Recent Developments in Multiple Comparison Procedures
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 47 (2004) 48–64

Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case

Manfred Horn[1] **and Charles W. Dunnett**[2]

Friedrich Schiller University and Mcmaster University

Abstract: Our paper compares the any-pair power, all-pairs power and per-pair power of several test procedures that control either the familywise error rate (FWE) or the false discovery rate (FDR). Our investigations are restricted to one-sided many-one comparisons of normal distributions. Some of the investigated procedures make use of this known special structure and some do not. The numbers k of hypotheses we considered were 5, 10 and 100. The statements with small and large k were quite similar. We found that all methods except one do not essentially differ in their any-pair powers. The most remarkable differences between the test procedures can be observed concerning their all-pairs powers.

We also compared the sample sizes that are necessary for the different procedures to ensure a specified power and investigated their dependency on the number k of hypotheses. With specified per-pair or any-pair power, the FWE controlling methods need similar sample sizes as the FDR controlling methods. With specified all-pairs power, the sample sizes for four of the six FDR controlling procedures studied increase with k at noticeably lower rates than for the FWE controlling procedures.

1. Introduction

Most tests for multiple comparisons are traditionally designed to control the type I familywise error rate (FWE) at level α, i.e., they guarantee that FWE $\leq \alpha$. FWE is the probability $\mathrm{P}(V \geq 1)$, where V denotes the number of true hypotheses erroneously rejected. As an alternative to FWE control, Benjamini and Hochberg (1995) introduced the false discovery rate (FDR). FDR is the expectation $\mathrm{E}(V/R)$, where R denotes the total number of hypotheses rejected. If R is 0, V/R is defined to be 0.

As $\mathrm{E}(V/R) \leq \mathrm{P}(V \geq 1)$, FDR control is less stringent than FWE control and hence promises higher powers.

The simplest FWE controlling method is the Bonferroni method. The step-down (SD) procedure of Holm (1979) is a stepwise version of the Bonferroni method which also controls the FWE and which is more powerful. Simes (1986) proposed a modified Bonferroni procedure for the test of the overall hypothesis which was used by Hochberg (1988) to derive a step-up (SU) procedure. With these methods, the p-values of the test statistics for testing the hypotheses are compared with critical bounds which are fractions of α.

[1]Institute of Medical Statistics, Friedrich Schiller University, D-07740 Jena, Germany. e-mail: horn@imsid.uni-jena.de

[2]Department of Mathematics and Statistics, Mcmaster University, Hamilton, Ontario L8S4K1, Canada. e-mail: dunnett@mcmaster.ca

Keywords and phrases: multiple comparison procedures, false discovery rate, familywise error rate, any-pair power, all-pairs power, per-pair power, average power, sample size, comparisons with control.

AMS 2000 subject classifications: 62F03, 62J15.

Several FDR controlling procedures use the p-values in a similar way. The method of Benjamini and Hochberg (1995) is a SU procedure which is also based on the method of Simes (1986). Sarkar (2002) proposed to use the critical p-value bounds of Benjamini and Hochberg (1995) in a SD procedure. Other SD procedures are those of Benjamini and Liu (1999) and Benjamini and Liu (2001). A modification of the method of Benjamini and Hochberg (1995) is the SU procedure of Benjamini and Yekutieli (2001). Another modification is the SU procedure of Kwong, Holland and Cheung (2002).

Note that the method of Simes (1986) and with it the SU procedure of Hochberg (1988) originally were derived for independent test statistics. Sarkar (1998) showed that Simes' inequality also holds under so-called positive dependency. Hence, Hochberg's method which is based on Simes' method controls the FWE under positive dependency, which is given, for example, in one-sided many-one comparisons of normal variables with known variance which is the case we will consider. Also, the FDR controlling methods of Benjamini and Hochberg (1995), Benjamini and Liu (1999), and Sarkar (2002) were originally based on independent test statistics (in contrast to the procedures of Benjamini and Yekutieli (2001) and Benjamini and Liu (2001)). They also are valid under positive dependency, see Sarkar (2002). The same applies to the method of Kwong, Holland and Cheung (2002).

The aim of this paper is to compare the powers and required sample sizes for FWE and FDR controlling methods. We will restrict our considerations to many-one comparisons in the normal case, so that we can include in our study the FWE controlling SD procedure of Dunnett and Tamhane (1991), the FWE controlling SU procedure of Dunnett and Tamhane (1992), and the FDR controlling SD procedure of Troendle (2000). All of these last mentioned methods assume known fixed positive dependency among the normally distributed test statistics, a condition we shall refer to as 'known positive normality' for brevity. Naturally, the procedures which make use of the known positive normality of the many-one problem have better performance. In Table 1, we list all methods considered in this paper together with their characteristics and their abbreviations used in the remainder of this paper.

Table 1: Test procedures and their characteristics

Reference	Abbreviation	SD/SU	Control of	Restriction
Holm (1979)	Holm	SD		none
Hochberg (1988)	Hoch	SU	FWE	positive dependency
Dunnett, Tamhane (1991)	DT91	SD		known positive normality
Dunnett, Tamhane (1992)	DT92	SU		known positive normality
Benjamini, Hochberg (1995)	BH95	SU		positive dependency
Sarkar (2002)	Sark	SD		positive dependency
Benjamini, Liu (1999)	BL99	SD		positive dependency
Benjamini, Liu (2001)	BL01	SD	FDR	none
Benjamini, Yekutieli (2001)	BY01	SU		none
Kwong, Holland, Cheung (2002)	KH02	SU		known positive normality
Troendle (2000)	Troe	SD		known positive normality

2. Critical values

Assume $X_0, X_1, \ldots, X_k$ are normally distributed random variables with expectations $\mu_0, \mu_1, \ldots, \mu_k$ and common variance σ^2. X_0 represents a control against which k treatments are to be compared. We only consider one-sided comparisons. Then the null and alternative hypotheses are $H_i : \mu_i \leq \mu_0, H_{Ai} : \mu_i > \mu_0$ $(i = 1, \ldots, k)$. We restrict our investigations to the case of equal sample sizes $n_0 = n_1 = \cdots = n_k = n$. Then the test statistics have the form $t_i = \sqrt{2/n}(\bar{x}_i - \bar{x}_0)/s$, where $\bar{x}_i$ and $\bar{x}_0$ are the sample means of the observations of treatment i and of the control group, respectively, and s^2 the estimate of σ^2 with $\nu = (k+1)(n-1)$ degrees of freedom.

