exceed with probability P*.

$$n_{[1]} = \text{Min}(n_1, \ldots., n_k)$$

$\lambda_{[k]p}$ = largest of the characteristic roots of $\Sigma_1, \ldots., \Sigma_k$.

In design problems, one would normally take

$$n_1 = n_2 = \dots. = n_k = \text{smallest integer} \geq \chi^2_{kp}\ (P^*)\lambda_{[k]p}/(\delta^*)^2.$$

However, if unequal sample sizes $n_1, n_2, \dots., n_k$ have already been taken, the smallest δ^* for which the probability requirement is satisfied can be calculated by

$$\delta^* \geq \left(\frac{\chi^2_{kp}\ (P^*)\lambda_{[k]p}}{n_{[1]}}\right)^{\frac{1}{2}}$$

6.1.4 SELECTING THE BEST WHEN $\Sigma_1, \dots., \Sigma_k$ UNKNOWN, UNEQUAL

Let Π_i be $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$ for $i = 1, 2, \dots., k$. Assume $p \geq 1$ and $\Sigma_1, \dots., \Sigma_k$ are unknown and they are $p \times p$ positive definite matrices. $g(\underset{\sim}{\mu}_1, \dots., \underset{\sim}{\mu}_k)$ is an experimenter specified function with range space $\{1, 2, \dots., k\}$. $g(\underset{\sim}{\mu}_1, \dots., \underset{\sim}{\mu}_k) = t$ denotes among $\underset{\sim}{\mu}_1, \dots., \underset{\sim}{\mu}_k$, $\underset{\sim}{\mu}_t$ is preferred.

No single stage procedure R for this problem can satisfy the probability requirement $P(CS/R) \geq P^*$ whenever $d_B(\underset{\approx}{\mu}) \geq \delta^*$. The Heteroscedastic Method by Dudewicz and Bishop (1979) is used to modify the procedure $R_{mvp}(\Sigma)$ of the case described in Chapter 6.1.2 into a procedure R_{HM} to solve this problem. The procedure R_{HM} is specified by a sampling rule and a terminal decision rule.

Sampling Rule for R_{HM}

Select $z > 0$ and an integer $n > p$ and a $p \times p$ positive definite matrix (α_{rs}). Take observations from populations Π_i $(i = 1, 2, \dots., k)$ as follows:

Take n initial observations $\underset{\sim}{X}_1^{(i)}, \dots., \underset{\sim}{X}_n^{(i)}$ where

$$\underset{\sim}{X}_j^{(i)} = \left(X_{1j}^{(i)}, X_{2j}^{(i)}, \ldots\ldots, X_{pj}^{(i)}\right)', \quad j = 1, 2, \ldots\ldots, n$$

Compute
$$\overline{X}_c^{(i)} = \frac{1}{n} \sum_{j=1}^{n} X_{cj}^{(i)}$$

$$s_{cd}^{(i)} = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_{cj}^{(i)} - \overline{X}_c^{(i)}\right)\left(X_{dj}^{(i)} - \overline{X}_d^{(i)}\right)$$

$$c,d = 1, 2, \ldots\ldots, p.$$

Define the positive integer N_i by

$$N_i = \mathrm{Max}\left\{n + p^2, \left[z^{-1} \sum_{c,d=1}^{p} \alpha_{cd}\, s_{cd}^{(i)}\right] + 1\right\}$$

where [q] denotes the largest integer less than q, and select $p(p\times N_i)$ matrices

$$A_{ir} = \begin{bmatrix} a_{ir_{11}} & \ldots\ldots & a_{ir_{1N_i}} \\ \vdots & & \\ a_{ir_{p1}} & & a_{ir_{pN_i}} \end{bmatrix} \quad r = 1,2,\ldots\ldots,p$$

in such a way that

a) $a_{ir_{c1}} = \ldots\ldots = a_{ir_{cn}}$

b) $A_{ir}\, \eta_i = \varepsilon_r$ where η_i is the $N_i\times 1$ vector $(1 \ldots\ldots 1)'$ and ε_r is the $p\times 1$ vector whose r th element is 1 and all the other elements zero.

c) $A_i A_i' = z(\alpha^{rs}) \otimes s_i^{cd}$, where $A_i' = (A_{i1}', A_{i2}', \ldots\ldots, A_{ip}')$. $\otimes$ denotes the direct product and (b^{cd}) denotes the inverse of the matrix (b_{cd}), r, c = 1, 2, ..., p.

Next, take N_i - n additional observations $\underset{\sim}{X}^{(i)}_{n+1}$, ..., $\underset{\sim}{X}^{(i)}_{N_i}$ and compute

$$\overset{\Delta}{X}^{(i)}_{r} = \sum_{c=1}^{p} \sum_{j=1}^{N_i} a_{ir_{cj}} \, X^{(i)}_{cj}, \quad r = 1, 2, \ldots., p.$$

For Π_i construct the p-dimensional vector $\overset{\Delta}{\underset{\sim}{X}}_i$

$$\hat{\underset{\sim}{X}}_i = \left(\overset{\Delta}{X}^{(i)}_1, \ldots., \overset{\Delta}{X}^{(i)}_p\right), \quad i = 1, 2, \ldots., k.$$

Terminal Decision Rule for R_{HM}

The same decision as $R_{mvp}(\Sigma)$ is taken when a sample size n per population was taken and had $\Sigma/n = z(\alpha^{rs})$ and observed

$$(\overline{\underset{\sim}{X}}_1, \overline{\underset{\sim}{X}}_2, \ldots., \overline{\underset{\sim}{X}}_k) = (\hat{\underset{\sim}{X}}_1, \hat{\underset{\sim}{X}}_2, \ldots., \hat{\underset{\sim}{X}}_k).$$

i.e. select $\Pi_{g(\hat{\underset{\sim}{X}}_1, \hat{\underset{\sim}{X}}_2, \ldots., \hat{\underset{\sim}{X}}_k)}$.

Selection procedure R_{HM} satisfies the probability requirement $P(CS/R) \geq P^*$ whenever $d_B(\underset{\approx}{\mu}) \geq \delta^*$. $z > 0$ is chosen so that

$$P\left[\sum_{i=1}^{k} (\hat{\underset{\sim}{X}}_i - \underset{\sim}{\mu}_i)'(\hat{\underset{\sim}{X}}_i - \underset{\sim}{\mu}_i) \leq (\delta^*)^2\right] = P^*$$

This is very complicated. Therefore, for large n, an approximate solution is given.

As $n \to \infty$ the $z > 0$ which solves the above, approaches a solution when $(\hat{\underset{\sim}{X}}_1, \hat{\underset{\sim}{X}}_2, \ldots., \hat{\underset{\sim}{X}}_k)$ is replaced by $(\underset{\sim}{Y}_1, \ldots., \underset{\sim}{Y}_k)$ where $\underset{\sim}{Y}_1, \ldots., \underset{\sim}{Y}_k$ are independent random variables and $\underset{\sim}{Y}_i = N_p(\underset{\sim}{\mu}_i, zp(\alpha^{rs}))$. The probability density function of a p variate normal population is defined in Chapter 3.1.

CHAPTER 7

7. THE COMPLETE RANKING OF MULTIVARIATE POPULATIONS

In this chapter an application pertaining to New Zealand's overseas trade is used to describe the Multivariate Ranking of populations according to a linear combination of their means. The properties of the Multivariate Normal Distribution and the theoretical aspects of the procedures related to the Ranking of Populations are discussed at first.

