Listy Biometryczne - Biometrical Letters
Vol. 35(1998), No. 2, 83-93

On the robustness of multiple comparison procedures

Daniel Rabczenko[1] and Wojciech Zieliński[2]

[1]National Institute of Hygiene, Department of Medical Statistics,
Chocimska 29, PL–00–791 Warszawa
[2]Department of Mathematical Statistics and Experimentation,
University of Agriculture, Rakowiecka 26/30, PL–02–568 Warszawa

e–mail: daniel@galaxy.medstat.pl, wojtek.zielinski@omega.sggw.waw.pl

SUMMARY

The paper concerns investigations on the robustness against non-normality of multiple comparison procedures. The probability of obtaining a division of means which is compatible with the true division is used as the robustness criterion. Results of Monte Carlo experiments suggest that the probability of correct decision increases with the proportion of observations coming from the contaminating distribution.

KEY WORDS: multiple comparisons, simultaneous inference, ANOVA, robustness.

1. Introduction

Consider a problem of testing the hypothesis

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

of equality of means of k normal distributions. Fisher (1935) proposed an F–test for this hypothesis. This test is known as a one–way analysis of variance and is based on the following statistic

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(N-k)}.$$

Here X_{ij} denotes the j–th observation from the i–th distribution ($j = 1, \ldots, n_i$), $\bar{X}_i$ is the arithmetic mean of all observations from the i–th distribution, $\bar{X}$ stands for the arithmetic mean of all observations and $N = n_1 + \cdots + n_k$. The hypothesis is rejected if $F > F^{\alpha}_{k-1,\nu}$, where $F^{\alpha}_{k-1,\nu}$ is a critical value of the F distribution and

$\nu = N - k$. In the case of rejecting H_0 the question arises which of the means may be considered as equal. This problem is known as a problem of multiple comparisons. Zieliński (1994) showed that the F–test is not robust against non-normality. Similar investigations are of interest for multiple comparison procedures.

There are many different multiple comparison procedures. The most frequently used in applications are simultaneous confidence intervals of Tukey and of Scheffé, multiple tests of Newman–Keuls and of Duncan. The above-mentioned and many other procedures of multiple comparisons are described in Miller (1982) and Hochberg and Tamhane (1988). Procedures of multiple comparisons may give different homogenous groups. The question is which division is nearest to "reality". The probability of obtaining a division of a set of means consistent with reality is considered as the criterion of goodness of a procedure. This probability will be called the probability of the correct decision.

There were some simulation studies of this probability for different procedures (cf. Zieliński 1991) but always in the case of normality. In what follows we show results of simulation in a non-normal case.

Suppose that our observations usually follow a normal distribution but occasionally there may appear an "outlier". This means that we observe a random variable Z such that

$$Z = \begin{cases} Z_1 & \text{with probability } 1-\varepsilon, \\ Z_2 & \text{with probability } \varepsilon, \end{cases}$$

where Z_1 is normally distributed $N(\mu, \sigma^2)$, Z_2 is normally distributed $N(\mu, \tau^2)$ and Z_1 and Z_2 are independent. Such contaminated normal distributions were suggested by Tukey (1960). Tukey proposed the following model

$$T(\varepsilon) = (1-\varepsilon)N(\mu, \sigma^2) + \varepsilon N(\mu, \tau^2).$$

This model is referred to as the "Tukey contaminated model".

We are interested in how much the probability of obtaining the correct division changes if observations are drawn not from a normal distribution but from a distribution $T(\varepsilon)$. For obvious reasons we assume that $0 \leq \varepsilon \leq 0.5$.

2. Procedures

We consider the following procedures of multiple comparisons: simultaneous confidence intervals of Tukey and of Scheffé, Newman–Keuls and Duncan multiple hypothesis test, and the procedure W based on a decision theoretic approach. In what follows we deal with the balanced case, i.e. $n_1 = \cdots = n_k = n$.

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 homogenous groups is obtained.

Scheffé simultaneous confidence intervals

Scheffé simultaneous confidence intervals have the following form:

$$P\left\{\mu_{i_1}-\mu_{i_2}\in\left(\bar{X}_{i_1}-\bar{X}_{i_2}\pm s\sqrt{\frac{2}{n}(k-1)F^{\alpha}_{k-1,\nu}}\right),\forall i_1,i_2=1,\ldots,k,\ i_1\neq i_2\right\}=1-\alpha,$$

where $F^{\alpha}_{k-1,\nu}$ is a critical value of the F distribution. Conclusions are made in the same manner as for Tukey's simultaneous confidence intervals.

