On Identification of the Number of Best Treatments Using the Newman-Keuls Test

Samuel S. Wu[*,1,2], **Weizhen Wang**[3], and **David H. Annis**[4]

[1] Division of Biostatistics, University of Florida, Gainesville, FL 32610, USA.
[2] Department of Statistics, Tianjin University of Finance and Economics, Tianjin 300222, P.R. China
[3] Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435, USA
[4] Wachovia Bank, 201 North Tryon Street, 22nd Floor, Charlotte, NC 28288-0040, USA

Received 5 September 2007, revised 14 February 2008, accepted 28 June 2008

Summary

In this paper, we provide a stochastic ordering of the Studentized range statistics under a balanced one-way ANOVA model. Based on this result we show that, when restricted to the multiple comparisons with the best, the Newman–Keuls (NK) procedure strongly controls experimentwise error rate for a sequence of null hypotheses regarding the number of largest treatment means. In other words, the NK procedure provides an upper confidence bound for the number of best treatments.

Key words: Analysis of variance; Experimentwise error rate; Multiple comparison with the best; Newman–Keuls test; Step-down procedure; Stochastic ordering.

1 Introduction

Consider a balanced one-way analysis of variance model, with independent observations $X_{ij} \sim N(\mu_i, \sigma^2)$, $1 \le i \le k, 1 \le j \le n$. By convention, we assume that larger means are preferable to smaller means. In this paper, we are concerned with the inference about the number of best treatments whose means equal $\mu_{(k)} = \max_{1 \le i \le k} \mu_i$.

First, we formulate a sequence of null hypotheses concerning the number of largest treatment means. Without loss of generality, we assume each of k treatments satisfies $\mu_i \le 0$ with at least one treatment achieving equality (i.e., $\mu_{(k)} = 0$). Let

$$N = \text{the number of } \mu_i \text{ being zero (i.e. the number of the best treatments)}. \tag{1}$$

Conversely the number of inferior treatments is $k - N$. For each integer $m \in [2, k]$, consider testing:

$$H_{0,m} : N \ge m \text{ vs. } H_{A,m} : N \le m - 1 \tag{2}$$

and define a corresponding parameter configuration for each $H_{0,m}$,

$$\mu_m = (\mu_1, ..., \mu_k) =: (0, \cdots, 0, -\infty, \cdots, -\infty)\,, \tag{3}$$

where the first m components are zero. Finally define

$$\mathcal{B} = \{H_{0,m} : 2 \le m \le k\}\,, \tag{4}$$

as the set containing all null hypotheses of interest in this paper.

Notice that $H_{0,m_2} \subset H_{0,m_1}$ for any $m_1 < m_2$; thus B is closed under intersection. Therefore, we may apply the closed test procedure (Marcus, Peritz, and Gabriel, 1976) to conduct simultaneous tests for all hypotheses concerning N, the exact number of largest treatment means.

[*] Corresponding author: e-mail: samwu@biostat.ufl.edu, Phone: +001 352 265 8035 x 86544, Fax: +001 352 265 8047

In particular the Newman–Keuls (NK) step-down test (Keuls, 1952), adapted from all pairwise comparisons to the set of multiple comparisons with the best, can be described as follows. Let $\hat{\mu}_i = \sum_{j=1}^{n} X_{ij}/n$ and $\hat{\sigma}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \hat{\mu}_i)^2/(k(n-1))$ be the familiar estimates of respective treatment means and common variance. Note that each $\hat{\mu}_i$ has a normal sampling distribution, while $k(n-1)\,\hat{\sigma}^2/\sigma^2$ is χ^2_ν-distributed with $\nu = k(n-1)$ degrees of freedom. Denote by T_i the Studentized statistics $\sqrt{n}\,\hat{\mu}_i/\hat{\sigma}, 1 \le i \le k$ and, for any integer $m \in [2, k]$, define the m range statistic

$$W_m = T_{[k]} - T_{[k-m+1]} = \frac{\sqrt{n}(\hat{\mu}_{[k]} - \hat{\mu}_{[k-m+1]})}{\hat{\sigma}}, \tag{5}$$

where $[1], \ldots, [k]$ are the random integers satisfying $\hat{\mu}_{[1]} \le \ldots \le \hat{\mu}_{[k]}$. The NK procedure proceeds as follows:

- **Step 1** Test $H_{0,k} : N \ge k$ vs. $H_{A,k} : N \le k-1$ by comparing the k range W_k to q_k, the critical value proposed by Tukey (1953) for all pairwise comparisons (see Appendix 1 for details). If W_k is no larger than q_k, we fail to reject $H_{0,k}$ and conclude none of the treatment means are significantly different and stop. Otherwise we reject $H_{0,k}$ and move to
- **Step 2** Test $H_{0,k-1} : N \ge k-1$ vs. $H_{A,k-1} : N \le k-2$ by comparing the $k-1$ range W_{k-1} with q_{k-1}. Again, if W_{k-1} is no larger than q_{k-1}, we fail to reject $H_{0,k-1}$ and conclude that the $k-1$ largest means are not significantly different and stop. Otherwise we reject $H_{0,k-1}$ and move to the next step.
  ⋮
- **Step $k-1$** Test $H_{0,2} : N \ge 2$ vs. $H_{A,2} : N \le 1$ by comparing W_2 with q_2. If W_2 is no larger than q_2, we fail to reject $H_{0,2}$ and conclude that the two largest means are not significantly different and stop. Otherwise, conclude that there is a single largest mean and stop.

In this article, we show that the NK procedure strongly controls experimentwise error rate for $\mathcal{B}$, the sequence of null hypotheses regarding the number of largest treatment means. In practice, it would be of interest to identify as many inferior treatments as possible under the condition that there is at most α probability of claiming any of the best treatments to be inferior; or equivalently select a subset of treatments such that it contains *all* of the best treatments with a pre-specified minimum probability of $1-\alpha$. Finner and Giani (1994) conjectured that the NK procedure may be the best among step-down procedures in identifying inferior treatments. Note a Type I error regarding $\mathcal{B}$ is an error in identifying the set of inferior treatments. Therefore, our result on strong control of experimentwise error rate for test of N is a necessary condition for the conjecture to hold. However, we are unable to prove the stronger statement that the NK procedure controls the Type I error of claiming any of the best treatments to be inferior.

Various types of selection procedures have been developed in the literature, pioneered by the works of Bechhofer (1954) and Gupta (1956). The indifference zone formulation (Bechhofer, 1954) requires that the probability of selecting the best population is at least $1-\alpha$ whenever there is only one best treatment and it exceeds other means by a particular threshold representing the quantity of indifference to the user. The subset selection formulation (Gupta, 1956) chooses a nonempty subset of the populations so that the selected subset (of random size) includes *one* of the best population with a high probability. If d_{k-1} is the critical value used in Dunnett's one-sided multiple comparison with control method (see Appendix 1 for definition), then Gupta's method identifies the subset $I = \{[i] : T_{[i]} < T_{[k]} - d_{k-1}\}$ as consisting of inferior treatments. More recent developments include enhanced two-stage selection procedures (Chick and Inoue, 2001) and fully sequential procedures (Kim and Nelson, 2001; Chen and Kelton, 2005). However, none of these procedures controls the probability of including *all* best treatments, which is accomplished by the four methods introduced below.

Lam (1986) proposed some *single-step* procedures to identify a subset of "good" populations whose means are within ε of the best, i.e. $\mu_i \ge \mu_{[k]} - \varepsilon$. This requires the probability of including all "good" treatments to be no less than $1-\alpha$. In the special case of selecting the best populations ($\varepsilon = 0$), the procedure identifies the subset $I = \{[i] : T_{[i]} < T_{[k]} - q_k\}$ as consisting of inferior treatments.

From the perspective of multiple comparisons, Edward and Hsu (1983) provided simultaneous confidence intervals for the difference between each treatment and the best. These intervals may be used to identify the inferior treatments.[1] More specifically, if $|d|_{k-1}$ denotes the critical value used in Dunnett's two-sided multiple comparison with control method (also see Appendix 1 for details), then the unconstrained multiple comparison with the best (UMCB) method of Edward and Hsu (which we abbreviate as EH) identifies inferior treatments as $I = \{[i] : T_{[i]} < \min G - |d|_{k-1}\}$, where $G = \{T_{[j]} : T_{[j]} > T_{[k]} - |d|_{k-1},\ 1 \le j \le k\}$. That is, it first selects a subset of good treatments denoted by G and subsequently designates those treatments at least $|d|_{k-1}$ worse than the worst element in G as inferior.

Broström (1981) and Finner and Giani (1994) considered a step-down subset selection procedure. With critical values $c_{k,m}, 2 \le m \le k$, defined in Appendix 1, their procedure (which we abbreviate as BFG) identifies inferior treatment as follows:

- **Step 1** Start with $W_k = T_{[k]} - T_{[1]}$. If $W_k \le c_{k,k}$, then conclude that there is no inferior treatment and stop; otherwise, conclude that treatment $[1]$ is inferior and go to step 2.
- **Step 2** If $W_{k-1} = T_{[k]} - T_{[2]} \le c_{k,k-1}$, then stop; otherwise, conclude that treatments $[1]$ and $[2]$ are inferior and go to step 3.