The most popular FWE controlling stepwise procedure for testing the hypotheses H_i is the SD procedure DT91 which utilizes the multivariate t-distribution. The FDR controlling procedure Troe is also SD and utilizes the multivariate t-distribution of the test statistics. Troe can be considered as the FDR controlling analogue of DT91. Troendle (2000) provided some critical constants c_i $(i = 1, \ldots, k)$ and an explanation how to calculate them. He also proposed a SU procedure, however we will investigate only his SD procedure.

The Bonferroni method and the Bonferroni like methods mentioned in Section 1 do not utilize the multivariate t-distribution of the test statistics. They compare the ordered p-values $p_{(1)} \leq \cdots \leq p_{(k)}$, which are obtained from t-tests for the hypotheses H_i, with critical bounds $\gamma_1, \ldots, \gamma_k$ which are fractions of α, see Table 2. For example, the SD procedure Holm compares $p_{(1)}$ with $\gamma_1 = \alpha/k$ in the first step, $p_{(2)}$ with $\gamma_2 = \alpha/(k-1)$ in the second step, ... , $p_{(j)}$ with $\gamma_j = \alpha/(k-j+1)$ in the j-th step. The SU procedure Hoch compares the p-values with the same γ-values in the reverse order, starting with $p_{(k)}$. Similarly, both the SU procedure BH95 and the SD procedure Sark compare $p_{(j)}$ with the same bound $\alpha j/k$.

Note that a step-up procedure has a power which equals or exceeds the power of a step-down procedure which uses the same critical bounds. Hence, Hochberg's method is more powerful than Holm's method and Benjamini and Hochberg's method is more powerful than Sarkar's method. Furthermore, it can be seen from Table 2 that the critical bounds γ_i of BH95 are greater than those of BY01 by the factor $1 + 1/2 + \cdots + 1/k$. Thus, BH95 has a higher power than BY01. This could also be expected from the fact that BY01 is a modification of BH95 which does not require independence or positive dependency.

The SU procedure of KH02 is an improvement of BH95. It uses $\gamma_1, \ldots, \gamma_{k-1}$ of BH95 and $\gamma_k = r^* \alpha$ $(r^* \geq 1)$. Tables of r^* are given in Kwong, Holland and Cheung (2002).

Table 2: Critical bounds γ_j for the ordered p-values $p_{(j)}$ $(p_{(1)} \leq \cdots \leq p_{(k)})$ of different FWE and FDR controlling procedures

Method	Control of	SD/SU	γ_j
Holm	FWE	SD	$\alpha/(k-j+1)$
Hoch	FWE	SU	$\alpha/(k-j+1)$
BH95	FDR	SU	$\alpha j/k$
Sark	FDR	SD	$\alpha j/k$
BL99	FDR	SD	1-[1-min$\{1, \alpha k/(k-j+1)\}]^{1/(k-j+1)}$
BL01	FDR	SD	min$[1, \alpha k/(k-j+1)^2]$
BY01	FDR	SU	$\alpha j/[k(1 + 1/2 + \cdots + 1/k)]$

Table 3: Critical constants for the different test procedures (one-sided tests, $k = 5$, $\alpha = 0.05$, $\nu = \infty$)

	Method	c_1	c_2	c_3	c_4	c_5
	DT91	1.6449	1.9163	2.0621	2.1603	2.2338
FWE ≤ 0.05	DT92	1.6449	1.9330	2.0708	2.1651	2.2367
	Holm, Hoch	1.6449	1.9600	2.1280	2.2414	2.3263
	Troe*	0.6745	1.3420	1.6586	1.9252	2.2338
	BH95, Sark	1.6449	1.7507	1.8808	2.0537	2.3263
FDR ≤ 0.05	BL99	0.6745	1.5174	1.9020	2.1443	2.3187
	BL01	0.6745	1.5341	1.9145	2.1539	2.3263
	BY01	2.0160	2.1079	2.2221	2.3756	2.6213
	KH02	1.0446	1.7507	1.8808	2.0537	2.3263

* *100,000 simulations were used for determining the critical constants of Troe.*

When we consider the ordered test statistics $t_{(1)}, \ldots, t_{(k)}$, then the smallest p-value $p_{(1)}$ is associated with the largest t-value $t_{(k)}$, $p_{(2)}$ is associated with $t_{(k-1)}$, $\ldots, p_{(i)}$ is associated with $t_{(k-i+1)}$. Hence, comparing $p_{(i)}$ with γ_i is the same as comparing $t_{(i)}$ with $c_i = t_{\nu,1-\gamma_{k-i+1}}$, i.e., with the $(1-\gamma_{k-i+1})$-quantile of the central univariate t-distribution.

We will restrict our investigations to $\nu = \infty$ which corresponds to the case of known σ. We expect results for $\nu < \infty$ to be very similar.

The power of a multiple test procedure will be high if its critical constants c_i are small. Table 3 provides for $k = 5$, $\alpha = 0.05$, $\nu = \infty$ the critical constants for all methods that we want to compare in this paper. The smallest c_i's are those of Troe, so that Troe can be expected to have the highest power. The differences between BL99 and BL01 are small.

Note that the FDR controlling methods Troe, BL99, BL01 and KH02 have a strange property. It is possible that these methods reject hypotheses that cannot be rejected by the t-test. This can be concluded from the critical values c_i in Table 3 which in some cases are below the unadjusted critical value 1.6449. This can also be concluded from the critical bounds γ_i in Table 2. For example, with $\alpha = 0.05$ and $k = 5$ we obtain for BL01 the bound $\gamma_5 = 5(0.05) = 0.25$. This means the hypothesis tested in the fifth step would be rejected if its unadjusted p-value is 0.24, say. If k is large, we even may have $\gamma_k = 1$ which means that any hypothesis will be rejected in the last step of this step-down procedure no matter how large its unadjusted p-value is. Benjamini and Liu (1999) mention that in practical cases where this property is not wanted one may add the constraint not to reject hypotheses if the corresponding p-values are larger than a prespecified value. This will not inflate the FDR, however it will decrease the power.

