Didelez, Pigeot, Walter:

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

Sonderforschungsbereich 386, Paper 185 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

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

VANESSA DIDELEZ, IRIS PIGEOT, and PATRICIA WALTER
University of Munich, Institute of Statistics, Ludwigstrasse 33,
D–80539 Munich, Germany

Abstract

This paper aims at constructing stepwise test procedures based on the Bonferroni–Holm principle for a multi–way ANOVA. Especially for the two–way ANOVA it is shown, that the procedures keep the multiple level α. These theoretical results are supplemented by a simulation study to compare the multiple procedures regarding two power concepts and to learn about which of the introduced procedures is the best.

Key words: Adjustment for multiplicity, Bonferroni–Holm procedure, multiple test problem, multi–way ANOVA, stepwise 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. Such detailed comparisons call for a multiple test procedure, which captures 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. Particularly 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 if non–standard situations such as unequal sample sizes or linear contrasts instead of pairwise comparisons are investigated.

Multiple tests in the context of a two or multi–way ANOVA, however, has not been paid so much attention up to now, so that for this case only very few procedures are known, as the method of HARTLEY (1955) or OTTESTAD (1960, 1970).
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. Since our proposals are mainly based on a modification of the Bonferroni–Holm procedure, they can be easily extended to applications in a multi–way layout. They are defined as stepwise test procedures and thus more powerful than their simultaneous counterparts. It is additionally investigated if the proposed test procedures keep the multiple level α, where it can be shown that two of our proposals fulfil this property whereas the third modification does not. The procedures are then compared with respect to their power by means of Monte–Carlo experiments based on the simultaneous power (MAURER & MELLEIN, 1988) and the relative frequency of correctly rejected false hypotheses.

2 Multiple tests in a two–way ANOVA

The multiple test procedures which will be introduced in Section 2.2 are based on the Bonferroni–Holm approach. This general principle for constructing stepwise test procedures allows for the application of any suitable level α test. Thus, our procedures are not restricted to the classical Gaussian case as introduced in Section 2, but also apply to nonparametric tests. The simulation study (Section 3) is nevertheless restricted to the classical situation, i.e. F–tests are used to check overall hypotheses and multiple t–tests for all pairwise comparisons.

2.1 Basic notations

For convenience, let us briefly introduce the classical two–way ANOVA setting. The statistical model reads as

$$Y_{kln} = \mu + \alpha_k + \beta_l + (\alpha\beta)_{kl} + \epsilon_{kln}, \quad k = 1, ..., K, \quad l = 1, ..., L, \quad n = 1, ..., N,$$

where the error terms ϵ_{kln} are assumed as i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. The main effect of factor A on level k and the main effect of factor B on level l are denoted

as α_k and β_l, the interaction effect of factor A and B on levels (k,l) as $(\alpha\beta)_{kl}$, and the grand mean as μ.
The latter is estimated via the arithmetic mean of all observations, i.e.

$$\hat{\mu} = \frac{1}{KLN} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} Y_{kln} = \overline{Y}_{...}.$$

The maximum likelihood estimators of the two main effects α_k and β_l are given as deviation of the mean on the corresponding factor level from the grand mean, i.e.

$$\hat{\alpha}_k = \overline{Y}_{k..} - \overline{Y}_{...} \quad \text{with} \quad \overline{Y}_{k..} = \frac{1}{LN} \sum_{l=1}^{L} \sum_{n=1}^{N} Y_{kln},$$

$$\hat{\beta}_l = \overline{Y}_{.l.} - \overline{Y}_{...} \quad \text{with} \quad \overline{Y}_{.l.} = \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} Y_{kln}.$$

The ML estimator of the interaction effect $(\alpha\beta)_{kl}$ reads as

$$\widehat{(\alpha\beta)}_{kl} = \overline{Y}_{kl.} - \overline{Y}_{k..} - \overline{Y}_{.l.} + \overline{Y}_{...} \quad \text{with} \quad \overline{Y}_{kl.} = \frac{1}{N} \sum_{n=1}^{N} Y_{kln}.$$

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 interaction $A \times B$. In detail, the intersection hypothesis w.r.t. to factor A is denoted as H_0^A with

$$H_0^A : \mu_{1.} = \mu_{2.} = ... = \mu_{K.}$$

and has to be tested against

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

where $\mu_{i.}$ is the mean on level i of factor A. 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 influence on Y. For factor A, we have in total $\frac{1}{2} \cdot K(K-1)$ pairwise comparisons of the type

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

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 can be applied in a straightforward manner 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, irrespectively 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 $n_* = \left[\frac{1}{2} \cdot K(K-1) + \frac{1}{2} \cdot L(L-1) + \frac{1}{2} \cdot KL(KL-1)\right]$. 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 here not tested explicitly.