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

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

where $q^{\alpha}_{k,\nu}$ is a critical value of the studentized range. 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, than 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.

Multiple test of Duncan

The Duncan procedure differs from the Newman–Keuls procedure in choosing critical values. Instead of $q^{\alpha}_{m,\nu}$, $q^{\alpha}_{k,\nu}$ is taken. This means that in the Duncan procedure the critical value is the same for all the tested hypotheses while in the Newman–Keuls procedure it depends on the number of compared means.

W procedure

The W procedure is based on the F distribution. 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 $S(p, \mathcal{J}^*) < s^2(k-p)F^{\alpha}_{k-p,\nu}$, where $F^{\alpha}_{k-p,\nu}$ is a critical value of the F distribution with $(k-p, \nu)$ degrees of freedom. In such a way we obtain a division $\mathcal{J}^*$ of a set of means into p disjoint homogenous groups.

3. Criterion

There are many different criteria for comparing procedures of multiple comparisons. For example, for simultaneous confidence intervals the length of individual confidence intervals are compared. Because we want to compare different procedures, a criterion which can be applied to all procedures is needed. Such a criterion is the probability of obtaining a division of the set of means which is compatible with the true division.

Let s be a real division of a set $\{\mu_1, \ldots, \mu_k\}$ and let $d_\xi(\mathbf{X})$ be a division obtained as a result of applying a procedure ξ of multiple comparisons (here $\mathbf{X}$ denotes the set of all observations). Note that s is a subset of k–dimensional real space. Our criterion is

$$P_F\{d_\xi(\mathbf{X}) = s | \{\mu_1, \ldots, \mu_k\} \in s\},$$

where F is the joint distribution of $\mathbf{X}$. This probability depends not only on division s but also on values of means. Its analytical computation for all sets of means is impossible, so it was estimated in a Monte Carlo experiment.

The probability of the correct decision was estimated for all the procedures mentioned in Section 2. The procedure is better if the probability is higher.

We are interested in the robustness of the probability of the correct decision of the procedure ξ against the contaminated distribution $T(\varepsilon)$. Let $P_\xi(s,\varepsilon)$ be the mean probability of the correct decision for the state s when the underlying distribution is $T(\varepsilon)$, i.e.

$$P_\xi(s,\varepsilon) = \int \cdots \int P_{F(\varepsilon)}\{d_\xi(\mathbf{X}) = s | \{\mu_1, \cdots, \mu_k\} \in s\} d\mu_1 \ldots d\mu_k.$$

We assume that X_{ij} follows the distribution $(1-\varepsilon)N(\mu_i, \sigma^2) + \varepsilon N(\mu_i, \tau^2)$ and $F(\varepsilon)$ is the joint distribution of all observations.

The robustness at the state s of a procedure ξ can be measured by

$$r_\xi(s,\varepsilon) = \frac{P_\xi(s,\varepsilon)}{P_\xi(s,0)}.$$

The procedure ξ_1 will be called more robust than ξ_2 if

$$\sup_{0\le\varepsilon\le 0.5} |1 - r_{\xi_1}(s,\varepsilon)| \le \sup_{0\le\varepsilon\le 0.5} |1 - r_{\xi_2}(s,\varepsilon)|.$$

Note that $r_\xi > 1$ if the probability of the correct decision is greater under distribution $T(\varepsilon)$ than under the normal distribution, which may be interpreted as a positive behaviour of the procedure ξ. Hence, the robustness may be measured by $\sup_{0\le\varepsilon\le 0.5}(1 - r_\xi(s,\varepsilon))$. This expression shows how much a procedure ξ looses under $T(\varepsilon)$ in comparison to the normal distribution.

Another way of measuring deviations of probability of the correct decision for the state s of the procedure ξ is the efficacy parameter defined by

$$\frac{\sup_{0\le\varepsilon\le 0.5} P_\xi(s,\varepsilon) - \inf_{0\le\varepsilon\le 0.5} P_\xi(s,\varepsilon)}{\int_0^{0.5} P_\xi(s,\varepsilon) d\varepsilon}.$$

The efficacy parameter measures oscillations of the probability of the correct decision relatively to the mean level of the probability. This parameter is an analogue of a coefficient of variability of a random variable.

