Statistica Sinica **7**(1997), 957-972

ON STEP-UP TESTS FOR COMPARING SEVERAL TREATMENTS

W. Liu

University of Southampton

Abstract: Consider a one-way layout in which independent observations X_i have distributions $F(x - \mu_i)$ with μ_i unknown, $1 \leq i \leq k$. The basic problem is to test the family of subset hypotheses about μ_i. The step-up test scheme of Welsch (1977) is studied in detail. A general result is given on how to determine the critical values so that the type I familywise error rate is strongly controlled at a preassigned level α, and a particular set of critical values is recommended. The interrelationship between the step-up tests and the closed tests of Marcus, Peritz and Gabriel (1976) is illuminated. This enables us to construct a closed test which is uniformly more powerful than a step-up test. Monte Carlo simulation is carried out to compare the power of the two new tests with that of Welsch's (1977) test and several other multiple tests. The two new tests turn out to be preferable. The one-way ANOVA setting is given special attention, and the critical values of the new step-up test are tabulated.

Key words and phrases: Multiple tests, power, type I familywise error rate.

1. Introduction

Suppose that X_i, $1 \leq i \leq k$, are independent random observations with distributions $F(x - \mu_i)$, where the μ_i are unknown parameters. Consider the problem of testing the family of subset hypotheses $\{H_P : \forall\ P \subseteq K \text{ with } |P| \geq 2\}$, where $K = \{1, \ldots, k\}$, $|A|$ denotes the number of elements in set A, and H_P is the hypothesis that all the μ_i, $i \in P$ are equal. There is only one step-up test for this family of hypotheses in the literature, which is proposed by Welsch (1977) and denoted as WELUP in the paper.

For m given numbers a_i, $i \in I \subset \{1, 2, \ldots\}$ where $|I| = m$, let the ordered a_i values be denoted as $a_{[1]_I} \leq a_{[2]_I} \leq \cdots \leq a_{[m]_I}$. So $X_{[1]_K} \leq X_{[2]_K} \leq \cdots \leq X_{[k]_K}$ are the ordered random observations. Hochberg and Tamhane (1987, pp.124) describe Welsch's (1977) step-up test, based on the critical values $\xi_2 \leq \xi_3 \leq \cdots \leq \xi_k$, as follows:

Step 1. Begin by testing all the "gaps" or "2-ranges", $X_{[i+1]_K} - X_{[i]_K}$ $(1 \leq i \leq k-1)$, by comparing them with the critical value ξ_2. If $X_{[i+1]_K} - X_{[i]_K} > \xi_2$, then declare that gap as significant and the corresponding pair of treatments as different. Also declare all sets of treatments containing that pair as heterogeneous and all the p-ranges containing that gap as significant by implication without further tests. Proceed to step 2 if at least two adjacent gaps are not declared significant.

Step 2. In general, test a q-range $X_{[i+q-1]_K} - X_{[i]_K}$ $(1 \leq i \leq k-q+1, 2 \leq q \leq k)$, using the critical value ξ_q, if that q-range is not declared significant by implication at an earlier step. If $X_{[i+q-1]_K} - X_{[i]_K} > \xi_q$, then declare that q-range significant and the pair of treatments corresponding to $X_{[i+q-1]_K}$ and $X_{[i]_K}$ different. Also declare all sets of treatments containing that subset of q treatments heterogeneous and all the p-ranges containing that q-range with $p > q$ significant by implication without further tests. Continue in this manner until no ranges remain to be tested that are not already declared significant.

One important question is how to determine the values of ξ_i so that the type I familywise error (FWE) rate is strongly controlled at a preassigned level α, i.e. the probability of a false rejection of any H_P is at most α, irrespective of which and how many of the H_P are true. WELUP uses a set of ξ_i values which is based on the Bonferroni inequality. In this paper a general result is established on how to determine the critical values ξ_i so that the type I FWE rate is strongly controlled at level α. A particular set of critical values is recommended, and the corresponding test turns out to be slightly more powerful than WELUP. These results are contained in Section 2. In Section 3 the interrelationship between the step-up tests and the closed tests of Marcus, Peritz and Gabriel (1976) is investigated. It is shown that a step-up test is equivalent to a closed test. This result enables us to construct a closed test that is uniformly more powerful than a step-up test. The important setting of normal distributions with a common unknown variance is dealt with in Section 4. To compare the power of the two new tests proposed in this paper with that of WELUP and several other multiple tests, a Monte Carlo simulation study is carried out. The results of this study are reported in Section 5. Finally, Section 6 contains some closing remarks.

Before ending this section, we define some notation which is used throughout this paper. A multiple subset hypothesis $H_{\mathbf{P}}$ is of the form $H_{\mathbf{P}} = \cap_{i=1}^{r} H_{P_i}$ where $\mathbf{P} = (P_1, \ldots, P_r)$ is a "partition" of the set $K = \{1, \ldots, k\}$, that is, the P_i are disjoint subsets of K with $|P_i| = p_i \geq 2$ such that $\sum_{i=1}^{r} p_i \leq k$; here r, the number of homogeneous subsets, is between 1 and $k/2$. In order to discuss the strong control of type I FWE rate and closed tests, it is necessary to consider multiple subset hypotheses.

For $P \subseteq K$ with $|P| = p \geq 2$ and $2 \leq q \leq p$, define

$$R_q(X_i : i \in P) = \max_{1 \leq j \leq p-q+1} (X_{[j+q-1]_P} - X_{[j]_P}) \text{ and}$$
$$\{M(X_i : i \in P) \leq (c_2, \ldots, c_p)\} = \cap_{q=2}^{p} \{R_q(X_i : i \in P) \leq c_q\},$$

where the c_i are given constants.

2. Determination of the Critical Values

The choice of the critical values $\xi_i (2 \le i \le k)$ can be made in terms of the probabilities

$$\alpha_p = 1 - \mathrm{P}_0\{M(X_i : 1 \le i \le p) \le (\xi_2, \ldots, \xi_p)\}, \quad p = 2, \ldots, k, \tag{2.1}$$

where the subscript 0 indicates that the probability is evaluated under $\mu_1 = \cdots = \mu_p = 0$. The following lemma is fundamental in determining the values of α_p $(2 \le p \le k)$ so that the type I FWE rate is strongly controlled at level α.

