Statistical Papers manuscript No.
(will be inserted by the editor)

Modifications of the Bonferroni–Holm procedure for a multi–way ANOVA

Vanessa Didelez[1], **Iris Pigeot**[2], **Patricia Walter**[3] ⋆

[1] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

[2] Bremen Institute of Prevention Research and Social Medicine, Linzer Str. 8-10, 28359 Bremen, Germany

[3] Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany

The date of receipt and acceptance will be inserted by the editor

Abstract We present step–wise test procedures based on the Bonferroni–Holm principle for multi–way ANOVA–type models. It is shown for two plausible modifications that the multiple level α is preserved. These theoretical results are supplemented by a simulation study, in a two–way ANOVA setting, to compare the multiple procedures with respect to their simultaneous power and the relative frequency of correctly rejected false hypotheses.

⋆ Financial support of the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

Key words Adjustment for multiplicity, Bonferroni–Holm procedure, multiple test problem, multi–way ANOVA, step–wise procedures.

1 Introduction

If several hypotheses are to be tested simultaneously in the context of a single statistical experiment, the classical test theory does not account for the multiplicity of the test decisions. For example the classical F–test in a one–way analysis of variance is only able to show overall significant differences among the population means but it cannot specify them. More detailed comparisons require a multiple test procedure to capture the complexity of the statistical problem and the multiplicity of possibly wrong decisions.

Multiple tests are often applied in the context of multiple pairwise comparison in the setting of an analysis of variance. In particular, for the case of a balanced one–way layout numerous procedures have been developed and improved by various suggestions, for instance with less restrictive adjustments of the size of the individual tests. The corresponding multiple tests can still be used after appropriate modifications in non–standard situations such as unequal sample sizes or linear contrasts.

Multiple tests in the context of a two or multi–way ANOVA, however, has not been paid much attention so far, so that for this case only few procedures are known, e.g. the method of Hartley (1955) or Ottestad (1960, 1970). Further, the procedure discussed in Bauer et al. (1998) can be adapted for this situation, as we will show below.

In this paper, multiple test procedures are derived, in particular for a two–way ANOVA, which are less conservative than for instance a procedure obtained from a Bonferroni adjustment of simultaneous tests originally proposed for a one–way layout. As our proposals are mainly based on a modification of the Bonferroni–Holm procedure, they can easily be extended to applications in a multi–way layout. They are defined as step–wise test procedures and are thus more powerful than their simultaneous counterparts. The underlying idea is to consider subfamilies of null hypotheses, for which a 'local' test of multiple level $\tilde{\alpha}$ exists which is obtained using a Bonferroni–(Holm–)type split (cf. Bauer et al., 1998). In addition, it is investigated whether the proposed test procedures keep the multiple level α. It can be shown that two of our proposals fulfill this property whereas the third modification does not. Nevertheless, all modifications are discussed since they can all be encountered in practice. The procedures are then compared with respect to their power by means of Monte–Carlo experiments based on the simultaneous power (Maurer and Mellein, 1988) and the relative frequency of correctly rejected false hypotheses.

2 Multiple tests in a two–way ANOVA

The multiple test procedures introduced in Section 2.2 are based on the Bonferroni–Holm approach. This general principle for constructing step–wise test procedures allows the application of any suitable level α test. Thus, our procedures are not restricted to the classical Gaussian case as introduced in Section 2.1 nor do they require a balanced design. In fact any multi–way layout where comparisons of different levels within subclasses are of interest can be tackled with the proposed procedures and, if desirable or required, non–parametric tests could be used. However, for the sake of simplicity we restrict the exposition to the classical two–way ANOVA situation and the simulation study (Section 3) is based on F–tests for the overall hypotheses and multiple t–tests for the pairwise comparisons.

2.1 Basic notations

For convenience, let us briefly recall the classical two–way ANOVA setting, where

$$Y_{kln} = \mu + \alpha_k + \beta_l + (\alpha\beta)_{kl} + \epsilon_{kln}, \tag{1}$$

for $k = 1, ..., K$, $l = 1, ..., L$, $n = 1, ..., N_{kl}$, where N_{kl} are the frequencies of combinations k in factor A and l in factor B, and the error terms ϵ_{kln} are assumed to be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. The parameters α_k and β_l are the main effects of factor A and B, respectively, $(\alpha\beta)_{kl}$ is the interaction effect, and μ the grand mean.

The family of hypotheses to be tested in this set–up mainly consists of three intersection hypotheses concerning the main and interaction effects as well as the hypotheses of all pairwise comparisons within the factors A, B, and the interactions $A \times B$. For example, the intersection hypothesis w.r.t. factor A is denoted as H_0^A with

$$H_0^A : \alpha_1 = \alpha_2 = ... = \alpha_K$$

and has to be tested against

$$H_1^A : \exists\, j, k \in \{1, ..., K\},\ j \neq k : \alpha_j \neq \alpha_k.$$

The intersection hypotheses H_0^B and H_0^{AB} are defined analogously. The multiple pairwise comparisons are used to identify those factor levels which actually differ regarding their effect on Y. For factor A, we have in total $\frac{1}{2}K(K-1)$ pairwise comparisons of the type

$$H_0^{A(jk)} : \alpha_j = \alpha_k \qquad \text{vs} \qquad H_1^{A(jk)} : \alpha_j \neq \alpha_k, \qquad 1 \leq j < k \leq K.$$