3. Powers

We consider configurations where some hypotheses are false and assume that $\mu_i - \mu_0 = \Delta$ holds for the false hypotheses, and $\mu_i - \mu_0 = 0$ holds for the true hypotheses. The value of Δ/σ is specified. $\Delta > 0$ denotes the practical relevant difference $\mu_i - \mu_0$.

In multiple comparisons, the power can be defined in different ways. The probability of rejecting at least one of the false hypotheses is called any-pair power, and the probability of rejecting all false hypotheses is called all-pairs power, see

Ramsey (1978). If we consider a single false hypothesis, then the probability to reject it is called per-pair power, see Einot and Gabriel (1975). Liu (1997), Troendle (2000) and Kwong, Holland and Cheung (2002) preferred in their power comparisons the average power which is $\mathrm{E}(Z_m/m) = \mathrm{E}(Z_m)/m$, where m denotes the number of false hypotheses and Z_m the random number of rejected false hypotheses. They estimated the average power in simulations by the proportion of false hypotheses that were rejected. The average power can be calculated in the following way. Since $\mathrm{P}(Z_m = t) = \mathrm{P}(Z_m \geq t) - \mathrm{P}(Z_m \geq t+1)$ for $t = 1, \ldots, m-1$ and $\mathrm{P}(Z_m \geq m) = \mathrm{P}(Z_m = m)$, we obtain for given m the expectation

$$\begin{aligned}
\mathrm{E}(Z_m) &= \sum_{t=1}^{m} t\mathrm{P}(Z_m = t) \\
&= \mathrm{P}(Z_m \geq 1) - \mathrm{P}(Z_m \geq 2) + 2\mathrm{P}(Z_m \geq 2) - 2\mathrm{P}(Z_m \geq 3) \\
&\quad + \cdots + (m-1)\mathrm{P}(Z_m \geq m-1) - (m-1)\mathrm{P}(Z_m \geq m) + m\mathrm{P}(Z_m \geq m) \\
&= \sum_{t=1}^{m} \mathrm{P}(Z_m \geq t).
\end{aligned}$$

We have a program called POWERN which calculates the probabilities $\mathrm{P}(Z_m \geq t)$ for each procedure considered here, so that we obtain $\mathrm{E}(Z_m)$ by calculating their sum. The original version of this program was developed for the power calculations in Dunnett, Horn, Vollandt (2001) and can be downloaded from *www.bioinf.uni-hannover.de/mcp_home/*. We extended it to include the FDR procedures as well as FWE procedures.

As $\mu_i - \mu_0 = \Delta$ holds for all m false hypotheses the per-pair power of each false hypothesis has the same value, say p. Then $\mathrm{E}(Z_m) = mp$ and with it $\mathrm{E}(Z_m/m) = p$. This means that in our considerations the average power is identical with the per-pair power.

Tables 4, 5 and 6 contain the probabilities $\mathrm{P}(Z_m \geq t)$ of rejecting at least t of m false hypotheses of the different procedures in one-sided many-one comparisons in the normal case for $k = 5$, $n = 6$, $\alpha = 0.05$, $\Delta/\sigma = 2$, σ known i.e., $\nu = \infty$. These probabilities are the any pair powers if $t = 1$, see Table 4, and the all-pairs powers

Table 4: Any-pair power with different numbers m of false hypotheses ($k = 5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$). The maximum value of each column is underlined

	Method	m				
		1	2	3	4	5
FWE $\leq .05$	DT91	<u>.891</u>	<u>.963</u>	<u>.982</u>	.989	.993
	DT92	.890	<u>.963</u>	<u>.982</u>	<u>.990</u>	.993
	Holm	.872	.954	.977	.986	.990
	Hoch	.872	.954	.977	.986	.991
FDR $\leq .05$	Troe	<u>.891</u>	<u>.963</u>	<u>.982</u>	.989	.993
	BH95	.873	.959	.981	<u>.990</u>	.993
	Sark, BL01	.872	.954	.977	.986	.990
	BL99	.874	.955	.977	.986	.991
	BY01	.800	.919	.957	.973	.982
	KH02	.873	.959	.981	<u>.990</u>	<u>.995</u>

Table 5: All-pairs power with different numbers m of false hypotheses ($k = 5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$). The maximum value of each column is underlined

	Method	m					Minimum
		1	2	3	4	5	Value*
	DT91	<u>.891</u>	.838	.818	.826	.868	$.818^{3}$
FWE $\leq .05$	DT92	.890	.837	.817	.825	.879	$.817^{3}$
	Holm	.872	.816	.798	.812	.863	$.798^{3}$
	Hoch	.872	.816	.799	.817	.879	$.799^{3}$
	Troe	<u>.891</u>	<u>.888</u>	<u>.907</u>	<u>.937</u>	<u>.974</u>	$\underline{.888}^{2}$
	BH95	.873	.865	.868	.873	.879	$.865^{2}$
FDR $\leq .05$	Sark	.872	.860	.864	.870	.877	$.860^{2}$
	BL99	.874	.841	.856	.904	.959	$.841^{2}$
	BL01	.872	.838	.853	.902	.957	$.838^{2}$
	BY01	.800	.775	.768	.768	.771	$.768^{4}$
	KH02	.873	.865	.868	.874	.968	$.865^{2}$

* *Note: Number shown as superscript is the least favorable m, i.e. the number of false hypotheses for which the all-pairs power is minimum.*

Table 6: Probability to reject at least t of m false hypotheses for different configurations $(m; t)$ with $t \leq m$ ($k = 5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$). The maximum value of each column is underlined

	Method	$m; t$					
		3;2	4;2	4;3	5;2	5;3	5;4
	DT91	.937	.967	.924	.980	.959	.927
FWE $\leq .05$	DT92	.938	.968	.927	.982	.965	.938
	Holm	.924	.959	.912	.974	.951	.918
	Hoch	.926	.961	.918	.978	.961	.936
	Troe	<u>.961</u>	<u>.980</u>	<u>.968</u>	.988	.984	.978
	BH95	.954	.978	.954	.988	.978	.955
FDR $\leq .05$	Sark	.947	.972	.948	.983	.972	.950
	BL99	.938	.967	.943	.980	.979	.962
	BL01	.936	.966	.942	.979	.969	.961
	BY01	.903	.947	.897	.967	.942	.895
	KH02	.954	.978	.954	<u>.992</u>	<u>.987</u>	<u>.980</u>