Lemma 2.1. *Suppose that $\xi_2 \le \cdots \le \xi_k$. Then under $H_{\mathbf{P}} = \cap_{i=1}^r H_{P_i}$ we have*

$$\sup P_{H_{\mathbf{P}}} \{\textit{at least one } H_{P_i} \textit{ is rejected}\} = 1 - \prod_{i=1}^r (1 - \alpha_{p_i}), \tag{2.2}$$

where the sup *is taken over all the possible values of the μ_i under $H_{\mathbf{P}}$, and $p_i = |P_i|$.*

Proof. The key is to observe that under the assumption $\xi_2 \le \cdots \le \xi_k$

$$\{M(X_j : j \in P_i) \le (\xi_2, \ldots, \xi_{p_i})\} \subseteq \{H_{P_i} \text{ is accepted}\}.$$

It follows therefore that

$$\begin{aligned} &\mathrm{P}_{H_{\mathbf{P}}}\{\text{at least one } H_{P_i} \text{ is rejected}\} \\ &= 1 - \mathrm{P}_{H_{\mathbf{P}}}\{\cap_{i=1}^r (H_{P_i} \text{ is accepted})\} \\ &\le 1 - \mathrm{P}_{H_{\mathbf{P}}}\{\cap_{i=1}^r (M(X_j : j \in P_i) \le (\xi_2, \ldots, \xi_{p_i}))\} = 1 - \prod_{i=1}^r (1 - \alpha_{p_i}), \end{aligned} \tag{2.3}$$

where the equality in (2.3) follows from the independence of the X_i, and the last equality follows from the definition of α_p in (2.1). Also note that the only inequality above becomes equality when all the μ_j not in the same P_i spread infinitely apart from each other. The proof is thus completed.

The monotonicity assumption of the ξ_i in the lemma is noteworthy. It can be shown that $\xi_2 \le \xi_3$ is necessary for (2.2), though it is unknown whether $\xi_3 \le \cdots \le \xi_k$ are also necessary for (2.2). Lehmann and Shaffer (1977) show that for the step-down tests the monotonicity of the critical values is not only sufficient but also necessary for a result similar to (2.2).

From lemma 2.1, in order to control strongly the type I FWE rate at level α, the critical values ξ_i should be chosen so that

$$\max_{p_1, \ldots, p_r} \Big[1 - \prod_{i=1}^r (1 - \alpha_{p_i})\Big] \le \alpha, \tag{2.4}$$

where the max is taken over all sets of integers $p_1, \ldots, p_r$ satisfying $p_i \geq 2, \sum_{i=1}^{r} p_i \leq k$ and $1 \leq r \leq k/2$. We can therefore determine, firstly, a set of α_p $(2 \leq p \leq k)$ from (2.4) and, then, the ξ_i $(2 \leq i \leq k)$ from (2.1).

The specification of α_p from (2.4) is discussed by several authors when studying the step-down tests. Ryan (1960) suggests $\alpha_p = \alpha p/k$ $(2 \leq p \leq k)$. Tukey (1953) proposes $\alpha_p = \alpha p/k$ $(2 \leq p \leq k-2)$ and $\alpha_{k-1} = \alpha_k = \alpha$. The Tukey-Welsch specification (Hochberg and Tamhane (1987), pp.69) uses $\alpha_p = 1-(1-\alpha)^{p/k}$ $(2 \leq p \leq k-2)$ and $\alpha_{k-1} = \alpha_k = \alpha$. Lehmann and Shaffer (1979) suggest another specification of α_p which is optimal under a certain criterion.

For step-up tests, an intuitively promising specification of α_p is the one that minimizes ξ_2, then for fixed ξ_2 minimizes ξ_3, etc. This is because step-up tests conclude significance by implication, starting from the 2-ranges to the k-range. To find this specification of α_p, we proceed sequentially: first determine the maximum α_2 from (2.4), then for this α_2 determine the maximum α_3 from (2.4), etc.

The maximum α_2, denoted as α_2^*, can be determined from those constraints in (2.4) that involve α_2 only and are given by $1-(1-\alpha_2)^r \leq \alpha$ for all $1 \leq r \leq k/2$. So $\alpha_2^* = 1-(1-\alpha)^{1/[k/2]}$, where $[x]$ denotes the integer part of x.

After fixing $\alpha_2 = \alpha_2^*$, the maximum α_3, denoted as α_3^*, can be determined from those constraints in (2.4) that involve α_2 and α_3 only and are given by

$$1-(1-\alpha_2)^{r_2}(1-\alpha_3)^{r_3} \leq \alpha \quad \text{for all } r_2 \geq 0, r_3 \geq 1, 2r_2+3r_3 \leq k.$$

Replacing α_2 by α_2^*, we find that α_3 should satisfy

$$\alpha_3 \leq 1-(1-\alpha)^{(1-r_2/[k/2])r_3^{-1}} \quad \text{for all } r_2 \geq 0, r_3 \geq 1, 2r_2+3r_3 \leq k,$$

from which α_3^* can be determined.

Continuing in this manner, we see that the maximum α_p $(4 \leq p \leq k)$, denoted as α_p^*, after fixing $\alpha_i = \alpha_i^*$, $2 \leq i \leq p-1$, can be determined. In fact, when k is even, then α_p^* is the same as the Tukey-Welsch specification, which is given by $\alpha_p^* = 1-(1-\alpha)^{p/k}$ $(2 \leq p \leq k-2)$ and $\alpha_{k-1}^* = \alpha_k^* = \alpha$. When k is odd, α_p^* is the same as the optimal specification of Lehmann and Shaffer (1979), which is given by $\alpha_p^* = 1-(1-\alpha)^{[p/2]/[k/2]}$. Welsch (1977) suggests the specification $\alpha_p' = \alpha p/k$ $(2 \leq p \leq k-2)$ and $\alpha_{k-1}' = \alpha_k' = \alpha$, which is based on the Bonferroni inequality and hence does not need the independence of the X_i.