⋮

- **Step k-1** If $W_2 = T_{[k]} - T_{[k-1]} \le c_{k,2}$, then conclude that treatments $[1], \ldots, [k-2]$ are inferior; otherwise, conclude that treatments $[1], \ldots, [k-1]$ are inferior.

Finner and Giani (2001) and Finner, Giani and Strassburger (2006) also studied selection of good treatments, providing least favorable parameter configurations.

Based upon an acceptance set approach, Hayter (2007) proposed a new step-down procedure with the primary goal of identifying as many treatments as possible to be strictly inferior to the best treatments. Instead of $c_{k,m}$, his procedure uses critical values $w_{k,m}$ as defined in Appendix 1. Generally, we have $c_{k,m} > w_{k,m}$, $\forall\, 2 \le m \le k$ (which are known to hold when $k \le 5$), thus the Hayter procedure is more powerful than the BFG method. The Hayter procedure also allows construction of a set of simultaneous confidence intervals which indicate *how* inferior each treatment could be relative to the best one.

All procedures listed above may be used to conduct simultaneous test of $\mathcal{B}$ as follows: reject hypotheses $H_{0,m}$ for $k - |I| + 1 \le m \le k$ but accept $H_{0,m}$ for $m \le k - |I|$, where $|I|$ is number of identified inferior treatments. Since all methods (Lam, EH, BFG and Hayter) excluding the Gupta procedure control the Type I error of claiming any of the best treatments to be inferior, they all control the experimentwise error rate for the simultaneous test of $\mathcal{B}$. However, because $q_k > c_{k,m} > q_m$ and $q_k > w_{k,m} > q_m$, $\forall\, 2 \le m \le k-1$, the Lam, BFG and Hayter procedures are always less powerful than the NK method for the simultaneous test regarding the number of best treatments.

The remainder of the article is organized as follows. In Section 2, we state the main results: first we establish a monotone property of the range statistics, which identifies the least favorable distributions in the null hypotheses; then we prove that the NK test strongly controls experimentwise error rate. An illustrative example is presented in Section 3 and some simulation results are given in Section 4. The technical details are deferred to the Appendices.

2 Main Results

2.1 A monotone property

In this section, we establish a stochastic ordering of the Studentized range statistics as the μ's change. Theorem 1 below gives our first result. The proof is deferred to Appendix 2.

[1] The constrained multiple comparison with the best (CMCB) method of Hsu (1984) may also be used to identify the inferior treatments. It identifies the same inferior treatments as Gupta's (1956) method. See Hsu (1996) for more details.

Theorem 2.1 *For any $2 \leq m \leq k$, the random variable W_m is stochastically largest at μ_m among $H_{0,m}$. Hence, the rejection region $R_m = \{W_m > q_m\}$ is a level-α test for $H_{0,m}$.*

2.2 Strong control of the experimentwise error rate

To conduct the simultaneous tests for $\mathcal{B} = \{H_{0,m} : 2 \leq m \leq k\}$,

$$\text{reject } H_{0,m} \text{ (i.e., assert } H_{A,m}) \text{ if } R_j \text{ is true for all } m \leq j \leq k. \tag{6}$$

Notice two facts: (i) $\mathcal{B}$ is closed under intersection, and (ii) for each $2 \leq m \leq k$, R_m is a level-α test for $H_{0,m}$ by Theorem 1. Therefore, the experimentwise error rate is no greater than α by the closed test procedure proposed by Marcus, Peritz, and Gabriel (1976). In summary, we have our second result.

Theorem 2.2 *The simultaneous tests (6) for $\mathcal{B}$ (i.e., using the NK procedure for MCB) control the experimentwise error rate at α in the strong sense.*

Equivalently, the above result can be stated in terms of upper confidence bound for the number of best treatments N. More formally, we summarize this as the following Corollary.

Corollary 2.3 *Let $\hat{N}$ be the largest index m such that $H_{0,m}$ is not rejected by the NK procedure, that is,*

$$\hat{N} = \max\{\{1\} \cup \{i : i \in [2,k], W_i \leq q_i\}\}. \tag{7}$$

Then $\hat{N}$ is a proper upper confidence bound for the number of best treatments N, i.e., $P(\hat{N} \geq N) \geq 1 - \alpha$.