7.1 THE PROPERTIES OF THE MULTIVARIATE NORMAL DISTRIBUTION

When $X_1, X_2, \ldots., X_p$ follow a p variate normal distribution with mean

$$\overline{X}_c = \frac{\sum_{j=1}^{n} X_{cj}}{n}, \quad c = 1, 2, \ldots., p$$

and variance-covariance matrix

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & & & \\ s_{p1} & & & s_{pp} \end{bmatrix}$$

where

$$s_{cd} = \frac{\sum_{j=1}^{n} (X_{cj} - \overline{X}_c)(X_{dj} - \overline{X}_d)}{(n - 1)}$$

$$c = 1, \ldots., p, \quad d = 1, \ldots., p$$

the linear combination or weighted score,

$$L = b_1X_1 + b_2X_2 + \dots. + b_pX_p$$

where b_1, b_2, ..., b_p are real constants, has a univariate normal distribution with mean θ given by,

$$\theta = b_1\mu_1 + b_2\mu_2 + \dots. + b_p\mu_p$$

and variance Var(L) given by,

$$\mathrm{Var}(L) = \sum_{c=1}^{p} \sum_{d=1}^{p} b_c\, b_d\, \sigma_{cd}.$$

For k different multivariate populations the i th population has mean vector

$$\underset{\sim}{\mu}_i = (\mu_1^{(i)}, \mu_2^{(i)}, \dots., \mu_p^{(i)}),$$

variance-covariance matrix Σ_i with elements $\sigma_{cd}^{(i)}$ and the inverse $\sum_i^{-1}$ with elements σ_i^{cd} for i = 1, 2, ..., k. i.e. $\mu_c^{(i)}$ is the mean of the c th component of the i th population. $\sigma_{cd}^{(i)}$ is the covariance between the c th and the d th component for the i th population.

A random sample of n observations from each of the k multivariate populations will be nk, p tuples of measurements. Let $X_{cj}^{(i)}$ denote the j th measurement on the c th component of the i th population. Then the data consists of kpn measurements. i.e. n observations from k populations with p measurements in each observation.

7.1.1 PRESENTATION OF SAMPLE DATA CONSISTING OF n OBSERVATIONS FOR EACH OF THE k POPULATIONS HAVING p COMPONENTS

population i	1	2	k
component c of the p variables	1, 2, ..., p		1, 2, ..., p
observation j=1	$X_{11}^{(1)}, X_{21}^{(1)}, \ldots, X_{p1}^{(1)}$		$X_{11}^{(k)}, X_{21}^{(k)}, \ldots, X_{p1}^{(k)}$
j=2			
⋮			
j=n	$X_{1n}^{(1)}, \quad \ldots \quad , X_{pn}^{(1)}$		$X_{1n}^{(k)}, \quad \ldots \quad , X_{pn}^{(k)}$
sample mean of component c	$\overline{X}_1^{(1)}, \quad \ldots \quad , \overline{X}_p^{(1)}$		$\overline{X}_1^{(k)}, \quad \ldots \quad , \overline{X}_p^{(k)}$

For each of the i = 1, 2, ..., k populations an estimate of the mean θ_i is given by the linear combination $\overline{L}_i$, where

$$\overline{L}_i = b_1\overline{X}_1^{(i)} + b_2\overline{X}_2^{(i)} + \ldots. + b_p\overline{X}_p^{(i)} \quad .$$

These average scores for each population can then be ordered as

$$\overline{L}_{[1]} \leqslant \overline{L}_{[2]} \leqslant \ldots. \leqslant \overline{L}_{[k]}.$$

Since $\overline{X}_c^{(i)}$ for c = 1, 2, ..., p follow the multivariate normal distribution, $\overline{L}_i$ has a univariate normal distribution with mean θ_i and variance

$$\text{Var}(\overline{L}_i) = \frac{1}{n} \sum_{c=1}^{p} \sum_{d=1}^{p} b_c \, b_d \, \sigma_{cd}^{(i)}$$

$$= \text{Var}(L_i)/n$$

The $\sigma_{cd}^{(i)}$ in

$$\mathrm{Var}(\overline{L}_i) = \frac{1}{n} \sum_{c=1}^{p} \sum_{d=1}^{p} b_c \, b_d \, \sigma_{cd}^{(i)}$$

are replaced by $s_{cd}^{(i)}$ to calculate the estimate of $\mathrm{Var}(L_i)$. If the population variance-covariance matrices Σ_i, $i = 1, 2, \ldots., k$ are unknown but assumed to have a common value Σ then this common value is estimated by S, where $S = (S_1 + S_2 + \ldots. + S_k)/k$ and $\sigma_{cd}^{(i)}$ in the equation for $\mathrm{Var}(\overline{L}_i)$ is replaced by the corresponding entry in S for $i = 1, 2, \ldots., k$.

7.2 THE COMPLETE RANKING OF k POPULATION MEANS

Although the problem is mentioned in Bechhofer (1954), considerable progress in this field was made after Carroll and Gupta (1977) published a paper on the problem of completely ordering (ranking) k ($\geq$ 3) populations according to their means.

Two procedures that can be used to completely order k univariate population means when

a) the common variance is known

b) the common variance is unknown

are given below.

7.2.1 RANKING OF k NORMAL POPULATIONS ACCORDING TO THEIR MEANS WHEN THE COMMON VARIANCE IS KNOWN

The k normal populations with common known variance σ^2 are

$$N(\mu_1, \sigma^2), N(\mu_2, \sigma^2), \ldots., N(\mu_k, \sigma^2).$$

The ordered μ are denoted by

$$\mu_{[1]} \leqslant \mu_{[2]} \leqslant \ldots\ldots \leqslant \mu_{[k]}.$$

As a first step towards ranking the means a "distance" measure must be defined. The distance between each of the successive pairs of ordered μ values are

$$\begin{aligned} \delta_1 &= \mu_{[k]} - \mu_{[k-1]} \\ \delta_2 &= \mu_{[k-1]} - \mu_{[k-2]} \\ &\vdots \\ \delta_{k-1} &= \mu_{[2]} - \mu_{[1]} \end{aligned}$$

The problem is to construct a procedure for ranking the populations such that the probability of a completely correct ranking is at least some specified value P* whenever each of the distances δ_i is greater than a common threshold value δ^*. i.e.

$$\delta_1 \geqslant \delta^*,\ \delta_2 \geqslant \delta^*,\ \ldots\ldots,\ \delta_{k-1} \geqslant \delta^*.$$

The ranking procedure is then to take a sample of n observations from each of the k populations, compute the k sample means and order them as

$$\overline{X}_{[1]} \leqslant \overline{X}_{[2]} \leqslant \ldots\ldots \leqslant \overline{X}_{[k]}\ ,$$

where $\overline{X}_{[k]}$ is the largest sample mean.

When designing a fixed sample size experiment to rank the k populations δ^* and P* will have to be specified in order to determine the common sample size n. The table given (T 7.1) lists values

$\tau = \delta^* \sqrt{n}/\sigma$ that satisfies the probability requirement corresponding to a given k and P* and is taken from Gibbons, Olkin and Sobel (1977). Therefore,

$$n = \left(\frac{\tau\sigma}{\delta^*}\right)^2$$

7.2.2 TWO STAGE PROCEDURE TO RANK k NORMAL POPULATION MEANS WHEN THE COMMON VARIANCE IS UNKNOWN

In the first stage, a sample of n observations is taken from each of the k populations and the k sample variances $s_1^2, s_2^2, \ldots., s_k^2$ are calculated. The pooled sample variance is given by

$$s^2 = \frac{(n-1)s_1^2 + (n-1)s_2^2 + \ldots. + (n-1)s_k^2}{k(n-1)}$$

The degree of freedom for s^2 is $\nu = k(n-1)$.

In the second stage, a second sample of size N-n is taken from each of the k populations. The value of N is obtained from

$$N = \text{Max}\left(n, \left\{\frac{2s^2h^2}{\delta^{*2}}\right\}^+\right)$$

where $\{a\}^+$ means smallest integer equal to or greater than a. δ^* is the threshold value such that

$$\mu_{[k]} - \mu_{[k-1]} = \delta_1 \geq \delta^*$$

$$\mu_{[k-1]} - \mu_{[k-2]} = \delta_2 \geq \delta^*$$

$$\vdots$$

$$\mu_{[2]} - \mu_{[1]} = \delta_{k-1} \geq \delta^*$$

The probability of a completely correct ranking is to be at least P* whenever the above (k-1) inequalities between successive means hold jointly.