For a given specification of α_p, either α_p^* or α_p', the corresponding critical values ξ_i can be solved sequentially from (2.1): first ξ_2, then ξ_3, etc. The ξ_i found in this way may not satisfy the monotonicity assumption in Lemma 2.1 however.

To overcome this problem, we replace ξ_i by ξ_{i-1} if $\xi_i < \xi_{i-1}$, and then go on to solve for ξ_{i+1}. It is clear that a step-up test using this set of "modified" critical values still strongly controls the type I FWE rate at level α. The step-up test based on α_p^* is denoted as NEWUP hereafter. WELUP uses the specification α_p'.

Table 1 in Section 4 contains the critical values ξ_i of NEWUP when $F(\cdot)$ is the standard normal distribution. To calculate these ξ_i, it is necessary to evaluate the probability $\mathrm{P}_0\{M(X_j : 1 \le j \le p) \le (\xi_2, \ldots, \xi_p)\}$. This probability is calculated using the expressions given in Liu (1996a) when $p = 3$ and estimated by simulation based on 1,000,000 experiments when $p \ge 4$.

3. Closed Tests

To compare tests of the family of subset hypotheses, the following two definitions, which are similar to those in Liu (1996b), are useful. Two tests are called *equivalent* if they always reach the same decisions, either rejections or acceptances, on all the subset hypotheses H_P. Test A is said to *dominate* test B if test A always rejects at least those H_P rejected by test B. Dominance implies at least as powerful. In this section, we first show that a step-up test, which uses critical values $\xi_2 \le \cdots \le \xi_k$ and strongly controls the type I FWE rate at level α, is equivalent to a closed test. By using this result we demonstrate how to construct a closed test which not only dominates but also is uniformly more powerful than the step-up test.

A closed test (Marcus, Peritz and Gabriel (1976)) has two essential ingredients. The first is a size α test of each multiple subset hypothesis $H_{\mathbf{P}}$. The second is a systematic way of making a decision on each subset hypothesis H_P: reject H_P if and only if all the $H_{\mathbf{P}}$ that imply H_P are rejected by the corresponding tests; $H_{\mathbf{P}} = \cap_{i=1}^r H_{P_i}$ implies H_P if $P \subseteq P_i$ for some $1 \le i \le r$. It is known that a closed test strongly controls the type I FWE rate at α.

Define a closed test which tests each $H_{\mathbf{P}} = \cap_{i=1}^r H_{P_i}$ in the following way

$$\text{Accept } H_{\mathbf{P}} \iff M(X_j : j \in P_i) \le (\xi_2, \ldots, \xi_{p_i}) \quad \forall\ 1 \le i \le r, \tag{3.1}$$

where $p_i = |P_i|$. This test of $H_{\mathbf{P}}$ is clearly of size α from (2.4), and so the closed test strongly controls the type I FWE rate at level α. Now we have

Theorem 3.1. *The closed test defined above is equivalent to the step-up test using critical values $\xi_2 \le \cdots \le \xi_k$.*

Proof. Let (i) denote the index of the population associated with $X_{[i]_K}$, so that $\mu_{(i)}$ is the population mean associated with $X_{[i]_K}$, $1 \le i \le k$. Suppose H_P is accepted by the step-up test. We shall construct an H_Q which implies H_P and is accepted by the corresponding test used in the closed test, and so H_P

itself is accepted by the closed test. Note that H_P can be expressed in the form $H_P : \mu_{(i_1)} = \mu_{(i_2)} = \cdots = \mu_{(i_p)}$, where $X_{[i_1]_K} \leq X_{[i_2]_K} \leq \cdots \leq X_{[i_p]_K}$ are the ordered values of those X_i with $i \in P$, and so $1 \leq i_1 < i_2 < \cdots < i_p \leq k$. Note that, if $H_P : \mu_{(i_1)} = \mu_{(i_2)} = \cdots = \mu_{(i_p)}$ is accepted by the step-up test, then $H_Q : \mu_{(i_1)} = \mu_{(i_1+1)} = \cdots = \mu_{(i_2)} = \mu_{(i_2+1)} = \cdots = \mu_{(i_p)}$ must also be accepted by the step-up test, and so

$$M(X_j : j \in Q) \leq (\xi_2, \ldots, \xi_{i_p - i_1 + 1}). \tag{3.2}$$

Now it is clear that H_Q implies H_P, and that H_Q is accepted by the corresponding test used in the closed test because of (3.2). Consequently, H_P itself is accepted by the closed test.

Next, we show that, if H_P is rejected by the step-up test, then it is also rejected by the closed test. For this it is sufficient to show that all the H_Q that imply H_P are rejected by the corresponding tests used in the closed test. As before, H_P can be written as $H_P : \mu_{(i_1)} = \mu_{(i_2)} = \cdots = \mu_{(i_p)}$. Since H_P is rejected by the step-up test, the following must be false for all the Q' satisfying $P \subseteq Q' \subseteq \bar{P} = \{(j) : i_1 \leq j \leq i_p\}$:

$$M(X_j : j \in Q') \leq (\xi_2, \ldots, \xi_{q'}), \tag{3.3}$$

where $q' = |Q'|$, otherwise we would have $M(X_j : j \in \bar{P}) \leq (\xi_2, \ldots, \xi_{i_p - i_1 + 1})$. Consequently $H_{\bar{P}}$, and hence H_P, would be accepted by the step-up test, which contradicts the assumption that H_P is rejected by the step-up test. Since (3.3) is false, $H_{Q'}$ is rejected by the corresponding test used in the closed test.

