On Rank Based Multiple Comparisons for the Balanced Sample Sizes

František Rublík

Institute of Measurement Science, Slovak Academy of Sciences
Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic
E-mail: umerrubl@savba.sk

Abstract. *It is shown that neither the Nemenyi rule based on the joint ranking, nor the Dwass, Steel, Critchlow-Fligner rule based on the pairwise ranking, is preferable to each other, if the number of the underlying populations is $k = 3$ or $k = 4$, the sample sizes from all sampled populations are the same and the rules are used in conjunction with the Kruskal-Wallis test. The assessment of the multiple comparisons rules is based on the simulation estimates of their effectiveness and reliability, expressed in terms of condional probabilities of good and wrong decision.*

Keywords: *Multiple comparisons procedure; Joint ranking; Pairwise ranking; Kruskal-Wallis test.*

1. Introduction

Consider the independent random samples with sample size n

$$X_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n}) \sim F_i(x)\,, \qquad i = 1, \dots, k\,, \tag{1}$$

i.e., the ith sample is drawn from the population with the continuous distribution function $F_i(x)$. Let $(R_{1,1}, \dots, R_{1,n}, R_{2,1}, \dots, R_{2,n}, \dots, R_{k,1}, \dots, R_{k,n})$ denote the ranks of the pooled sample $(X_1, \dots, X_k)$. The rank based multiple comparisons (MC) method declares the ith and the jth population to be different (or does not distinguish them) according to their ranks. Suppose that

$$F_1 = \dots = F_{k-1} = F\,, \qquad F_k = G\,, \qquad F \neq G\,, \tag{2}$$

where F, G are continuous distribution functions. Since the quality of MC procedure can be assessed by means of the probability of declaring equal distributions as different, the probability of the wrong decision

$$P_n = P\Big[\text{ the } i\text{th and the } j\text{th populations are declared as different for some } i < k,\, j < k\Big] \tag{3}$$

has been studied in [5]; here the index n is to stress that, as is postulated in (1), the sample size for the samples from all the involved distributions equals n, i.e., the experiment is carried out in the scheme of the balanced sampling. Hovever, (3) does not take into account the fact that the MC procedure is usually used in conjunction with the Kruskal-Wallis test (KWT). In this report we include simulation estimates of the quality of the MC procedures from the point of view of behaviour of the MC method used after rejection the hypothesis of equality of sampled populations by the KWT.

2. Particular methods of multiple comparisons and the assessment of their quality

In the approach worked out by Nemenyi in [3] the MC procedure is based on the joint ranking. Let

$$S_i = R_{i1} + \dots + R_{in}\,, \qquad i = 1, \dots, k \tag{4}$$

denote the sum of ranks belonging to the ith sample. The Nemenyi procedure declares as different the ith and the jth distribution if

$$|S_i - S_j| > nq_k^\alpha \sqrt{\frac{k(kn+1)}{12}}\,, \tag{5}$$

where q_k^α denotes the upper α critical point of the maximum modulus of k independent standardnormal variables, i.e.,

$$P\Big(\max_{1 \le i,j \le k} |y_i - y_j| > q_k^{(\alpha)} \Big| \mathcal{L}(y) = N_k(\mathbf{0}, \mathbf{I}_k) \Big) = \alpha\,.$$

Thus under the assumption (2), if (5) holds for some $i \neq j$ which are smaller than k, then this procedure falsely declares the ith and the jth distribution to be different. The limiting probability

$$\alpha(F,G) = \lim_{n\to\infty} P_n \tag{6}$$

of this error is in this case

$$\alpha(F,G) = \lim_{n\to\infty} P\Big[\max_{i,j \le k-1} |S_i - S_j| > nq_k^\alpha \sqrt{\frac{k(kn+1)}{12}} \Big]\,, \tag{7}$$

the distribution function G is included into this notation because it influences the distribution of the ranks of the pooled sample. According to table 3.1 of [5] the maximum values $M(\alpha)$ of $\alpha(F,G)$, where F, G range over the set of all continuous functions, are for $\alpha = 0.05$ and $\alpha = 0.1$ as follows.

k	3	4	5	6	7	8	9	10
M(0.05)	0.0512	0.0643	0.0682	0.0690	0.0688	0.0682	0.0674	0.0667
M(0.1)	0.0877	0.1123	0.1208	0.1240	0.1250	0.1250	0.1245	0.1238

(8)

The Nemenyi procedure is used in the scheme

$$F_i(x) \equiv F(x - \mu_i)\,, \qquad i = 1, \ldots, k \tag{9}$$

where $\mu_1, \ldots, \mu_k$ are real numbers. In this case the supremum of (6)

$$\alpha(F) = \sup\{\, \alpha(F,G);\;\; G(x) \equiv F(x-c)\,,\;\; c \textit{ is a real number}\}\,.$$

According to Theorem 6.1 and the argument on p. 83 of [5], if F possesses a density f with respect to the Lebesque measure and $\log(f)$ is concave, then $\alpha(F) \le \alpha$.