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

$$\varphi_{(i)} = \prod_{j=1}^{i} \tilde{\varphi}_{(j)}, \quad i = 1, ..., n_*, \tag{2.1}$$

where

$$\tilde{\varphi}_{(j)} = \begin{cases} 0 & > \\ \text{for } p_{(j)} & \dfrac{\alpha}{(n-j+1)}, \quad j = 1, ..., n_*, \\ 1 & \leq \end{cases} \tag{2.2}$$

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 2.1

The BH procedure according to (2.1) and (2.2) 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

$$\varphi_i = \begin{cases} 0 & \text{if } p_i > \alpha/3, \\ 1 & \text{if } p_i \leq \alpha/3, \end{cases} \qquad i \in \{A, B, A\times B\}, \tag{2.3}$$

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

$$\tilde{\varphi}_{i(k)} = \begin{cases} 0 & \text{if } p_{i(k)} > \dfrac{\alpha/3}{n_i - k + 1}, \\ 1 & \text{if } p_{i(k)} \leq \dfrac{\alpha/3}{n_i - k + 1}, \end{cases} \qquad k = 1, ..., n_i. \tag{2.4}$$

Here, $\tilde{\varphi}_{i(j)}$ represent the individual tests 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 2.2

The BHM I procedure according to (2.3) and (2.4) keeps the multiple level α.

Proof

Let I_0 be given as the set of indices of the true intersection hypotheses with $|I_0| = m_0$ and I_i the set of indices of true null hypotheses w.r.t. the pairwise comparisons within factor i and $|I_i|$ their number, $i \in \{A, B, A \times B\}$. Furthermore, φ_{ij} denotes

the corresponding unordered test and p_{ij} the unordered p–values for the elementary hypotheses. The rank of p_{ij} within $\{p_{ij}; j = 1, ..., n_i\}$ for fixed i is given as $R(p_{ij})$. The proof is divided into two steps. In the first step, we consider the situation that an intersection hypothesis H_0^i, $i \in \{A, B, A \times B\}$, is true. The probability for a false rejection within this factor is bounded by $\alpha/3$, i.e.

$$P\left[(\varphi_i = 1) \cup \left(\bigcup_{j \in I_i} \{\varphi_{ij} = 1\}\right)\right] = P(\varphi_i = 1) \leq \alpha/3,$$

since

$$\bigcup_{j \in I_i} \{\varphi_{ij} = 1\} \subseteq \{\varphi_i = 1\} \cap \bigcup_{j \in I_i} \{\tilde{\varphi}_{ij} = 1\}. \tag{2.5}$$

In the second step, we consider the situation that an intersection hypothesis H_0^i, $i \in \{A, B, A \times B\}$, is false. The probability for rejecting a possibly true pairwise comparison belonging to this factor is bounded as follows. Equation (2.5) yields that

$$\begin{aligned}
P\left(\bigcup_{j \in I_i} \{\varphi_{ij} = 1\}\right) &\leq P\left[(\varphi_i = 1) \cap \left(\bigcup_{j \in I_i} \{\tilde{\varphi}_{ij} = 1\}\right)\right] \\
&\leq P\left(\bigcup_{j \in I_i} \{\tilde{\varphi}_{ij} = 1\}\right) = 1 - P\left(\bigcap_{j \in I_i} \{\tilde{\varphi}_{ij} = 0\}\right) \\
&= 1 - P\left(\bigcap_{j \in I_i} \{\tilde{\varphi}_{i(R(p_{ij}))} = 0\}\right) \\
&= 1 - P\left(\tilde{\varphi}_{i(\min_{j \in I_i} R(p_{ij}))} = 0\right) = P\left(\tilde{\varphi}_{i(\min_{j \in I_i} R(p_{ij}))} = 1\right) \\
&= P\left(\exists\, s \in I_i : p_{is} \leq \frac{\alpha/3}{n_i - \min_{j \in I_i} R(p_{ij}) + 1}\right) \\
&\leq \sum_{s \in I_i} P\left(p_{is} \leq \frac{\alpha/3}{n_i - \min_{j \in I_i} R(p_{ij}) + 1}\right) \\
&\leq \sum_{s \in I_i} P\left(p_{is} \leq \frac{\alpha/3}{|I_i|}\right) \leq \sum_{s \in I_i} \frac{\alpha/3}{|I_i|} \quad = \quad \alpha/3.
\end{aligned}$$

These two steps imply that the probability of falsely rejecting at least one of the true hypotheses can be calculated as

$$\begin{aligned} P\left(\bigcup_{i\in I_0}\{\varphi_i=1\}\cup\bigcup_{i\notin I_0}\bigcup_{j\in I_i}\{\varphi_{ij}=1\}\right) &\leq \sum_{i\in I_0}P\left(\varphi_i=1\right)+\sum_{i\notin I_0}P\left(\bigcup_{j\in I_i}\{\varphi_{ij}=1\}\right) \\ &\leq \sum_{i\in I_0}\alpha/3+\sum_{i\notin I_0}\alpha/3=\alpha. \end{aligned}$$