The pairwise interaction comparisons are given by

$$H_0^{AB(jk,lm)} : (\alpha\beta)_{jk} = (\alpha\beta)_{lm} \qquad \text{vs} \qquad H_1^{AB(jk,lm)} : (\alpha\beta)_{jk} \neq (\alpha\beta)_{lm},$$

for $1 \leq j < l \leq K$, $1 \leq k < m \leq L$. Other choices of the individual hypotheses about the interactions are possible and depend on the respective application and interpretation. With the above choice, we consider the general case of analysing *any* kind of differences among the interactions. In practical applications, however, it will often be sensible to reduce these to a smaller number of hypotheses being of main interest. For the sake of simplicity, the hypotheses of pairwise comparisons are in the following consecutively numbered as $H_0^{A(j)}$ with $j = 1, ..., \frac{K(K-1)}{2}$ and $H_0^{B(j)}$, $H_0^{AB(j)}$ analogously.

2.2 Modifications of the Bonferroni–Holm procedure

As a first proposal, we consider the original Bonferroni–Holm procedure which is straightforward to apply not only in the case of a one–way ANOVA but also in ANOVA settings with more than one factor.

To use the Bonferroni–Holm procedure in a two–way ANOVA the p–values of the pairwise comparisons, only, are considered, irrespective of the particular factor or interaction to which they belong. These p–values are ordered such that $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(n_*)}$ with $n_* = [\frac{1}{2}K(K-1) + \frac{1}{2}L(L-1) + \frac{1}{2}KL(KL-1)]$. The corresponding null hypotheses are denoted as $H_0^{(1)}, H_0^{(2)}, ..., H_0^{(n_*)}$. The Bonferroni–Holm procedure rejects intersection hypotheses whenever at least one of the elementary hypotheses of the pairwise comparisons forming the intersection is rejected. In contrast to the procedures presented below the intersection hypotheses are not tested explicitly.

The BH procedure is given as $(\varphi_i;\ i = 1, ..., n_*)$ with step–wise tests

$$\varphi_{(i)} = \prod_{j=1}^{i} \tilde{\varphi}_{(j)}, \qquad i = 1, ..., n_*, \tag{2}$$

where

$$\tilde{\varphi}_{(j)} = \left\{ \begin{array}{lll} 0 & & > \\ & \text{for } \; p_{(j)} & \dfrac{\alpha}{(n-j+1)}, \quad j = 1, ..., n_*, \\ 1 & & \leq \end{array} \right. \tag{3}$$

and $\tilde{\varphi}_{(j)}$ are the individual tests for the elementary hypotheses ordered according to the ordered p–values. For procedures of this type, the following result originally derived by Holm (1977, 1979) holds.

Theorem 1

The BH procedure according to (2) and (3) keeps the multiple level α.

Since the Bonferroni–Holm procedure is applied to the pairwise comparisons w.r.t. both factors and all interactions, the first adjusted significance level is given by $\frac{\alpha}{[K(K-1)+L(L-1)+KL(KL-1)]/2}$. This may obviously be very small which makes it in most applications difficult to reject the corresponding hypotheses.

Bonferroni–Holm Modification I (BHM I)

The second test procedure is a combination of the Bonferroni–Holm procedure and the simple Bonferroni adjustment applied to the intersection hypotheses. This implies that first, a suitable level $\alpha/3$ test for each of the intersection hypotheses H_0^A, H_0^B, and H_0^{AB} is performed. If one of these is rejected it is investigated which of the corresponding means differ significantly from each other using the Bonferroni–Holm procedure.

For a more formal description of this procedure let $p_i, i \in \{A, B, A \times B\}$, denote the p–values for the intersection hypotheses, and $p_{i(j)}$, $j = 1, ..., n_i$, the p–values for the corresponding pairwise comparisons such that $p_{i(1)} \leq ... \leq p_{i(n_i)}$ for each $i \in \{A, B, A \times B\}$, where $n_A = \frac{K(K-1)}{2}$, $n_B = \frac{L(L-1)}{2}$, $n_{A\times B} = \frac{KL(KL-1)}{2}$.

The BHM I procedure is then given as $\varphi = (\varphi_i, \varphi_{ij}; i \in \{A, B, A \times B\}, j \in \{1, ..., n_i\})$ with

$$\varphi_i = \begin{cases} 0 & \qquad > \\ \text{if} \quad p_i & \qquad \alpha/3, \qquad i \in \{A, B, A \times B\}, \\ 1 & \qquad \leq \end{cases} \tag{4}$$

and $\varphi_{i(j)} = \varphi_i \cdot \prod_{k=1}^{j} \tilde{\varphi}_{i(k)}$, $j = 1, ..., n_i$, with

$$\tilde{\varphi}_{i(k)} = \begin{cases} 0 & \qquad > \\ & \text{if} \;\; p_{i(k)} \qquad \dfrac{\alpha/3}{n_i - k + 1}, \\ 1 & \qquad \leq \end{cases} \qquad k = 1, ..., n_i. \tag{5}$$

Here, $\tilde{\varphi}_{i(j)}$ represents the individual test for the elementary hypotheses of the pairwise comparisons belonging to factor i and arranged according to the p–values. Concerning the size of this procedure, the following result can be shown.