The value of h is obtained from the Table T 7.2 which is taken from Freeman, Kuzmack and Maurice (1967).

The next step is to compute the k sample means using the entire sample of N observations and order them as

$$\overline{X}_{[1]} \leqslant \overline{X}_{[2]} \leqslant \ldots . \leqslant \overline{X}_{[k]}.$$

$\overline{X}_{[k]}$ is the largest sample mean.

With this procedure the probability of a correct ranking is guaranteed to be at least P* whenever $\mu_{[i+1]} - \mu_{[i]} \geqslant \delta^*$, $i = 1, 2, \ldots .,$ k-1 regardless of the true value of the unknown σ^2.

7.3 RANKING OF COUNTRIES ON THEIR TRADE PERFORMANCE WITH NEW ZEALAND - AN APPLICATION

In the example considered here, an attempt is made to rank several countries on their importance to New Zealand's trade using as variates, the percentages of New Zealand's seven major products exported to the respective countries.

Let the k populations under consideration be k countries, trading in the same products in similar proportions. Let n be the number of years for which the data is obtainable (in this case 6, 1976 to 1981), and let the p variates be the percentages of major items of New Zealand produce exported to the countries under consideration. These items are:

1) Beef and Veal
2) Lamb
3) Mutton
4) Cheese
5) Butter
6) Condensed, Evaporated and Dried Milk
7) Wool.

Therefore, $p = 7$ and $n = 6$.

In the calculations, instead of the actual figures, percentages of the items are used, as it makes more sense from an economic point of view. For example, it is more meaningful to say that in 1980 Japan imported 2.6% of New Zealand's Beef and Veal exports and 4.7% of New Zealand's Processed Milk exports, than $13.75 million worth of New Zealand's Beef and Veal exports and $10.15 million worth of New Zealand's Processed Milk exports. Also, this conforms to the assumption that the observations on each variate should be independent and identically distributed.

The data for this example was obtained from the Department of Statistics publication Report and Analysis of External Trade (1979/80 and 1980/81).

At first, in order to obtain a set of countries that imports three or more of the same main exports, it was necessary to obtain various combinations of countries and products from the data. A necessary requirement was that each country used in the study possess the same set of variates.

A set of countries and products suitable for this exercise is listed below.

Countries

Australia
China-Taiwan
Hong Kong
Japan
Malaysia
Philippines
Singapore

Products

Beef and Veal
Butter
Processed Milk Products
(Condensed, Evaporated and
Dried Milk)

Then for this combination, it was necessary to calculate for each country the mean exports of each product and the covariance matrix. A requirement in the ranking procedure is that the countries possess similar covariance matrices.

To verify the equality of the covariance matrices of the above data, the following test given in Box (1949) was performed.

Test for Homogeneity of Covariance Matrices

$$H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_k$$

where k = number of populations

Σ_i = covariance matrix of the i th population.

Test Statistic

$$\nu \ln |S| - \sum_{i=1}^{k} \nu_i \ln |S_i|$$

which is asymptotically chi squared with $\frac{1}{2}p(p+1)(k-1)$ degrees of freedom.

p = number of variates

n_i = number of observations from the i th population

$\nu_i = n_i - 1$, the degrees of freedom

$$\nu = \sum_{i=1}^{k} \nu_i$$

S_i = maximum likelihood estimator of Σ_i

S = maximum likelihood estimator of the common covariance matrix Σ, defined in Chapter 7.1.

i.e. S_i = sample covariance matrix of the i th population

$$S = \frac{1}{k}\left(\sum_{i=1}^{k} \text{sample covariance matrix of the i th population}\right)$$

$|S_i|$ = determinant of S_i

$|S|$ = determinant of S.

The data corresponding to the selected countries and products, the means of the products, the covariance matrices and the results of the homogeneity test are in the computer printout C 7.3.1.

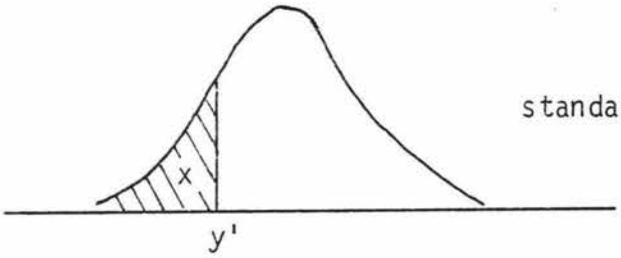
The value obtained for the chi squared test statistic is 130.4. This is a significant result. However, this is not unusual because the test in effect looks for any significant differences among the (in this case 9) sets of variances and covariances. In fact, it is very difficult not to get a significant result. Also, the test is asymptotic whereas here only 6 data values are considered.

To check the possibility of improving the results the data was transformed using,

a) Inverse Sine $Y_1 = \text{Sin}^{-1}\sqrt{x}$

b) Logit $Y_2 = \log_e \frac{x}{1-x}$

c) Probit $Y_3 = 5 + y'$ where $P(Z < y') = x$



Here x is a typical value of a proportion.

For the transformed data in each category the corresponding means, covariance matrices and the results of the homogeneity test are given in the computer printouts C 7.3.2 to C 7.3.4.

The values of the chi squared test statistic in the three categories are listed below.

a)	Inverse Sine	-	80.8
b)	Logit	-	132.3
c)	Probit	-	89.5

Although the chi squared values in all three categories are significant the chi squared values in categories (a) and (c) are better than the chi squared value obtained for the original data. The best result was obtained from the first transformation, Inverse Sine, and this was used in what follows.

Using the transformed data it was then necessary to calculate $\overline{L}_i$, an estimate of the mean θ_i for each country, i = 1, 2, ..., k.

$$\overline{L}_i = b_1\overline{X}_1 + b_2\overline{X}_2 + b_3\overline{X}_3$$

where $\overline{X}_1$ = the mean percentage of Beef and Veal imported by the country i over 6 years;

$\overline{X}_2$ = the mean percentage of Butter imported by country i over 6 years;

$\overline{X}_3$ = the mean percentage of Processed Milk Products imported by country i over 6 years.

The weights b_1, b_2 and b_3 were calculated as follows:

$$b_1 = \frac{b_v}{b_v + b_b + b_m}$$

$$b_2 = \frac{b_b}{b_v + b_b + b_m}$$

$$b_3 = \frac{b_m}{b_v + b_b + b_m}$$

where $b_v = \dfrac{\text{Value of Total Beef and Veal Exports}}{\text{Value of Total Exports}}$

$$b_b = \frac{\text{Value of Total Butter Exports}}{\text{Value of Total Exports}}$$

$$b_m = \frac{\text{Value of Total Milk Products Exports}}{\text{Value of Total Exports}}$$

Here the totals in each case were taken over the 6 years in question.

From the data of this exercise,

$$b_v = 0.190 \qquad b_b = 0.134 \qquad b_m = 0.075$$

Substituting the values of b_v, b_b and b_m

$$b_1 = 0.476 \qquad b_2 = 0.336 \qquad b_3 = 0.188$$

Substituting the values of b_1, b_2 and b_3 in the equation

$$\overline{L}_i = b_1\overline{X}_1^{(i)} + b_2\overline{X}_2^{(i)} + b_3\overline{X}_3^{(i)}, \quad (i = 1, 2, \ldots., 7)$$

the value $\overline{L}_i$ for each country was calculated.