4. Experiment

We are interested in how much the probability of obtaining a correct division changes if observations are drawn from a distribution

$$T(\varepsilon) = (1-\varepsilon)N(\mu, \sigma^2) + \varepsilon N(\mu, \tau^2).$$

The r-th moment δ_r of the distribution equals to $(1-\varepsilon)\delta'_r + \varepsilon\delta''_r$, where δ'_r, δ''_r are the r-th moments of $N(\mu, \sigma^2)$ and $N(\mu, \tau^2)$, respectively. Hence, the mean value of $T(\varepsilon)$ equals μ and its variance is $(1-\varepsilon)\sigma^2 + \varepsilon\tau^2$.

Our aim was to estimate the probability of a correct decision made by a procedure of multiple comparisons. So, for given means $\mu_1, \mu_2, \ldots, \mu_k$, k samples of size n were drawn from the distributions with appropriate means. Each of the investigated procedures of multiple comparisons was applied to the set of samples and a division of the set of means $\{\mu_1, \mu_2, \ldots, \mu_k\}$ was obtained. The obtained division was compared with given means. If the division agreed with the true one, we decided that the application of a procedure was successful. This procedure was repeated 1000 times and the probability of the correct decision was estimated as the proportion of correct decisions.

In the experiment, $n = 11$, $k = 6$ and $\varepsilon = 0.0, 0.01, 0.05, 0.1, 0.2, 0.4$ were taken. Variances σ^2 and τ^2 were chosen in such a way that the variance of the distribution was equal to 1.

In many papers it is pointed out that the size of the ANOVA test depends on the kurtosis of the underlying distribution. The kurtosis of the distribution $T(\varepsilon)$ equals to $3\{(1-\varepsilon)\sigma^4 + \varepsilon\tau^4 - 1\}$. Parameters of $T(\varepsilon)$ were chosen in such a manner that kurtosis of all the considered distributions was the same, and equal to 4 (for the normal distribution kurtosis is 0). Values of the parameters are shown below.

ε :	0.01	0.05	0.10	0.20	0.40
σ^2 :	0.88395	0.73509	0.61510	0.42265	0.05719
τ^2 :	12.48913	6.03322	4.46410	3.30940	2.41421

Since the probability of detecting the true division by a procedure depends neither on values of means nor on their order, but it does depend on differences between them, so $\mu_1 = 0$ and $\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_3 \leq \mu_5 \leq \mu_6$ were chosen. Hence, 11 configurations of mean values (divisions) were considered. Those configurations are shown in Table 1.

For the numerical experiment values of μ's were needed. Those values were chosen in such a manner that $\mu_i = 0(0.5)5$ and conditions of a division are satisfied. For example, for the division $(1-3, 4, 5, 6)$ the values of means were: $\mu_1 = \mu_2 = \mu_3 = 0$, $\mu_4 = 0.5(0.5)4$, $\mu_5 = \mu_4(0.5)4.5$, $\mu_6 = \mu_5(0.5)5$.

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

Table 1. Divisions of means used in the experiment

Symbol	Denotes division
$(1-6)$	$\mu_1 = \cdots = \mu_6$
$(1-5,6)$	$\mu_1 = \cdots = \mu_5, \mu_6$
$(1-4,5-6)$	$\mu_1 = \cdots = \mu_4, \mu_5 = \cdots = \mu_6$
$(1-3,4-6)$	$\mu_1 = \cdots = \mu_3, \mu_4 = \cdots = \mu_6$
$(1-4,5,6)$	$\mu_1 = \cdots = \mu_4, \mu_5, \mu_6$
$(1-3,4-5,6)$	$\mu_1 = \cdots = \mu_3, \mu_4 = \mu_5, \mu_6$
$(1-2,3-4,5-6)$	$\mu_1 = \mu_2, \mu_3 = \mu_3, \mu_5 = \mu_6$
$(1-3,4,5,6)$	$\mu_1 = \cdots = \mu_3, \mu_4, \mu_5, \mu_6$
$(1-2,3-4,5,6)$	$\mu_1 = \mu_2, \mu_3 = \mu_3, \mu_5, \mu_6$
$(1-2,3,4,5,6)$	$\mu_1 = \mu_2, \mu_3, \mu_3, \mu_5, \mu_6$
$(1,2,3,4,5,6)$	$\mu_1, \mu_2, \mu_3, \mu_3, \mu_5, \mu_6$

5. Results

Results of the Monte Carlo experiment for all configurations of means are shown in Table 2. In the table one may find the average probability (multiplied by 1000) of the correct decision and the relative probabilities with respect to the normal distribution (i.e. without outliers).