3 An Illustrative Example

In dose-response studies, after identifying effective doses that have mean response significantly better than the mean response of the control, it is of great interest to find the smallest dose beyond which no further beneficial effects are seen (International Conference on Harmonization, 1994). This is called *the maximum useful dose*, formally defined as $\text{MUD} = \min\{i : \mu_i \leq \min_{j \geq i} \mu_j\}$, where the treatment indices $1, \ldots, k$ correspond to increasing dose levels of a drug. A confidence lower bound for the MUD can be obtained by first selecting a subset that contains all best doses and then choosing the smallest dose among the selected. To illustrate the method, consider the dose-response data taken from Miyazaki et al. (2002), which reported the dose-response effects of pioglitazone on glycemic control, insulin sensitivity, and insulin secretion in patients with type 2 diabetes. Table 1 presents the mean and SD (SEM was reported in the original paper) of the change in plasma glucose concentration before and after 26 weeks of treatments, where the smaller means are preferable to larger means. The Bartlett's test fails to reject the hypothesis of equal variance, so the proposed procedure is applied under the equal variance assumption.

Table 1 The summary statistics for the change of plasma glucose concentration from Miyazaki et al. (2002).

Group	Pioglitazone dose (mg/day)	Sample		SD
i		size	mean	
1	Placebo	11	14	112.8
2	7.5	12	−14	65.8
3	15	13	−3	61.3
4	30	11	−70	63.0
5	45	11	−94	69.6

Table 2 The range statistics W_m and critical values $q_m, c_{k,m}, w_{k,m}$ in the case of $\alpha = 0.05$ and $k = 5$ for the dose-response data.

m	W_m	Corresponding dose	q_m	$c_{k,m}$	$w_{k,m}$
5	4.6979	Placebo	3.9885	3.9885	3.9885
4	4.1200	15	3.7467	3.8782	3.8557
3	3.5547	7.5	3.4082	**3.7334**	**3.6406**
2	1.0440	30	**2.8310**	3.5224	3.1967

Compared to the Placebo group, patients treated with 30 and 45 mg/day of pioglitazone have significantly more decreases in mean plasma glucose concentration, indicating the beneficial effect for the two highest doses. To identify the maximum useful dose, we compare all groups with the observed best treatment, the 45 mg/day group. The proposed procedure requires equal sample size across groups. However, since the sample sizes are approximately equal, the range statistics W_m are evaluated based on $\sqrt{2}(\hat{\mu}_{[k-m+1]} - \hat{\mu}_5)/\hat{\sigma}/\sqrt{1/n_5 + 1/n_{[k-m+1]}}$ using the pooled SD estimate of 76.25 with 53 degrees of freedom. Table 2 presents W_m along with the critical values $q_m, c_{k,m}, w_{k,m}$. From the table we see that the NK procedure rejects hypotheses $H_{0,m}$ for $3 \leq m \leq 5$ at $\alpha = 0.05$ and accepts hypothesis $H_{0,2}$. Therefore, we conclude that $N = 2$. Intuitively, since the two largest dose levels ($i = 4$ and 5) had the smallest sample means, their true means are likely to be the smallest. Hence we estimate the maximum useful dose of Pioglitazone as 30 mg/day. On the other hand, the Lam, BFG and Hayter methods would only select the Placebo and 15 mg/day as inferior treatments and claim $N = 3$, i.e., accept 7.5 mg/day as not significantly different from the best. As for the EG procedure, since $|d|_4 = 3.553$, it would first select the two highest doses as good treatments, and then only designate the Placebo as the inferior treatment through comparisons with the 30 mg/day group.

In addition, Strassburger, Bretz and Finner (2007) proposed an ordered multiple comparison with the best procedure that is also valid in the case of unbalanced design, and their procedure yields that 45 mg/day is a lower confidence bound for the MUD.

4 Some Simulation Results

Simulation studies were conducted to assess the effectiveness of different selection procedures: 1) EH: the CMCB method of Edward and Hsu (1983); 2) Lam: Lam's single step method; 3) BFG: the step-down procedure of Broström (1981) and Finner and Giani (1994); 4) Hayter: the step-down method proposed by Hayter (2007) and 5) NK: the NK procedure. The Gupta method was excluded from simulation study because it does not control the Type I error of claiming any of the best treatments to be inferior, even though the procedure can be modified to meet this condition (see Strassburger, Bretz and Finner, 2007). We considered scenarios such that the number of treatments, k, is between three and eight, while the number of best treatments, N, ranges from one to three. It is assumed that the means corresponding to inferior treatments are equally spaced between 0 and 0.8. For example, $\boldsymbol{\mu} = (-0.8, -0.6, -0.4, -0.2, 0, 0)$ when $k = 6$ and $N = 2$. The nominal Type I error rate is set at 0.05, and the within group variance is assumed fixed at $\sigma = 1$.