Australia	$\bar{L}_1 = 0.0618$
China-Taiwan	$\bar{L}_2 = 0.0884$
Hong Kong	$\bar{L}_3 = 0.1072$
Japan	$\bar{L}_4 = 0.1635$
Malaysia	$\bar{L}_5 = 0.1464$
Philippines	$\bar{L}_6 = 0.1484$
Singapore	$\bar{L}_7 = 0.1429$

The theory of the two stage ranking procedure discussed earlier is used in this example as the common variance is unknown and the sample size is small. However the theory is somewhat modified in this case. Here, the value of N is fixed, (in this case N = 6). Therefore when using the theory it is necessary to work backwards and calculate the value of δ^* for a predetermined h and P*,

where $$\delta^* = \left(\frac{2s^2h^2}{N}\right)^{\frac{1}{2}}$$

and $$s^2 = \frac{1}{6}\sum_{c=1}^{3}\sum_{d=1}^{3} b_c\, b_d\, \sigma_{cd}.$$

The average covariance matrix

$$S = \begin{bmatrix} 1.943\text{E-}4 & 1.069\text{E-}4 & 8.565\text{E-}5 \\ 1.069\text{E-}4 & 1.210\text{E-}3 & 6.309\text{E-}4 \\ 8.565\text{E-}5 & 6.309\text{E-}4 & 1.104\text{E-}3 \end{bmatrix}$$

is an estimate of each element corresponding to σ_{cd}. Hence

$$s^2 = 5.7668\text{E-}05$$

Using the table given in the theory for P* = 0.95, the value of h can be estimated by extrapolation.

Here $h \simeq 2.4$ for $N = 6$ and $k = 7$. Then

$$\delta^* = \left(\frac{2 \times 5.7668E\text{-}5 \times 2.4^2}{6}\right)^{\frac{1}{2}}$$

$$= .0105$$

$$\simeq .01$$

Therefore probability of a completely correct ranking will be at least .95 (= P^*) whenever the difference between successive means is at least .01 (= δ^*).

The seven means for the countries can be ordered from largest to smallest as follows:

1	Japan	$\overline{L}_4$	.1635
2	Philippines	$\overline{L}_6$	.1484
3	Malaysia	$\overline{L}_5$	.1464
4	Singapore	$\overline{L}_7$	.1429
5	Hong Kong	$\overline{L}_3$	.1072
6	China-Taiwan	$\overline{L}_2$	.0884
7	Australia	$\overline{L}_1$	.0618

7.4 COMMENTS ON THE APPLICATION

Real life examples that have been solved using Multivariate Ranking Procedures are practically non-existent. Therefore, the importance of the problem discussed in this chapter is stressed. However, the procedure used has several drawbacks.

1) A necessary requirement of the method is that the populations should have the same set of variates. i.e. in this example only the countries that import the same New Zealand major

products can be used. This is not very practical in real life situations where it may be necessary to compare the trade performance of a country that imports products A, B and C with a country that imports products B, D and E. Therefore, it would be ideal if the variates could be weighted in some way that all countries are included in the analysis, whether they trade in certain products or not. At present there is no such solution for a problem of this nature.

2) In this procedure there is no obvious way in which the coefficients b_i are chosen. They are picked on the judgement of the experimenter.

3) Since it is a requirement of this method that the Σ_i should be equal, it makes the problem less meaningful as some of the countries that could have been used had to be left out.

4) The value of δ^* calculated here is doubtful as the tables available to calculate this are not accurate.

TABLE T 7.1 The probability of a correct complete ordering of k normal populations with respect to means for given values of τ

					k				
τ	2	3	4	5	6	7	8	9	10
0.00	.500	.167	.041	.008	.001	.000	.000	.000	.000
0.10	.528	.196	.056	.014	.003	.001	.000	.000	.000
0.20	.556	.228	.077	.023	.006	.002	.000	.000	.000
0.30	.584	.263	.101	.036	.012	.004	.001	.000	.000
0.40	.611	.299	.130	.052	.020	.008	.003	.001	.000
0.50	.638	.337	.162	.074	.033	.014	.006	.003	.001
0.60	.664	.376	.192	.100	.050	.025	.012	.006	.003
0.70	.690	.416	.237	.132	.073	.040	.022	.012	.006
0.80	.714	.456	.279	.168	.101	.060	.036	.021	.013
0.90	.738	.496	.324	.208	.134	.086	.055	.035	.022
1.00	.760	.536	.369	.252	.172	.117	.080	.054	.037
1.10	.782	.574	.415	.298	.214	.154	.110	.079	.057
1.20	.802	.612	.461	.346	.260	.195	.146	.110	.082
1.30	.821	.647	.506	.395	.308	.240	.187	.146	.114
1.40	.839	.681	.550	.444	.358	.288	.232	.187	.151
1.50	.855	.714	.593	.492	.408	.338	.281	.233	.193
1.60	.871	.744	.633	.539	.458	.390	.332	.282	.240
1.70	.885	.772	.671	.584	.507	.441	.384	.334	.290
1.80	.898	.797	.707	.626	.555	.492	.436	.386	.342
1.90	.910	.821	.740	.667	.601	.541	.488	.439	.396
2.00	.921	.843	.770	.704	.644	.589	.538	.492	.450
2.10	.931	.862	.798	.739	.684	.633	.586	.543	.503
2.20	.940	.880	.824	.771	.722	.676	.632	.592	.554
2.30	.948	.856	.847	.800	.756	.715	.675	.638	.603
2.40	.955	.910	.867	.826	.788	.750	.715	.681	.649
2.50	.961	.923	.886	.850	.816	.783	.752	.721	.692
2.60	.967	.934	.902	.871	.841	.812	.785	.758	.732
2.70	.972	.944	.916	.890	.864	.839	.815	.791	.768
2.80	.976	.952	.929	.906	.884	.863	.842	.821	.801
2.90	.978	.960	.940	.921	.902	.883	.865	.847	.830
3.00	.983	.966	.950	.933	.917	.901	.886	.871	.856

TABLE T 7.1 (continued)

					k				
τ	2	3	4	5	6	7	8	9	10
3.10	.986	.972	.958	.944	.931	.917	.904	.891	.878
3.20	.988	.977	.965	.953	.942	.931	.920	.909	.898
3.30	.990	.981	.971	.961	.952	.942	.933	.924	.915
3.40	.992	.984	.976	.968	.960	.953	.945	.937	.930
3.50	.993	.987	.980	.974	.968	.961	.955	.948	.942
3.60	.995	.989	.984	.979	.973	.968	.963	.958	.953
3.70	.996	.991	.987	.983	.978	.974	.970	.966	.962
3.80	.996	.993	.990	.986	.983	.979	.976	.972	.969
3.90	.997	.994	.991	.989	.986	.983	.981	.978	.975
4.00	.998	.996	.993	.991	.989	.987	.985	.982	.980
4.10	.998	.996	.995	.993	.991	.990	.988	.986	.984
4.20	.999	.997	.996	.994	.993	.992	.990	.989	.988
4.30	.999	.998	.997	.996	.995	.994	.992	.991	.990
4.40	.999	.998	.997	.997	.996	.995	.994	.993	.992

TABLE T 7.2 Values of h to determine the common sample size in the second stage for attaining P* = .95 for a correct complete ordering of k normal populations with respect to means when n is the common size of the first stage sample.

	k			
n	3	4	5	6†
10	2.053	2.21	2.29	2.28
20	2.002	2.16	2.25	2.24
30	1.988	2.15	2.24	2.23
40	1.981	2.14	2.24	2.22
50	1.977	2.14	2.23	2.22
60	1.975	2.14	2.23	2.22
70	1.973	2.14	2.23	2.22
80	1.971	2.13	2.23	2.21
90	1.970	2.13	2.23	2.21
100	1.969	2.13	2.23	2.21
200	1.965	2.13	2.22	2.21
500	1.963	2.12	2.22	-

