Listy Biometryczne - Biometrical Letters
Vol. 35(1998), No. 1, 67-76

Comparison of a probability of the correct decision of multiple comparison procedures

Wojciech Zieliński

Department of Mathematical Statistics and Experimentation,
University of Agriculture, Rakowiecka 26/30, PL–02–568 Warszawa

e–mail: wojtek.zielinski@omega.sggw.waw.pl

SUMMARY

Four procedures of multiple comparisons are compared with respect to the probability of the correct decision. Among the procedures there are two classical ones (Tukey and Newman–Keuls), the FTP procedure of Caliński and Corsten and a new procedure W. On the basis of a Monte Carlo experiment it is shown that none of the procedures is uniformly the best.

KEY WORDS: multiple comparisons, simultaneous inference, ANOVA.

1. Introduction

Consider k normal populations $N(\mu_i, \sigma^2)$, $i = 1, \ldots, k$. The problem is to verify the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k.$$

This problem is known since the thirties, when Fisher (1935) developed his analysis of variance. In the analysis of variance the decision is *investigated means are equal* (H_0 not rejected) or *at least one mean is different from the others* (H_0 rejected). The second case is less informative from the practical point of view. The practician would like to know which of the means may be considered as equal, i.e. he would like to divide the set $\{\mu_1, \ldots, \mu_k\}$ into subsets of equal means. Such subsets are called *homogeneous groups*. Several procedures of dividing the set of means into homogeneous groups are known. The most famous and widely used in practical applications are the procedures of *Tukey, Scheffé, Bonfferroni* and the *Least Significant Difference*. Each procedure may give different homogeneous groups. The question is which of the divisions is nearest to reality, i.e. which procedure gives "the best" division.

Almost all procedures of multiple comparisons are described in Miller (1982) and Hochberg and Tamhane (1988). These procedures may be divided into three main groups: simultaneous confidence intervals, simultaneous hypothesis testing and "others". The problem is the comparison of procedures and selection of "the best" one. For example, simultaneous confidence intervals are frequently compared with respect to the length of individual intervals. But such a comparison is not good (Zieliński, 1990). Moreover, there is no reasonable criterion of comparison of procedures of different types. A proposition of such a criterion may be found in Zieliński (1991). The criterion is the probability of the correct decision, i.e. the probability of obtaining a division of a set of means consistent with reality.

DEFINITION (Zieliński 1991). The subset $\{\mu_{i_1}, \ldots, \mu_{i_m}\}$ is called a homogeneous group if $\mu_{i_1} = \cdots = \mu_{i_m}$ and no other mean is equal to μ_{i_1}.

The aim of a multiple comparison procedure is to divide the set $\{\mu_1, \ldots, \mu_k\}$ of means into homogeneous groups on the basis of a set of observations $\{X_{ij} : j = 1, \ldots, n,\ i = 1, \ldots, k\}$. If the obtained division is equal to the real one we say that the procedure made the correct decision. We are interested in the probability of the correct decision, and a procedure with a higher probability is better. In what follows, four procedures are compared with respect to that criterion.

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

2. Procedures

Consider four procedures: simultaneous confidence intervals of Tukey, the Newman–Keuls multiple hypothesis test, cluster analysis procedure of Caliński and Corsten (F-test Procedure, FTP) and a new procedure W. The first two procedures are classical procedures and they were chosen with respect to results of Zieliński (1991). Those procedures may give non-disjoint homogeneous groups. The two latter procedures divide a set of means into disjoint homogeneous groups.

Let $\nu = k(n-1)$ and

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij},\ i = 1, \ldots, k, \qquad s^2 = \frac{1}{\nu}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2.$$

Tukey's simultaneous confidence intervals

Tukey's simultaneous confidence intervals have the following form:

$$P\left\{\mu_{i_1}-\mu_{i_2}\in\left(\bar{X}_{i_1}-\bar{X}_{i_2}\pm q^{\alpha}_{k,\nu}\frac{s}{\sqrt{n}},\right),\ \text{for all } i_1,i_2=1,\ldots,k,\ i_1\neq i_2\right\}=1-\alpha,$$

where $q^{\alpha}_{k,\nu}$ is a critical value of the studentized range. If zero is in the confidence interval for $\mu_{i_1}-\mu_{i_2}$, then those two means are considered as equal. Applying that rule to all confidence intervals, a division into homogeneous groups is obtained.