Now suppose that $Q'' \supseteq P$ and that $Q'' \cap (\bar{P})^c$ is not empty, where $(\bar{P})^c$ denotes the complement of $\bar{P}$ in K. Then the following must be false:

$$M(X_j : j \in Q'') \leq (\xi_2, \ldots, \xi_{q''}),$$

where $q'' = |Q''|$, since (3.3) is false for $Q' = Q'' \cap \bar{P}$. Hence, $H_{Q''}$ is rejected by the corresponding test used in the closed test. Combining the two cases above, we have therefore proved that all the H_Q satisfying $Q \supseteq P$ are rejected by the corresponding tests in the closed test. The proof is thus completed.

Now, assuming that $k \geq 4$ and that the X_i are continuous random variables, we construct a closed test which not only dominates but is uniformly more powerful than the step-up test. For this it suffices to construct a closed test which dominates and is uniformly more powerful than the closed test defined in (3.1). This can be easily achieved by noting that some tests in (3.1) have size strictly less than α. We define a closed test that tests $H_{\mathbf{P}} = \cap_{i=1}^{r} H_{P_i}$ in the following way,

$$\text{if } r > 1: \text{ accept } H_{\mathbf{P}} \iff M(X_j : j \in P_i) \le (\xi_2, \ldots, \xi_{p_i}) \ \ \forall 1 \le i \le r;$$
$$\text{if } r = 1: \text{ accept } H_{\mathbf{P}} \iff M(X_j : j \in P_1) \le \lambda_{p_1}(\xi_2, \ldots, \xi_{p_1}),$$

where λ_{p_1} is a constant satisfying

$$\mathrm{P}_0\{M(X_j : 1 \le j \le p_1) \le \lambda_{p_1}(\xi_2, \ldots, \xi_{p_1})\} = 1 - \alpha. \tag{3.4}$$

By noting from (2.4) that

$$\mathrm{P}_0\{M(X_j : 1 \le j \le p) \le (\xi_2, \ldots, \xi_p)\} = 1 - \alpha_p \ge 1 - \alpha, \quad 2 \le p \le k,$$

the value of λ_{p_1} satisfying (3.4) must be no larger than one. In particular, the value of λ_2 must be strictly less than one since $k \ge 4$ and so

$$\begin{aligned} &\mathrm{P}_0\{M(X_j : 1 \le j \le 2) \le \xi_2\} = 1 - \alpha_2 \\ &\ge 1 - \alpha_2^* = (1-\alpha)^{1/[k/2]} \ge (1-\alpha)^{1/2} > 1 - \alpha. \end{aligned}$$

It is clear that this closed test strongly controls the type I FWE rate at level α since each individual test is of size α. It always rejects those H_P rejected by the closed test defined in (3.1) and, with a positive probability, it rejects some H_P (with $|P| = 2$) accepted by the closed test in (3.1) since $\lambda_2 < 1$ and the X_i are continuous. This closed test is therefore uniformly more powerful than the step-up test, which is equivalent to the closed test in (3.1). The closed test derived from NEWUP in this way is denoted as NEWCL hereafter. NEWCL is uniformly more powerful than NEWUP when $k \ge 4$ and the X_i are continuous.

When $k \ge 5$, even more powerful closed tests can be constructed by exploring all those tests in (3.1) that have size strictly less than α. The practical usefulness of these closed tests might be limited however, since they need more critical values and are much more difficult to perform than the corresponding step-up test.

For testing the family of subset hypotheses, the first step-down test was proposed independently by Newman (1939) and Keuls (1952) and known as the Newman-Keuls (NK) test. While the NK test does not strongly control the type I FWE rate for $k \ge 4$, various suggestions on modifying the critical values are made so that the resulting tests – NK type tests – strongly control the type I FWE rate (Hochberg and Tamhane (1987), pp.66-71). For instance, Welsch (1977) suggests basing the critical values on the Tukey-Welsch specification; the resulting test is denoted as NKDOWN in this paper. NKDOWN can be regarded as corresponding to NEWUP. It can be shown that a NK type test is also equivalent to a closed test. Peritz (1970) suggests a closed test that dominates and is uniformly more powerful than a given NK type test. This test is studied in detail by Begun and Gabriel (1981). A slightly revised version, which is based on the Tukey-Welsch specification and denoted as PERCL hereafter, is studied by Ramsey (1978). PERCL can be regarded as corresponding to NEWCL. It can be shown that there exist closed tests that dominate and are uniformly more powerful than PERCL when $k \geq 5$ and X_i are continuous.

4. Normal Distribution

In this section we assume that $F(x)$ is the normal distribution function with mean zero and variance σ^2/n, where σ^2 is an unknown parameter and n is a known constant which is usually the sample size. Suppose that an estimate of σ^2, S^2, is available which is independent of the X_i and distributed as a $\sigma^2\chi^2_\nu/\nu$ random variable. In this situation, a step-up test using critical values $\xi_2 \leq \cdots \leq \xi_k$ follows the same steps 1 and 2 as before, except that a q-range $X_{[i+q-1]_K} - X_{[i]_K}$ $(2 \leq q \leq k, 1 \leq i \leq k-q+1)$ is compared with $\xi_q S/\sqrt{n}$.

To determine the critical values ξ_i so that the step-up test strongly controls the type I FWE rate at level α, let

$$\alpha_p = 1 - \mathrm{P}_0\{M(X_j : 1 \leq j \leq p) \leq (\xi_2, \ldots, \xi_p)S/\sqrt{n}\}, \quad 2 \leq p \leq k. \tag{4.1}$$

Then, for $H_{\mathbf{P}} = \cap_{i=1}^r H_P$, we have

$$\mathrm{P}_{H_{\mathbf{P}}}\{\text{at least one } H_{P_i} \text{ is rejected}\} \leq 1 - \prod_{i=1}^r (1 - \alpha_{p_i}).$$

This can be proved in a way similar to Lemma 2.1, except that the equality in (2.3) in the proof of Lemma 2.1 should be replaced by "$\leq$", this following from Kimball's (1951) inequality. Consequently, the specifications of α_p discussed in Section 2 can still be used to ensure that the type I FWE rate is strongly controlled at level α, e.g. Welsch's (1977) specification is α'_p as before, and our recommendation is again α^*_p. For a given specification of α_p, the critical values ξ_p can be solved sequentially from (4.1) with the same treatment as before to enforce the monotonicity of the ξ_i. It is also clear that the results of Section 3 can be generalized to the current setting.