Table 7: Per-pair power with different numbers m of false hypotheses ($k = 5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$). The maximum value of each column is underlined

	Method	m				
		1	2	3	4	5
	DT91	.891	.901	.912	.927	.945
FWE $\leq .05$	DT92	.890	.900	.912	.927	.951
	Holm	.872	.885	.899	.917	.939
	Hoch	.872	.897	.901	.921	.949
	Troe	.<u>891</u>	.<u>926</u>	.<u>950</u>	.<u>969</u>	.983
	BH95	.873	.912	.934	.949	.959
FDR $\leq .05$	Sark	.872	.907	.929	.944	.955
	BL99	.874	.898	.924	.950	.972
	BL01	.872	.896	.922	.949	.971
	BY01	.800	.847	.876	.896	.911
	KH02	.873	.912	.934	.949	.<u>984</u>

if $t = m$, see Table 5. Table 7 contains the per-pair powers which are calculated as described above from the values in Tables 4, 5 and 6.

Note that with each procedure the any-pair power and the per-pair power are both monotonically increasing in m, and their minimum values at $m = 1$ coincide. The all-pairs power is not monotone. Its minimum is with different methods at different values of m , see Table 5, last column.

Table 4 shows that no procedure dominates the other methods concerning the any-pair power over all 5 configurations. However, the any-pair power differences between most methods are very small. Only BY01 shows distinctly smaller values.

The power values from Table 4 are plotted in Figure 1 for some selected methods. Figure 2 is the corresponding plot for $k = 10$. Note, that there are no differences between DT91 and Troe, and also between Holm, Sark and BL01.

Table 5 shows that Troe has the highest all-pairs power at all configurations. This could be expected from the critical values in Table 3. However, among the remaining procedures no one dominates the others over all configurations. BY01 is clearly worse than all other methods, even than Holm and Hoch.

The power values from Table 5 are plotted in Figure 3 for all methods except BY01. Figure 4 is the corresponding plot for $k = 10$. In Figure 5 are the all-pairs powers of five methods for $k = 100$. Figures 3, 4 and 5 show that the test procedures are very different in their power behavior. For example, the minimum and maximum values of the all-pairs power of DT91 differ strongly, whereas the all-pairs power of BH95 does not essentially change with m. The values of BH95 and KH02 coincide except at $m = k$ where the all-pairs power of KH02 is essentially higher. BL99 and BL01, BH95 and Sark, DT91 and DT92 or Holm and Hoch differ only slightly. (This is the reason why we omitted every second method in Figure 5.)

Table 7 shows that Troe dominates concerning the per-pair powers, except at $m = k = 5$ where KH02 dominates. Again, BL99 and BL01 differ only slightly, and BY01 is clearly the worst procedure. Figures 6, 7 and 8 demonstrate for $k = 5$, 10 and 100, respectively, how the per-pair powers of the different procedures change with increasing m.

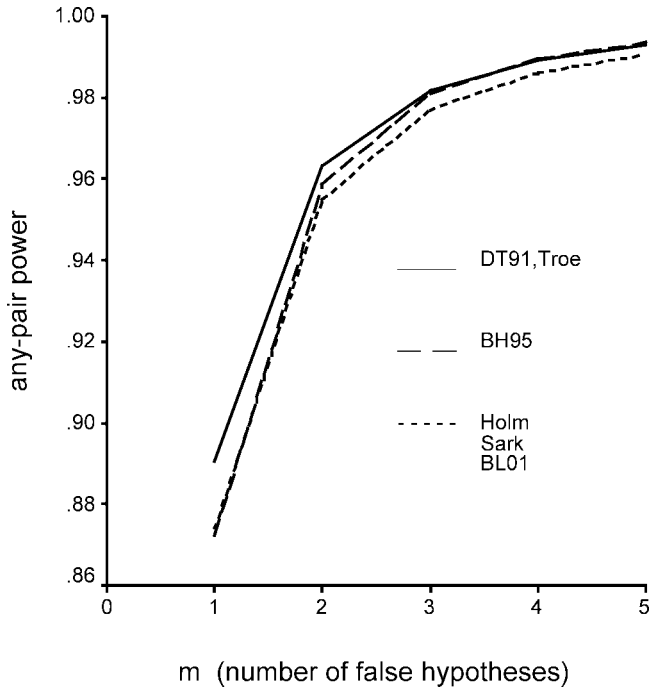

Figure 1: Any-pair powers for $k = 5$, $m = 1(1)5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

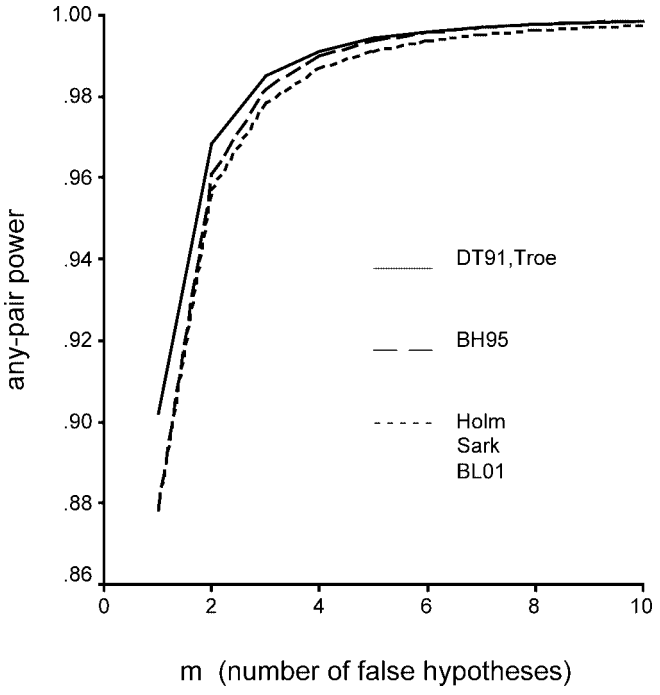

Figure 2: Any-pair powers for $k = 10$, $m = 1(1)10$, $n = 7$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

