Biometrical Journal **43** (2001) 5, 571–580

Nonparametric All-Pairs Multiple Comparisons

Markus Neuhäuser

Department of Biometry
Byk Gulden Pharmaceuticals
Konstanz, Germany

Frank Bretz

Unit of Bioinformatics
University of Hannover
Hannover, Germany

Summary

Nonparametric all-pairs multiple comparisons based on pairwise rankings can be performed in the one-way design with the Steel-Dwass procedure. To apply this test, Wilcoxon's rank sum statistic is calculated for all pairs of groups; the maximum of the rank sums is the test statistic. We provide exact calculations of the asymptotic critical values (and *P*-values, respectively) even for unbalanced designs. We recommend this asymptotic method whenever large sample sizes are present. For small sample sizes we recommend the use of the new statistic according to Baumgartner, Weiss, and Schindler (1998, *Biometrics* **54**, 1129–1135) instead of Wilcoxon's rank sum for the multiple comparisons. We show that the resultant procedure can be less conservative and, according to simulation results, more powerful than the original Steel-Dwass procedure. We illustrate the methods with a practical data set.

Key words: Pairwise rankings; Steel-Dwass procedure; Multivariate normal distribution; Baumgartner-Weiß-Schindler statistic

1. Introduction

In this paper we consider nonparametric all-pairs multiple comparisons in the one-way analysis of variance. A one-way design with k ($k > 2$) groups is often used in practice, and it occurs frequently that the normal assumption is not tenable. The observations within each sample are assumed to be independent and identically distributed. Independence between the samples is also assumed.

In the location-shift model the continuous distribution functions $F_1, F_2, \ldots, F_k$ of the k groups are the same except perhaps for a change in their locations: $F_i(t) = F(t - \theta_i)$, $i = 1, \ldots, k$. Under the null hypothesis H_0 there are no differences between the distribution functions, that is, $\theta_1 = \ldots = \theta_k$. The aim of multi-

 0323-3847/01/0509-0571 $ 17.50+.50/0

ple comparisons between all possible pairs of populations or treatments is to identify the pairs with $\theta_i \neq \theta_j$. Let n_i be the sample size of group i, $i = 1, \ldots, k$, and $X_{i1}, \ldots, X_{in_i}$ the i-th random sample. The total sample size $\sum_{i=1}^{k} n_i$ is denoted by N.

In a one-way design nonparametric multiple comparisons between the pairs of groups can be performed with two different approaches. One approach uses the method of jointly ranking all N observations. The other approach is based on the method of pairwise rankings; see MILLER (1981, chapter 4) or CRITCHLOW and FLIGNER (1991) for a discussion of both approaches. One important drawback of the joint ranking approach is that the outcome of the comparison between two groups depends upon the observations in the other samples. Consequently, for the same values $X_{i1}, \ldots X_{in_i}$ and $X_{j1}, \ldots X_{jn_j}$, the comparison i versus j can be significant in one experiment and not significant in another. MILLER (1981) recommended the pairwise rank approach, and in this paper we only consider this approach.

In order to perform all-pairs multiple comparisons, STEEL (1960) and DWASS (1960) proposed the use of the maximum of the $\binom{k}{2}$ two-sample Wilcoxon statistics. This Steel-Dwass method, which possesses several desirable properties (MORLEY, 1982), is based on pairwise rankings. It was generalized for the case of unequal sample sizes by CRITCHLOW and FLIGNER (1991).

Let

$$W_{ij} = \frac{S_{ij} - n_i(n_i + n_j + 1)/2}{\sqrt{n_i n_j (n_i + n_j + 1)/24}}$$

be the standardized Wilcoxon statistic (multiplied by $\sqrt{2}$), where S_{ij} is the rank sum associated with the i-th sample when ranked with the j-th sample for each pair (i, j) with $1 \leq i < j \leq k$ (a formal definition of S_{ij} is given below in Section 2).