For the specification α^*_p, Table 1 presents the critical values ξ_i for $\alpha = 0.05$, $k = 3(1)10$ and selected values of ν. As before, the probability $\mathrm{P}_0\{M(X_j : 1 \leq j \leq p) \leq (\xi_2, \ldots, \xi_p)S/\sqrt{n}\}$ is calculated using the expressions given in Liu (1996a) when $p = 3$ and estimated by simulation based on 1,000,000 experiments when $p \geq 4$. The standard error of the estimate of the probability is less than 0.0005, and the critical values given in Table 1 are expected to be accurate to the digits given.

Table 1. Critical values ξ_i $(2 \leq i \leq k)$ of the step-up test NEWUP for $\alpha = 0.05$

	i	$k = 10$	$k = 9$	$k = 8$	$k = 7$	$k = 6$	$k = 5$	$k = 4$	$k = 3$
$\nu = 5$	2	5.67	5.36	5.36	4.98	4.98	4.46	4.46	3.63
	3	6.35	6.96	6.01	6.51	5.59	5.93	4.61	5.04
	4	6.60	6.96	6.25	6.51	5.81	5.93	5.38	
	5	6.76	6.96	6.39	6.51	5.81	5.93		
	6	6.86	6.96	6.50	6.51	6.05			
	7	6.91	6.96	6.50	6.51				
	8	6.95	6.96	6.58					
	9	6.95	6.96						
	10	6.99							
$\nu = 7$	2	4.93	4.70	4.70	4.41	4.41	4.01	4.01	3.34
	3	5.48	5.87	5.23	5.55	4.92	5.13	4.17	4.47
	4	5.71	5.87	5.45	5.55	5.13	5.13	4.79	
	5	5.85	5.87	5.59	5.55	5.13	5.13		
	6	5.96	5.87	5.70	5.55	5.38			
	7	6.02	5.94	5.70	5.59				
	8	6.09	5.94	5.82					
	9	6.09	6.00						
	10	6.16							
$\nu = 10$	2	4.46	4.28	4.28	4.04	4.04	3.71	3.71	3.15
	3	4.93	5.21	4.74	4.97	4.50	4.64	3.88	4.11
	4	5.14	5.21	4.95	4.97	4.70	4.64	4.42	
	5	5.28	5.28	5.08	5.03	4.70	4.66		
	6	5.38	5.28	5.18	5.03	4.93			
	7	5.45	5.39	5.18	5.12				
	8	5.52	5.39	5.32					
	9	5.52	5.46						
	10	5.61							
$\nu = 15$	2	4.15	4.00	4.00	3.80	3.80	3.51	3.51	3.01
	3	4.56	4.77	4.40	4.58	4.20	4.31	3.68	3.86
	4	4.75	4.77	4.59	4.58	4.39	4.31	4.15	
	5	4.89	4.89	4.72	4.67	4.39	4.37		
	6	4.98	4.89	4.81	4.67	4.62			
	7	5.05	5.00	4.81	4.79				
	8	5.11	5.00	4.95					
	9	5.11	5.08						
	10	5.21							
$\nu = 20$	2	4.01	3.87	3.87	3.68	3.68	3.42	3.42	2.95
	3	4.39	4.57	4.25	4.40	4.06	4.16	3.58	3.75
	4	4.56	4.57	4.42	4.40	4.23	4.16	4.01	
	5	4.70	4.70	4.55	4.51	4.23	4.24		
	6	4.78	4.70	4.63	4.51	4.46			
	7	4.85	4.81	4.63	4.62				
	8	4.92	4.81	4.78					
	9	4.92	4.90						
	10	5.01							
$\nu = 25$	2	3.93	3.79	3.79	3.62	3.62	3.36	3.36	2.91
	3	4.29	4.46	4.16	4.30	3.98	4.07	3.53	3.68
	4	4.47	4.46	4.33	4.30	4.16	4.07	3.95	
	5	4.59	4.59	4.45	4.41	4.16	4.16		
	6	4.68	4.59	4.54	4.41	4.38			
	7	4.75	4.71	4.54	4.53				
	8	4.81	4.71	4.68					
	9	4.81	4.79						
	10	4.90							

Table 1. (Continued)