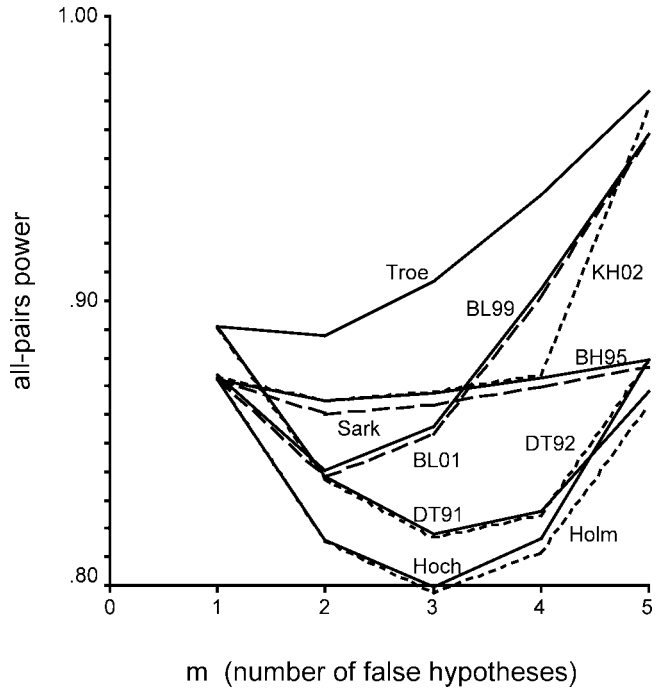

Figure 3: All-pairs powers for $k = 5$, $m = 1(1)5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

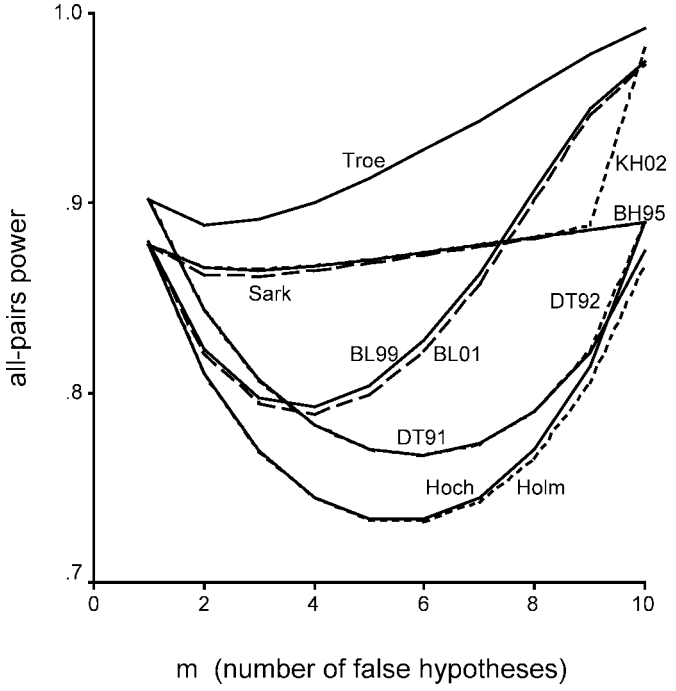

Figure 4: All-pairs powers for $k = 10$, $m = 1(1)10$, $n = 7$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

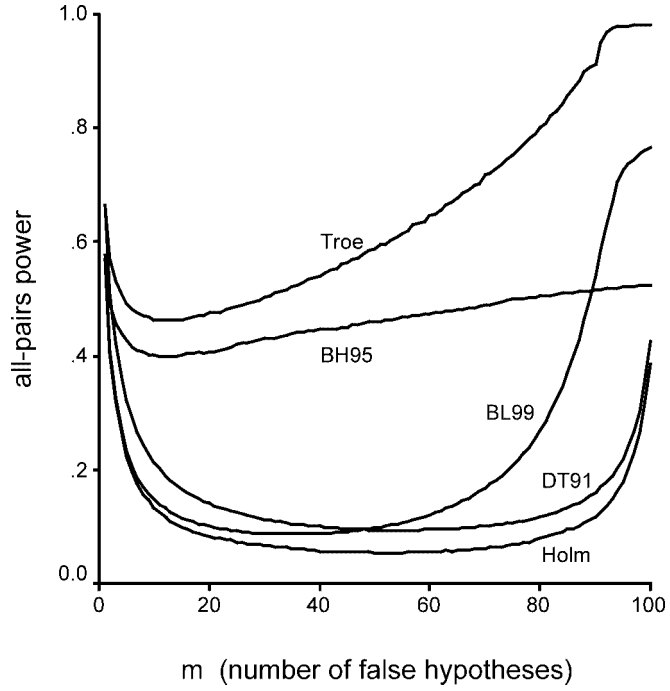

Figure 5: All-pairs powers of five methods for $k = 100$, $m = 1(1)100$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

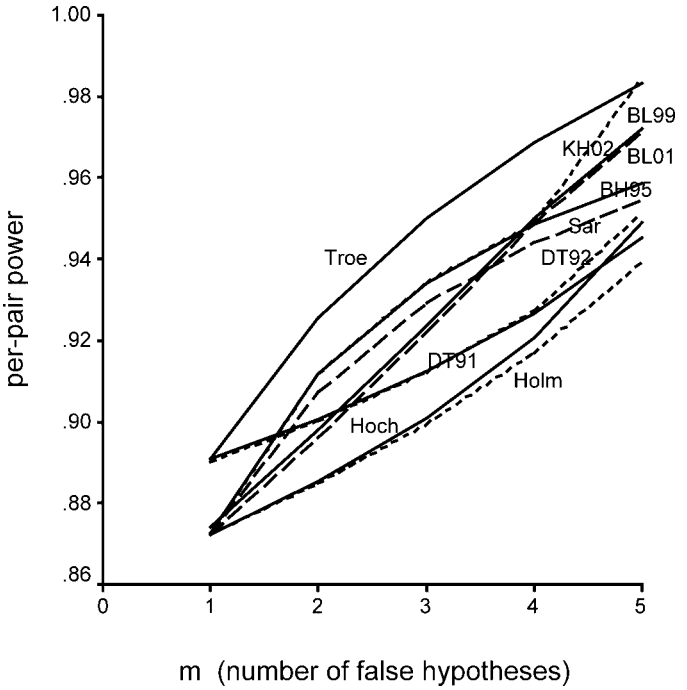

Figure 6: Per-pair powers for $k = 5$, $m = 1(1)5$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