The maximum $\max_{i<j} |W_{ij}|$ is the test statistic of the Steel-Dwass procedure which declares $\theta_i \neq \theta_j$ for each pair with $|W_{ij}| \geq w_\alpha$, where w_α is chosen so that

$$\Pr_{H_0}(\max_{i<j} |W_{ij}| \geq w_\alpha) \leq \alpha\,.$$

CRITCHLOW and FLIGNER (1991) tabulated w_α values for small sample sizes and provided a large sample approximation. However, their method only approximates the critical values in cases of unbalanced sample sizes. A second approximation for large balanced sample sizes compares $|W_{ij}|$ with the upper percentage point of the studentized range (see, for example, HOCHBERG and TAMHANE (1987, p. 244)). In Section 2, we provide exact calculations of the asymptotic critical values even for unbalanced designs. Instead of Wilcoxon's rank sum, the nonparametric two-sample test recently proposed by BAUMGARTNER, WEISS, and SCHINDLER (1998) can be used for the multiple comparisons (see Section 3). The two resultant Steel-Dwass procedures are compared in Section 4. In Section 5, the use of the multiple comparison procedures is demonstrated with a practical data set.

2. Exact calculation of asymptotic critical values

We first notice that the joint distribution under H_0 of the $k(k-1)/2$–component vector $\boldsymbol{W} = (W_{12}, W_{13}, \ldots, W_{k-1,k})$ follows asymptotically a multivariate normal distribution with the expectation vector $\boldsymbol{O}$ and a given correlation matrix $\boldsymbol{R}$. The entries of $\boldsymbol{R}$ have been determined in the balanced case by MILLER (1981, chapter 4), see also CRITCHLOW and FLIGNER (1991). In the general unbalanced case the entries of $\boldsymbol{R}$ are obtained as follows.

For each quadruple $(i,j) \times (l,m)$ we are interested in determining $\rho_{(i,j),(l,m)} = \mathrm{Corr}\,(S_{ij}, S_{lm})$ for $S_{ij} = \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} I(X_{ia} - X_{jb}) + \dfrac{n_i(n_i+1)}{2}$, where the indicator function $I(x) = 1$ for $x > 0$ and $I(x) = 0$ otherwise for $i,j,l,m = 1, \ldots, k,\ i < j,\ l < m$. Since $\mathrm{Var}\,(S_{ij}) = n_i n_j (n_i + n_j + 1)/12$ and $\mathrm{Cov}\,(S_{ij}, S_{im}) = n_i n_j n_m/12$, one obtains after some algebraic transformations for $i < j$ and $l < m$

$$\rho_{(i,j),(l,m)} = \begin{cases} \sqrt{\dfrac{n_j n_m}{(n_i + n_j + 1)\,(n_i + n_m + 1)}} & \text{for } i = l,\ j \neq m \\[2ex] \sqrt{\dfrac{n_i n_l}{(n_i + n_j + 1)\,(n_l + n_j + 1)}} & \text{for } i \neq l,\ j = m \\[2ex] -\sqrt{\dfrac{n_i n_m}{(n_i + n_j + 1)(n_j + n_m + 1)}} & \text{for } j = l \\[2ex] 0 & \text{otherwise}\,. \end{cases}$$

Note that in the balanced case $n = n_i = n_j = n_l = n_m$ the entries reduce to the common correlations $\dfrac{n}{2n+1}, \dfrac{-n}{2n+1}$ and 0.

With these results we are able to perform the exact calculation of the asymptotic P-values or critical values. Adequate integration routines, such as that of GENZ (1992), enable the user to obtain the values within a reasonable time for practical accuracies on any Pentium II or higher. We note that the actual $k(k-1)/2$-dimensional integration problem reduces to a $(k-1)$–dimensional one if appropriate prior transformations are applied, such as those of GENZ and KWONG (2000). SAS/IML and Fortran programs for the computation of multivariate normal probabilities are available from the authors upon request.