Theorem 2

The BHM I procedure according to (4) and (5) keeps the multiple level α.

As the proof of this thorem is essentially based on the Bonferroni inequality (cf. Appendix) it has to be expected that the nominal multiple level of this test can become smaller than α. Thus, despite of the Bonferroni–Holm adjustment being applied separately to each factor as well as for the interactions, the procedure may be rather conservative.

Bonferroni–Holm Modification II (BHM II)

The second modification of the Bonferroni–Holm procedure is similar to the BHM I procedure, with the only, but important, difference that the levels of the three tests of the intersection hypotheses are not simply determined by the Bonferroni inequality. They now depend on the results of the previous tests according to a second Bonferroni–Holm adjustment, such that the whole test may be regarded as a nested procedure.

Therefore, the p–values of the tests of the three intersection hypotheses are ordered such that $p_{(1)} \leq p_{(2)} \leq p_{(3)}$. This modification leads to a less conservative procedure since only the smallest p–value is now compared to $\alpha/3$. If it is larger than the adjusted level of significance, the procedure stops, and all intersection hypotheses as well as all hypotheses for the pairwise comparisons cannot be rejected. Otherwise those pairwise comparisons have to be tested, whose intersection yields the rejected intersection hypothesis. This has to be done according to a Bonferroni–Holm procedure with multiple level $\alpha/3$. As soon as a p–value for a pairwise comparison exceeds the corresponding level of significance, this particular Bonferroni–Holm procedure stops, and the whole procedure continues with the next intersection hypothesis, where $p_{(2)}$ is compared with $\alpha/2$.

Thus, the whole procedure stops if and only if one of the intersection hypotheses cannot be rejected or all hypotheses are rejected. In contrast, failing to reject one of the pairwise comparisons only implies that the inner Bonferroni–Holm procedure stops, without testing any further pairwise comparisons, but the procedure continues with the examination of the next intersection hypothesis. However, it does not keep the multiple level α, because apart from false decisions on the first level of the intersection hypotheses a type I error can also be committed on the second level when carrying out the pairwise comparisons.

The above procedure can, however, be improved so as to keep the multiple level, namely if the procedure does not only stop as soon as one of the intersection hypotheses cannot be rejected, but also if one of the elementary hypotheses of the pairwise comparisons has to be retained.

For a formal description of this BHM II test, let $p_i, i \in \{A, B, A \times B\}$, denote the p–values for the intersection hypotheses and $p_{(i)}$ the corresponding ordered p–values. The ordered p–values for the pairwise comparisons are given as $p_{(i)(j)}$ with $j = 1, ..., n_{(i)}$, where $n_{(i)} = n_{\overline{R}(i)}$ and $\overline{R}(i) \in \{A, B, A \times B\}$ is the anti–rank.

The BHM II procedure is given as $(\varphi_i,\ \varphi_{ij};\ i = 1, 2, 3,\ j = 1, ..., n_i)$ with the step–wise tests

$$\varphi_{(i)} = \tilde{\varphi}_{(i)} \cdot \prod_{j=1}^{i-1} \left[\tilde{\varphi}_{(j)} \prod_{k=1}^{n_{(j)}} \tilde{\varphi}_{(j)(k)} \right] \text{ and} \tag{6}$$

$$\varphi_{(i)(j)} = \varphi_{(i)} \cdot \prod_{k=1}^{j} \tilde{\varphi}_{(i)(k)}, \tag{7}$$

where

$$\tilde{\varphi}_{(i)} = \left\{ \begin{array}{lll} 0 & & > \\ & \text{if } p_{(i)} & \dfrac{\alpha}{3-i+1}, \\ 1 & & \leq \end{array} \right. \qquad i = 1, 2, 3, \tag{8}$$

and

$$\tilde{\varphi}_{(i)(j)} = \left\{ \begin{array}{lll} 0 & & > \\ & \text{if } p_{(i)(j)} & \dfrac{\alpha/(3-i+1)}{n_{(i)} - j + 1}, \\ 1 & & \leq \end{array} \right. i = 1, 2, 3,\ j = 1, ..., n_i. \tag{9}$$

Here, $\tilde{\varphi}_{(i)}$ and $\tilde{\varphi}_{(i)(j)}$, respectively, denote the individual tests for the intersection and elementary hypotheses arranged according to the corresponding p–values. For $i = 1$, $\prod_{j=1}^{i-1} [\tilde{\varphi}_{(j)} \prod_{k=1}^{n_{(j)}} \tilde{\varphi}_{(j)(k)}]$ is defined as 1.

Theorem 3

The BHM II procedure according to (6) – (9) keeps the multiple level α.

For the proof we essentially refer to Bauer et al. (1998) as detailed in the appendix. Like the BHM I procedure, but in other situations, the BHM II procedure may be rather conservative as will be discussed below.