† The last digit of each entry for k = 6 is of questionable accuracy because we expect all the entries to increase as k increases for fixed n.

```
C 7.3.1  RESULTS USING DIRECT DATA
         =========================


NAME OF THE COUNTRY : AUSTRALIA
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    0.200      0.000     0.100
1977    0.600      0.000     0.300
1978    0.600      0.000     0.500
1979    0.400      0.000     0.600
1980    0.900      0.700     1.700
1981    1.200      0.700     1.400
        =====      =====     =====
MEAN    0.650      0.233     0.767


   COVARIANCE MATRIX - S(1)
   ========================

          BF&VL        BUTTER       MILK
BF&VL     1.270E-1     1.120E-1     1.920E-1
BUTTER    1.120E-1     1.307E-1     2.193E-1
MILK      1.920E-1     2.193E-1     4.067E-1

DETERMINANT OF COVARIANCE MATRIX S(1) =  1.539E-04


NAME OF THE COUNTRY : CHINA - TAIWAN
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    0.600      0.100     5.100
1977    0.200      0.200     6.200
1978    0.100      0.400     7.300
1979    0.200      0.500     6.700
1980    0.300      0.300     3.300
1981    0.400      0.600     4.200
        =====      =====     =====
MEAN    0.300      0.350     5.467


   COVARIANCE MATRIX - S(2)
   ========================

          BF&VL        BUTTER       MILK
BF&VL      3.200E-2   -1.200E-2   -1.600E-1
BUTTER    -1.200E-2    3.500E-2    1.000E-2
MILK      -1.600E-1    1.000E-2    2.371E+0

DETERMINANT OF COVARIANCE MATRIX S(2) =  1.453E-03
```

```
NAME OF THE COUNTRY : HONG KONG
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    1.700      0.400     0.700
1977    2.000      0.600     0.600
1978    1.700      0.600     0.800
1979    1.400      0.600     0.600
1980    1.800      0.800     0.600
1981    2.100      1.400     0.600
        =====      =====     =====
MEAN    1.783      0.733     0.650


   COVARIANCE MATRIX - S(3)
   ========================

             BF&VL         BUTTER        MILK
BF&VL       6.167E-2     5.467E-2    -5.000E-3
BUTTER      5.467E-2     1.227E-1    -1.200E-2
MILK       -5.000E-3    -1.200E-2     7.000E-3

DETERMINANT OF COVARIANCE MATRIX S(3) =  2.665E-05



NAME OF THE COUNTRY : JAPAN
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    3.000      2.700    10.900
1977    2.700      2.900     6.800
1978    4.100      0.200     5.400
1979    2.800      1.100     6.700
1980    2.600      0.200     4.700
1981    3.400      0.100     4.700
        =====      =====     =====
MEAN    3.100      1.200     6.533


   COVARIANCE MATRIX - S(4)
   ========================

             BF&VL         BUTTER        MILK
BF&VL       3.200E-1    -3.260E-1    -2.720E-1
BUTTER     -3.260E-1     1.672E+0     2.394E+0
MILK       -2.720E-1     2.394E+0     5.435E+0

DETERMINANT OF COVARIANCE MATRIX S(4) =  7.971E-01
```

```
NAME OF THE COUNTRY : MALAYSIA
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    0.400      2.500    18.800
1977    0.300      2.400    13.200
1978    0.300      1.400    18.200
1979    0.200      0.900    14.100
1980    0.300      1.700    10.800
1981    0.300      2.800    13.600
        =====      =====     =====
MEAN    0.300      1.950    14.783

   COVARIANCE MATRIX - S(5)
   ========================

          BF&VL         BUTTER        MILK
BF&VL     4.000E-3      3.200E-2      9.400E-2
BUTTER    3.200E-2      5.390E-1      6.500E-2
MILK      9.400E-2      6.500E-2      9.610E+0

DETERMINANT OF COVARIANCE MATRIX S(5) =  6.490E-03


NAME OF THE COUNTRY : PHILIPPINES
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    0.500      1.200    12.900
1977    0.700      2.000    12.700
1978    0.400      1.700    14.500
1979    0.400      1.000    12.700
1980    0.500      2.000    16.500
1981    0.400      2.700    13.200
        =====      =====     =====
MEAN    0.483      1.767    13.750

   COVARIANCE MATRIX - S(6)
   ========================

          BF&VL         BUTTER        MILK
BF&VL     1.367E-2      7.333E-3     -2.500E-2
BUTTER    7.333E-3      3.787E-1      2.240E-1
MILK     -2.500E-2      2.240E-1      2.271E+0

DETERMINANT OF COVARIANCE MATRIX S(6) =  1.063E-02
```

```
NAME OF THE COUNTRY : SINGAPORE
===================

YEAR   BF&VL    BUTTER   MILK
====   =====    ======   ====

1976   1.900    1.500    6.100
1977   1.500    1.400    4.600
1978   1.600    0.900    4.400
1979   1.300    0.800    2.800
1980   1.900    1.700    4.900
1981   2.100    2.500    5.200
       =====    =====    =====
MEAN   1.717    1.467    4.667

   COVARIANCE MATRIX - S(7)
   ========================

           BF&VL       BUTTER      MILK
BF&VL      8.967E-2    1.607E-1    2.667E-1
BUTTER     1.607E-1    3.787E-1    4.107E-1
MILK       2.667E-1    4.107E-1    1.191E+0

DETERMINANT OF COVARIANCE MATRIX S(7) =  2.832E-03


      S = (S1+S2+S3+S4+S5+S6+S7) / 7

AVERAGE COVARIANCE MATIX - (S)
==============================

           BF&VL       BUTTER      MILK
BF&VL      9.257E-2    4.095E-3    1.295E-2
BUTTER     4.095E-3    4.652E-1    4.730E-1
MILK       1.295E-2    4.730E-1    3.041E+0

DETERMINANT OF AVERAGE COVARIANCE MATRIX =  1.102E-01

  35 Ln DET (S)  =  -77.2

   5 Ln DET (S1) =  -43.9
   5 Ln DET (S2) =  -32.7
   5 Ln DET (S3) =  -52.7
   5 Ln DET (S4) =   -1.1
   5 Ln DET (S5) =  -25.2
   5 Ln DET (S6) =  -22.7
   5 Ln DET (S7) =  -29.3
```

