ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

Robert Pietrzykowski, Wojciech Zieliński***

A NEW PROCEDURE OF MULTIVARIATE MULTIPLE COMPARISONS

Abstract. In the paper a statistical procedure for dividing a set $\{\mu_1, ..., \mu_k\}$ of vector of means of k normal p-variate distributions into homogenous groups is proposed. It appears that proposed procedure has a high probability of correct decision.

Key words: multivariate normal distribution, MANOVA, multiple comparisons.

1. INTRODUCTION

Consider k normal p-variate distributions $N_p(\mu_1, \Sigma)$, ..., $N_p(\mu_k, \Sigma)$. The problem is to divide a set of vectors of means $\{\mu_1, ..., \mu_k\}$ into homogeneous groups on the basis of k samples $\mathbf{X}_{ij}$, $j = 1, ..., n_j$, $i = 1, ..., k$. A subset $\{\mu_{i_1}, ..., \mu_{i_m}\}$ is called a homogeneous group if $\mu_{i_1} = \cdots = \mu_{i_m}$ any of remaining vectors equals μ_{i_1} W. Zieliński (1991). In univariate case, i.e. $p = 1$ several procedures of dividing the set of means into homogenous groups are known. The commonly used in practical applications are Tukey, Scheffé, Bonfferroni and *Least Significance Difference*. In the multivariate case there are a lot of different procedures which divide a set of means into groups. Some of them are described in P. R. Krishnaiah (1966), C. R. Rao (1973), H. Ahrens and J. Läuter (1979), T. Caliński *et al.* (1979), T. Caliński and M. Lejeune (1998), M. Krzyśko (2000). Those procedures mainly are based on distances of sample means, but none of them is a full analog of one dimensional procedures. In what follows there is a proposition of a procedure which is an extension of one dimensijonal M. Zieliński (1998) procedure to multivariate case.

* PhD., Department of Mathematical Statistics and Experimentation, University of Agriculture of Warsow, e-mail: pietrzyk@dela.sggw.waw.pl.

** Prof., Department of Mathematical Statistics and Experimentation, University of Agriculture of Warsow, e-mail: wojtek.zielinski@omega.sggw.waw.pl.

In the paper we restrict ourselves to the simplest situation. We assume, that we have samples with the same number of observations and that the samples are independent. Also we assume ewuality of covariauce matrices of compared distributions.

2. PROCEDURE

Let $\boldsymbol{\mu}_i = [\mu_{i1}, ..., \mu_{ip}]$ $(i = 1, ..., k)$ and let $\mathbf{X}_{ij} = [X_{i1j}, ..., X_{ipj}]$ $(j = 1, ..., n_i,$ $i = 1, ..., k)$. Let $N = \sum_{i=1}^{k} n_i$ denotes the number of all observed vectors and

$$\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij} = [\bar{X}_{i1}, ..., \bar{X}_{ip}], \quad \bar{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{X}_{ij} = [\bar{X}_1, ..., \bar{X}_p] \tag{1}$$

denote sample mean for i-th sample and overall sample mean respectively.

Let $\mathcal{J}(r) = \{I_1, ..., I_r\}$ be a division of a set of $\{1, ..., k\}$ and let

$$\mathbf{H}_{\mathcal{J}(r)} = [h_{ij}]_{i,j=1,...,p} \tag{2}$$

where

$$h_{ij} = \frac{1}{K-1} \sum_{m=1}^{r} \sum_{l \in I_m} n_l (\bar{X}_{li} - \bar{X}_{I_m,i})(\bar{X}_{lj} - \bar{X}_{I_m,j}) \tag{3}$$

$$\bar{X}_{I_m,i} = \frac{1}{\sum_{j \in I_m} n_j} \sum_{j \in I_m} \sum_{l=1}^{n_j} X_{li} \tag{4}$$

The mean $\bar{X}_{I_m,i}$ is the sample mean of i-th variate in the group I_m. Let $\mathbf{E} = [e_{ij}]_{i,j=1,...,p}$ be a standard matrix of random errors:

$$e_{ij} = \frac{1}{N-k} \sum_{m=1}^{k} \sum_{l=1}^{n_m} (X_{mil} - \bar{X}_{mi})(X_{mjl} - \bar{X}_{mj}) \tag{5}$$