For each combination of k and N, we evaluated the sample size required by each procedure so that its probability of completely correct selection (excluding all inferior treatments while selecting all best treatments) is 80% based on 100,000 simulations, see Table 3. Note that, in many cases especially when $k = 7$ or $k = 8$, the 80% completely correct selection goal would require sample sizes too large to be useful for usual experiments, in such cases it may be more appropriate to require 80% probability of excluding all inferior treatment. Also presented in Table 3 are the efficiencies of the NK procedure relative to the other methods, where the relative efficiency is defined as the ratio of sample sizes needed by any other procedure and the NK method to achieve 80% probability of completely correct

Table 3 Sample sizes needed by the five selection procedures to ensure that P (Completely Correct Selection) = 0.8. The numbers in parenthesis are relative efficiencies defined as the ratio of sample sizes needed by any other procedure and the NK method.

N	Procedures	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
1	EH	127 (1.30)	303 (1.37)	562 (1.42)	909 (1.48)	1344 (1.52)	1875 (1.55)
	Lam	127 (1.30)	327 (1.48)	639 (1.62)	1065 (1.73)	1621 (1.83)	2632 (2.18)
	BFG	114 (1.16)	283 (1.28)	533 (1.35)	870 (1.41)	1294 (1.46)	1807 (1.50)
	Hayter	108 (1.10)	259 (1.17)	483 (1.22)	782 (1.27)	1163 (1.31)	1612 (1.34)
	NK	98	221	395	615	885	1206
2	EH	39 (1.56)	166 (1.60)	382 (1.63)	704 (1.69)	1123 (1.73)	1650 (1.76)
	Lam	25 (1.00)	109 (1.05)	274 (1.17)	523 (1.26)	865 (1.33)	1520 (1.62)
	BFG	25 (1.00)	104 (1.00)	249 (1.06)	460 (1.11)	747 (1.15)	1103 (1.18)
	Hayter	25 (1.00)	104 (1.00)	245 (1.04)	449 (1.08)	718 (1.10)	1060 (1.13)
	NK	25	104	235	416	650	938
3	EH		44 (1.76)	181 (1.76)	420 (1.82)	767 (1.85)	1221 (1.89)
	Lam		25 (1.00)	108 (1.05)	259 (1.12)	484 (1.17)	929 (1.44)
	BFG		25 (1.00)	104 (1.01)	244 (1.06)	444 (1.07)	709 (1.10)
	Hayter		25 (1.00)	103 (1.00)	240 (1.04)	439 (1.06)	699 (1.08)
	NK		25	103	231	415	645

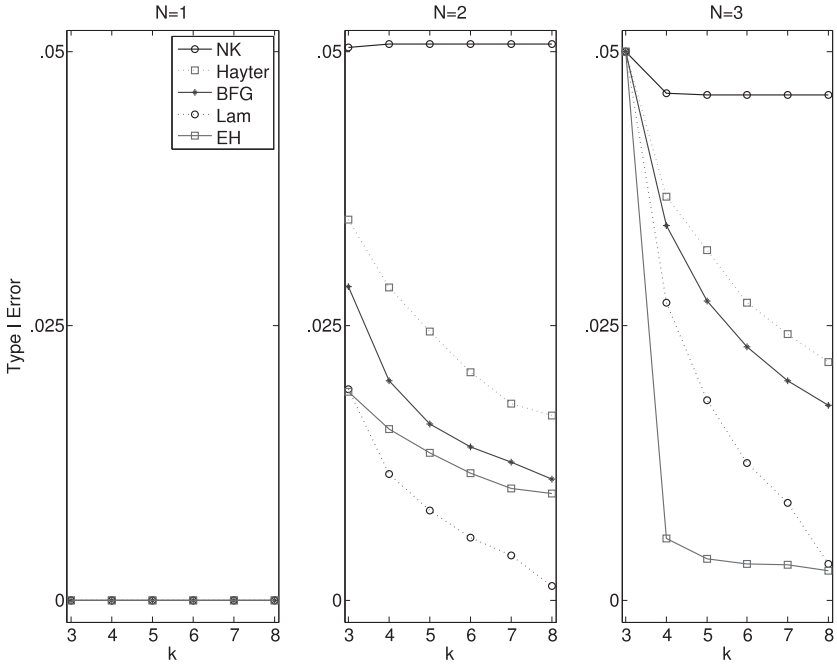

Figure 1 Comparison of the five selection procedures on Type I error of claiming any of the best treatments to be inferior. All procedures control the Type I error. When $N = 1$ these procedures are overly conservative – they almost never identify the besttreatment as inferior. In addition, the figure shows that the NK procedure is less conservative than the Hayter method, which is less conservative than the BFG method.

