MEASUREMENT SCIENCE REVIEW, Volume 1, Number 1, 2001

ON PERFORMANCE OF MULTIPLE COMPARISON METHODS USED IN CONJUNCTION WITH THE KRUSKAL-WALLIS TEST

František Rublík

Institute of Measurement Science SAS, Dúbravská cesta 9, 84219 Bratislava, Slovakia
E-mail: umerrubl@savba.sk

Abstract. It is shown that the best choice of the multiple comparison method is the exact one and that the Conover method possesses a high risk because it overvalues differences between populations ranks. The results of simulations show that in the balanced case the asymptotic approximation of the exact critical constant is good even for small joint sample sizes.

Let for $i = 1, \ldots, k$ the random samples $X_{i1}, X_{i2}, \ldots, X_{in_i}$ from $X_i = \mu_i + \varepsilon$, where ε has a continuous distribution function, be independent. Let

$$T_{n_1,\ldots,n_k} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{S_i^2}{n_i} - 3(n+1)\,, \qquad n = n_1 + \ldots + n_k\,,$$

be the Kruskal-Wallis test statistic, i.e., $S_i = R_{i1} + \ldots + R_{in_i}$ denotes the sum of ranks belonging to the ith sample. Further, let for constant $h(\alpha, n_1, \ldots, n_k)$ the equality

$$P(T_{n_1,\ldots,n_k} \geq h(\alpha, n_1, \ldots, n_k)) = \alpha \tag{1}$$

be fulfilled whenever the hypothesis

$$\mu_1 = \ldots = \mu_k \tag{2}$$

holds. The Kruskal-Wallis test KWT rejects (2) if $T_{n_1,\ldots,n_k} \geq h(\alpha, n_1, \ldots, n_k)$, and accepts it otherwise. If the exact value of the constant h fulfilling (1) is not available, then the approximation $h(\alpha, n_1, \ldots, n_k) = \chi^2_{k-1}(\alpha)$ by the critical α value of the chi-square distribution is used and the null hypothesis (2) is rejected whenever $T_{n_1,\ldots,n_k} > \chi^2_{k-1}(\alpha)$. There are several methods of multiple comparisons used for detecting different populations after (2) is rejected with the KWT.

Put

$$D_{i,i^*} = \frac{\left|\dfrac{S_i}{n_i} - \dfrac{S_{i^*}}{n_{i^*}}\right|}{\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_{i^*}}}}\,. \tag{3}$$

It is proved in [4] that if $T_{n_1,\ldots,n_k} \leq t$, then $\max_{i,i^*} D_{i,i^*} \leq \sqrt{t\frac{n(n+1)}{12}}$ (for the proof cf. also p. 133 of [5]). Hence if the KWT rejects (2), it is logical to consider responsible for the

rejection the populations violating this inequality. Therefore after the rejecting (2) the ith and the jth population are declared to be different, if

$$D_{i,j} \geq \sqrt{h(\alpha, n_1, \ldots, n_k)\frac{n(n+1)}{12}}\,. \tag{4}$$

This method is in [3] called the **conservative method**, because under validity of (2)

$$P\left(\max_{i,i^*} D_{i,i^*} > \sqrt{h(\alpha, n_1, \ldots, n_k)\frac{n(n+1)}{12}}\right) \leq \alpha\,, \tag{5}$$

and typically, the left hand side of this inequality is visibly below its upper limit.

Suppose that the constant $c_2(\alpha, n_1, \ldots, n_k)$ is such that

$$P\left(\max_{i,i^*} D_{i,i^*} \geq c_2(\alpha, n_1, \ldots, n_k)\right) = \alpha \tag{6}$$

whenever (2) holds. After rejecting the null hypothesis (2) by the KWT, the ith and the jth population are declared to be different if

$$D_{i,j} \geq c_2(\alpha, n_1, \ldots, n_k)\,. \tag{7}$$

This method is in the further text refered as the **exact method**.

Another method is presented in [2] on p. 231. After the rejection of (2) by the KWT the ith and the jth population are declared to be different if

$$D_{i,j} > t_{1-\alpha/2}(n-k)\sqrt{\frac{n(n+1)}{12}\frac{(n-1-T_{n_1,\ldots,n_k})}{n-k}}\,, \tag{8}$$

where $t_{1-\alpha/2}(n-k)$ denotes the $1-\alpha/2$ quantile of the Student t distribution with $n-k$ degrees of freedom and $T_{n_1,\ldots,n_k}$ is the Kruskal-Wallis test statistic. This method will be refered as the **Conover method**.