Let $\mathcal{J}_r^* = \{I_1^*, ..., I_r^*\}$ be a division into r disjoint subsets such taht

$$\operatorname{tr} \mathbf{H}_{\mathcal{J}^*(r)} \mathbf{E}^{-1} = \min_{\mathcal{J}(r)} \operatorname{tr} \mathbf{H}_{\mathcal{J}(r)} \mathbf{E}^{-1} \tag{6}$$

Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_{k-1})$ be a sequence of numbers such that $\alpha_r \in (0, 1)$. Procedure starts with $r = 1$ and r is increased till

$$\operatorname{tr} \mathbf{H}_{\mathcal{J}^*(r)} \mathbf{E}^{-1} \leqslant c F(\alpha_r, a, b) \tag{7}$$

where $F(\alpha_r, a, b)$ is the critical value of the F distribution with (a, b) degress of fredom:

$$a = pr;\ b = 4 + \frac{a+2}{B-1};\ B = \frac{(N-k+r-p-1)(N-k-1)}{(N-k-p-3)(N-k-p)};$$

$$c = \frac{a(b-2)}{b(N-k-p-1)} \tag{8}$$

If the procedure stops and if the inequality $\operatorname{tr} \mathbf{H}_{\mathcal{J}^*(r)} \mathbf{E}^{-1} \leqslant cF(\alpha_r, a, b)$ holds we decide that we have the following division of the vectors of means:

$$\{\{\boldsymbol{\mu}_i : i \in I_1^*\}, \ldots, \{\boldsymbol{\mu}_i : i \in I_r^*\}\}, \tag{9}$$

otherwise

$$\{\{\boldsymbol{\mu}_1\}, \ldots, \{\boldsymbol{\mu}_k\}\}. \tag{10}$$

3. CRITERION

Let $\Theta = \{\theta_1, \theta_2, \ldots,\}$ denotes the set of all possible divisions of the set of vectors of means into homogenous groups. Elements of the set Θ are disjoint subsets of $(\mathbf{R}^p)^k$ and for every $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) \in (\mathbf{R}^p)^k$ there exists only one $\theta \in \Theta$ such that $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) \in \theta$. Note that Θ is a finite set. The elements of the set Θ are commonly called "states of nature".

The aim of any multiple comparison procedure is to "detect" the true state of nature. Let $\mathscr{D}$ be a set of all dicisions which can be made on the basis of observations. The elements of the set $\mathscr{D}$ are called "decisions". We assume that $\mathscr{D} \supseteq \Theta$.

We define the loss function in the following manner:

$$L(d, \theta) = \begin{cases} 0, & \text{if } d = \theta, \\ 1, & \text{if } d \neq \theta, \end{cases} \quad \text{for } d \in \mathscr{D} \text{ and } \theta \in \Theta \tag{11}$$

This loss function gives penalty of one when our decision is not correct.

If we denote by $\mathscr{X}$ the space of all observations, than the functijon $\delta : \mathscr{X} \to \mathscr{D}$ is called a "decision rule". Any of the above mentioned procedures of multiple comparisons may be described as a decision rule.

A decision rule δ is characterized by its risk function, i.e., average loss. Let $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) \in \theta$. Then the risk function of the rule δ equals

$$R_\delta(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = P_{(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)}\{\delta(\mathbf{X}) \neq \theta\}. \tag{12}$$

Note that in general the risk depends on the distances of vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$.

The risk of the rule δ is the probability of the false decision. This probability should be as small as possible. In our investigation we are interested in a probability of a correct decision which is equal to $1 - R_\delta$.

We are interested in the probability of the correct decision of the described procedure. The probability is very difficult to calculate even in the "simple" case of $k = 3$. So we perform a Monte Carlo experiment to estimate the probability.

4. MONTE CARLO EXPERIMENT

To estimate the probability of the correct decision, a Monte Carlo experiment was performed. In the experiment $p = 4$, $k = 8$, $n = 30$ were taken. Means were taken in the ranges

$$\boldsymbol{\mu}_1 \in (-3; 3);\ \boldsymbol{\mu}_2 \in (-2; 2);\ \boldsymbol{\mu}_3 \in (-2.5; 2.5);\ \boldsymbol{\mu}_4 \in (0.5; 0.5)$$

and the coveriance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.35 & 0.35 & 0.35 \\ & 1 & 0.35 & 0.35 \\ & & 1 & 0.35 \\ & & & 1 \end{bmatrix}$$

Parameter α was chosen as $\alpha_1 = \cdots = \alpha_{k-1} = 0.05$. Hence the following critical values were obtained in Tab. 1.