2.3 Comparison of the procedures

There is a crucial difference between the BH procedure and the BHM I as well as the BHM II method. While the intersection hypotheses for the factors A, B and the interaction $A \times B$ are explicitly tested in the latter two procedures, they are only implicitly tested in the BH procedure.

Let for instance the test of H_0^{AB} have the smallest p–value. If now one of the hypotheses related to the interaction cannot be rejected, then the BHM II procedure stops without testing any of the pairwise comparisons related to the main effects of A and B. Using the BH procedure, however, one might have the chance to reject some of the pairwise hypotheses of the two main effects. The BHM I procedure also allows for testing pairwise comparisons related to the factors A and B, even if some of the pairwise interaction hypotheses turn out to be non–significant, since here the two factors and the interaction are treated separately.

As mentioned earlier, the BH procedure might result in very small adjusted p–values, if many elementary hypotheses are to be tested. But this is also the case for the other procedures. Consider again the situation that $p_{(A\times B)}$

is the smallest p–value of the intersection hypotheses. Then, the smallest p–value of the BHM II pairwise comparisons is compared with $\frac{\alpha/3}{KL(KL-1)/2}$, which is even smaller than the smallest of the BH procedure. However, if $p_{(A\times B)}$ is not the smallest p–value then the adjusted values will be larger. The smallest possible adjusted level of the BHM I procedure is $\frac{\alpha/3}{KL(KL-1)/2}$, too. However, the adjusted significance levels that the two smallest p–values of factor A and B have to be compared with are greater for the BHM II procedure than for the BHM I method. This is because the three intersection hypotheses are interconnected not simply by the Bonferroni inequality, but according to the Bonferroni–Holm principle.

Another aspect of multiple test procedures besides committing errors of type I concerns the possibility that their components may lead to overall decisions which are not free of contradictions. Comparing the above procedures w.r.t. the concepts of coherence and consonance introduced by Gabriel (1969) it is obvious that all three procedures are coherent by construction, but only the original Bonferroni–Holm procedure is also consonant whereas the BHM I and BHM II procedures may yield non–consonant decisions.

3 Simulation

In the previous section, it was shown that the Bonferroni–Holm procedure and two of its modifications, namely BHM I and BHM II, keep the multiple level α and thus also control the per–comparison error rate. To get an idea, which of these three test procedures is best regarding its power, a small simulation study is performed, with 1000 simulation runs carried out for each constellation.

The comparison is based on the simultaneous power, briefly denoted as power I in the following, as analogue to the multiple level, and on the proportion of correctly rejected false hypotheses, briefly denoted as power II, corresponding to the per–comparison error rate.

3.1 Design

The simulation study is based on model (1) assuming normality for the error terms, homogeneity of variances, and a balanced design. For each factor we have three levels, i.e. $K = L = 3$. This results in three pairwise comparisons for each factor and in 36 hypotheses concerning all possible interaction comparisons. The individual tests are performed as F–tests for the intersection hypotheses and as t–tests for the pairwise comparisons.

The multiple level α is fixed at 5%, which results in $5.95 \cdot 10^{-4}$ as adjusted significance level in the first step of the BH procedure. If $p_{(A \times B)}$ is the smallest p–value of the three intersection hypotheses, the smallest p–value of the pairwise comparisons using the BHM I or BHM II procedure is compared with $2.31 \cdot 10^{-4}$, which is even smaller than the one of the BH procedure as noted above. The adjusted significance levels, with which the two smallest p-values of the tests for the pairwise comparisons within factors A and B are compared afterwards, are larger using the BHM II procedure with $4.17 \cdot 10^{-3}$ and $8.33 \cdot 10^{-3}$ than using the BHM I procedure with $2.78 \cdot 10^{-3}$.

Using the polar Marsaglia procedure (Moeschlin et al., 1995) normally distributed random numbers are generated. The sample size N is fixed at 100 and the grand mean μ is 0 without loss of generality. Regarding the variance, another parameter is important to judge the power of the different multiple tests: the smallest difference of two (non equal) means denoted by δ. Different values of δ allow us to get an idea of the capacity of the various procedures to detect small differences in the means. It seems reasonable not to look at δ and σ separately, but to use a combined measure, i.e. δ/σ. Thus, the actual value of σ is no longer of particular interest. It is therefore fixed at 1, but varying values of δ/σ are considered ranging from 0.03 to 0.90 with a step width of 0.03. The obtained Monte–Carlo results are only reported for the most interesting cases.

Three constellations of true and false elementary hypotheses are investigated. First, all elementary hypotheses, i.e. those belonging to the two factors and to the interaction, are true. Second, they are all false, and in the third case they are partially true and false.