3. The Baumgartner-Weiß-Schindler statistic

Recently, BAUMGARTNER et al. (1998) presented a novel nonparametric two-sample test. Let $R_1 < \ldots < R_{n_i}$ and $H_1 < \ldots < H_{n_j}$ denote the combined-samples ranks (in increasing order of magnitude) in the two considered groups i and j. The proposed test statistic is $B_{ij} = \frac{1}{2} \cdot (B_X^{ij} + B_Y^{ij})$, where

$$B_X^{ij} = \frac{1}{n_i} \sum_{r=1}^{n_i} \frac{\left(R_r - \frac{n_i + n_j}{n_i} \cdot r\right)^2}{\frac{r}{n_i + 1} \cdot \left(1 - \frac{i}{n_i + 1}\right) \cdot \frac{n_j(n_i + n_j)}{n_i}}$$

and

$$B_Y^{ij} = \frac{1}{n_j} \sum_{s=1}^{n_j} \frac{\left(H_s - \frac{n_i + n_j}{n_j} \cdot s\right)^2}{\frac{s}{n_j + 1} \cdot \left(1 - \frac{j}{n_j + 1}\right) \cdot \frac{n_i(n_i + n_j)}{n_j}} .$$

This test uses the square norm of the difference between the empirical distributions weighted by its variance. Large values of B_{ij} indicate that $\theta_i \neq \theta_j$ (for details concerning the motivation behind this statistic, see the paper by BAUMGARTNER et al.).

BAUMGARTNER et al. (1998) used the asymptotic distribution of B_{ij} for the comparison against other nonparametric tests and their simulation results indicated that the new rank test is at least as powerful as the Wilcoxon test. Instead of an asymptotic test, exact permutation tests can be performed by generating the entire permutation null distribution of a rank-based test statistic. The P-value of a permutation test is the proportion of permutations yielding a statistic as supportive or more supportive of the alternative than the original observed test statistic (see GOOD (1994) for details).

Comparing exact tests, the test based on the statistic B_{ij} has a less conservative size and is, according to simulation results, more powerful than the Wilcoxon test (NEUHÄUSER, 2000). According to these results it may be useful to replace the Wilcoxon statistic in the Steel-Dwass procedure by the new statistic of BAUMGARTNER et al. (1998). Therefore, we propose to use $\max_{i<j} B_{ij}$ as the test statistic. Thus, we conclude $\theta_i \neq \theta_j$ for each pair with $B_{ij} \geq b_\alpha$, where b_α is chosen so that

$$\Pr_{H_0}\left(\max_{i<j} B_{ij} \geq b_\alpha\right) \leq \alpha .$$

Please note that, in a design with unequal sample sizes, the statistics B_{ij} have to be standardized prior to maximization.

We determined b_α values by the permutation approach. However, in case of three or more groups (and/or large sample sizes) the number of permutations and, consequently, the resulting computer time can be enormous. Therefore, the permutation tests were performed simulation-based by taking a simple random sample from all the possible permutations. To be precise, we randomly generated 50 000 permutations to estimate the null distributions. This approach was used to determine both b_α and w_α values since CRITCHLOW and FLIGNER (1991) tabulated w_α

for $k = 3$ and $n_i \leq 7$ only. It should be noted that a SAS/IML program for simulation-based nonparametric k-sample tests was given by BERRY (1995).

In the following section, the Steel-Dwass procedures with the two-sample statistics of Wilcoxon and Baumgartner et al. are compared for small sample sizes with regard to size and power. The following abbreviations will be used: SDW for the original Steel-Dwass test based on W_{ij}, and SDB for the new Steel-Dwass test using the statistic B_{ij}.

4. Comparison of the procedures SDW and SDB

We compared the tests SDW and SDB based on $\max_{i<j} |W_{ij}|$ and $\max_{i<j} B_{ij}$, respectively, in a Monte Carlo simulation study performed with SAS version 6.12.