selection. Table 3 shows that the NK procedure is more efficient than all other methods, as implied by the theoretical results. It is worth noting that the efficiencies tend to increase in k for a fixed N, and decrease in N for a fixed k (except for the EH method). For example, when $k = 8$, the efficiencies of the NK procedure relative to the Hayter method are 134%, 113%, 108%, respectively as N goes from 1 to 3. Furthermore, the EH method becomes worse than the Lam method for the cases where $N > 1$.

Figure 1 compares the five selection procedures on frequency of Type I error (i.e. of claiming any of the best treatments to be inferior). It should be kept in mind that the comparison is based on different sample sizes (as presented in Table 3) for the procedures. The figure indicates that all procedures control the Type I error but are overly conservative when there is only one best treatment and the inferior treatments are equally spaced between zero and 0.8 times standard deviations from the best. These procedures seldom identify the best treatment as inferior if sample sizes are large enough to guarantee that the probability of completely correct selection is 80%. In addition, the figure shows that the NK procedure is less conservative than the Hayter method, which, in turn, is less conservative than the BFG method, consistent with the theoretical results.

5 Discussion

In this paper, we provide a stochastic ordering of the Studentized range statistics under a balanced one-way ANOVA model. Based on this result we show that, when restricted to the multiple comparisons with the best, the Newman–Keuls (NK) procedure strongly controls experimentwise error rate for a sequence of null hypotheses regarding the number of largest treatment means. It is worth mentioning that, if we have some prior knowledge or believe that there are at most $k_0 < k$ largest means, we may conduct the simultaneous tests for the subset $\mathcal{B}_{k_0} = \{H_{0,m} : 2 \le m \le k_0\}$ of $\mathcal{B}$.

Our hypotheses consider N, the number of best treatments. However, the Lam, BFG, EH, and Hayter methods all strongly control experimentwise error of claiming any of the best treatments to be inferior. In other words, they are concerned with the multiple testing problem of testing the family of null hypotheses $\mathcal{H} = \{H_i : 1 \le i \le k\}$, where $H_i = \{\mu_i = \mu_{(k)}\}$ is the hypothesis that treatment i is a best treatment. A connection between $\mathcal{B}$ and the closure of $\mathcal{H}$ can be established as follows. Let $H_J = \cap_{j\in J} H_j = \{\boldsymbol{\mu} : \mu_{(k)} = \min_{j\in J} \mu_j\}$, then $H_{0,m}$ can be expressed as $H_{0,m} = \cup_{J:|J|=m} H_J$. Thus, the test that rejects $H_{0,m}$ if $W_m > q_m$ can be interpreted as a test for the union of all intersection hypotheses H_J with $|J| = m$. In addition, the hypotheses in $\mathcal{B}$ can also be written in terms of ranked means ($H_{0,m} = \{\boldsymbol{\mu} : \mu_{(k-m+1)} = \mu_{(k)}\}$). It would be an equally interesting problem to show that the NK test also strongly control experimentwise error with regard to $\mathcal{H}$, i.e., when we determine $\hat{N}$, we claim that not only $N \le \hat{N}$, but also $\mu_{[1]}, \ldots, \mu_{[k-\hat{N}]}$ are the treatments inferior to the best. This problem is very challenging and deserves further study. Nevertheless, if one makes a Type I error regarding N, then an error is made for identifying the set of inferior treatments. Therefore, a strong control of experimentwise error rate on N is a necessary condition for that on the set of inferior treatments.

Corollary 2.3 provided an upper confidence bound for the number of best treatments N. It remains unknown whether the confidence bound $\hat{N}$ can still be improved and whether it is a median unbiased estimate for N in the case of $\alpha = 0.5$. It may also be of interest to construct a lower confidence bound or a two-sided confidence interval for N.