Let us denote the number of true elementary hypotheses belonging to the factors A, B and the interaction $A \times B$ as $|I_i|$ as above, the number of false elementary hypotheses as $|\overline{I}_i|, i \in \{A, B, A \times B\}$. If some of the elementary hypotheses of the interaction are false, there are different possibilities for the number of true and false hypotheses. We decide to report only the cases $|I_{A\times B}| = 12$ or 5. For all other situations with $|I_{A\times B}| < 12$, the results tend to be of the same order of magnitude. For $|I_{A\times B}| \geq 18$, however, the results

Table 1 Power I and power II for the situation of main effects for exactly two levels of each factor A and B and no interactions.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.15	0.006	0.156	0.000	0.139	0.000	0.073
0.18	0.028	0.306	0.000	0.255	0.006	0.153
0.21	0.074	0.436	0.000	0.333	0.011	0.244
0.24	0.188	0.565	0.000	0.394	0.022	0.357
0.27	0.383	0.755	0.000	0.466	0.138	0.531
0.30	0.590	0.859	0.000	0.488	0.270	0.664
0.33	0.730	0.919	0.000	0.499	0.459	0.786
0.36	0.858	0.964	0.000	0.500	0.644	0.881
0.39	0.929	0.985	0.000	0.500	0.781	0.934
0.42	0.982	0.994	0.000	0.500	0.892	0.968
0.45	0.997	0.999	0.000	0.500	0.942	0.984

are quite different especially concerning the most powerful test. Only in the case described in Table 4 the results obtained for $|I_{A\times B}| \geq 18$ are in general of similar size as those obtained for $|I_{A\times B}| \leq 12$. Some selected simulation results are summarized in Tables 1–8.

3.2 Results

Level of Significance

The situation of homogeneity of means and of no interaction effects is mainly considered to assess the nominal multiple level achieved by the proposed procedures. In the simulation, we observe a multiple level of significance of 3.7% for the BHM I and II procedure and a value of 3.5% for the BH procedure. Thus, the problem already addressed above, that the nominal level can be clearly below α, in fact occurs. All procedures are conservative with the BH procedure slightly more conservative than the others.

For the nominal per–comparison error rate we get a value of 0.22% using the BHM I and II procedure and a value of 0.14% using the BH method. Again, the latter is most conservative. Note that the nominal multiple level and the nominal per–comparison error rate are also kept with designs different from the one chosen here.

Power

The simultaneous power depends substantially more on the size of the differences in the means than the power II. To achieve a simultaneous power larger than zero, δ/σ has to be at least – with a few exceptions – 0.15 if all elementary hypotheses concerning the interaction terms are true. Otherwise δ/σ must be larger than 0.27. For a positive power II, however, we only need the differences in the means to be 0.03 times the standard deviation.

Regarding the remaining simulation results, let us point out that there is no simple answer to the question which of the procedures is best with regard to its power. One should be aware of the fact that the performances of the test procedures heavily depend on the true parameter values. But

Table 2 Power I and power II for the situations of no (one) true null hypothesis for the main effects of factor A, one (no) for the main effects of factor B, and 12 true (in brackets 5) null hypotheses for the interactions. The results are the same for both constellations of factors A and B.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.39	0.040	0.850	0.000	0.721	0.101	0.879
	(0.114)	(0.941)	(0.000)	(0.763)	(0.242)	(0.957)
0.42	0.137	0.890	0.000	0.762	0.232	0.919
	(0.143)	(0.958)	(0.000)	(0.781)	(0.314)	(0.970)
0.45	0.194	0.923	0.000	0.808	0.381	0.946
	(0.359)	(0.977)	(0.000)	(0.798)	(0.540)	(0.980)
0.48	0.364	0.951	0.000	0.842	0.539	0.968
	(0.548)	(0.983)	(0.000)	(0.806)	(0.727)	(0.990)
0.51	0.493	0.970	0.000	0.857	0.664	0.981
	(0.793)	(0.994)	(0.000)	(0.822)	(0.850)	(0.997)
0.54	0.644	0.980	0.000	0.873	0.797	0.989
	(0.824)	(0.995)	(0.000)	(0.840)	(0.922)	(0.998)
0.57	0.784	0.989	0.000	0.886	0.855	0.993
	(0.880)	(0.997)	(0.000)	(0.869)	(0.954)	(0.999)
0.60	0.859	0.994	0.000	0.896	0.937	0.997
	(0.934)	(0.998)	(0.000)	(0.900)	(0.973)	(0.999)
0.63	0.902	0.996	0.000	0.895	0.959	0.998
	(0.981)	(0.999)	(0.000)	(0.914)	(0.992)	(1.000)
0.66	0.972	0.999	0.000	0.901	0.992	1.000
	(0.987)	(1.000)	(0.000)	(0.922)	(0.993)	(1.000)

Table 3 Power I and power II for the situations of three (one) true null hypotheses for the main effects of factor A, one (three) for the main effects of factor B, and no interactions. The results are the same for both constellations.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.15	0.084	0.184	0.084	0.184	0.031	0.089
0.18	0.135	0.275	0.135	0.275	0.025	0.121
0.21	0.301	0.447	0.301	0.447	0.108	0.238
0.24	0.443	0.614	0.443	0.614	0.213	0.379
0.27	0.595	0.717	0.595	0.717	0.550	0.513
0.30	0.721	0.832	0.721	0.832	0.464	0.633
0.33	0.881	0.933	0.881	0.933	0.668	0.790
0.36	0.934	0.965	0.934	0.965	0.809	0.889
0.39	0.961	0.985	0.961	0.985	0.890	0.931
0.42	0.976	0.992	0.976	0.992	0.940	0.966
0.45	1.000	1.000	1.000	1.000	0.992	0.998

additional information for instance due to subject–matter knowledge may help to reach a decision. The results are now given in more detail.