Table 1 presents critical values w_α and b_α for SDW and SDB. With regard to the small sample size 5 per group the null distributions are too discrete for small nominal α-levels. For $k = 7$ or 8 even $\alpha = 0.1$ is not possible; therefore, these configurations ($k \geq 7$ and $n_i = 5\ \forall i$) are not displayed in Table 1. For $n_i = 5$,

Table 1
Critical values w_α and b_α and sizes for the procedures SDW and SDB (based on simple random samples of 50 000 permutations)

k	n_i	nominal α	w_α	$\Pr_{H_0}\left(\max_{i<j} \lvert W_{ij}\rvert \geq w_\alpha\right)$	b_α	$\Pr_{H_0}\left(\max_{i<j} B_{ij} \geq b_\alpha\right)$
3	5[a]	0.1	3.102	0.082	3.013	0.082
		0.05	3.397	0.043	3.742	0.043
		0.025	3.693	0.022	4.894	0.022
4	5	0.1	3.397	0.076	3.742	0.076
		0.05	3.693	0.040	4.894	0.040
5	5	0.1	3.693	0.063	4.894	0.063
6	5	0.1	3.693	0.088	4.894	0.088
3	10	0.1	2.993	0.091	2.791	0.100
		0.05	3.314	0.049	3.497	0.050
		0.025	3.635	0.024	4.209	0.025
		0.01	4.062	0.008	5.200	0.010
4	10	0.1	3.314	0.087	3.364	0.100
		0.05	3.635	0.045	4.109	0.050
		0.025	3.955	0.021	4.896	0.025
		0.01	4.276	0.009	5.942	0.010
5	10	0.1	3.528	0.089	3.843	0.100
		0.05	3.849	0.044	4.644	0.050
		0.025	4.169	0.019	5.434	0.025
		0.01	4.383	0.010	6.353	0.010

Table 1 (continued)

k	n_i	nominal α	w_α	$\Pr_{H_0}\left(\max_{i<j} \lvert W_{ij}\rvert \geq w_\alpha\right)$	b_α	$\Pr_{H_0}\left(\max_{i<j} B_{ij} \geq b_\alpha\right)$
6	10	0.1	3.635	0.097	4.202	0.100
		0.05	3.955	0.047	5.037	0.050
		0.025	4.276	0.020	5.788	0.025
		0.01	4.490	0.010	6.758	0.010
7	10	0.1	3.849	0.079	4.520	0.100
		0.05	4.062	0.047	5.339	0.050
		0.025	4.383	0.019	6.117	0.025
		0.01	4.597	0.009	7.093	0.010
8	10	0.1	3.849	0.099	4.805	0.100
		0.05	4.169	0.045	5.593	0.050
		0.025	4.383	0.024	6.353	0.025
		0.01	4.704	0.008	7.495	0.010

[a] Values for $k = 3$ and $n_i = 5$ are based on the entire permutation null distribution

Table 2
Simulated any-pair power of the tests SDW and SDB for different distributions ($n_i = 10\ \forall i, \alpha = 0.05$)

k	Location shift $t_1, \ldots, t_k$ [a]		$F_1(t) \sim N(0,1)$	$F_1(t) \sim U(0,1)$	$F_1(t) \sim$ Cauchy	$F_1(t) \sim Ex(1)$	$F_1(-t) \sim Ex(1)$	$F_1(t) \sim \chi^2_2$
3	0, 0, 1.5	SDW	0.87	0.86	0.72	0.88	0.73	0.81
		SDB	0.87	0.85	0.79	0.92	0.81	0.87
	0, 0.75, 1.5	SDW	0.77	0.76	0.65	0.74	0.73	0.66
		SDB	0.77	0.75	0.70	0.80	0.82	0.71
4	0, 0, 0, 1.2	SDW	0.66	0.64	0.56	0.75	0.52	0.65
		SDB	0.67	0.64	0.65	0.81	0.60	0.71
	0, 0.4, 0.8, 1.2	SDW	0.51	0.48	0.48	0.53	0.52	0.44
		SDB	0.52	0.48	0.55	0.58	0.61	0.49
	0, 0, 1.2, 1.2	SDW	0.76	0.72	0.67	0.72	0.71	0.63
		SDB	0.76	0.72	0.75	0.77	0.79	0.68
	0, 0.6, 0.6, 1.2	SDW	0.48	0.45	0.44	0.50	0.49	0.42
		SDB	0.48	0.45	0.51	0.55	0.58	0.46
8	0, …, 0, 1.4	SDW	0.75	0.74	0.60	0.89	0.49	0.80
		SDB	0.76	0.74	0.67	0.92	0.55	0.84
	0, 0.2, 0.4, 0.6,	SDW	0.62	0.57	0.63	0.62	0.62	0.53
	0.8, 1, 1.2, 1.4	SDB	0.63	0.56	0.69	0.66	0.69	0.56