The left-hand side of the inequality (5) depends also on the values of the samples X_t , $t \notin \{i,j\}$, so that the comparison of the ith and the jth population is affected by the observations from the remaining populations. This property does not have the approach based on the pairwise ranking.

Let $i \neq j$ and $R_1^{(i,j)}, \ldots, R_{n_j}^{(i,j)}$ denote the ranks of $X_{j,1}, \ldots, X_{j,n}$ in the sample $(X_{i,1}, \ldots, X_{i,n}, X_{j,1}, \ldots, X_{j,n})$ pooled from X_i, X_j. Compute

$$S_{i,j} = \sum_{b=1}^{n} R_b^{(i,j)}\,, \qquad S_{i,j}^* = \frac{S_{i,j} - \dfrac{n(2n+1)}{2}}{\sqrt{\dfrac{n^2(2n+1)}{24}}}\,. \tag{10}$$

Let $1 \le i < j \le k$. The DSCF (Dwass, Steel, Critchlow-Fligner) rule is

$$\textit{declare} \quad F_i \neq F_j \quad \textit{if} \quad |S_{i,j}^*| > q_k^\alpha\,. \tag{11}$$

We remark that here the condition $i < j$ is imposed only to achieve smaler computational complexity, because $|S_{i,j}^*| = |S_{j,i}^*|$. In difference from the Nemenyi method in this pairwise setting the comparison of the ith and the jth population is not affected by the observation from the remaining populations. Moreover, in the setting (2) and (9) the limit of the probability P_n in (3) as $n \to \infty$ does not exceed α. For these reasons the rule (11) is included into the monograph [2] in contradistinction to [1], where only the classical rule (5) is used.

As can be seen from (8), the size of exceeding the nominal α is from the limiting point of view acceptable and the natural question is what will be the effect of this discordance in the finite sample case. The concordance of (7) with α is mainly of aesthetic value, because the MC method is used in conjunction with the KWT in such a way that the MC rule is applied only when the KWT rejects the null hypothesis (cf. (1))

$$F_1 = \ldots = F_k\,. \tag{12}$$

The rationale for this is the opinion that the violation of (12) is caused by those populations for which the quantities (4) differ in some significant way. However, what matters from the statistical point of view, is the effect of such a procedure. Since the MC method is used in conjunction with KWT, in accordance with the approach proposed in [4] we shall assess the performance of the MC method by the conditional probability of the good and the wrong decision as follows.

Label A the random event that the KWT rejects (12). Let *gd* be the conditional probability that the particular multiple comparisons method makes good decision in the sense that it declares at least one pair of different populations as being different, the conditioning is made with respect to A and everything (including the KWT) is carried out at $\boldsymbol{\alpha = 0.05}$. Thus

$$\text{gd} = P(\textit{ method correctly detects at least one pair of different populations} \mid A\,) \tag{13}$$

and the detection of different pairs is carried out with the particular method at $\alpha = 0.05$. Similarly, the symbol *wd* denotes the conditional probability of the wrong decision, i.e.,

$$wd = P(\textit{ method wrongly declares at least one pair of identical populations as different} \mid A\,) \tag{14}$$

and the detection of different pairs is carried out with the particular method at $\alpha = 0.05$.

3. Some simulation results

The effect of the rule (5), which will be referred to as NR (Nemenyi rule) and the effect of the DSCF rule (11) were investigated by means of simulation estimates of gd and wd, obtained in each case from $N = 20000$ trials, n is the sample size defined in (1), μ_j the location parameter of the jth population and $\sigma_j = \sigma$ (for normal and and Cauchy populations) is the scale parameter of the jth population (i.e., all k sampled populations have the same scale parameter). The boldface typed numbers indicate better performance, i.e., the greater value of the probability *gd* of the good decision and the smaller value of the probability *wd* of the wrong decision.