$$35 \text{ Ln DET } (S) - \sum_{i=1}^{7} 5 \text{ Ln DET } (S_i) = 130.4$$

```
C 7.3.2  RESULTS WHEN DATA WAS TRANSFORMED USING INVERSE
         ===============================================
         SINE CONVERSION
         ===============
```

```
NAME OF THE COUNTRY : AUSTRALIA
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976    0.045     0.000     0.032
1977    0.078     0.000     0.055
1978    0.078     0.000     0.071
1979    0.063     0.000     0.078
1980    0.095     0.084     0.131
1981    0.110     0.084     0.119
        =====     =====     =====
MEAN    0.078     0.028     0.081

   COVARIANCE MATRIX - S(1)
   ========================

           BF&VL         BUTTER        MILK
BF&VL      5.243E-4      8.178E-4      7.502E-4
BUTTER     8.178E-4      1.871E-3      1.474E-3
MILK       7.502E-4      1.474E-3      1.426E-3

DETERMINANT OF COVARIANCE MATRIX S(1) =  6.163E-11
```

```
NAME OF THE COUNTRY : CHINA - TAIWAN
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976    0.078     0.032     0.228
1977    0.045     0.045     0.252
1978    0.032     0.063     0.274
1979    0.045     0.071     0.262
1980    0.055     0.055     0.183
1981    0.063     0.078     0.206
        =====     =====     =====
MEAN    0.053     0.057     0.234

   COVARIANCE MATRIX - S(2)
   ========================

           BF&VL         BUTTER        MILK
BF&VL      2.608E-4     -1.124E-4     -3.501E-4
BUTTER    -1.124E-4      2.900E-4      2.383E-5
MILK      -3.501E-4      2.383E-5      1.217E-3

DETERMINANT OF COVARIANCE MATRIX S(2) =  4.290E-11
```

```
NAME OF THE COUNTRY : HONG KONG
===================


YEAR    BF&VL      BUTTER     MILK
====    =====      ======     ====

1976    0.131      0.063      0.084
1977    0.142      0.078      0.078
1978    0.131      0.078      0.090
1979    0.119      0.078      0.078
1980    0.135      0.090      0.078
1981    0.145      0.119      0.078
        =====      =====      =====
MEAN    0.134      0.084      0.081


   COVARIANCE MATRIX - S(3)
   ========================

             BF&VL         BUTTER        MILK
BF&VL      9.016E-5     1.070E-4    -1.063E-5
BUTTER     1.070E-4     3.565E-4    -4.137E-5
MILK      -1.063E-5    -4.137E-5     2.557E-5

DETERMINANT OF COVARIANCE MATRIX S(3) =  4.284E-13


NAME OF THE COUNTRY : JAPAN
===================


YEAR    BF&VL      BUTTER     MILK
====    =====      ======     ====

1976    0.174      0.165      0.336
1977    0.165      0.171      0.264
1978    0.204      0.045      0.235
1979    0.168      0.105      0.262
1980    0.162      0.045      0.219
1981    0.185      0.032      0.219
        =====      =====      =====
MEAN    0.176      0.094      0.256


   COVARIANCE MATRIX - S(4)
   ========================

             BF&VL         BUTTER        MILK
BF&VL      2.498E-4    -4.676E-4    -1.423E-4
BUTTER    -4.676E-4     3.973E-3     2.325E-3
MILK      -1.423E-4     2.325E-3     1.968E-3

DETERMINANT OF COVARIANCE MATRIX S(4) =  4.010E-10
```

```
NAME OF THE COUNTRY : MALAYSIA
==================


YEAR    BF&VL      BUTTER     MILK
====    =====      ======     ====

1976    0.063      0.159      0.448
1977    0.055      0.156      0.372
1978    0.055      0.119      0.441
1979    0.045      0.095      0.385
1980    0.055      0.131      0.335
1981    0.055      0.168      0.378
        =====      =====      =====
MEAN    0.055      0.138      0.393


   COVARIANCE MATRIX - S(5)
   ========================

             BF&VL          BUTTER         MILK
BF&VL        3.458E-5       1.217E-4       1.104E-4
BUTTER       1.217E-4       7.847E-4       3.239E-5
MILK         1.104E-4       3.239E-5       1.897E-3

DETERMINANT OF COVARIANCE MATRIX S(5) =  1.463E-11



NAME OF THE COUNTRY : PHILIPPINES
==================


YEAR    BF&VL      BUTTER     MILK
====    =====      ======     ====

1976    0.071      0.110      0.367
1977    0.084      0.142      0.364
1978    0.063      0.131      0.391
1979    0.063      0.100      0.364
1980    0.071      0.142      0.418
1981    0.063      0.165      0.372
        =====      =====      =====
MEAN    0.069      0.132      0.379


   COVARIANCE MATRIX - S(6)
   ========================

             BF&VL          BUTTER         MILK
BF&VL        6.438E-5       2.496E-5      -2.190E-5
BUTTER       2.496E-5       5.595E-4       1.435E-4
MILK        -2.190E-5       1.435E-4       4.585E-4

DETERMINANT OF COVARIANCE MATRIX S(6) =  1.448E-11
```

```
NAME OF THE COUNTRY : SINGAPORE
==================

YEAR   BF&VL     BUTTER   MILK
====   =====     ======   ====

1976   0.138     0.123    0.250
1977   0.123     0.119    0.216
1978   0.127     0.095    0.211
1979   0.114     0.090    0.168
1980   0.138     0.131    0.223
1981   0.145     0.159    0.230
       =====     =====    =====
MEAN   0.131     0.119    0.216

   COVARIANCE MATRIX - S(7)
   ========================

            BF&VL        BUTTER       MILK
BF&VL       1.358E-4     2.566E-4     2.638E-4
BUTTER      2.566E-4     6.354E-4     4.583E-4
MILK        2.638E-4     4.583E-4     7.378E-4

DETERMINANT OF COVARIANCE MATRIX S(7) =  4.391E-12


      S = (S1+S2+S3+S4+S5+S6+S7) / 7

AVERAGE COVARIANCE MATIX - (S)
==============================

            BF&VL        BUTTER       MILK
BF&VL       1.943E-4     1.069E-4     8.565E-5
BUTTER      1.069E-4     1.210E-3     6.309E-4
MILK        8.565E-5     6.309E-4     1.104E-3

DETERMINANT OF AVERAGE COVARIANCE MATRIX =  1.723E-10

  35 Ln DET (S)   =  -786.9

   5 Ln DET (S1) =  -117.5
   5 Ln DET (S2) =  -119.3
   5 Ln DET (S3) =  -142.4
   5 Ln DET (S4) =  -108.2
   5 Ln DET (S5) =  -124.7
   5 Ln DET (S6) =  -124.8
   5 Ln DET (S7) =  -130.8

                     7
  35 Ln DET (S) -    ∑   5 Ln DET (Si) =  80.8
                    i=1
```