Table 1

Critical values

r	a	b	c	$F(\alpha_r, a, b)$	$cF(\alpha_r, a, b)$
1	4	229.0000000	0.017467	2.411066	0.042115
2	8	324.8255159	0.035025	1.966947	0.068893
3	12	396.4590164	0.052597	1.776629	0.093445
4	16	452.0349345	0.070173	1.665917	0.116902
5	20	496.4083770	0.087751	1.591755	0.139678
6	24	532.6563615	0.105330	1.537818	0.161978
7	28	562.8235294	0.122910	1.496415	0.183924

In case of $k = 8$ there are 21 possibilities of dividing a set of vectors of means into disjoint homogenous groups. All possible states of nature are shown in the Tab. 2. Notation $(i_1, i_2, ..., i_m)$ means m groups with $i_1, i_2, ..., i_m$ vectors. It is assumed,j that $i_1 \leqslant i_2 \leqslant \cdots \leqslant i_m$ and $i_1 + i_2 + \cdots + i_m = 8$. For example, $(1,2,5)$ means the division into three homogeneous groups: $\{\mu_1\}$, $\{\mu_2, \mu_3\}$, $\{\mu_4, \mu_5, \mu_6, \mu_7, \mu_8\}$, i.e. $\mathcal{J} = \{\{1\}, \{2,3\}, \{4,5,6,7,8\}\}$.

Table 2

States of nature for $k = 8$

r	State of nature
1	(8)
2	(1, 7), (2, 6), (3, 5), (4, 4)
3	(1, 1, 6), (1, 2, 5), (1, 3, 4), (2, 2, 4), (2, 3, 3)
4	(1, 1, 1, 5), (1, 1, 2, 4), (1, 1, 3, 3), (1, 2, 2, 3), (2, 2, 2, 2)
5	(1, 1, 1, 1, 4), (1, 1, 1, 2, 3), (1, 1, 1, 2, 2)
6	(1, 1, 1, 1, 1, 3), ((1, 1, 1, 1, 2, 2)
7	(1, 1, 1, 1, 1, 1, 2)

In a Monte Carlo xperiment one has to choose values of means. It is diffcult to make a "planned" experiment in the sense of choosing the mean values. This values were taken in a random way. For example, for the state (1, 7) there were randomly generated:

$$t_{11},\ t_{12} \in (-3;\ 3);\ t_{21},\ t_{22} \in (-2;\ 2);\ t_{31},\ t_{32} \in (-2.5;\ 2.5);\ t_{41},\ t_{42} \in (-0.5;\ 0.5)$$

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
t_{11}	t_{12}	t_{12}	t_{12}	t_{12}	t_{12}	t_{12}	t_{12}
t_{21}	t_{22}	t_{22}	t_{22}	t_{22}	t_{22}	t_{22}	t_{22}
t_{31}	t_{32}	t_{32}	t_{32}	t_{32}	t_{32}	t_{32}	t_{32}
t_{42}	t_{42}	t_{42}	t_{42}	t_{42}	t_{42}	t_{42}	t_{42}

Such procedure was applied 100 times for each state.

At each generated point $(\mu_1, ..., \mu_8)$ there were made 1000 drawn of eight samples of thirty from normal populations with means μ_i respectively. To each sample the procedure was applied and the number of division consistent with "reality" was noted.

5. RESULTS

Results are presented on Fig. 1 and 2 for chosen configurations of means. For a given state of nature the generated configurations of vectors were ordered due to the estimated value of probability of the correct decision. On x axis there is a number (divided by 100) of generated configuration and on y axis there is an approprate value of the estimated probability. In the Fig. 1 results for the state (1, 7) are presented and in the Fig. 2 appropriate results for (1, 1, 1, 1, 1, 1, 2) state. The results for other states were similar.