First we present some results concerning $k = 3$ populations.

	Normal distribution, $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1.25$, $\sigma = 1$							
n	10		20		30		50	
P(A)	0.074		0.108		0.144		0.216	
Method	NR	DSCF	NR	DSCF	NR	DSCF	NR	DSCF
gd	**0.744**	0.719	**0.823**	0.788	**0.847**	0.831	**0.905**	0.897
wd	**0.190**	0.203	0.153	0.155	0.118	0.120	**0.081**	0.082

	Cauchy distribution, $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1.8$, $\sigma = 1$							
n	10		20		30		50	
P(A)	0.137		0.248		0.372		0.570	
Method	NR	DSCF	NR	DSCF	NR	DSCF	NR	DSCF
gd	**0.850**	0.815	**0.910**	0.878	**0.935**	0.911	**0.957**	0.948
wd	**0.110**	0.125	**0.062**	0.068	**0.042**	0.047	**0.030**	0.034

	χ^2 distribution with 2 degrees of freedom, $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1.25$							
n	10		20		30		50	
P(A)	0.067		0.09		0.11		0.165	
Method	NR	DSCF	NR	DSCF	NR	DSCF	NR	DSCF
gd	0.658	**0.697**	**0.764**	0.762	0.805	**0.821**	0.845	**0.868**
wd	0.306	**0.273**	0.222	**0.193**	0.189	**0.162**	0.134	**0.118**

The following tables concern the case of $k = 4$ sampled populations.

	Normal distribution $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1.35$, $\mu_4 = 1.35$, $\sigma = 1$							
n	10		20		30		50	
P(A)	0.107		0.201		0.310		0.496	
Method	NR	DSCF	NR	DSCF	NR	DSCF	NR	DSCF
gd	**0.763**	0.703	**0.821**	0.780	**0.846**	0.823	**0.875**	0.863
wd	**0.121**	0.138	**0.079**	0.089	**0.055**	0.059	**0.035**	0.037

	Normal distribution $\mu_1 = 1$, $\mu_2 = 1.3$, $\mu_3 = 1.3$, $\mu_4 = 1.7$, $\sigma = 1$							
n	10		20		30		50	
P(A)	0.183		0.389		0.576		0.826	
Method	NR	DSCF	NR	DSCF	NR	DSCF	NR	DSCF
gd	**0.851**	0.777	**0.919**	0.886	**0.948**	0.933	**0.979**	0.972
wd	**0.040**	0.042	**0.019**	0.022	**0.016**	0.018	**0.011**	0.012

	χ^2 distribution 1 degree of freedom $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1$, $\mu_4 = 1.35$							
n	10		20		30		50	
P(A)	0.185		0.409		0.609		0.877	
Method	NR	DSCF	NR	DSCF	NR	DSCF	NR	DSCF
gd	0.776	**0.820**	0.896	**0.926**	0.941	**0.961**	0.978	**0.985**
wd	0.174	**0.135**	0.083	**0.061**	0.061	**0.044**	0.042	**0.031**

4. Discussion

Because of the space limitation only a few simulation results are included, but with the aim to choose results representing the variety of the behaviour of the underlying MC methods under various possibilities of the alternative. From the results one can see that both for $k = 3$ and $k = 4$ each of the considered NR and DSCF methods has the property that in some cases it exhibits large value (14) of the probability of the wrong decision *wd* while in some other cases it has the probability (13) of the good decision *gd* visibly greater than the other method. Therefore taking into account these results (and also the simulations not included into this report) one may conclude that there is no clear reason for prefering either NR or DSCF multiple comparisons method in the case $k = 3$ and $k = 4$.

References

[1] Holländer, M. and Wolfe, D. A. Nonparametric Statistical Methods. J. Wiley, New York, 1973.

[2] Holländer, M. and Wolfe, D. A. Nonparametric Statistical Methods. J. Wiley, New York, 1999.

[3] Miller, R. G. Simultaneous Statistical Inference. Springer Verlag, Berlin 1981.

[4] Rublík, F. On performance of multiple comparison methods used in conjunction with the Kruskal-Wallis test. *Measurement 2001, Proceedings, 3rd International Conference on Measurement,* Smolenice 2001, 36 – 40.

[5] Voshaar, J. H. O. (k-1)-mean significance levels of nonparametric multiple comparisons procedures. *The Annals of Statistics*, 8(1980), 75 – 86.