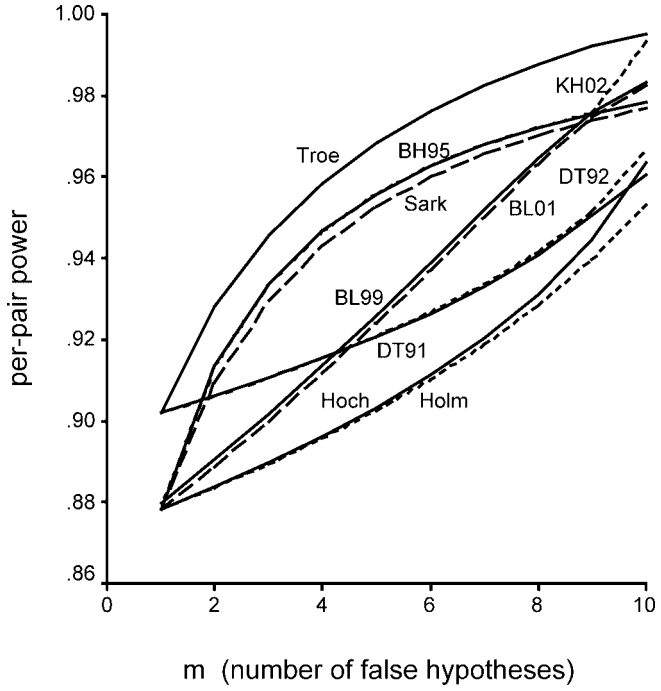

Figure 7: Per-pair powers for $k = 10$, $m = 1(1)10$, $n = 7$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

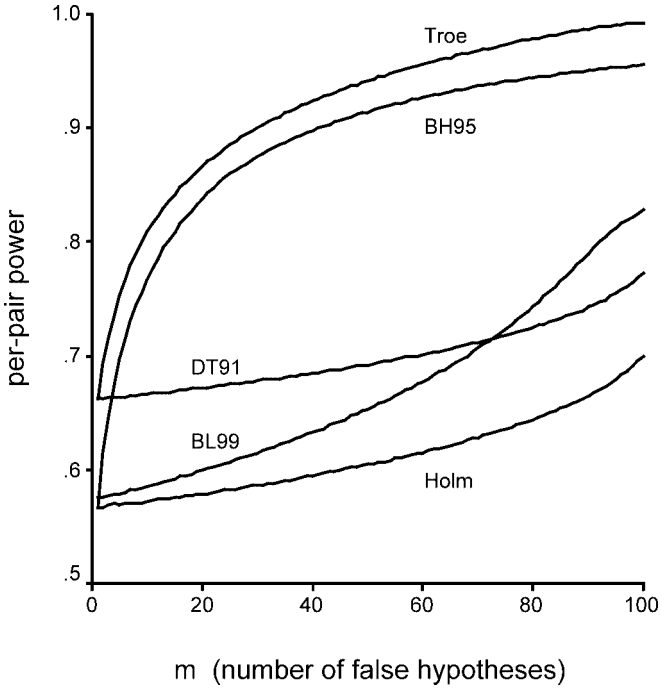

Figure 8: Per-pair powers of five methods for $k = 100$, $m = 1(1)100$, $n = 6$, $\Delta/\sigma = 2$, $\nu = \infty$, $\alpha = 0.05$

4. Sample sizes

The determination of sample sizes that guarantee a specified power $1-\beta$ in multiple comparisons is complicated, especially with SD and SU procedures. Without prior knowledge of the number of false hypotheses, one must look for the least favorable configuration (LFC) where the power attains its minimum. As already mentioned, for any procedure we studied the minimum any-pair power and the minimum per-pair power coincide and both occur at $m = 1$. In contrast, the procedures differ in the values of m at which their all-pairs powers attain the minimum, see Figures 3, 4 and 5. The sample size must be sufficiently large that this minimum is at least $1 - \beta$. The program POWERN was developed in order to determine the minimum sample size that guarantees a specified power value $1-\beta$. It finds out the LFC of SD and SU procedures, i.e., the configuration $(m; t)$ at which the power is minimum. In Dunnett, Horn, Vollandt (2001), tables of $\lambda = \sqrt{n}\Delta/\sigma$ at the LFC were given for the FWE controlling procedures DT91 and DT92 for $\nu = \infty$ and various k, α and all-pairs power values $1 - \beta$. We now calculated for $5 \leq k \leq 14$, $\alpha = 0.05$, $\nu = \infty$ values of $\lambda^2 = n\Delta^2/\sigma^2$ for specified minimum per-pair or any-pair power $1-\beta = 0.8$, see Table 8, and for specified minimum all-pairs power $1-\beta = 0.8$, see Table 9.

The values of λ^2 permit a fast determination of the sample size n needed for the control group and each treatment group, see our example. Note that this sample size is a lower bound for the sample size needed with unknown σ^2. In most cases there is only a small difference to the sample size with unknown σ^2 (sometimes there is no difference, due to the rounding of $\lambda^2\sigma^2/\Delta^2$).

Example 1. Assume an experimenter wants to compare $k = 12$ treatments with a control, either with the FWE controlling procedure DT91 or with the FDR controlling procedure Troe. He wants to detect with probability $1 - \beta = 0.8$ all differences $\mu_i - \mu_0 \geq \sigma$. The sizes of the control sample and the k treatment samples are required to have the same size n. In Table 9, we find $\lambda^2 = 31.185$ for DT91 and $\lambda^2 = 24.099$ for Troe. Hence, after rounding to the next largest integer, we obtain the sample sizes $n = 32$ for DT91 and $n = 25$ for Troe.