A striking result is that the simultaneous power of the BHM II procedure is exactly zero whenever at least two of the intersection hypotheses but not all of the associated pairwise hypotheses are false (cf. Tables 1, 2). Since this procedure stops as soon as one of the elementary hypotheses cannot be rejected, the false hypotheses belonging to the other factor will always

Table 4 Power I and power II for the situations of no main effects of the factors A and B, and 12 (in brackets 5) true null hypotheses for the interactions.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.39	0.037	0.800	0.037	0.800	0.085	0.848
	(0.083)	(0.932)	(0.083)	(0.932)	(0.154)	(0.945)
0.42	0.104	0.870	0.104	0.870	0.200	0.902
	(0.218)	(0.955)	(0.218)	(0.955)	(0.284)	(0.963)
0.45	0.216	0.906	0.216	0.906	0.344	0.930
	(0.400)	(0.972)	(0.400)	(0.972)	(0.488)	(0.978)
0.48	0.367	0.939	0.367	0.939	0.473	0.956
	(0.585)	(0.985)	(0.585)	(0.985)	(0.657)	(0.988)
0.51	0.485	0.957	0.485	0.957	0.593	0.970
	(0.696)	(0.989)	(0.696)	(0.989)	(0.757)	(0.991)
0.54	0.617	0.971	0.617	0.971	0.741	0.982
	(0.834)	(0.994)	(0.834)	(0.994)	(0.869)	(0.995)
0.57	0.774	0.985	0.774	0.985	0.848	0.991
	(0.872)	(0.996)	(0.872)	(0.996)	(0.919)	(0.997)
0.60	0.859	0.992	0.859	0.992	0.894	0.994
	(0.957)	(0.998)	(0.957)	(0.998)	(0.963)	(0.999)
0.63	0.921	0.996	0.921	0.996	0.958	0.998
	(0.963)	(0.999)	(0.963)	(0.999)	(0.978)	(0.999)
0.66	0.946	0.998	0.946	0.998	0.978	0.999
	(0.995)	(1.000)	(0.995)	(1.000)	(0.997)	(1.000)

Table 5 Power I and power II for the situation of no main effects of the factors A and B and all possible interactions present.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.24	0.000	0.856	0.000	0.856	0.000	0.868
0.27	0.025	0.898	0.025	0.898	0.006	0.904
0.30	0.133	0.932	0.133	0.932	0.095	0.933
0.33	0.350	0.959	0.350	0.959	0.245	0.957
0.36	0.530	0.976	0.530	0.976	0.421	0.973
0.39	0.648	0.986	0.648	0.986	0.563	0.983
0.42	0.774	0.992	0.774	0.992	0.659	0.987
0.45	0.878	0.996	0.878	0.996	0.810	0.994
0.48	0.954	0.999	0.954	0.999	0.894	0.997

be retained which yields the above phenomenon. In addition, its power II can never reach 1 in these situations since the BHM II procedure can reject all false elementary hypotheses within one factor, but not those within the other one if it stops when not rejecting some true elementary hypotheses. As illustrated in Table 1, the power II, for instance, cannot exceed 50% if there are exactly two false elementary hypotheses per factor regarding the main effects (and no interactions).

Another general result is that both modifications, BHM I and BHM II, have the same power I and II when exactly one intersection hypothesis is false (cf. Tables 3, 4, 5). This seems plausible as both procedures would typically

Table 6 Power I and power II for the situation of all three main effects of factor A and B being present and no interactions.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.15	0.000	0.495	0.003	0.336	0.000	0.284
0.18	0.006	0.641	0.026	0.423	0.000	0.402
0.21	0.064	0.755	0.088	0.538	0.000	0.500
0.24	0.190	0.842	0.254	0.693	0.006	0.586
0.27	0.361	0.894	0.445	0.797	0.032	0.678
0.30	0.652	0.950	0.711	0.902	0.177	0.795
0.33	0.802	0.973	0.843	0.955	0.383	0.866
0.36	0.882	0.985	0.906	0.973	0.600	0.925
0.39	0.965	0.996	0.974	0.994	0.763	0.957
0.42	0.991	0.999	0.992	0.997	0.886	0.980
0.45	0.997	1.000	0.998	0.999	0.942	0.993

start by testing this intersection hypothesis using the same local level of significance.

In the case that there are no interactions, the power of the Bonferroni–Holm procedure is usually the worst (cf. Tables 1, 3, 6). This is because the BHM I and II procedures start with an adjusted significance level for the pairwise comparisons of the main effects of $\frac{\alpha/3}{3-i+1}$ which is much larger than the one of the BH procedure with $\frac{\alpha}{42-i+1}$ for $1 \leq i \leq 3$. The bad performance of the Bonferroni–Holm procedure, here, is due to the much higher number of elementary hypotheses for the interactions than for the main effects together with all these interaction hypotheses being true. In a situation where the subsets of elementary hypotheses are of equal size one might expect results that are more favourable for the BH procedure. Further, if there are no interactions and the power I of the BHM II procedure is not zero, BHM II is usually better than BHM I w.r.t. power I but worse regarding power II (e.g. Table 6) so that no clear ranking of these two modifications can be established for these constellations.