Analysis of the probability of making a correct decision by procedures in dependence on the value of ε shows that the probability increases with ε. It may be interpreted as a positive behviour: the procedures are better when outliers are present. It is easier to detect the true division of means when the number of outliers increases. Hence, the investigated procedures may be considered as robust against the presence of outliers. It is clear that

$$\max\{r_\xi(s,\varepsilon) : \varepsilon \in \{0, 0.01, 0.05, 0.10, 0.20, 0.40\}\} = r_\xi(s, 0.40)$$

for all divisions s [except $(1-6)$]. So, robustness of the procedure may be estimated by the numbers given in the last column of Table 2 .

Among the investigated procedures the W procedure seems to be the best for two reasons. One reason is that $r_W(s, 0.40)$ is the smallest, hence it is the most robust procedure. The second reason is that the average probability of the correct decision is the highest. There are only three exceptions when the Newman–Keuls procedure is better than the W procedure.

Table 2. Results of Monte Carlo simulations for different values of ε and different divisions of means

	Average probability						Relative to $\varepsilon = 0$					
ε :	0.00	0.01	0.05	0.10	0.20	0.40	0.00	0.01	0.05	0.10	0.20	0.40
Division $(1-6)$												
W	957.00	939.00	945.00	947.00	945.00	955.00	1	0.981	0.988	0.990	0.988	0.998
Tukey	944.00	936.00	943.00	952.00	949.00	952.00	1	0.992	0.999	1.009	1.005	1.009
Scheffé	987.00	972.00	980.00	986.00	989.00	986.00	1	0.985	0.993	0.999	1.002	0.999
N–K	944.00	936.00	943.00	952.00	949.00	952.00	1	0.992	0.999	1.009	1.005	1.009
Duncan	990.00	979.00	983.00	993.00	990.00	992.00	1	0.989	0.993	1.003	1.000	1.002
Division $(1-5,6)$												
W	791.20	800.20	792.40	792.70	810.30	834.40	1	1.011	1.002	1.002	1.024	1.055
Tukey	706.00	728.50	715.50	717.50	743.70	776.50	1	1.032	1.014	1.016	1.053	1.100
Scheffé	676.00	707.50	690.00	700.30	723.60	764.20	1	1.047	1.021	1.036	1.070	1.130
N–K	762.10	774.80	769.30	766.60	784.10	813.80	1	1.017	1.009	1.006	1.029	1.068
Duncan	732.70	756.50	744.60	750.60	769.30	804.10	1	1.033	1.016	1.024	1.050	1.097
Division $(1-4,5-6)$												
W	817.80	830.10	823.50	823.10	840.00	855.20	1	1.015	1.007	1.007	1.027	1.046
Tukey	690.40	708.80	693.10	699.70	717.70	759.10	1	1.027	1.004	1.014	1.040	1.100
Scheffé	657.20	680.10	668.00	673.60	695.70	742.00	1	1.035	1.016	1.025	1.059	1.129
N–K	697.60	714.40	704.60	702.60	723.60	749.20	1	1.024	1.010	1.007	1.037	1.074
Duncan	708.10	723.60	709.40	715.90	733.00	775.10	1	1.022	1.002	1.011	1.035	1.095
Division $(1-3,4-6)$												
W	826.60	835.30	826.20	836.80	846.80	869.70	1	1.011	1.000	1.012	1.024	1.052
Tukey	687.00	701.10	689.60	703.20	717.00	764.00	1	1.021	1.004	1.024	1.044	1.112
Scheffé	656.20	674.40	661.20	674.50	692.40	744.70	1	1.028	1.008	1.028	1.055	1.135
N–K	697.60	709.40	698.90	707.10	717.60	751.40	1	1.017	1.002	1.014	1.029	1.077
Duncan	702.00	713.60	704.10	716.70	732.30	775.40	1	1.017	1.003	1.021	1.043	1.105
Division $(1-4,5-6)$												
W	387.67	428.40	406.38	418.84	454.47	531.78	1	1.105	1.048	1.080	1.172	1.372
Tukey	264.67	304.29	283.60	296.00	332.20	418.00	1	1.150	1.072	1.118	1.255	1.579
Scheffé	187.82	223.78	204.87	219.13	253.20	344.56	1	1.191	1.091	1.167	1.348	1.835
N–K	427.56	467.62	445.31	457.80	494.44	567.69	1	1.094	1.042	1.071	1.156	1.328
Duncan	310.93	350.27	327.58	341.47	379.53	465.04	1	1.127	1.054	1.098	1.221	1.496
Division $(1-3,4-5,6)$												
W	467.29	500.93	479.02	495.31	525.31	597.91	1	1.072	1.025	1.060	1.124	1.280
Tukey	212.93	251.62	228.98	248.42	280.38	367.29	1	1.182	1.075	1.167	1.317	1.725
Scheffé	145.24	180.42	158.96	175.49	207.96	293.68	1	1.242	1.094	1.208	1.432	2.022
N–K	342.29	380.07	358.60	375.51	405.20	480.44	1	1.110	1.048	1.097	1.184	1.404
Duncan	243.71	283.67	258.47	279.67	312.13	398.44	1	1.164	1.061	1.148	1.281	1.635