	i	$k=10$	$k=9$	$k=8$	$k=7$	$k=6$	$k=5$	$k=4$	$k=3$
$\nu=30$	2	3.88	3.75	3.75	3.58	3.58	3.33	3.33	2.89
	3	4.23	4.39	4.10	4.24	3.93	4.01	3.49	3.64
	4	4.40	4.39	4.27	4.24	4.10	4.01	3.90	
	5	4.51	4.52	4.39	4.35	4.10	4.11		
	6	4.61	4.52	4.48	4.35	4.32			
	7	4.68	4.64	4.48	4.47				
	8	4.74	4.64	4.61					
	9	4.74	4.72						
	10	4.83							
$\nu=35$	2	3.84	3.71	3.71	3.55	3.55	3.30	3.30	2.87
	3	4.19	4.34	4.06	4.19	3.90	3.98	3.47	3.61
	4	4.35	4.34	4.23	4.19	4.07	3.98	3.87	
	5	4.47	4.47	4.34	4.30	4.07	4.07		
	6	4.55	4.47	4.43	4.30	4.28			
	7	4.62	4.59	4.43	4.42				
	8	4.69	4.59	4.56					
	9	4.69	4.67						
	10	4.78							
$\nu=40$	2	3.81	3.69	3.69	3.52	3.52	3.29	3.29	2.86
	3	4.15	4.30	4.03	4.16	3.87	3.95	3.45	3.59
	4	4.33	4.30	4.20	4.16	4.04	3.95	3.85	
	5	4.43	4.43	4.31	4.27	4.04	4.04		
	6	4.52	4.43	4.39	4.27	4.24			
	7	4.59	4.55	4.39	4.39				
	8	4.65	4.55	4.53					
	9	4.65	4.64						
	10	4.73							
$\nu=50$	2	3.77	3.65	3.65	3.49	3.49	3.26	3.26	2.84
	3	4.11	4.25	3.99	4.11	3.83	3.91	3.42	3.55
	4	4.27	4.25	4.15	4.11	3.99	3.91	3.81	
	5	4.39	4.39	4.27	4.24	4.00	4.01		
	6	4.47	4.39	4.35	4.24	4.21			
	7	4.54	4.51	4.35	4.35				
	8	4.59	4.51	4.48					
	9	4.59	4.59						
	10	4.68							
$\nu=60$	2	3.75	3.63	3.63	3.47	3.47	3.24	3.24	2.83
	3	4.08	4.22	3.96	4.08	3.81	3.88	3.40	3.53
	4	4.24	4.22	4.12	4.08	3.97	3.88	3.79	
	5	4.35	4.36	4.24	4.20	3.98	3.98		
	6	4.44	4.36	4.32	4.20	4.18			
	7	4.51	4.48	4.32	4.32				
	8	4.56	4.48	4.45					
	9	4.56	4.55						
	10	4.65							
$\nu=120$	2	3.69	3.58	3.58	3.42	3.42	3.20	3.20	2.80
	3	4.01	4.14	3.90	4.01	3.75	3.82	3.36	3.48
	4	4.17	4.14	4.05	4.01	3.91	3.82	3.73	
	5	4.27	4.28	4.16	4.13	3.91	3.92		
	6	4.35	4.28	4.24	4.13	4.11			
	7	4.42	4.39	4.24	4.25				
	8	4.48	4.39	4.37					
	9	4.48	4.47						
	10	4.57							
$\nu=\infty$	2	3.63	3.52	3.52	3.38	3.38	3.16	3.16	2.77
	3	3.94	4.06	3.83	3.94	3.69	3.75	3.32	3.44
	4	4.09	4.06	3.98	3.94	3.84	3.75	3.68	
	5	4.20	4.20	4.09	4.06	3.86	3.86		
	6	4.27	4.20	4.17	4.06	4.04			
	7	4.33	4.31	4.17	4.17				
	8	4.39	4.31	4.29					
	9	4.39	4.39						
	10	4.48							

5. Power Comparisons

In this section we use simulation to compare the powers of NKDOWN, WELUP, NEWUP, PERCL and NEWCL. Tukey's (1953) (studentised) range test, denoted as RANGE, is also included for its simplicity and availability of confidence intervals. All these tests strongly control the type I FWE rate at level α. From Section 3, it is known that PERCL and NEWCL dominate NKDOWN and NEWUP, respectively. It is also known that NKDOWN dominates RANGE.

Several notions of power are available in the literature. Carmer and Swanson (1973) and Welsch (1977) use the overall power, which is the probability of rejecting all false subset hypotheses. Ramsey (1978) introduces the any-pair power which is the probability of rejecting at least one false pairwise hypothesis, and the all-pair power which is the probability of rejecting all false pairwise hypotheses. Einot and Gabriel (1975) propose the P-subset power, which, for a given subset $P \subseteq K$, is the probability of rejecting the subset hypothesis H_P when it is false. In general, for arbitrary μ_i, a multiple test has a collection of $2^k - k - 1$ P-subset powers, one for each subset P with $|P| \geq 2$. It is most unlikely that for two competing tests, e.g. NKDOWN and NEWUP, one would dominate the other for all P-subset powers. On the other hand, if the P-subset power of a particular subset P is of concern, then it might be argued whether a multiple test is needed at all. The any-pair power overemphasizes the sensitivity to the largest difference between the μ_i. The overall and all-pair powers do not differentiate among rejecting none, some, or most of the false hypotheses, as long as not all false hypotheses are rejected.

We use the average power as the criterion of our power comparisons. For a given configuration of μ_i, let M denote the number of false pairwise hypotheses. Let the random variable m denote the number of false pairwise hypotheses being rejected by a given multiple test. The average power of this test is then defined as the expectation $E(m/M)$.

The distribution function $F(\cdot)$ is assumed to be normal with a known variance σ^2. (We also tried unknown σ^2 with $\nu = 20$ and similar results were observed.) The configurations of μ_i considered are the maximum range and minimum range configurations of Ramsey (1978). Let

$$f = \Big(\frac{1}{k}\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2/\sigma^2\Big)^{1/2} \qquad \text{where } \bar{\mu} = \frac{1}{k}\sum_{i=1}^{k}\mu_i$$

and $\delta = \text{mean}_{i,j}|\mu_i - \mu_j|/\sigma$ over all i, j such that $\mu_i \neq \mu_j$. Without loss of generality, assume the μ_i sum to zero and are monotonically increasing in i.

Then, for fixed f, the maximum range configuration is given by

$$\mu_1 = -\sqrt{k/2}f,\ \mu_2 = \cdots = \mu_{k-1} = 0,\ \mu_k = \sqrt{k/2}f,$$

and the minimum range configuration is given by

$$\begin{aligned}
&\mu_1 = \cdots = \mu_{k/2} = -f,\ \mu_{k/2+1} = \cdots = \mu_k = f \quad \text{for even } k;\\
&\mu_1 = \cdots = \mu_{(k+1)/2} = -\sqrt{(k-1)/(k+1)}f,\\
&\mu_{(k+3)/2} = \cdots = \mu_k = \sqrt{(k+1)/(k-1)}f \quad \text{for odd } k.
\end{aligned}$$

The other experimental conditions are $\alpha = 0.05$, $k = 3, 4, 5$ and 6. The value of f is varied systematically to cover the full range of power. For each given set of experimental conditions, $N = 100,000$ experiments are simulated. For the ith experiment, m_i is calculated and $E(m/M)$ is estimated by $\sum_{i=1}^{N} m_i/(MN)$.

Results for $k = 3, 4, 5$ and 6 are given in Tables 2,3,4 and 5, respectively. Note that, when $k = 3$, NKDOWN and PERCL are the same, and WELUP, NEWUP and NEWCL are the same. When $k = 6$, Table 6 also contains the power at the equal spaced configuration.