Multiple test of Newman–Keuls

The Newman–Keuls procedure is based on testing hypothesis $H_{i_1,\ldots,i_m}:\mu_{i_1}=\cdots=\mu_{i_m}$ for all sets of indices $\{i_1,\ldots,i_m\}$, $m=k,k-1,\ldots,2$, which are subsets of $\{1,\ldots,k\}$. Hypothesis $H_{i_1,\ldots,i_m}$ is rejected if

$$\frac{\sqrt{n}}{s}\left\{\max\{\bar{X}_i:i\in\{i_1,\ldots,i_m\}\}-\min\{\bar{X}_i:i\in\{i_1,\ldots,i_m\}\}\right\}\geq q^{\alpha}_{m,\nu}.$$

If hypothesis $H_{i_1,\ldots,i_m}$ is not rejected, then the decision is: $\mu_{i_1}=\cdots=\mu_{i_m}$.

The Newman–Keuls procedure is a stepwise one. It starts with $m=k$ and m is decreased. In the first step hypothesis $H_{1,\ldots,k}$ (which is usually noted as H_0) is verified. If the hypothesis is rejected, then the procedure goes to the second step, otherwise it stops and equality of all means is claimed. The second step consists of testing k subhypotheses $\mu_1=\cdots=\mu_{i-1}=\mu_{i+1}=\cdots=\mu_k$, $i=1,\ldots,k$, of $H_{1,\ldots,k}$. If an i–th hypothesis is rejected, then $k-1$ subhypotheses are tested, or else the set $\{\mu_1,\ldots,\mu_{i-1},\mu_{i+1},\ldots,\mu_k\}$ is said to be a homogeneous group and none of the subhypotheses is tested. Next steps consist in testing all the appropriate subhypotheses of the hypothesis rejected in the previous step. The procedure stops if there is nothing left to test.

Caliński-Corsten FTP procedure

The FTP procedure is based on an idea of cluster analysis. Let $\mathcal{J}=\{I_1,\ldots,I_p\}$ be a division of $\{1,\ldots,k\}$ into disjoint subsets. For $\mathcal{J}$ let

$$S(p,\mathcal{J})=n\sum_{i=1}^{p}\sum_{j\in I_i}(\bar{X}_j-\bar{X}_{I_i})^2,$$

where

$$\bar{X}_{I_i}=\frac{1}{nk_i}\sum_{j\in I_i}\sum_{l=1}^{n}X_{jl}=\frac{1}{k_i}\sum_{j\in I_i}\bar{X}_j$$

and k_i is the number of elements of I_i. Let $\mathcal{J}^*$ be a division into p disjoint subsets such that $S(p,\mathcal{J}^*)$ is minimal among $S(p,\mathcal{J})$. The procedure starts with $p=1$ and p

is increased till $S(p, \mathcal{J}^*) < s^2(k-1)F^{\alpha}_{k-1,\nu}$, where $F^{\alpha}_{k-1,\nu}$ is a critical value of the F distribution with $(k-1, \nu)$ degrees of freedom. In this way we obtain a division $\mathcal{J}^*$ of a set of means into p disjoint homogeneous groups.

W procedure

The W procedure is very similar to the FTP procedure. It differs from the previous one in selection of critical values. In the W procedure the number p of homogeneous groups is increased till $S(p, \mathcal{J}^*) < s^2(k-p)F^{\alpha}_{k-p,\nu}$.