Table 2. Continued

	Average probability						Relative to $\varepsilon = 0$					
ε :	0.00	0.01	0.05	0.10	0.20	0.40	0.00	0.01	0.05	0.10	0.20	0.40
Division $(1-2, 3-4, 5-6)$												
W	500.56	530.20	512.89	529.04	555.62	626.67	1	1.059	1.025	1.057	1.110	1.252
Tukey	199.24	238.18	209.98	229.89	264.42	353.36	1	1.195	1.054	1.154	1.327	1.773
Scheffé	131.69	166.71	145.00	160.02	194.64	279.49	1	1.266	1.101	1.215	1.478	2.122
N–K	310.36	344.49	322.31	335.91	368.47	442.24	1	1.110	1.039	1.082	1.187	1.425
Duncan	224.33	264.58	237.04	255.69	291.09	376.82	1	1.179	1.057	1.140	1.298	1.680
Division $(1-3, 4, 5, 6)$												
W	106.44	136.03	118.46	132.43	160.80	242.93	1	1.278	1.113	1.244	1.511	2.282
Tukey	31.37	46.77	37.34	44.89	64.63	124.08	1	1.491	1.191	1.431	2.061	3.956
Scheffé	8.32	16.09	10.76	14.33	25.28	64.50	1	1.935	1.294	1.722	3.040	7.756
N–K	155.59	190.21	168.20	183.81	217.57	305.97	1	1.223	1.081	1.181	1.398	1.967
Duncan	53.28	74.63	62.30	72.23	96.00	166.62	1	1.401	1.169	1.356	1.802	3.128
Division $(1-2, 3-4, 5, 6)$												
W	139.74	173.23	152.77	167.47	199.65	289.65	1	1.240	1.093	1.198	1.429	2.073
Tukey	20.60	33.99	26.19	31.02	46.60	100.00	1	1.650	1.271	1.506	2.262	4.854
Scheffé	5.06	10.42	7.18	9.53	17.18	50.13	1	2.059	1.420	1.886	3.397	9.911
N–K	110.62	139.93	121.98	135.46	162.61	245.53	1	1.265	1.103	1.225	1.470	2.220
Duncan	34.91	52.86	42.18	48.94	68.50	131.38	1	1.514	1.208	1.402	1.962	3.763
Division $(1-2, 3, 4, 5, 6)$												
W	18.59	30.22	22.78	27.51	41.90	92.64	1	1.626	1.226	1.480	2.254	4.984
Tukey	0.62	1.71	1.09	1.58	3.70	18.01	1	2.754	1.762	2.546	5.977	29.085
Scheffé	0.05	0.13	0.07	0.12	0.45	4.38	1	2.700	1.400	2.600	9.500	92.000
N–K	30.93	46.81	37.07	44.47	63.60	127.05	1	1.513	1.198	1.438	2.056	4.107
Duncan	2.65	5.43	3.60	4.85	9.97	34.86	1	2.052	1.358	1.833	3.764	13.167
Division $(1, 2, 3, 4, 5.6)$												
W	4.14	8.44	5.66	7.63	14.00	44.77	1	2.040	1.366	1.845	3.385	10.817
Tukey	0.02	0.02	0.03	0.02	0.14	1.72	1	1.000	1.400	1.200	7.000	86.600
Scheffé	0.00	0.00	0.00	0.00	0.00	0.22	–	–	–	–	–	–
N–K	4.14	8.44	5.65	7.64	14.00	44.76	1	2.038	1.365	1.845	3.384	10.815
Duncan	0.06	0.18	0.12	0.14	0.56	5.22	1	3.067	1.933	2.333	9.467	87.733