If there is a considerable amount of interactions, however, the Bonferroni–Holm procedure is usually the most powerful (cf. Tables 2, 4, 7). A few ambiguous situations occur when all interactions are present with no main effects in one ore both factors (cf. Tables 5, 8) but the power II of BHM I and of the original Bonferroni–Holm then still seem to be very similar.

4 Discussion

From the above simulation results it becomes obvious that no simple and general rule can be given for one of the procedures being the best one. Such a rule does not even exist if it is restricted to particular situations since the performance of the tests heavily depends on the true parameter constellation. It would of course be helpful to have some further knowledge of the empirical situation before choosing a test procedure. Typically, such an information is, however, not known in advance. Without going into details, one possible way–out might be to perform preliminary tests in order to reach

Table 7 Power I and power II for the situations of no (all) main effects of factor A, all (no) main effects of factor B, and 12 (in brackets 5) true null hypotheses for the interactions. The results are the same for both constellations.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.39	0.051	0.840	0.078	0.841	0.119	0.881
	(0.093)	(0.937)	(0.000)	(0.828)	(0.181)	(0.951)
0.42	0.131	0.882	0.151	0.896	0.231	0.914
	(0.189)	(0.961)	(0.000)	(0.850)	(0.354)	(0.972)
0.45	0.245	0.923	0.282	0.927	0.398	0.946
	(0.402)	(0.975)	(0.003)	(0.865)	(0.583)	(0.983)
0.48	0.383	0.950	0.417	0.951	0.520	0.965
	(0.524)	(0.983)	(0.010)	(0.875)	(0.644)	(0.988)
0.51	0.507	0.966	0.556	0.968	0.671	0.979
	(0.690)	(0.990)	(0.060)	(0.889)	(0.773)	(0.993)
0.54	0.639	0.979	0.686	0.981	0.760	0.987
	(0.847)	(0.996)	(0.247)	(0.914)	(0.884)	(0.997)
0.57	0.797	0.984	0.800	0.986	0.896	0.995
	(0.893)	(0.997)	(0.474)	(0.945)	(0.942)	(0.998)
0.60	0.839	0.991	0.861	0.992	0.895	0.994
	(0.927)	(0.998)	(0.758)	(0.974)	(0.951)	(0.999)
0.63	0.926	0.995	0.926	0.996	0.956	0.998
	(0.972)	(0.999)	(0.912)	(0.991)	(0.992)	(1.000)
0.66	0.971	0.999	0.986	0.998	0.990	0.999
	(1.000)	(1.000)	(0.984)	(0.997)	(1.000)	(1.000)

Table 8 Power I and power II for the situations of no (all) main effects of factor A, all (no) main effects of factor B, and all interactions present. The results are the same for both constellations.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.24	0.000	0.835	0.000	0.789	0.000	0.853
0.27	0.021	0.897	0.022	0.811	0.032	0.932
0.30	0.094	0.932	0.099	0.850	0.135	0.943
0.33	0.244	0.961	0.257	0.893	0.240	0.962
0.36	0.616	0.980	0.622	0.945	0.633	0.983
0.39	0.676	0.987	0.677	0.956	0.641	0.988
0.42	0.770	0.993	0.773	0.972	0.757	0.994
0.45	0.889	0.996	0.890	0.985	0.883	0.996
0.48	0.938	0.998	0.939	0.992	0.912	0.998
0.51	0.973	0.999	0.975	0.997	0.962	0.999

a decision about the final test procedure. Such an approach can be regarded as an adaptive procedure where the final multiple test depends on the given data. When using such an adaptive procedure it needs to be checked, again, whether the multiple level is being kept and how the simultaneous power or power II behave. To summarize, the results of Section 3 may be regarded as rough hints when confronted with the problem of selecting an adequate test.

Furthermore, it has to be mentioned that the three procedures introduced in this paper are not optimal, since none of them fully exhausts the significance level of 5%. The question arises whether improvements can be achieved by a more specific determination of the adjusted levels, as for instance those proposed by Shaffer (1986) or Royen (1987) exploiting logical dependencies among the null hypotheses and/or using different test statistics (cf. Royen, 1988, 1990, Finner, 1988, Bergmann and Hommel, 1988). Since the proposed procedures of the Bonferroni–Holm type are generally applicable they can be easily modified accounting for the approaches presented by the authors listed above.

Let us point out that another approach could be based on a Scheffé–type procedure (Scheffé, 1953). The family of null hypotheses that we are investigating in the two–way layout can in fact be regarded as contrasts in a one–way layout with $K \times L$ levels of one combined factor. However, as the Scheffé procedure ensures the multiple level for *all* contrasts, not only for those of specific interest, we expect it to perform worse than the above Bonferroni–Holm modifications which are designed to find the 'deviant' main effects and interactions. Further simulations are required to corroborate this and especially to quantify the difference in performance.