3. Criterion

Let $\mathcal{S} = \{s_1, s_2, \ldots\}$ denote the set of all possible divisions of the set of means into homogeneous groups. Elements of the set $\mathcal{S}$ are disjoint subsets of $\mathbf{R}^k$ and for $(\mu_1, \ldots, \mu_k) \in \mathbf{R}^k$ there exists only one $s \in \mathcal{S}$ such that $(\mu_1, \ldots, \mu_k) \in s$. Note that $\mathcal{S}$ is a finite set. The elements of the set $\mathcal{S}$ are commonly called *states of nature.*

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

We define the loss function in the following manner

$$L(d, s) = \begin{cases} 0, & \text{if } d = s \\ 1, & \text{if d} \neq s \end{cases} \qquad \text{for } d \in \mathcal{D} \text{ and } s \in \mathcal{S}.$$

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

If we denote by $\mathcal{X}$ the space of all observations, than the function $\delta : \mathcal{X} \to \mathcal{D}$ 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 $(\mu_1, \ldots, \mu_k) \in s$. Then the risk function of the rule δ equals to

$$R_\delta(\mu_1, \ldots, \mu_k) = P_{(\mu_1, \ldots, \mu_k)}\{\delta(\mathbf{X}) \neq s\}.$$

Note that in general the risk depends on the differences between the values of means $(\mu_1, \ldots, \mu_k)$. For example, if we assume $k = 3$ and $\sigma^2 = 1$, then it is easier to make a misclassification for $\mu_1 = \mu_2 = 1$, $\mu_3 = 1.1$ than for $\mu_1 = \mu_2 = 1$, $\mu_3 = 5$, though both cases belong to the same state of nature. Only in the case $\mu_1 = \cdots = \mu_k$ the risk does not depend on the values of means.

The risk of the rule δ is the probability of the false decision. This probability should be as small as possible. In our investigations we are interested in the probability of the correct decision which equals $1 - R_\delta$. We are going to compare this probability for the four mentioned procedures.

The probability of the correct decision is very difficult to calculate even in the "simple" case of $k = 3$. Therefore, we performed a Monte Carlo experiment to estimate the probability.

4. Experiment

To compare the probability of the correct decision a Monte Carlo experiment was performed. In the experiment means of ten normal distributions were compared on the basis of samples of size eight. Thus, $k = 10$, $n = 8$ and $\sigma^2 = 1$ were taken.

It is obvious that the probability of the correct decision does not depend on the values of compared means but only on differences between them, so $\mu_1 = 0$ and $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{10}$ were taken. Hence, there are 42 possibilities of dividing a set of means into disjoint homogeneous groups. All the possible states of nature are shown in Table 1. Notation $(i_1, i_2, \ldots, i_m)$ indicates m groups with $i_1, i_2, \ldots, i_m$ means. It is assumed that $i_1 \geq i_2 \geq \cdots \geq i_m$ and $i_1 + i_2 + \cdots + i_m = 10$. For example, $(4, 3, 2, 1)$ indicates the division into four homogeneous groups: $\{\mu_1, \mu_2, \mu_3, \mu_4\}$, $\{\mu_5, \mu_6, \mu_7\}$, $\{\mu_8, \mu_9\}$ and $\{\mu_{10}\}$.

Table 1. States of nature for ten means

Number of groups	State
10	(1,1,1,1,1,1,1,1,1,1)
9	(2,1,1,1,1,1,1,1,1)
8	(2,2,1,1,1,1,1,1), (3,1,1,1,1,1,1,1)
7	(2,2,2,1,1,1,1), (3,2,1,1,1,1,1), (4,1,1,1,1,1,1)
6	(2,2,2,2,1,1), (3,2,2,1,1,1), (3,3,1,1,1,1), (4,2,1,1,1,1), (5,1,1,1,1,1)
5	(2,2,2,2,2), (3,2,2,2,1), (3,3,2,1,1), (4,2,2,1,1) (4,3,1,1,1), (5,2,1,1,1), (6,1,1,1,1)
4	(3,3,2,2), (4,2,2,2), (3,3,3,1), (4,3,2,1), (5,2,2,1) (4,4,1,1), (5,3,1,1), (6,2,1,1), (7,1,1,1)
3	(4,3,3), (4,4,2), (5,3,2), (6,2,2), (5,4,1), (6,3,1), (7,2,1), (8,1,1)
2	(5,5), (6,4), (7,3), (8,2), (9,1)