C 7.3.3 RESULTS WHEN DATA WAS TRANSFORMED USING LOGIT CONVERSION

```
NAME OF THE COUNTRY : AUSTRALIA
===================


YEAR   BF&VL     BUTTER    MILK
====   =====     ======    ====

1976  -6.213     0.000    -6.907
1977  -5.110     0.000    -5.806
1978  -5.110     0.000    -5.293
1979  -5.517     0.000    -5.110
1980  -4.701    -4.955    -4.057
1981  -4.411    -4.955    -4.255
       =====     =====     =====
MEAN  -5.177    -1.652    -5.238

   COVARIANCE MATRIX - S(1)
   ========================

             BF&VL        BUTTER        MILK
BF&VL      4.021E-1    -1.231E+0     5.915E-1
BUTTER    -1.231E+0     6.547E+0    -2.145E+0
MILK       5.915E-1    -2.145E+0     1.098E+0

DETERMINANT OF COVARIANCE MATRIX S(1) =  2.093E-01


NAME OF THE COUNTRY : CHINA - TAIWAN
===================


YEAR   BF&VL     BUTTER    MILK
====   =====     ======    ====

1976  -5.110    -6.907    -2.924
1977  -6.213    -6.213    -2.717
1978  -6.907    -5.517    -2.541
1979  -6.213    -5.293    -2.634
1980  -5.806    -5.806    -3.378
1981  -5.517    -5.110    -3.127
       =====     =====     =====
MEAN  -5.961    -5.808    -2.887

   COVARIANCE MATRIX - S(2)
   ========================

             BF&VL        BUTTER        MILK
BF&VL      3.932E-1    -1.855E-1    -1.294E-1
BUTTER    -1.855E-1     4.415E-1     6.689E-3
MILK      -1.294E-1     6.689E-3     1.025E-1

DETERMINANT OF COVARIANCE MATRIX S(2) =  7.173E-03
```

```
NAME OF THE COUNTRY : HONG KONG
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976   -4.057    -5.517    -4.955
1977   -3.892    -5.110    -5.110
1978   -4.057    -5.110    -4.820
1979   -4.255    -5.110    -5.110
1980   -3.999    -4.820    -5.110
1981   -3.842    -4.255    -5.110
        =====     =====     =====
MEAN   -4.017    -4.987    -5.036


   COVARIANCE MATRIX - S(3)
   ========================

              BF&VL         BUTTER        MILK
BF&VL        2.126E-2      3.427E-2     -3.588E-3
BUTTER       3.427E-2      1.782E-1     -2.358E-2
MILK        -3.588E-3     -2.358E-2      1.500E-2

DETERMINANT OF COVARIANCE MATRIX S(3) =  3.091E-05



NAME OF THE COUNTRY : JAPAN
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976   -3.476    -3.585    -2.101
1977   -3.585    -3.511    -2.618
1978   -3.152    -6.213    -2.863
1979   -3.547    -4.499    -2.634
1980   -3.623    -6.213    -3.009
1981   -3.347    -6.907    -3.009
        =====     =====     =====
MEAN   -3.455    -5.154    -2.706


   COVARIANCE MATRIX - S(4)
   ========================

              BF&VL         BUTTER        MILK
BF&VL        3.147E-2     -1.276E-1     -1.204E-2
BUTTER      -1.276E-1      2.181E+0      4.323E-1
MILK        -1.204E-2      4.323E-1      1.176E-1

DETERMINANT OF COVARIANCE MATRIX S(4) =  1.287E-03
```

```
NAME OF THE COUNTRY : MALAYSIA
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976  -5.517    -3.664    -1.463
1977  -5.806    -3.705    -1.883
1978  -5.806    -4.255    -1.503
1979  -6.213    -4.701    -1.807
1980  -5.806    -4.057    -2.111
1981  -5.806    -3.547    -1.849
       =====     =====     =====
MEAN  -5.826    -3.988    -1.769

   COVARIANCE MATRIX - S(5)
   ========================

             BF&VL          BUTTER        MILK
BF&VL        4.925E-2      7.673E-2      2.074E-2
BUTTER       7.673E-2      1.929E-1      2.325E-3
MILK         2.074E-2      2.325E-3      6.051E-2

DETERMINANT OF COVARIANCE MATRIX S(5) =  1.427E-04


NAME OF THE COUNTRY : PHILIPPINES
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976  -5.293    -4.411    -1.910
1977  -4.955    -3.892    -1.928
1978  -5.517    -4.057    -1.774
1979  -5.517    -4.595    -1.928
1980  -5.293    -3.892    -1.621
1981  -5.517    -3.585    -1.883
       =====     =====     =====
MEAN  -5.349    -4.072    -1.841

   COVARIANCE MATRIX - S(6)
   ========================

             BF&VL          BUTTER        MILK
BF&VL        4.934E-2      1.315E-2     -3.055E-3
BUTTER       1.315E-2      1.382E-1      1.459E-2
MILK        -3.055E-3      1.459E-2      1.484E-2

DETERMINANT OF COVARIANCE MATRIX S(6) =  8.571E-05
```

```
NAME OF THE COUNTRY : SINGAPORE
==================


YEAR   BF&VL      BUTTER   MILK
====   =====      ======   ====

1976  -3.944    -4.185   -2.734
1977  -4.185    -4.255   -3.032
1978  -4.119    -4.701   -3.079
1979  -4.330    -4.820   -3.547
1980  -3.944    -4.057   -2.966
1981  -3.842    -3.664   -2.903
      =====     =====    =====
MEAN  -4.061    -4.280   -3.043


   COVARIANCE MATRIX - S(7)
   ========================

            BF&VL        BUTTER       MILK
BF&VL       3.323E-2     6.773E-2     4.239E-2
BUTTER      6.773E-2     1.818E-1     8.412E-2
MILK        4.239E-2     8.412E-2     7.532E-2

DETERMINANT OF COVARIANCE MATRIX S(7) =  3.064E-05


      S = (S1+S2+S3+S4+S5+S6+S7) / 7

AVERAGE COVARIANCE MATIX - (S)
==============================

            BF&VL        BUTTER       MILK
BF&VL       1.400E-1    -1.931E-1     7.237E-2
BUTTER     -1.931E-1     1.409E+0    -2.326E-1
MILK        7.237E-2    -2.326E-1     2.119E-1

DETERMINANT OF AVERAGE COVARIANCE MATRIX =  2.543E-02

  35 Ln DET (S)   = -128.5

   5 Ln DET (S1) =    -7.8
   5 Ln DET (S2) =   -24.7
   5 Ln DET (S3) =   -51.9
   5 Ln DET (S4) =   -33.3
   5 Ln DET (S5) =   -44.3
   5 Ln DET (S6) =   -46.8
   5 Ln DET (S7) =   -52.0
```

$$35 \text{ Ln DET (S)} - \sum_{i=1}^{7} 5 \text{ Ln DET (Si)} = 132.3$$

```
C 7.3.4  RESULTS WHEN DATA WAS TRANSFORMED USING PROBIT
         ==============================================
         CONVERSION
         ==========


NAME OF THE COUNTRY : AUSTRALIA
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976    7.880     8.090     8.090
1977    7.510     8.090     7.750
1978    7.510     8.090     7.580
1979    7.650     8.090     7.510
1980    7.370     7.460     7.120
1981    7.260     7.460     7.200
        =====     =====     =====
MEAN    7.530     7.880     7.542

   COVARIANCE MATRIX - S(1)
   ========================

            BF&VL        BUTTER       MILK
BF&VL       4.724E-2     5.418E-2     6.858E-2
BUTTER      5.418E-2     1.058E-1     9.618E-2
MILK        6.858E-2     9.618E-2     1.282E-1

DETERMINANT OF COVARIANCE MATRIX S(1) =  4.465E-05


NAME OF THE COUNTRY : CHINA - TAIWAN
===================


YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976    7.510     8.090     6.640
1977    7.880     7.880     6.540
1978    8.090     7.650     6.450
1979    7.880     7.580     6.500
1980    7.750     7.750     6.840
1981    7.650     7.510     6.730
        =====     =====     =====
MEAN    7.793     7.743     6.617

   COVARIANCE MATRIX - S(2)
   ========================

            BF&VL        BUTTER       MILK
BF&VL       4.115E-2    -1.901E-2    -1.975E-2
BUTTER     -1.901E-2     4.575E-2     1.453E-3
MILK       -1.975E-2     1.453E-3     2.211E-2

DETERMINANT OF COVARIANCE MATRIX S(2) =  1.679E-05
```

```
NAME OF THE COUNTRY : HONG KONG
===================


YEAR    BF&VL     BUTTER   MILK
====    =====     ======   ====

1976    7.120     7.650    7.460
1977    7.050     7.510    7.510
1978    7.120     7.510    7.410
1979    7.200     7.510    7.510
1980    7.100     7.410    7.510
1981    7.040     7.200    7.510
        =====     =====    =====
MEAN    7.105     7.465    7.485


   COVARIANCE MATRIX - S(3)
   ========================

            BF&VL        BUTTER       MILK
BF&VL      3.350E-3     4.550E-3    -4.500E-4
BUTTER     4.550E-3     2.271E-2    -2.750E-3
MILK      -4.500E-4    -2.750E-3     1.750E-3

DETERMINANT OF COVARIANCE MATRIX S(3) =  7.824E-08


NAME OF THE COUNTRY : JAPAN
===================


YEAR    BF&VL     BUTTER   MILK
====    =====     ======   ====

1976    6.880     6.930    6.230
1977    6.930     6.900    6.490
1978    6.740     7.880    6.610
1979    6.910     7.290    6.500
1980    6.940     7.880    6.670
1981    7.030     8.090    6.670
        =====     =====    =====
MEAN    6.905     7.495    6.528


   COVARIANCE MATRIX - S(4)
   ========================

            BF&VL        BUTTER       MILK
BF&VL      9.070E-3     4.510E-3     3.110E-3
BUTTER     4.510E-3     2.732E-1     7.349E-2
MILK       3.110E-3     7.349E-2     2.762E-2

DETERMINANT OF COVARIANCE MATRIX S(4) =  1.829E-05
```

```
NAME OF THE COUNTRY : MALAYSIA
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    7.650      6.960     5.890
1977    7.750      6.980     6.120
1978    7.750      7.200     5.910
1979    7.880      7.370     6.080
1980    7.750      7.120     6.240
1981    7.750      6.910     6.100
        =====      =====     =====
MEAN    7.755      7.090     6.057


   COVARIANCE MATRIX - S(5)
   ========================

             BF&VL         BUTTER        MILK
BF&VL        5.350E-3      9.880E-3      3.940E-3
BUTTER       9.880E-3      3.056E-2      5.600E-4
MILK         3.940E-3      5.600E-4      1.787E-2

DETERMINANT OF COVARIANCE MATRIX S(5) =  7.446E-07



NAME OF THE COUNTRY : PHILIPPINES
===================


YEAR    BF&VL      BUTTER    MILK
====    =====      ======    ====

1976    7.580      7.260     6.130
1977    7.460      7.050     6.140
1978    7.650      7.120     6.060
1979    7.650      7.330     6.140
1980    7.580      7.050     5.970
1981    7.650      6.930     6.120
        =====      =====     =====
MEAN    7.595      7.123     6.093


   COVARIANCE MATRIX - S(6)
   ========================

             BF&VL         BUTTER        MILK
BF&VL        5.550E-3      1.900E-3     -5.600E-4
BUTTER       1.900E-3      2.191E-2      3.047E-3
MILK        -5.600E-4      3.047E-3      4.547E-3

DETERMINANT OF COVARIANCE MATRIX S(6) =  4.715E-07
```

```
NAME OF THE COUNTRY : SINGAPORE
===================

YEAR    BF&VL     BUTTER    MILK
====    =====     ======    ====

1976    7.070     7.170     6.550
1977    7.170     7.200     6.690
1978    7.140     7.370     6.710
1979    7.230     7.410     6.910
1980    7.070     7.120     6.650
1981    7.040     6.960     6.630
        =====     =====     =====
MEAN    7.120     7.205     6.690


   COVARIANCE MATRIX - S(7)
   ========================

             BF&VL        BUTTER       MILK
BF&VL        5.280E-3     1.024E-2     7.680E-3
BUTTER       1.024E-2     2.755E-2     1.428E-2
MILK         7.680E-3     1.428E-2     1.472E-2

DETERMINANT OF COVARIANCE MATRIX S(7) =  1.421E-07


     S = (S1+S2+S3+S4+S5+S6+S7) / 7

AVERAGE COVARIANCE MATIX - (S)
==============================

             BF&VL        BUTTER       MILK
BF&VL        1.671E-2     9.464E-3     8.936E-3
BUTTER       9.464E-3     7.535E-2     2.661E-2
MILK         8.936E-3     2.661E-2     3.097E-2

DETERMINANT OF AVERAGE COVARIANCE MATRIX =  2.288E-05

  35 Ln DET (S)  =  -374.0

   5 Ln DET (S1) =   -50.0
   5 Ln DET (S2) =   -55.0
   5 Ln DET (S3) =   -81.8
   5 Ln DET (S4) =   -54.5
   5 Ln DET (S5) =   -70.6
   5 Ln DET (S6) =   -72.8
   5 Ln DET (S7) =   -78.8

                        7
  35 Ln DET (S) -       ∑   5 Ln DET (Si) =  89.5
                       i=1
```