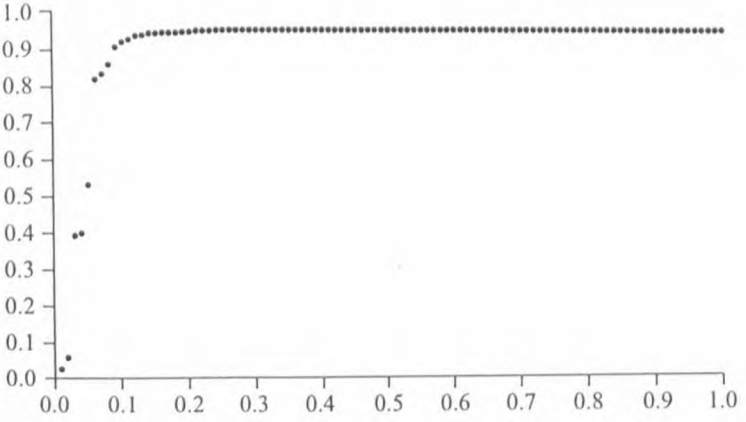

Fig. 1. Results for (1, 7)

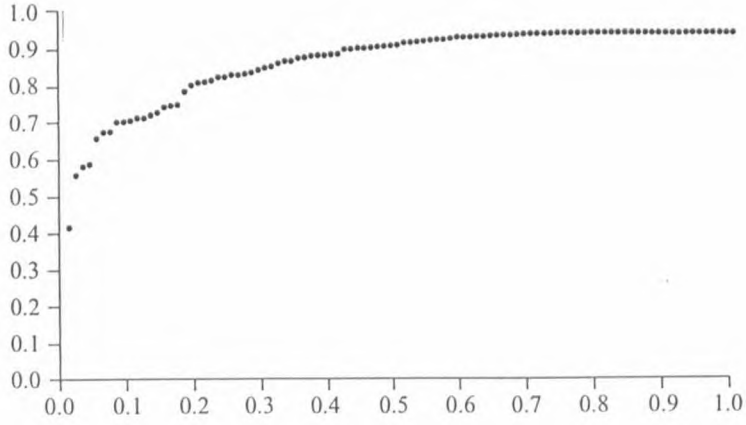

Fig. 2. Results for (1, 1, 1, 1, 1, 1, 2)

On the basis of simulations following conclusions may be formulated.

1. In general the procedure has a quite high probability of a correct decision. In average, about 80% of configurations of vectors of means were correctly deteted with probability at least 0.90.

2. Results may be interpreted in the folowing way. If, for example, the procedure give a division (1, 7) then with confidence at least 90% we may be almost sure (more than 95%) that obtained division coincide with reality.

3. We made investigations only for one case ($k = 8$ and $p = 4$), but it may be expected, that for other k's and p's results will be similar. Investigations for other k and p are in progress.

REFERENCES

Ahrens H., Läuter J. (1979), *Wielowymiarowa analiza wariancji*, PWN, Warszawa.

Caliński T., Lejeune M. (1998), *Dimensionality in MANOVA Tested by Closed Testing Procedure*, "Journal of Multivariate Analysis", **65**, 181–194.

Caliński T., Dyczkowski A., Sitek M. (1979), *Procedury testów jednoczesnych w wielozmiennej analizie wariancji*, "Matematyka Stosowana", **14**, 5–31.

Krishmaiah P. R. (1966), *Multivariate Analysis*, Academic Press, New York–London.

Krzyśko M. (2000), *Wielowymiarowa analiza statystyczna*, Uniwersytet im. Adama Mickiewicza, Poznań.

Rao C. Radhakrishna (1973), *Linear Statistical Inference and Its Applicatijons*, Wiley & Sons, New York.

Zieliński W. (1991), *Nowa procedura porównań wielokrotnych*, Wydawnictwo SGGW, Warszawa.

Zieliński W. (1998), *On a Procedure of Multiple Comparisons*, "Biometrical Letters", **35**, 67–76.

Robert Pietrzykowski, Wojciech Zieliński

NOWA WIELOWYMIAROWA PROCEDURA PORÓWNAŃ WIELOKROTNYCH

W pracy rozważano problem podziału wektorów średnich pochodzących z k populacji o wielowymiarowych rozkładach normalnych na grupy jednorodne. Jako kryterium jakości procedury porównań wielokrotnych przyjęto prawdopodobieństwo podjęcia poprawnej decyzji (PCD), to znaczy prawdopodobieństwo uzyskania podziału zgodnego z rzeczywistym układem badanych zbiorowości. Takie podejście do problemu jest pewnym rozwinięciem prac W. Zielińskiego (1991, 1998) dla przypadku jednowymiarowego. Jako statystykę testową wykorzystano statystykę T^2 Lawleya–Hotellinga. W pracy przedstawiono wyniki badań symulacyjnych dla ośmiu populacji czterocechowych.