In the Monte Carlo experiment one has to choose values of means. It is difficult to make a "planned" experiment in the sense of choosing mean values. Hence, values of means $\mu_2, \ldots, \mu_{10}$ were taken randomly according to the uniform distribution (μ_1 was always zero). For example, for the state $(5, 4, 1)$ we generated two random numbers, say $0 < z_1 < z_2$, and the values of $\mu_1 = \cdots = \mu_5 = 0$, $\mu_6 = \cdots = \mu_9 = z_1$ and $\mu_{10} = z_2$ were taken. Such a procedure was applied 1000 times for each state.

At each generated point $(\mu_1, \ldots, \mu_{10})$, 1000 sets of ten samples of size eight were drawn from normal populations with means μ_i. To each sample all four procedures were applied and it was noted if the obtained division was consistent with "reality".

For generating random numbers from the uniform distribution, a 32–bit multiplicative generator was applied. This generator was written by the author. To obtain normally distributed random numbers the algorithm of Box and Muller (1958) was applied.

5. Results

The probability of the correct decision depends on the differentation of means. A classical measure of such a differentation is the non-centrality parameter $\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2$, where $\bar{\mu}$ is the arithmetic mean of $\mu_1, \ldots, \mu_k$. But in the case of the probability of the correct decision this parameter is rather useless (Fig. 1). One can see high irregularities in the dependence of the probability of a correct decision on the non-centrality parameter. Therefore, presentation of results is made in a different way, namely in relation to the probability of the correct decision of the W procedure. Figures 2, 3 and 4 present results of simulation for three states of nature: $(9,1)$, $(7,1,1,1)$ and $(3,3,1,1,1,1)$, respectively. On the x–axis one can find probabilities of a correct decision for W procedure, and on the y–axis differences between probabilities for W and FTP, Newman–Keuls and Tukey procedures, respectively. If, for example, probability of a correct decision for W procedure is 0.40, FTP — 0.28, Newman–Keuls — 0.30 and Tukey — 0.10, then in the figure we have three points with coordinates $(0.40, -0.12)$, $(0.40, -0.10)$ and $(0.40, -0.30)$.

Figures 2, 3 and 4 show comparisons of the Tukey, Newman–Keuls and FTP procedures with the W procedure. If the appropriate line is above zero then the procedure is better than the W procedure. If the line is below zero, then the respective procedure is worse than W. In Figure 2 results for the state $(9,1)$ (two homogeneous groups of nine and one mean, respectively) are shown. One can see that the W procedure is better than Tukey and Newman–Keuls and is comparable with the FTP procedure. In Figures 3 and 4 we may see that the W procedure is, in general, better than the FTP and Tukey procedures. In the case $(7,1,1,1)$ the Newman-Keuls and W procedures are comparable, but in the situation $(3,3,1,1,1,1)$ the W procedure may be considered as the best one. For other states of nature figures are very similar to the presented ones.