CHAPTER 8

8. CONCLUSIONS

An overview of Multivariate Ranking and Selection procedures and an example using Multivariate Ranking has been discussed in the preceding chapters. This field of study is relatively new and many improvements can be made to the theory in the years to follow. Although there are several methods available to select the 'best' population, methods dealing with the complete ranking of populations are scarce.

It was noticed that in the methods discussed in Chapter 4 and Chapter 5 that a scalar function θ_i had to be found to select the populations. However, there is no indication as to how a clear and meaningful choice of that scalar function is made for a given situation.

Also, it is evident that the P(CS) expressions are quite difficult to evaluate and the tabulated values of such expressions are not always available.

The method by Dudewicz and Taneja (1981) on the Multivariate Solution to the Multivariate Ranking and Selection problem explained in Chapter 6 is difficult to apply in practice as defining the function g is not easy.

Here again the importance of the example on Multivariate Ranking discussed in Chapter 7 is stressed as solved examples of that nature are practically non-existent.

BIBLIOGRAPHY

ALAM, K. and M.H. Rizvi (1966)
Selection from multivariate populations. Annals of the Institute of Statistical Mathematics 18: 307-318.

ALAM, K., M.H. Rizvi and H. Solomon (1975)
Selection of largest multiple correlation coefficients: exact sample size case. Annals of Statistics 4: 614-620.

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

BARTLETT, M.S. and D.G. Kendall (1946)
The statistical analysis of variance heterogeneity and the logarithmic transformation. Journal of the Royal Statistical Society Supplement 8: 128-138.

BECHHOFER, R.E. (1954)
A single sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics 25: 16-39.

BECHHOFER, R.E., J. Kiefer and M. Sobel (1968)
Sequential identification and ranking procedures. Chicago, University of Chicago Press.

BOX, G.E.P. (1949)
A general distribution theory for a class of likelihood criteria. Biometrika 36: 317-346.

CARROLL, R.J. and S.S. Gupta (1977)
On the probabilities of ranking of k populations with applications. Journal of Statistical Computation and Simulation 5: 145-157.

CHOW, Y.S. and H. Robbins (1965)
On the asymptotic theory of fixed-width sequential confidence intervals for the mean. Annals of Mathematical Statistics 36: 457-462.

DUDEWICZ, E.J. and T.A. Bishop (1979)
The heteroscedastic method. p. 183-203. In Rustagi, J.S. ed. Optimizing Methods in Statistics. Academic Press, New York.

DUDEWICZ, E.J. and V.S. Taneja (1981)
A multivariate solution to the multivariate ranking and selection problem. Communications in Statistics Series A: 1849-1868.

FREEMAN, H., A. Kuzmack and R. Maurice (1967)
Multivariate t and the ranking problem. Biometrika 54: 305-308.

FRISCHTAK, R.M. (1973)
Statistical multiple decision procedures for some multivariate selection problems. Thesis, Ph.D., New York, Cornell University, 102 p.

GIBBONS, J.D., I. Olkin and M. Sobel (1977)
Selecting and ordering populations. New York, Wiley. 569 p.

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

GOVINDARAJULU, Z., and A.P. Gore (1971)
Selection procedures with respect to measures of association. p. 313-345. In Gupta, S.S. and J. Yackel eds. Statistical Decision Theory and Related Topics. New York, Academic Press.

GUPTA, S.S. (1956)
On a decision rule for a problem in ranking means. Thesis, Ph.D., Chapel Hill, University of North Carolina.

GUPTA, S.S. (1963)
On a selection and ranking procedure for gamma populations. Annals of the Institute of Statistical Mathematics 14: 199-216.

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

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

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

GUPTA, S.S. and S. Panchapakesan (1979)
Multiple decision procedures. New York, Wiley. 573 p.

GUPTA, S.S. and M. Sobel (1962)
On the smallest of several correlated F statistics. Biometrika 49: 509-523.

GUPTA, S.S. and W.J. Studden (1970)
On some selection and ranking procedures with applications to multivariate populations. p. 327-338. In Bose, R.C. and others ed. Essays in Probability and Statistics. Chapel Hill, University of North Carolina Press.

HOEL, P.G. (1937)
A significance test for component analysis. Annals of Mathematical Statistics 8: 149-158.

HOEL, D.G. and M. Mazumdar (1968)
An extension of Paulson's selection procedure. Annals of Mathematical Statistics 39: 2067-2074.

PAULSON, E. (1964)
A sequential procedure for selecting the population with the largest mean from k normal populations. Annals of Mathematical Statistics 35: 174-180.

RAO, C.R. (1965)
Linear statistical inference and its applications. New York, John Wiley & Sons.

REGIER, M.H. (1976)
Simplified selection procedures for multivariate normal populations. Technometrics 18: 483-489.

Report and Analysis of External Trade, 1979/80 and 1980/81.
New Zealand, Department of Statistics Publication. Wellington, Government Printer.

RIZVI, M.H. and H. Solomon (1973)
Selection of largest multiple correlation coefficients: asymptotic case. Journal of the American Statistical Association 68: 184-188. Corrigendum, 69, 288.

SRIVASTAVA, M.S. and V.S. Taneja (1972)
Some sequential procedures for ranking multivariate normal populations. Annals of the Institute of Statistical Mathematics 24: 455-464.

STARR, N. (1966)
The performance of a sequential procedure for the fixed-width interval estimation of the mean. Annals of Mathematical Statistics 37: 36-50.