Tables 8 and 9 permit to compare the different procedures. Table 8 and the corresponding Figure 9 show that, with specified per-pair or any-pair power, the smallest sample sizes are needed, simultaneously, for DT91 and Troe. These two method have identical λ^2-values. The second best method is DT92. Identical λ^2-values appear also for Holm, Hoch, BH95, KH02, Sark, and BL01. Hence, these methods need the same sample sizes. The worst method is BY01. These findings demonstrate that the FDR controlling methods are not superior to the FWE controlling methods concerning the sample sizes needed with specified per-pair or any-pair power.

In contrast to the results for the any-pair and per-pair power, we state that with specified all-pairs power, the FDR controlling methods except BY01 require smaller sample sizes than the FWE controlling procedures. Table 9 shows that Troe is the best method, followed by BH95 and KH02. This is what we expected because Troe utilizes the known multivariate t-distribution of the test statistics, similarly as DT91. For values of k studied in this paper, BY01 requires larger sample sizes than DT91 and DT92, and with $k \leq 9$ even larger sample sizes than Holm and Hoch.

The values of Table 9 are plotted in Figure 10. It can be seen that the λ^2-values of Troe, BH95, KH02 and Sark increase with noticeable lower rates than the λ^2-values of the remaining procedures. (This applies also to the worst method BY01.) This means that with increasing k the sample size required for specified all-pairs power less rapidly increases for these FDR controlling methods than for the

Table 8: Values of $\lambda^2 = n\Delta^2/\sigma^2$ to guarantee a specified per-pair or any-pair power $1-\beta$ ($n_0/n = 1$, $\nu = \infty$, $1-\beta = 0.8$)

Method	k											
	2	3	4	5	6	7	8	9	10	11	12	14
DT91	15.213	16.868	18.023	18.917	19.642	20.251	20.776	21.237	21.648	22.019	22.356	22.950
DT92	15.397	16.965	18.081	18.952	19.665	20.267	20.788	21.247	21.656	22.025	22.361	22.953
Holm	15.698	17.638	19.010	20.072	20.938	21.670	22.302	22.860	23.358	23.808	24.219	24.946
Hoch	15.698	17.638	19.010	20.072	20.938	21.670	22.302	22.860	23.358	23.808	24.219	24.946
Troe	15.219	16.868	18.023	18.917	19.642	20.251	20.776	21.237	21.648	22.019	22.356	22.950
BH95	15.698	17.638	19.010	20.072	20.938	21.670	22.302	22.860	23.358	23.808	24.219	24.946
KH02	15.698	17.638	19.010	20.072	20.938	21.670	22.302	22.860	23.358	23.808	24.219	24.946
Sark	15.698	17.638	19.010	20.072	20.938	21.670	22.302	22.860	23.358	23.808	24.219	24.946
BL99	15.637	17.557	18.919	19.975	20.837	21.566	22.196	22.752	23.249	23.699	24.109	24.834
BL01	15.698	17.638	19.010	20.072	20.938	21.670	22.302	22.860	23.358	23.808	24.219	24.946
BY01	17.638	20.525	22.495	23.984	25.176	26.169	27.018	27.759	28.416	29.006	29.541	30.481

Table 9: Values of $\lambda^2 = n\Delta^2/\sigma^2$ to guarantee a specified all-pairs power $1-\beta$ ($n_0/n = 1$, $\nu = \infty$, $\alpha = 0.05$, $1-\beta = 0.8$)

Method	k											
	2	3	4	5	6	7	8	9	10	11	12	14
DT91	16.072	19.103	21.311	23.186	24.764	26.132	27.360	28.439	29.444	30.335	31.185	32.679
DT92	15.828	19.238	21.408	23.249	24.814	26.161	27.383	28.452	29.453	30.340	31.188	32.680
Holm	16.163	19.671	21.928	24.112	25.727	27.332	28.591	29.863	30.904	31.950	32.843	34.508
Hoch	15.829	19.574	21.762	24.036	25.634	27.276	28.545	29.821	30.874	31.917	32.818	34.488
Troe	15.213	16.853	18.253	19.491	20.502	21.305	21.977	22.571	23.094	23.614	24.099	24.953
BH95	15.829	18.028	19.574	20.765	21.762	22.630	23.378	24.036	24.623	25.152	25.634	26.518
KH02	15.829	18.028	19.574	20.765	21.762	22.630	23.378	24.036	24.623	25.152	25.634	26.518
Sark	16.163	18.335	19.865	21.045	22.005	22.858	23.599	24.252	24.834	25.359	25.837	26.682
BL99	15.637	17.724	20.280	22.059	23.418	24.972	26.213	27.239	28.335	29.290	30.117	31.738
BL01	15.698	17.825	20.398	22.185	23.575	25.129	26.368	27.393	28.511	29.463	30.287	31.922
BY01	18.028	21.272	23.607	25.358	26.792	27.994	29.018	29.926	30.736	31.461	32.117	33.288

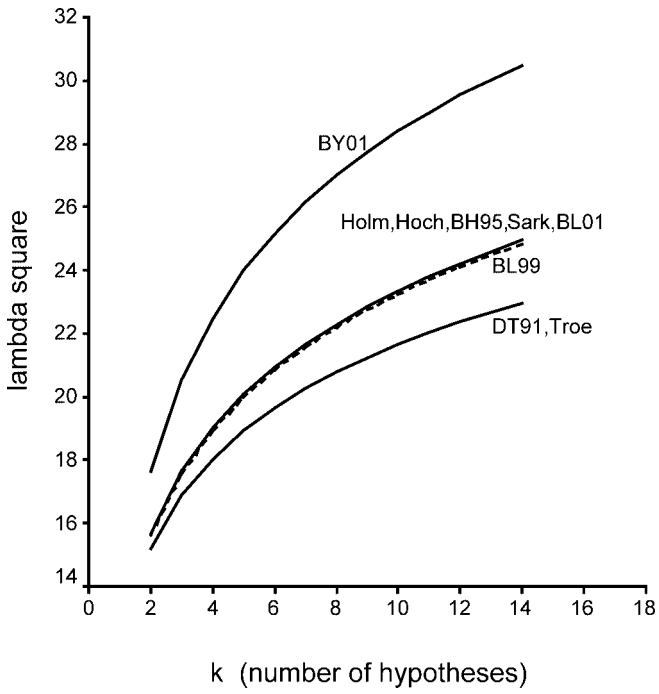

Figure 9: Values of $\lambda^2 = n(\Delta/\sigma)^2$ to guarantee a per-pair or any-pair power $1-\beta = 0.8$ for $k = 2(1)14$, $n_0/n = 1$, $\nu = \infty$