[a] $\theta_i = f \cdot t_i$ with $f = 1$ (standard normal distribution), $f = 3/10$ (uniform distribution on (0, 1)), $f = 2$ (Cauchy distribution), $f = 3/4$ (exponential distribution with $\lambda = 1$), or $f = 4/3$ (χ^2 distribution with d$f = 2$)

$i = 1, \ldots, k$, there is no difference in conservatism between the two procedures. However, for the sample size of 10 per group the test SDB is almost not conservative. Table 1 shows that SDB achieves a type I error rate which equals the nominal level for all considered values of α (0.1, 0.05, 0.025 and 0.01). By comparison, the procedure SDW is conservative since the exact null distribution of W_{ij} is much more discrete than that of B_{ij} (NEUHÄUSER, 2000).

For the power comparisons, two different experimentwise power criteria were used: the any-pair and the all-pairs power. The any-pair power is defined as the number of experiments in which at least one heterogeneous pair was found correctly divided by the number of experiments. The all-pairs power is the number of experiments in which all heterogeneous pairs were found divided by the number of experiments (MORLEY, 1982). Concerning the Monte Carlo power simulations, 10 000 simulation runs were generated for each particular configuration.

In the case of three groups we apply the closure principle in combination with the Steel-Dwass procedure. That is, if and only if the Steel-Dwass procedure declares $\theta_i \neq \theta_j$ for at least one pair (i, j), all further pairwise comparisons can be done with single two-sample tests of local level α. In case of $k > 3$ groups this nonparametric LSD test is not a closed testing procedure (see e.g. PIGEOT, 2000)

Table 3
Simulated all-pairs power of the tests SDW and SDB for different distributions ($n_i = 10\ \forall i$, $\alpha = 0.05$)

k	Location shift $t_1, \ldots, t_k$ [a]		$F_1(t) \sim N(0,1)$	$F_1(t) \sim U(0,1)$	$F_1(t) \sim$ Cauchy	$F_1(t) \sim Ex(1)$	$F_1(-t) \sim Ex(1)$	$F_1(t) \sim \chi^2_2$
3	0, 0, 1.5	SDW	0.74	0.72	0.53	0.61	0.68	0.51
		SDB	0.74	0.74	0.63	0.72	0.79	0.62
	0, 1.5, 3	SDW	0.71	0.68	0.44	0.59	0.58	0.48
		SDB	0.72	0.70	0.59	0.72	0.77	0.61
4	0, 0, 0, 2	SDW	0.79	0.80	0.32	0.45	0.67	0.33
		SDB	0.80	0.81	0.46	0.56	0.80	0.42
	0, 2, 4, 6	SDW	0.74	0.75	0.17	0.39	0.40	0.27
		SDB	0.74	0.76	0.31	0.53	0.59	0.38
	0, 0, 2, 2	SDW	0.74	0.75	0.26	0.49	0.49	0.37
		SDB	0.74	0.76	0.38	0.61	0.66	0.48
	0, 2, 2, 4	SDW	0.71	0.71	0.17	0.39	0.39	0.26
		SDB	0.70	0.72	0.30	0.52	0.57	0.37
8	0, ..., 0, 3	SDW	0.96	1.00	0.20	0.37	0.80	0.22
		SDB	0.96	1.00	0.25	0.46	0.86	0.29
	0, 0, 0, 0,	SDW	0.92	1.00	0.07	0.45	0.44	0.28
	3, 3, 3, 3	SDB	0.92	1.00	0.10	0.52	0.58	0.35

[a] $\theta_i = f \cdot t_i$ with $f = 1$ (standard normal distribution), $f = 3/10$ (uniform distribution on (0,1)), $f = 2$ (Cauchy distribution), $f = 3/4$ (exponential distribution with $\lambda = 1$), or $f = 4/3$ (χ^2 distribution with d$f = 2$)