6 Appendix 1. Definitions of Critical Values

Suppose $Z_i \sim N(0,1), 1 \le i \le k$ and $U \sim \sqrt{\chi^2_\nu/\nu}$ are independent random variables. For any given α, we define critical values $d_{\alpha,k-1,\nu}$, $|d|_{\alpha,k-1,\nu}$ and $q_{\alpha,k,\nu}$ such that:

$$
\begin{aligned}
&P\left(\max_{1\le i\le k-1} Z_i - Z_k \le d_{\alpha,k-1,\nu} U\right) = 1-\alpha\,,\\
&P\left(\max_{1\le i\le k-1} |Z_i - Z_k| \le |d|_{\alpha,k-1,\nu} U\right) = 1-\alpha\,,\\
&P\left(\max_{1\le i,j\le k} |Z_i - Z_j| \le q_{\alpha,k,\nu} U\right) = 1-\alpha\,. \qquad (8)
\end{aligned}
$$

These critical values are used in Dunnett's one-sided multiple comparison with control (MCC) method, Dunnett's two-sided MCC method and Tukey's all pairwise comparisons, respectively. When there is no ambiguity concerning α and ν, we suppress them notationally and denote these critical values by $d_{k-1}, |d|_{k-1}$ and q_k instead of the more cumbersome $d_{\alpha,k-1,\nu}, |d|_{\alpha,k-1,\nu}$ and $q_{\alpha,k,\nu}$. It is worth mentioning that for fixed α and ν, these critical values satisfy $d_{k-1} < |d|_{k-1} < q_k$ and they are increasing in k.

For the step-down procedures proposed by Broström (1981) and Finner and Giani (1994), we define critical values $c^{\alpha}_{k,m,\nu}$ such that $c^{\alpha}_{k,k,\nu} = q_{\alpha,k,\nu}$ and $c^{\alpha}_{k,m,\nu}$ satisfies

$$P\begin{pmatrix} \max_{1\le a\le k-m} Z_a - \min_{k-m+1\le b\le k} Z_b \le c^{\alpha}_{k,m,\nu} U \\ \max_{k-m+1\le i,j\le k} |Z_i - Z_j| \le c^{\alpha}_{k,m,\nu} U \end{pmatrix} = 1-\alpha\,. \tag{9}$$

Furthermore, for the Hayter method, the critical values $w^{\alpha}_{k,m,\nu}$ are iteratively defined by

$$P\begin{pmatrix} \max_{1\le a\le k-m} Z_a - \min_{k-m+1\le b\le k} Z_b \le w^{\alpha}_{k,m+1,\nu} U \\ \max_{k-m+1\le i,j\le k} |Z_i - Z_j| \le w^{\alpha}_{k,m,\nu} U \end{pmatrix} = 1-\alpha\,, \tag{10}$$

with initial values $w^{\alpha}_{k,k,\nu} = q_{\alpha,k,\nu}$. Once again, we suppress α and ν in the notation when there is no ambiguity, and denote these critical values by $c_{k,m}$ and $w_{k,m}$. Note that $c_{k,m}$ and $w_{k,m}$ must lie between q_m and q_k. Furthermore numerical investigations have shown that for common α and ν, $c_{k,m} > w_{k,m}$, $\forall 2 \le m \le k$ (this is true for all α and ν when $k \le 5$), which implies that the Hayter procedure is more powerful than the BFG method.

7 Appendix 2. Proof of Theorem 2.1

It suffices to show that for any $\boldsymbol{\mu} \in H_{0,m}$ and any constant c,

$$P\boldsymbol{\mu}_m(W_m \le c) \le P\boldsymbol{\mu}(W_m \le c)\,. \tag{11}$$

Let $X_i = \sqrt{n}\,\hat{\mu}_i/\sigma, 1 \le i \le k$ when the true treatment means are $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_k)$, while $Y_i = \sqrt{n}\,\hat{\mu}_i/\sigma, 1 \le i \le k$ when the true treatment means are $\boldsymbol{\mu}_m$. Since $\boldsymbol{\mu} \in H_{0,m}$, it has at least m components equal to zero. Without loss of generality we assume that $\mu_1 = \mu_2 = \cdots = \mu_m = 0$, hence $(X_1, \cdots, X_m)$ and $(Y_1, \cdots, Y_m)$ have the same multivariate standard normal distributions.