The robustness of procedures against non-normality may be also measured by the efficacy parameter

$$\frac{\max\{p_1, p_2, p_3, p_4, p_5, p_6\} - \min\{p_1, p_2, p_3, p_4, p_5, p_6\}}{\bar{p}},$$

where $p_1, \ldots, p_6$ are estimated probabilities of making a correct decision by a procedure for $\varepsilon = 0, 0.01, 0.05, 0.1, 0.2, 0.4$, respectively, and $\bar{p}$ is the average of $p_1, \ldots, p_6$.

The procedure is more robust when its efficacy is smaller. Values of efficacy are shown in Table 3. The parameter is very small for the division $(1-6)$ and much larger for the remaining divisions. It is interesting that we rather cannot see (except for the difference between the first division and others) any dependence between the values of the efficacy parameter and the division. Such a relation can be easily noticed in estimated probabilities of the correct decision. Hence, it may be concluded that relative oscillations of the probability of correct decisions are the same for all the procedures.

Table 3. Efficacy parameter

Division	Tukey	Scheffé	N–K	Duncan	W
$(1-6)$	0.038	0.008	0.038	0.007	0.034
$(1-5,6)$	3.860	3.848	3.858	3.864	3.847
$(1-4,5-6)$	3.848	3.847	3.841	3.851	3.839
$(1-3,4-6)$	3.870	3.872	3.851	3.875	3.857
$(1-4,5,6)$	3.862	3.863	3.851	3.864	3.854
$(1-3,4-5,6)$	3.860	3.860	3.864	3.861	3.867
$(1-2,3-4,5-6)$	3.860	3.855	3.866	3.863	3.867
$(1-3,4,5,6)$	3.877	3.875	3.879	3.879	3.878
$(1-2,3-4,5,6)$	3.873	3.870	3.880	3.871	3.880
$(1-2,3,4,5,6)$	3.878	3.879	3.875	3.877	3.876
$(1,2,3,4,5,6)$	3.871	3.871	3.871	3.871	3.871

6. Conclusions

1. Monte Carlo simulations showed that in the presence of outliers the probability of the correct decision of procedures of multiple comparisons is higher than in the normal case.
2. The results for a special case of six means ($k = 6$) were presented. Very similar results were obtained for $k = 4$ and $k = 5$. It may be expected that for other numbers of means results will be similar. It is difficult to find exact analytical results because formulas for appropriate probabilities are mathematically very complicated.
3. The Monte Carlo experiment was made under the assumption that the kurtosis of the underlying distribution equals 4. It may be very interesting to know how the considered probability depends on the kurtosis of the underlying distribution. Such simulations are in progress.

Acknowledgement

The authors would like to express their thanks to the referee for very detailed and helpful comments which allowed them to improve the paper. This paper was partially supported by the KBN Grant No. 5010 109 00 23.

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.

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.* 2nd ed., Springer Verlag.

Tukey J.W. (1960). A survey of sampling from contaminated distributions. In: *Contributions to Probability and Statistics*, I. Olkin (ed.), 448–485. Stanford University Press, Palo Alto, CA.

Zieliński R. (1994). One–way analysis of variance under Tukey contamination: a small sample case simulation study. In: *Proceedings of the International Conference on Linear Statistical Inference LINSTAT 93*, T. Caliński and R. Kala (eds.), 79-86. Kluwer Academic Publishers.

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 8 October 1998

O odporności procedur porównań wielokrotnych

Streszczenie

Praca dotyczy badania odporności procedur porównań wielokrotnych na nienormalność rozkładów. Jako kryterium odporności przyjmuje się prawdopodobieństwo otrzymania podziału średnich zgodnego z podziałem prawdziwym. Wyniki symulacji Monte Carlo wskazują iż prawdopodobieństwo podjęcia właściwej decyzji wzrasta wraz z proporcją obserwacji które pochodzą z rozkładu zanieczyszczającego dane.

Słowa kluczowe: porównania wielokrotne, wnioskowanie jednoczesne, ANOVA, odporność.