To assess the performance of these methods label A the random event that the KWT rejects (2). Let *gd* be the conditional probability that the particular 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 made at $\boldsymbol{\alpha = 0.05}$. Thus

$$\mathrm{gd} = P(\textit{ method correctly detects at least one pair of different populations} \,|\, A\,)$$

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.,

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

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

In the following table n_i denotes the sample size from the $N(\mu_i, 1)$ distribution and $P(A)$ the simulation estimate of this probability. Its value is included into the table to indicate the reliability of the simulation estimate of the conditional probabilities, because in the formula $P(B|A) = P(B \cap A)/P(A)$ in the case of the small value of $P(A)$ even a small discrepancy between $P(A)$ and its simulation estimate can destroy accuracy of the estimate of $P(B|A)$. For this reason the values of the means μ_i were chosen in such a way that the probabilities $P(A)$ are reasonably far from 0; since the exact value of the rejection constant h in the KWT is available only in a limited range of cases, the values of the sample sizes n_i are chosen so as the tables of h in [3] could be used.

The abbreviation *cons* means the conservative method and *cono* the Conover method, the simulation results of the table 1 were obtained from $N = 20000$ trials for each particular case.

Method	cons	exact	cono	cons	exact	cono	cons	exact	cono
$n_1, n_2, n_3,\ P(A)$	3, 3, 4, 0.33			3, 3, 5, 0.44			3, 4, 4, 0.38		
gd	0.50	0.80	0.99	0.62	0.85	1	0.68	0.89	1
wd	0.00	0.01	0.25	0.00	0.00	0.18	0.01	0.01	0.19
$n_1, n_2, n_3,\ P(A)$	3, 4, 5, 0.45			3, 5, 5, 0.49			4, 4, 5, 0.47		
gd	0.76	0.88	1	0.76	0.86	1	0.82	0.92	1
wd	0.00	0.01	0.15	0.01	0.01	0.12	0.00	0.01	0.13
$n_1, n_2, n_3,\ P(A)$	4, 5, 5, 0.51								
droz	0.82	0.95	1						
zroz	0.01	0.01	0.13						

Table 1. Simulation estimates of the performance of particular methods for $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 2.5$.

Since under validity of (2) the left hand side of (5) is expected to be visibly smaller than α, in these situations the constant in (4) will be larger than the constant in (7), which suggests that there should be some difference between the conservative and the exact method in favor of the exact one. This expected superiority is confirmed also by the simulation results presented in the previous table, because in all considered cases the exact method has clearly better conditional probability *gd* than the conservative method, which unambiguously compensates for the fact that in some cases there is a slight difference in the quantity *wd* in favor of the conservative method.

The disadvantage of the exact method is that it requires knowledge of special constants depending on the sample sizes, and the easiest way of obtaining them is by a simulation process. Since this could be either time consuming or (for a large part of users) difficult, an important competitor of this method is the Conover method, requiring only knowledge of the well known quantiles of the Student distribution. However, the simulation results

presented in the previous table suggest that this method exhibits a tendency to overestimate distinctness of the underlying populations and tends to declare identical populations as different strikingly more often than its considered competitors. Therefore, before using it one has to decide between the following options. The one is to use the Conover method to have a better chance to reveal any difference. However, if a safeguarding against declaring identical populations as different plays also a role, than for small or moderate sample sizes one should use rather the conservative method (usually based on the chi-square quantile), which although having worse performance in the sense of *gd* exhibits negligible values of *wd*.

The following method is presented in [4] for the balanced case. Let $q^{\alpha}_{k,\infty}$ denotes the constant for which

$$P\left(\max_{i,i^*}|y_i - y_{i^*}| \geq q^{\alpha}_{k,\infty}\,\Big|N_k(0, I_k)\right) = \alpha\,.$$

If $n_1 = n_2 = \ldots = n_k = n$ and the KWT rejects (2), then the ith and the jth population are declared to be different if

$$|S_i - S_j| \geq ca(n), \qquad ca(n) = n\sqrt{\frac{k(kn+1)}{12}}q^{\alpha}_{k,\infty}$$

(a more detailed explanation of this procedure is also in [5], pp. 133 - 137). This method is in the further text refered as the **approximative method**. The exact method can be in the balanced case expressed by means of the constant c_3 such that

$$P(\max_{i,i^*}|S_i - S_{i^*}| \geq c_3(\alpha, n_1, \ldots, n_k)) = \alpha\,, \tag{9}$$

its rule is that after the rejection of (2) the populations i, j are declared to be different, if

$$|S_i - S_j| \geq c_3(\alpha, n_1, \ldots, n_k)\,.$$

Obviously, the approximative method is an asymptotic approximation of the exact method. It is therefore logical to ask about precision of this asymptotic and whether the difference between these two methods should be considered as significant.

To investigate this problem we carried out simulations to obtain c_3. Since the random variable $\max_{i,j}|S_i - S_j|$ is discrete, in general one cannot find a constant c_3 fulfilling (9). Therefore we have chosen c_3 in such a way that the difference between the left-hand side of (9) and $\alpha = 0.05$ is as small as possible.