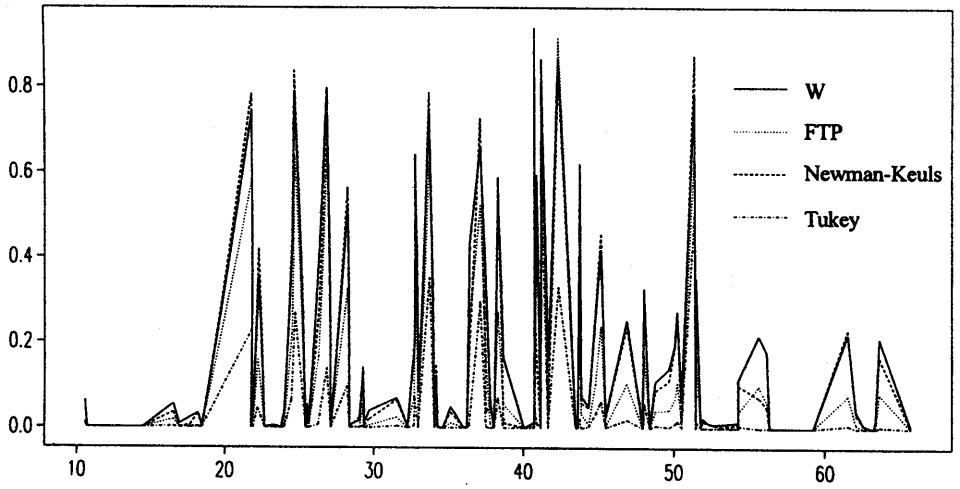

Figure 1. Probability of the correct decision vs. noncentrality parameter

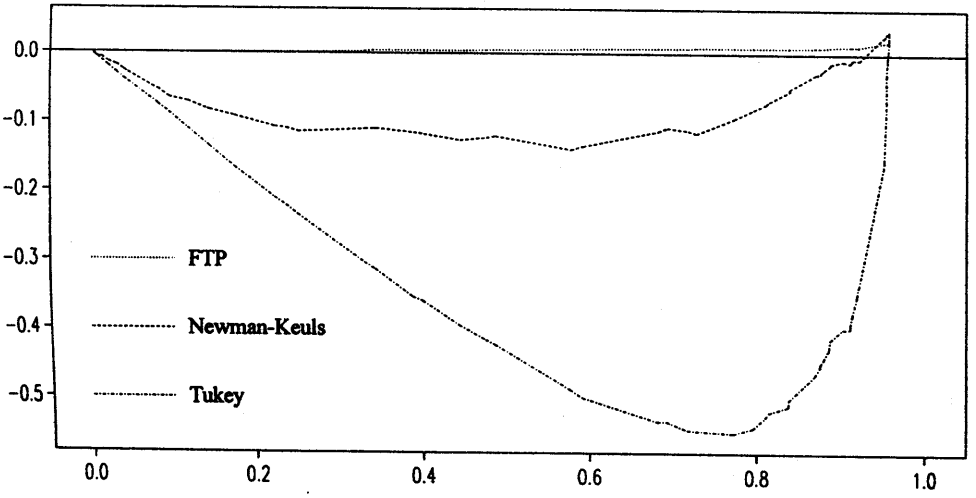

Figure 2. Probability of the correct decision for (9,1) vs. W procedure

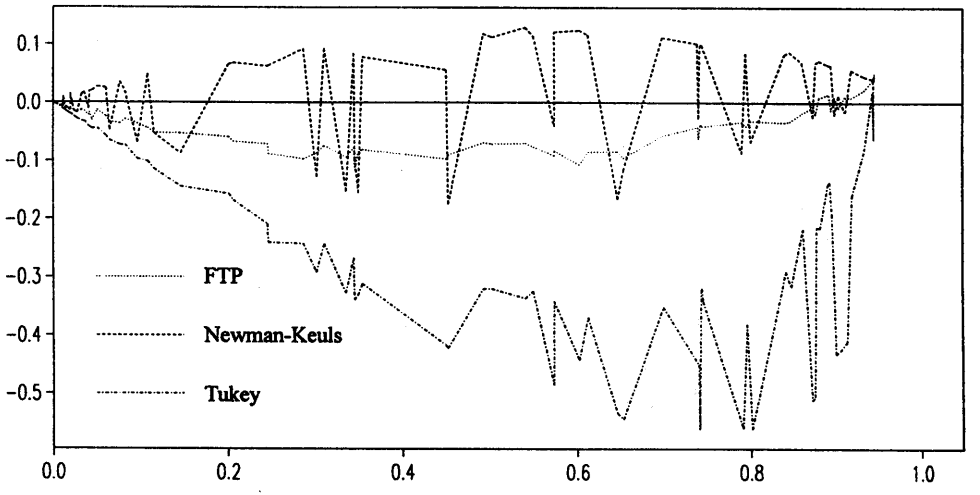

Figure 3. Probability of the correct decision for (7,1,1,1) vs. W procedure

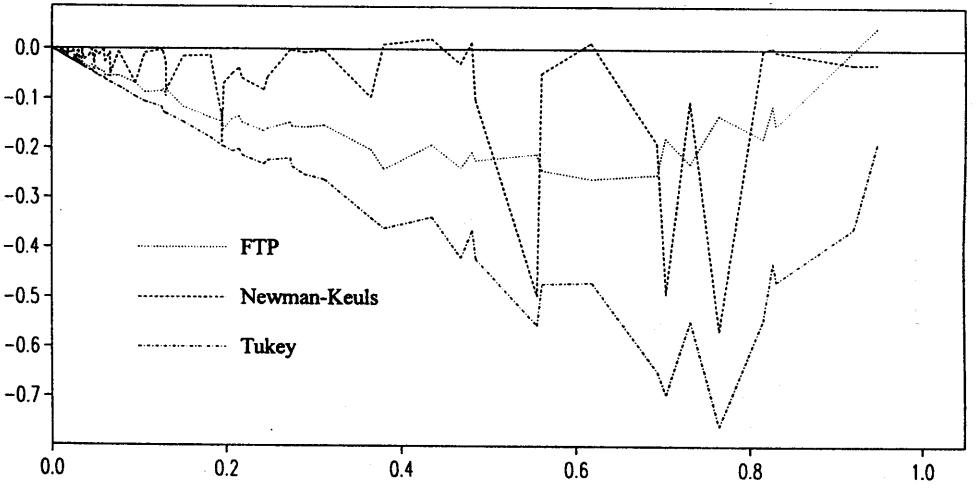

Figure 4. Probability of the correct decision for (3,3,1,1,1,1) vs. W procedure

6. Concluding remarks

1. There is no uniformly best procedure in the sense of the probability of the correct decision.
2. The W procedure is, in general, better than the other considered procedures. It should be remembered that this conclusion is based on the simulation results for ten means. But it may be expected that for other numbers of compared means results will be similar.
3. It may be interesting to change the loss function. The considered loss function assumes that all errors of inference are of equal importance. But it is rather natural to assume that some errors are of a higher weight than the others. Such investigations are in progress.
4. The given definition of homogeneous groups is very restrictive. From the practical point of view two means may be considered as equal if they "slightly" differ. One may define an ε–homogeneous group in the sense that two means μ_1 and μ_2 are in one group if $|\mu_1 - \mu_2| \leq \varepsilon$. Such groups may not be disjoint. From the mathematical point of