From these tables, the following can be observed. NEWUP is more powerful than WELUP most of the time, and WELUP is more powerful than NKDOWN almost all the time. The largest power difference among NKDOWN, WELUP and NEWUP is about 0.01. NEWCL is more powerful than PERCL almost all the time, with the largest power difference between them being about 0.01. PERCL and NEWCL are always more powerful than NKDOWN and NEWUP respectively, which agrees with the theoretical result. The largest power differences between PERCL and NKDOWN and between NEWCL and NEWUP are about 0.02. There are quite a few occasions when PERCL is less powerful than both WELUP and NEWUP, but NEWCL is almost always more powerful than NKDOWN. RANGE is always the least powerful test; the largest power difference is about 0.11 between RANGE and NEWCL, and about 0.10 between RANGE and NKDOWN.

It is clear from the present study that NEWCL is the most powerful multiple test, with PERCL a close runner-up. However, both these tests are genuine closed tests and so not easy to perform. It must be noted that NEWUP is a compelling choice: it is as easy to perform as WELUP and NKDOWN, and does not suffer a major sacrifice in power. Finally, RANGE does suffer from considerable power deficiencies, which should be weighed against its advantages such as simplicity and the availability of confidence intervals.

Table 2. Powers of three multiple tests for $k = 3$, $\alpha = 0.05$ and known σ^2

f	δ	RANGE	NKDOWN	NEWUP
a. Minimum range ($M = 2$)				
0.800	1.697	0.126	0.149	0.156
1.100	2.334	0.242	0.284	0.296
1.400	2.970	0.401	0.463	0.478
1.700	3.606	0.579	0.651	0.667
2.000	4.242	0.743	0.808	0.819
2.300	4.879	0.865	0.913	0.919
2.600	5.515	0.941	0.967	0.970
2.900	6.152	0.978	0.990	0.990
3.200	6.788	0.993	0.997	0.998
b. Maximum range ($M = 3$)				
1.000	1.633	0.136	0.167	0.174
1.500	2.449	0.298	0.360	0.366
2.000	3.265	0.469	0.560	0.561
2.500	4.083	0.611	0.712	0.712
3.000	4.899	0.732	0.825	0.825
3.500	5.716	0.836	0.906	0.906
4.000	6.532	0.913	0.956	0.956
4.500	7.348	0.960	0.983	0.983
5.000	8.165	0.984	0.994	0.994

Table 3. Powers of six multiple tests for $k = 4$, $\alpha = 0.05$ and known σ^2

f	δ	RANGE	NKDOWN	WELUP	NEWUP	PERCL	NEWCL
a. Minimum range ($M = 4$)							
0.800	1.600	0.075	0.090	0.092	0.093	0.091	0.094
1.100	2.200	0.155	0.186	0.190	0.192	0.190	0.196
1.400	2.800	0.278	0.332	0.338	0.340	0.341	0.349
1.700	3.400	0.435	0.510	0.516	0.518	0.526	0.534
2.000	4.000	0.603	0.685	0.689	0.691	0.706	0.712
2.300	4.600	0.754	0.825	0.828	0.829	0.847	0.850
2.600	5.200	0.867	0.917	0.918	0.919	0.933	0.935
2.900	5.800	0.938	0.966	0.966	0.967	0.976	0.976
3.200	6.400	0.975	0.988	0.988	0.989	0.993	0.993
b. Maximum range ($M = 5$)							
1.000	1.697	0.104	0.126	0.127	0.129	0.126	0.129
1.500	2.545	0.248	0.299	0.299	0.301	0.300	0.302
2.000	3.394	0.413	0.491	0.491	0.492	0.497	0.498
2.400	4.073	0.544	0.633	0.632	0.634	0.647	0.647
2.800	4.752	0.673	0.758	0.758	0.759	0.780	0.780
3.200	5.430	0.789	0.859	0.859	0.860	0.883	0.883
3.600	6.109	0.880	0.928	0.928	0.928	0.946	0.947
4.000	6.788	0.940	0.968	0.968	0.968	0.979	0.979
4.400	7.468	0.974	0.988	0.987	0.988	0.993	0.993

Table 4. Powers of six multiple tests for $k = 5$, $\alpha = 0.05$ and known σ^2

f	δ	RANGE	NKDOWN	WELUP	NEWUP	PERCL	NEWCL
			a. Minimum range ($M = 6$)				
0.800	1.633	0.057	0.068	0.068	0.064	0.068	0.066
1.100	2.245	0.126	0.151	0.152	0.144	0.153	0.149
1.400	2.858	0.239	0.284	0.287	0.274	0.291	0.285
1.700	3.470	0.391	0.457	0.461	0.449	0.469	0.468
2.000	4.082	0.562	0.639	0.644	0.637	0.654	0.658
2.300	4.695	0.722	0.795	0.800	0.799	0.808	0.815
2.600	5.307	0.848	0.902	0.905	0.907	0.910	0.917
2.900	5.920	0.928	0.961	0.962	0.964	0.964	0.968
3.200	6.532	0.971	0.987	0.987	0.989	0.988	0.990
			b. Maximum range ($M = 7$)				
0.900	2.033	0.070	0.084	0.085	0.082	0.084	0.082
1.300	2.936	0.168	0.201	0.202	0.197	0.201	0.198
1.700	3.840	0.297	0.354	0.356	0.351	0.356	0.357
2.100	4.743	0.441	0.519	0.522	0.520	0.525	0.532
2.500	5.647	0.594	0.680	0.683	0.686	0.691	0.704
2.900	6.550	0.740	0.818	0.821	0.826	0.830	0.843
3.300	7.454	0.856	0.912	0.914	0.919	0.921	0.929
3.700	8.357	0.932	0.965	0.966	0.969	0.969	0.973
4.100	9.261	0.973	0.988	0.989	0.990	0.990	0.992