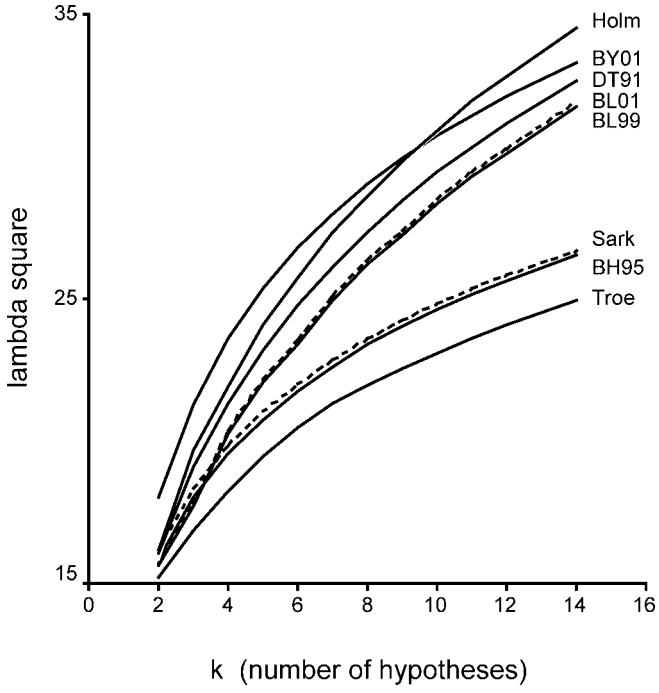

Figure 10: Values of $\lambda^2 = n(\Delta/\sigma)^2$ to guarantee an all-pairs power $1 - \beta = 0.8$ for $k = 2(1)14$, $n_0/n = 1$, $\nu = \infty$

FWE controlling procedures we studied. The fact that the advantage of these FDR methods over the FWE methods becomes greater with increasing k is an important and desirable property. The remaining two FDR controlling procedures BL99 and BL01 do not have this property.

Benjamini, Krieger and Yekutieli (2001) have proposed a promising improvement of the method of Benjamini and Hochberg (1995). However, our program for calculating λ cannot be used for their method which is a two staged SU procedure.

5. Discussion

With FWE controlling methods, the sample sizes needed to achieve a desired level of power increase, if the number k of hypotheses becomes large. FDR controlling methods were developed with the hope that they do not have this negative property to the same extend. We compared powers and sample sizes of different FWE and FDR controlling methods in multiple comparisons of k treatments with a control when the observations are normally distributed. Our calculations were for $k = 5, 10$ and 100. The statements with small and large k do not differ. There are no essential differences between the any-pair powers of the different methods, except for the FDR controlling SU procedure of Benjamini and Yekutieli (2001). However, concerning the all-pairs power and per-pair power the FDR controlling SD procedure of Troendle (2000) dominates. The FDR controlling SD procedures of Benjamini and Liu (1999) and Benjamini and Liu (2001) which do not much differ in their powers are worse than the best FWE controlling procedures if most hypotheses are true, and they are better than the FDR controlling SU procedure of Benjamini and Hochberg (1995) if most hypotheses are false. If all hypotheses are false the SU procedure of Kwong, Holland and Cheung (2002) shows rather high power values.

Concerning the sample size which is necessary to guarantee a specified all-pairs power, the advantage of the FDR controlling procedures of Troendle (2000), Benjamini and Hochberg (1995), Sarkar (2002) and Kwong, Holland, Cheung (2002) over all FWE controlling methods considered becomes greater with increasing k.

Acknowledgments

We are grateful to James Troendle for computing the critical constants of method Troe for $k = 5$ and $k = 10$ to a high level of accuracy. We are much obliged to Paul Somerville for doing the power simulations for $k = 100$ in Figures 5 and 8. We acknowledge the helpful comments and advises of the two referees. We thank the editors for valuable comments and suggestions.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300. MR1325392

[2] Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82**, 163–170. MR1736441

[3] Benjamini, Y. and Liu, W. (2001). A distribution-free multiple-test procedure that controls the false discovery rate. Manuscript available from FDR Website of Y. Benjamini.

[4] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188. MR1869245

[5] Benjamini, Y., Krieger, A. and Yekutieli, D. (2001). Two staged linear step up FDR controlling procedure. Manuscript available from FDR Website of Y. Benjamini.

[6] Dunnett, C. W. and Tamhane, A. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statist. Med.* **10**, 939–947.

[7] Dunnett, C. W. and Tamhane, A. (1992). A step-up multiple test procedure. *J. Amer. Statist. Assoc.* **87**, 162–170. MR1158635

[8] Dunnett, C. W., Horn, M. and Vollandt, R. (2001). Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. *J. Statist. Plann. Inference* **97**, 367–384. MR1861160

[9] Einot, I. and Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *J. Amer. Statist. Assoc.* **70**, 574–583.

[10] Holm, S. (1979). A simple sequentially rejective multiple testing procedure. *Scand. J. Statist.* **6**, 65–70. MR538597

[11] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802. MR995126

[12] Kwong, K. S., Holland, B. and Cheung, S. H. (2002). A modified Benjamini-Hochberg multiple comparisons procedure for controlling the false discovery rate. *J. Statist. Plann. Inference* **104**, 351–362. MR1906017

[13] Liu, W. (1997). On step-up tests for comparing several treatments. *Statistica Sinica* **7**, 957–972. MR1488653

[14] Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. Comments. *J. Amer. Statist. Assoc.* **73**, 479–487.

[15] Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: A proof of Simes' conjecture. *Ann. Statist.* **26**, 494–502. MR1626047

[16] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 239–257. MR1892663

[17] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754. MR897872

[18] Troendle, J. F. (2000). Stepwise normal theory multiple test procedures controlling the false discovery rate. *J. Statist. Plann. Inference* **84**, 139–158. MR1747501