By Lemma 2.1 in Liu (1995), $P(X_{(k)} - X_{(k-m+1)} \le c)$ is nondecreasing in $\mu_{(1)} = \min_{1\le i\le k}\mu_i$. Therefore $P(X_{(k)} - X_{(k-m+1)} \le c)$ achieves minimum at $\mu_{(1)} = -\infty$. Hence we have

$$P(Y_{(k)} - Y_{(k-m+1)} \le c) \le P(X_{(k)} - X_{(k-m+1)} \le c), \quad \forall\, c\,. \tag{12}$$

Secondly, if $\gamma(s)$ is the probability density function of $\hat{\sigma}/\sigma$,

$$\begin{aligned} P\boldsymbol{\mu}_m(W_m \le c) &= P\boldsymbol{\mu}_m\left(\frac{\sqrt{n}(\hat{\mu}_{(k)} - \hat{\mu}_{(k-m+1)})}{\hat{\sigma}} \le c\right) \\ &= P\boldsymbol{\mu}_m(\sqrt{n}\,\hat{\mu}_{(k)}/\sigma - \sqrt{n}\,\hat{\mu}_{(k-m+1)}/\sigma \le c\hat{\sigma}/\sigma) \\ &= P(Y_{(k)} - Y_{(k-m+1)} \le c\hat{\sigma}/\sigma) \\ &= \int \gamma(s)\, P(Y_{(k)} - Y_{(k-m+1)} \le cs)\,\mathrm{d}s. \end{aligned} \tag{13}$$

Similarly, when the true treatment means are $\boldsymbol{\mu}$,

$$\begin{aligned} P_{\boldsymbol{\mu}}(T_m \le c) &= P(X_{(k)} - X_{(k-m+1)} \le c\hat{\sigma}/\sigma) \\ &= \int \gamma(s)\, P(X_{(k)} - X_{(k-m+1)} \le cs)\,\mathrm{d}s\,. \end{aligned} \tag{14}$$

The proof is complete by noting that (12), (13) and (14) together imply (11).

Acknowledgements *The authors would like to thank the Editors and two anonymous referees for their insightful comments and valuable suggestions.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* **25**, 16–39.

Broström, G. (1981). On sequentially rejective subset selection procedures. *Communications in Statististics – Theory and Methods* **10**, 203–221.

Chen, E. J. and Kelton, W. D. (2005). Sequential selection procedures: Using sample means to improve efficiency. *European Journal of Operational Research* **166**, 133–153.

Chick, S. E. and Inoue, K. (2001). New two-stage and sequential procedures for selecting the best simulated system. *Operational Research* **49**, 732–743.

Edwards, D. G. and Hsu, J. C. (1983). Multiple comparisons with the best treatment. *Journal of the American Statistical Association* **78**, 965–971.

Finner, H. and Giani, G. (1994). Closed subset selection procedures for selecting good populations. *Journal of Statistical Planning and Inference* **38**, 179–200.

Finner, H. and Giani, G. (2001). Least favourable parameter configurations for a step-down subset selection procedure. *Biometrical Journal* **43**, 543–552.

Finner, H., Giani, G., and Strassburger, K. (2006). Partitioning principle and selection of good treatments. *Journal of Statistical Planning and Inference* **136**, 2053–2069.

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Ph.D. thesis, Institute of Statistics, University of North Carolina, Chapel Hill.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* **7**, 225–245.

Hayter, A. J. (2007). A combination multiple comparisons and subset selection procedure to identify treatments that are strictly inferior to the best. *Journal of Statistical Planning and Inference* **137**, 2115–2126.

Hsu, J. C. (1984). Ranking and selection and multiple comparisons with the best, *Design of Experiments: Ranking and Selection.* Marcel Dekker, New York.

Hsu, J. C. (1996). *Multiple Comparisons, Theory and Methods*. Chapman & Hall, New York.

International Conference on Harmonization (1994). *Dose-response information to support drug registration, Guideline E4.* London: Committee for Proprietary Medical Products, The European Agency for the Evaluation of Medical Products.

Kim, S. H. and Nelson, B. L. (2001). A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on modeling and computer simulation* **11**, 251–273.

Keuls, M. (1952). The use of the "Studentized range" in connection with an analysis of variance. *Euphytica* **1**, 112–122.

Lam, K. (1986). A new procedure for selecting good populations. *Biometrika* **73**, 201–206.

Liu W. (1995). On the construction of exact lower confidence bounds on the distance between any two ranked location parameters of K populations. *Communications in Statististics – Theory and Methods* **24**, 2073–2085.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Miyazaki, Y., Matsuda, M., and DeFronzo R. A. (2002). Dose-response effect of pioglitazone on insulin sensitivity and insulin secretion in type 2 diabetes. *Diabetes Care* **25**, 517–523.

Strassburger, K., Bretz, F., and Finner H. (2007). Ordered multiple comparisons with the best and their applications in dose-response studies. *Biometrics* **63**, 1143–1151.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Department of Statistics. Princeton University, Princeton, NJ.