The following table contains simulation estimates of the constants $K1$, $K2$ such that under validity of (2) both the inequalities $P1 = P(\max_{i,j}|S_i - S_j| \geq K1) \geq 0.05$, $P2 = P(\max_{i,j}|S_i - S_j| \geq K2) \leq 0.05$ hold and $K2 = K1 + 1$. The differences $d_1 = |P1 - 0.05|$, $d_2 = |P2 - 0.05|$ between the estimated probabilities and the nominal value are also included and the value of the konstant K for which this difference is minimal, is printed in bold and is used as the value of c_3 in (9). Further, together with the value of $q = q^{0.05}_{k,\infty}$

taken from the table A10 of [3], the values of $ca(n)$ and simulation estimates of $Pca = P(\max_{i,j} |S_i - S_j| \geq ca(n)) \geq 0.05)$, $da = |Pca - 0.05|$ are presented: n stands for the joint value $n_1 = \ldots = n_k = n$ of the sample sizes of the samples from k identical continuous populations. Here should be noted that the values of $ca(n)$ can be found in table XVIII.11 on p. 337 of [1].

The results of the next table show that the approximate constant $ca(n)$ well coincides with c_3 even for small values of the joint sample size n. In the overwhelming majority of the considered cases $ca(n)$ slightly overvalues the true magnitude of c_3. Therefore if the table 2 is not at hand or if the concerned values of k,n are not included into the table 2, it is advisable to round $ca(n)$ to the nearest smaller integer.

$q_{k,\infty}^{0.05}$	3.314			3.633			3.858		
		3			4			5	
	K1	K2	ca(n)	K1	K2	ca(n)	K1	K2	ca(n)
	P1	P2	Pca	P1	P2	Pca	P1	P2	Pca
	d_1	d_2	da	d_1	d_2	da	d_1	d_2	da
k		3			4			5	
n									
	15	16	15.7	21	**22**	22.7	**28**	29	29.9
3	0.0661	0.0286	0.0286	0.0728	0.0432	0.0243	0.0572	0.0358	0.0216
	0.0161	0.0214	0.0214	0.0228	0.0068	0.0257	0.0072	0.0142	0.0284
	23	**24**	23.9	33	**34**	34.6	**44**	45	45.6
4	0.0633	0.0462	0.0462	0.0613	0.0475	0.0359	0.0513	0.0413	0.0321
	0.0133	0.0038	0.0038	0.0133	0.0025	0.0141	0.0013	0.0087	0.0179
	32	**33**	33.1	**47**	48	48.1	61	**62**	63.5
5	0.0608	0.0494	0.0389	0.0527	0.0437	0.0357	0.0569	0.0494	0.0363
	0.0108	0.0006	0.0111	0.0027	0.0063	0.0143	0.0069	0.0006	0.0137
	42	**43**	43.3	61	**62**	62.9	80	**81**	83.2
6	0.0568	0.0480	0.0398	0.0537	0.0478	0.0421	0.0552	0.0489	0.0355
	0.0068	0.0020	0.0102	0.0037	0.0022	0.0079	0.0052	0.0011	0.0145
	53	**54**	54.4	77	**78**	79.1	102	**103**	104.6
7	0.0541	0.0486	0.0427	0.0535	0.0484	0.0406	0.0522	0.0484	0.0413
	0.0041	0.0014	0.0073	0.0035	0.0016	0.0094	0.0022	0.0016	0.0087
	66	67	66.3	94	**95**	96.4	**125**	126	127.6
8	0.0505	0.0453	0.0453	0.0546	0.0497	0.0423	0.0504	0.0472	0.0410
	0.0005	0.0047	0.0047	0.0046	0.0003	0.0077	0.0004	0.0028	0.0090
	78	79	78.9	113	**114**	114.8	149	**150**	152
9	0.0516	0.0484	0.0484	0.0517	0.0486	0.0455	0.0516	0.0493	0.0444
	0.0016	0.0016	0.0016	0.0017	0.0014	0.0045	0.0016	0.0007	0.0056

Table 2. Simulation estimates of the constant c_3 obtained from $N = 25000$ trials for each considered case.

Even though according to the previous table $ca(n)$ well coincides with c_3, simulations show that the use of c_3 yields results better than the approximative method with $ca(n)$, and that the difference in quality diminishes with the increase of n; because of the lack of space

further simulations are not included into this paper. Some more extensive simulations related to this matter can be found in [6].

References

[1] Anděl, J. *Matematická statistika*. SNTL, Praha, 1985.

[2] Conover, W. J. *Practical Nonparametric Methods*. J. Wiley, New York, 1980.

[3] Hollander, M. and Wolfe, D. A. *Nonparametric Statistical Methods*. J. Wiley, New York, 1973.

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

[5] Rublík, F. *Neparametrické metódy a štatistická kontrola akosti*. Skriptá, Univerzita Komenského, Bratislava, 1993.

[6] Rublík, F. Methods of multiple comparisons used after rejection of equality by means of the Kruskal-Wallis test. Technical Report, Institute of Measurement Science of the Slovak Academy of Sciences, Bratislava, January 2001.