As a last point to be made, it has to be examined how the three procedures behave w.r.t. their power, if they are used in the context of an ANOVA with more than two factors. Since the adjusted levels will then be even smaller, it is obvious that any rejection of a hypothesis becomes improbable for small differences. Other techniques based on modelling the correlation structure e.g. in the framework of a multivariate t–distribution and thus avoiding any adjustments may be more appropriate (cf. Bretz, 1999, and Bretz et al., 2001), although such an approach requires more specific distributional assumptions.

Finally, let us emphasize that the problems occurring when adjusting for multiplicity in a multi–way ANOVA point to the necessity to keep the number of hypotheses to be tested small. It could e.g. be thought about whether all pairwise interaction hypotheses are equally important or whether some of them could be discarded.

5 Appendix

Proof of Theorem 2 Consider first testing one intersection hypothesis, H_0^A say, together with the collection of the corresponding pairwise comparisons. With G_1 being the set containing the intersection hypothesis and G_2 the collection of pairwise comparisons, G_1 and G_2 can be regarded as two sets of partially ordered nullhypotheses as addressed in Maurer et al. (1995). It is therefore clear that our procedure ensures that the null hypotheses in G_1 and G_2 are tested at the multiple level $\alpha/3$.

Now, it follows immediately from the Bonferroni inequality that the whole set of null hypotheses, the three types of intersections and their corresponding pairwise comparisons, are being tested at the multiple level α.

Proof of Theorem 3 The proof essentially refers to the one given in Bauer et al. (1998). These authors consider the case of multi–dose experiments including an active control but it becomes clear from the proof of their Lemma 2 that their procedure is more general. It can be used whenever a collection of nullhypotheses that are to be tested can be partioned such that the subsets can be tested at a given local multiple level. It is not actually relevant which multiple test within the partitions is used to keep this local multiple level — in our case it is a Bonferroni–Holm procedure.

Furthermore, in our case, the null hypotheses are partioned naturally into the pairwise comparisons of the main effects for each factor and the comparisons of the interactions. The local multiple levels themselves are again chosen according to the Bonferroni–Holm idea and this ensures (by Lemma 2 of Bauer et al., 1998) the overall multiple level α.

References

1. Bauer P, Röhmel J, Maurer W, Hothorn L (1998) Testing strategies in multi-dose experiments including active control. Statist. in Med. **17**, 2133-2146.
2. Bergmann B, Hommel G (1988) Improvements of general multiple test procedures for redundant systems of hypotheses. In: Multiple hypotheses testing (Eds. Bauer P, Hommel G, Sonnemann E), 100-115. Springer Verlag, Berlin.
3. Bretz F (1999) Powerful modifications of Williams' test on trends. PhD–Thesis, University of Hannover.

4. Bretz F, Hayter AJ, Genz A (2001) Critical point and power calculations for the studentized range test for generally correlated means. J. Statist. Comput. and Simul. **71**, 85-99.

5. Finner H (1988) Abgeschlossene multiple Spannweitentests. In: Multiple hypotheses testing (Eds. Bauer P, Hommel G, Sonnemann E), 10-32. Springer Verlag, Berlin.

6. Gabriel KR (1969) Simultaneous test procedures – some theory of multiple comparisons. Ann. Math. Statist. **40**, 224-250.

7. Hartley HO (1955) Some recent developments in analysis of variance. Comm. Pure and Appl. Math. **8**, 47-72.

8. Holm S (1977) Sequentially rejective multiple test procedures. Statistical Research Report 1977-1, Institute of Mathematics and Statistics, University of Umeå.

9. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand. J. Statist. **6**, 65-70.

10. Maurer W, Hothorn LA, Lehmacher W (1995) Multiple comparisons in drug clinical trials and preclinical assays: a–priori ordered hypotheses. In: Biometrie in der chmisch–pharmazeutischen Industrie, Vol. 6 (Ed. Vollmer J), 3-18. Fischer Verlag, Stuttgart.

11. Maurer W, Mellein B (1988) On new multiple tests based on independent p–values and the assessment of their power. In: Multiple hypotheses testing (Eds. Bauer P, Hommel G, Sonnemann E), 48-66. Springer Verlag, Berlin.

12. Moeschlin O, Pohl C, Grycki E, Steinert F (1995) Statistik und Experimentelle Stochastik. Birkhäuser, Basel.

13. Ottestad P (1960) On the use of the F–test in cases in which a number of variance ratios are computed by the same error mean square. Science Reports

from the Agriculture College of Norway **39**, 1-8.

14. Ottestad P (1970) Statistical models and their experimental application. Griffin, London.

15. Royen T (1987) Eine verschärfte Holm–Prozedur zum Vergleich aller Mittelwertpaare. EDV in Medizin und Biologie **18**, 45-49.

16. Royen T (1988) The maximum range test – an improved step down procedure for the comparison of all pairs of means. EDV in Medizin und Biologie **19**, 58-63.

17. Royen T (1990) A probability inequality for ranges and its application to maximum range test procedures. Metrika **37**, 145-154.

18. Shaffer JP (1986) Modified sequentially rejective multiple test procedures. J. Amer. Statist. Assoc. **81**, 826-831.

19. Sheffé H (1953) A method for judging all contrasts in the analysis of variance. Biometrika **40**, 87-104.