Table 4

Simulated size and power of the tests SDW and SDB in case of heteroscedasticity based on normal distributions ($k = 3$, $n_1 = n_2 = n_3 = 10$, $\alpha = 0.05$)

μ_1, μ_2, μ_3 [a]	$\sigma_1, \sigma_2, \sigma_3$ [b]	SDW	SDB	SDW	SDB
		Size[c]			
0, 0, 0	1, 1, 2	0.054	0.070		
	1, 1.5, 2	0.053	0.065		
	1, 2, 2	0.054	0.074		
		Any-pair power		All-pairs power	
0, 0, 1.5	1, 1, 2	0.42	0.50	0.33	0.41
0, 0.75, 1.5	1, 1.5, 2	0.41	0.47	0.01	0.02
0, 1.5, 3	1, 1.5, 2	0.94	0.96	0.20	0.25
0, 1.5, 1.5	1, 2, 2	0.54	0.62	0.23	0.34

[a] Population means of the three groups

[b] Population standard deviations of the three groups

[c] Size is defined as the probability to erroneously decide that there is a difference in location between at least one pair (i, j)

and, consequently, we use the critical values of Table 1 for all $\binom{k}{2}$ pairwise comparisons.

Table 2 contains representative simulation results with regard to the any-pair power. The results indicate that the procedure SDB is at least as powerful as SDW. In case of the symmetric normal and uniform distributions both tests have a similar any-pair power. However, for the heavy-tailed Cauchy distribution and the investigated skew distributions the new test SDB is more powerful than SDW. Please note that the location shifts (in Tables 2 and 3) are adjusted with a distribution-specific factor f in order to bring the power for different distributions in a comparable range. The values of f are chosen on empirical grounds.

The Steel-Dwass procedure is especially useful with regard to the all-pairs power (MORLEY, 1982). Results concerning this power criterion are displayed in Table 3. According to the results given in this table the all-pairs power of SDB and SDW is similar in the case of normally and uniformly distributed data. However, in case of the Cauchy and the skew distributions, the all-pairs power of SDB is much higher than that of SDW.

In the location-shift model the distribution functions are the same except perhaps for a change in their locations. In practice, however, increasing locations are often accompanied by an increase in variability. This phenomenon is common in toxicological, medical and epidemiological studies (BLAIR and SAWILOWSKY, 1993, p. 2233). Table 4 shows the behaviour of the procedures SDW and SDB in case of heteroscedasticity (based on underlying normal distributions). The results indicate that in this case SDB is more powerful than SDW. However, the procedure SDB is also more anticonservative than SDW.

5. Example

The different procedures are applied to a practical data set. We consider the length of cuckoos' eggs deposited in nests of three different host species. These data (see Table 5) represent a subset of that presented by LATTER (1901).

We obtain the following values from the standardized Wilcoxon statistic: $W_{12} = -0.428$, $W_{13} = 2.993$, and $W_{23} = 3.314$. Following Section 2, the correlation coefficients are given by $\rho_{(1,2),(1,3)} = \rho_{(1,3),(2,3)} = 10/21$ and $\rho_{(1,2),(2,3)} = -10/21$. The resultant asymptotic critical value is 3.322 for $\alpha = 0.05$. Since the maximum test statistic $W_{23} = 3.314$ is smaller, there is no significant difference between any two groups when using the asymptotic multivariate normal distribution of the three Wilcoxon statistics. The corresponding *P*-value is calculated with 4 significant digits as 0.0507.

Using the statistic according to BAUMGARTNER et al. (1998) we obtain the following values: $B_{12} = 0.314$, $B_{13} = 2.864$, $B_{23} = 3.758$. At an α level of 0.05 we can declare groups 2 and 3 to be different with both exact procedures SDW and SDB. W_{23} equals the appropriate exact critical value $w_{0.049} = 3.314$ and B_{23} exceeds $b_{0.05} = 3.497$ (cf. Table 1).