Table 5. Powers of six multiple tests for $k = 6$, $\alpha = 0.05$ and known σ^2

f	δ	RANGE	NKDOWN	WELUP	NEWUP	PERCL	NEWCL
			a. Minimum range ($M = 9$)				
0.800	1.600	0.042	0.049	0.049	0.050	0.049	0.051
1.100	2.200	0.097	0.113	0.114	0.115	0.114	0.117
1.400	2.800	0.191	0.223	0.225	0.228	0.226	0.231
1.700	3.400	0.327	0.378	0.382	0.386	0.384	0.391
2.000	4.000	0.491	0.557	0.563	0.568	0.563	0.573
2.300	4.600	0.656	0.728	0.735	0.739	0.732	0.742
2.600	5.200	0.796	0.858	0.864	0.866	0.859	0.868
2.900	5.800	0.894	0.938	0.941	0.942	0.938	0.943
3.200	6.400	0.953	0.977	0.979	0.979	0.977	0.979
3.500	7.000	0.982	0.993	0.993	0.994	0.993	0.994
			b. Maximum range ($M = 9$)				
0.900	1.732	0.064	0.075	0.076	0.077	0.075	0.077
1.300	2.502	0.162	0.190	0.192	0.194	0.190	0.195
1.700	3.271	0.297	0.349	0.354	0.357	0.350	0.357
2.100	4.041	0.456	0.533	0.539	0.542	0.536	0.545
2.500	4.811	0.629	0.714	0.720	0.723	0.719	0.727
2.900	5.581	0.785	0.857	0.862	0.864	0.861	0.867
3.300	6.351	0.896	0.943	0.945	0.946	0.945	0.947
3.600	6.928	0.947	0.975	0.976	0.977	0.976	0.977
3.900	7.506	0.976	0.990	0.991	0.991	0.991	0.991
4.300	8.276	0.993	0.998	0.998	0.998	0.998	0.998

Table 5. (Continued)

f	δ	RANGE	NKDOWN	WELUP	NEWUP	PERCL	NEWCL
			b. Equally spaced means ($M = 15$)				
1.300	1.776	0.096	0.113	0.114	0.115	0.114	0.116
1.800	2.459	0.210	0.242	0.243	0.245	0.245	0.248
2.300	3.143	0.331	0.375	0.375	0.378	0.379	0.382
2.800	3.826	0.436	0.487	0.488	0.491	0.492	0.494
3.300	4.509	0.522	0.578	0.579	0.582	0.582	0.584
3.800	5.192	0.592	0.652	0.652	0.654	0.655	0.657
4.600	6.285	0.679	0.740	0.740	0.741	0.744	0.745
5.600	7.651	0.756	0.818	0.817	0.818	0.825	0.825
6.600	9.017	0.816	0.877	0.876	0.877	0.891	0.891
7.600	10.383	0.872	0.925	0.924	0.925	0.943	0.943

6. Closing Remarks

Step-up tests based on range statistics are studied in detail. Lemma 2.1 is fundamental in specifying the α_p's which in turn determine the critical values so that the type I FWE rate is strongly controlled at level α. The new step-up test NEWUP, which is based on the specification α_p^*, is slightly more powerful than Welsch's (1977) test WELUP and recommended.

The interrelationship between step-up tests and closed tests is described. This enables us to construct a closed test that is uniformly more powerful than a particular step-up test. Indeed, similar interrelationships exist between step-down tests and closed tests. The power superiority of these genuine closed tests should be weighed against the complexity of performing them.

It is difficult to decide which definition of power is more appropriate in general or in a particular situation. The motivation behind the definition of the average power is to address the all-or-nothing drawback of the all-pair and overall powers. The simulation results of the present study and Ramsey (1978) agree over the order of powers of the RANGE, NKDOWN and PERCL, but differ in the magnitudes of the power differences.

Finally, a general step-down test scheme is available (Einot and Gabriel (1975)), of which a NK type test is only a shortcut version. Different test statistics can be employed in this scheme to obtain different step-down tests. An analogous theory, however, is not available for step-up tests. Research in this direction is no doubt worthwhile.

Acknowledgements

I would like to thank Mr. B. J. R. Bailey for his help in preparing the paper, and the referees for comments which greatly improved the presentation of the paper.

References

Begun, J. M. and Gabriel, K. R. (1981). Closure of the Newman-Keuls multiple comparisons procedure. *J. Amer. Statist. Assoc.* **76**, 241-245.

Carmer, S. G. and Swanson, M. R. (1973). An Evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *J. Amer. Statist. Assoc.* **68**, 66-74.

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.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures.* Wiley, New York.

Keuls, M. (1952). The use of the 'studentised range' in connection with an analysis of variance. *Euphytical* **1**, 112-122.

Kimball, A. W. (1951). On dependent tests of significance in the analysis of variance. *Ann. Math. Statist.* **22**, 600-602.

Lehmann, E. L. and Shaffer, J. P. (1977). On a fundamental theorem in multiple comparisons. *J. Amer. Statist. Assoc.* **72**, 576-578.

Lehmann, E. L. and Shaffer, J. P. (1979). Optimum significance levels for multistage comparison procedures. *Ann. Statist.* **7**, 27-45.

Liu, W. (1996a). On some single-stage, step-down, and step-up procedures for comparing three normal means. *Comput. Statist. Data Anal.* **21**, 215-227.

Liu, W. (1996b). Multiple tests of a non-hierarchical finite family of hypotheses. *J. Roy. Statist. Soc. Ser. B* **58**, 455-461.

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.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika* **31**, 20-30.

Peritz, E. (1970). A note on multiple comparisons. Unpublished paper, Hebrew University.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *J. Amer. Statist. Assoc.* **73**, 479-487.

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances and other statistics. *Psychological Bulletin* **57**, 318-328.

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished paper, Department of Mathematics, Princeton University.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *J. Amer. Statist. Assoc.* **72**, 566-575.

Department of Mathematics, University of Southampton, Southampton, SO17 1BJ, England.

E-mail: wl@maths.soton.ac.uk

(Received November 1995; accepted November 1996)