view it is much harder to work with ε–homogeneous groups than with homogenous ones. It should be mentioned that all procedures which are applied in practice were constructed under a given definition of homogeneous groups.

References

Box G.E.P., Muller M.E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics* **29**, 610–611.

Caliński T., Corsten C.A. (1986). Clustering means in ANOVA by simultaneous testing. *Biometrics* **41**, 39–48.

Fisher R.A. (1935). *The Design of Experiments.* Edinburgh, Oliver and Boyd.

Hochberg Y., Tamhane A.C. (1988). *Multiple Comparison Procedures.* John Wiley & Sons.

Miller Jr. R.G. (1982). *Simultaneous Statistical Inference.* Springer Verlag, 2nd ed.

Zieliński W. (1990). Two remarks on the comparison of simultaneous confidence intervals. *Biometrical Journal* **32**, 717–719.

Zieliński W. (1991). Monte Carlo comparison of multiple comparison procedures. *Biometrical Journal* **34**, 291–296.

Received 25 April 1998; revised 10 August 1998

O pewnej procedurze porównań wielokrotnych

STRESZCZENIE

Praca przedstawia porównanie czterech procedur porównań wielokrotnych ze względu na prawdopodobieństwo podjęcia poprawnej decyzji. Wśród procedur są dwie klasyczne (Tukey'a i Newmana-Keulsa), procedura FTP Calińskiego i Corstena oraz nowa procedura *W*. Na podstawie symulacji Monte Carlo pokazano, że wśród porównywanych nie ma procedury jednostajnie najlepszej.

SŁOWA KLUCZOWE: porównania wielokrotne, wnioskowanie jednoczesne, ANOVA.