Since we have three groups and a difference is found for one pair, we can perform all further pairwise comparisons with single tests of local level $\alpha = 0.05$. That is, we apply the closure principle in combination with the Steel-Dwass procedure as mentioned above in Section 4. As a result, we additionally obtain a difference between the groups 1 and 3 (Wilcoxon test: $P = 0.0355$; Baumgartner et al. test: $P = 0.0368$). The comparison group 1 versus group 2 is far from being significant.

Table 5

Lengths of cuckoos' (*Cuculus canorus*) eggs deposited in nests of three different host species (a subset of the data presented by LATTER, 1901, p. 173)

Group	Host species	Lengths (in mm)				
1	Tree pipit	21.1	21.8	22.1	22.4	22.7
	(*Anthus trivialis*)	23.2	23.3	23.4	23.6	24.0
2	Hedge-sparrow	20.9	21.7	22.0	22.8	23.0
	(*Prunella modularis*)	23.1	23.5	23.8	23.9	25.0
3	Meadow pipit	19.6	20.1	20.6	21.6	21.9
	(*Anthus pratensis*)	22.2	22.3	22.5	22.6	22.9

6. Discussion

Even if equal sample sizes per group are planned, lack of balance often occurs on account of missing values. For this case our instructions enable the exact computation of asymptotic critical values and *P*-values of the Steel-Dwass procedure. We recommend the use of the asymptotic method for large sample sizes (say, $n_i > 10$ $\forall\, i = 1, \dots, k$ in more or less balanced cases). For small sample sizes we recommend the use of the new statistic of BAUMGARTNER et al. (1998) for the multiple comparisons and the application of permutation tests. The resultant new procedure SDB is superior to the original Steel-Dwass procedure SDW.

References

BAUMGARTNER, W., WEISS, P., and SCHINDLER, H., 1998: A nonparametric test for the general two-sample problem. *Biometrics* **54**, 1129–1135.

BERRY, J. J., 1995: A simulation-based approach to some nonparametric statistics problems. *Observations: The Technical Journal for SAS Software Users* **5**, 19–26.

BLAIR, R. C. and SAWILOWSKY, S., 1993: Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls. *Statistics in Medicine* **12**, 2233–2243.

CRITCHLOW, D. E. and FLIGNER, M. A., 1991: On distribution-free multiple comparisons in the one-way analysis of variance. *Communications in Statistics – Theory and Methods* **20**, 127–139.

DWASS, M., 1960: Some k-sample rank-order tests. In: I. OLKIN, S. G. GHURYE, W. HOEFFDING, W. G. MADOW and H. B. MANN (eds.): *Contributions to probability and statistics*. Stanford University Press, Stanford, 198–202.

GENZ, A., 1992: Numerical computation of the multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–150.

GENZ, A. and KWONG, K. S., 2000: Numerical evaluation of singular multivariate normal distributions. *Journal of Statistical Computation and Simulation* **68**, 1–21.

GOOD, P. I., 1994: *Permutation tests*. Springer, New York.

HOCHBERG, Y. and TAMHANE, A., 1987: *Multiple comparison procedures*. Wiley, New York.

LATTER, O. H., 1901: The egg of *Cuculus canorus*. *Biometrika* **1**, 164–176.

MILLER, R. G., 1981: *Simultaneous statistical inference* (2nd edition). Springer, New York.

MORLEY, C. L., 1982: A simulation study of the powers of three multiple comparison statistics. *Australian Journal of Statistics* **24**, 201–210.

NEUHÄUSER, M., 2000: An exact two-sample test based on the Baumgartner-Weiß-Schindler statistic and a modification of Lepage's test. *Communications in Statistics – Theory and Methods* **29**, 67–78.

PIGEOT, I., 2000: Basic concepts of multiple tests – A survey. *Statistical Papers* **41**, 3–36.

STEEL, R. G. D., 1960: A rank sum test for comparing all pairs of treatments. *Technometrics* **2**, 197–207.

Dr. MARKUS NEUHÄUSER
Byk Gulden Pharmaceuticals
Dept. of Biometry
Byk-Gulden-Str. 2
D-78467 Konstanz, Germany
E-mail: markus.neuhaeuser@byk.de

Received, August 2000
Revised, April 2001
Accepted, May 2001