□

Since parts of the above proof are based on the Bonferroni inequality, it has to be expected that the nominal multiple level of this test can become clearly smaller than α. That means that despite of the Bonferroni–Holm adjustment applied separately to each factor as well as for the interaction 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 with $\alpha/3$. 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, if one of the pairwise

comparisons cannot be rejected this 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.
This procedure, however, 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 in each case of the pairwise comparisons. Let us for instance assume that not all means are equal, but that the last pairwise comparison to be tested within the factors and intersections, respectively, is true but rejected. This error occurs at worst with a probability of $\alpha/3+\alpha/2+\alpha$, so that the multiple level α is exceeded.
The above procedure can, however, be improved such that it keeps 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 antirank.
The BHM II procedure is given as $(\varphi_i,\ \varphi_{ij};\ i = 1, 2, 3,\ j = 1, ..., n_i)$ with the stepwise tests

$$\varphi_{(i)} = \tilde{\varphi}_{(i)} \cdot \prod_{j=1}^{i-1} \left[\tilde{\varphi}_{(j)} \prod_{k=1}^{n_{(j)}} \tilde{\varphi}_{(j)(k)} \right] \text{ and} \tag{2.6}$$

$$\varphi_{(i)(j)} = \varphi_{(i)} \cdot \prod_{k=1}^{j} \tilde{\varphi}_{(i)(k)}, \tag{2.7}$$

where

$$\tilde{\varphi}_{(i)} = \begin{cases} 0 & > \\ \text{if } p_{(i)} & \dfrac{\alpha}{3-i+1}, \\ 1 & \leq \end{cases} \qquad i = 1, 2, 3, \tag{2.8}$$

and

$$\tilde{\varphi}_{(i)(j)} = \begin{cases} 0 & > \\ \text{if } p_{(i)(j)} & \dfrac{\alpha/(3-i+1)}{n_{(i)}-j+1}, \\ 1 & \leq \end{cases} \quad i = 1, 2, 3,\ j = 1, ..., n_i. \tag{2.9}$$

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 2.3

The BHM II procedure according to (2.6) – (2.9) keeps the multiple level α.

Proof

In addition to the notations used in the proof for the BHM I procedure, let p_{ij} be the unordered p–values for the elementary hypotheses. We have to show that the probability of rejecting one or more true hypotheses is bounded by α.
Since

$$\bigcup_{i\in I_0}\bigcup_{j\in I_i}\{\varphi_{ij}=1\} \subseteq \bigcup_{i\in I_0}\{\varphi_i=1\}, \tag{2.10}$$

the probability for a multiple type I error results in

$$\begin{aligned}
& P\left(\bigcup_{i\in I_0}\{\varphi_i=1\} \cup \bigcup_{i=1}^{3}\bigcup_{j\in I_i}\{\varphi_{ij}=1\}\right) \\
\overset{(2.10)}{=}\ & P\left(\bigcup_{i\in I_0}\{\varphi_i=1\} \cup \bigcup_{k\notin I_0}\bigcup_{j\in I_k}\{\varphi_{kj}=1\}\right) \\
=\ & 1-P\left(\bigcap_{i\in I_0}\{\varphi_i=0\} \cap \bigcap_{k\notin I_0}\bigcap_{j\in I_k}\{\varphi_{kj}=0\}\right) \\
=\ & 1-P\left[\left(\varphi_{(\min_{i\in I_0} R(p_i))}=0\right) \cap \left(\varphi_{(\min_{k\notin I_0} R(p_k))(\min_{j\in I_k} R(p_{kj}))}=0\right)\right]. \qquad (2.11)
\end{aligned}$$

Let us now distinguish two cases.

1. case: If $\min_{i\in I_0} R(p_i) > \min_{k\notin I_0} R(p_k)$ it holds that

$$\{\varphi_{(\min_{k\notin I_0} R(p_k))(\min_{j\in I_k} R(p_{kj}))}=0\} \subset \{\varphi_{(\min_{i\in I_0} R(p_i))}=0\},$$

and (2.11) is equal to $P\left(\varphi_{(\min_{i\in I_0} R(p_i))}=1\right)$.
Using now the same arguments as in the proof of Theorem 2.2, it follows that

$$P\left(\varphi_{(\min_{i\in I_0} R(p_i))}=1\right) \ \leq\ P\left(\tilde{\varphi}_{(\min_{i\in I_0} R(p_i))}=1\right) \leq \alpha.$$

2. case: If $\min_{i\in I_0} R(p_i) < \min_{k\notin I_0} R(p_k)$ we have

$$\{\varphi_{(\min_{i\in I_0} R(p_i))} = 0\} \subset \{\varphi_{(\min_{k\notin I_0} R(p_k))(\min_{j\in I_k} R(p_{kj}))} = 0\}.$$

It follows that (2.11) is equal to

$$P\left(\varphi_{(\min_{k\notin I_0} R(p_k))(\min_{j\in I_k} R(p_{kj}))} = 1\right) \leq P\left(\tilde{\varphi}_{(\min_{k\notin I_0} R(p_k))(\min_{j\in I_k} R(p_{kj}))} = 1\right).$$

Using again the same arguments as above we get $P\left(\tilde{\varphi}_{(\min_k R(p_k))(\min_{j\in I_k} R(p_{kj}))} = 1\right) \leq \alpha$. □

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 tested explicitly in the latter two procedures, they are tested only implicitly 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 already mentioned, 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 first one of the BH procedure. However, it has to be taken into account that a smallest p–value means that the intersection hypothesis is most unlikely. The chance that existing differences in the corresponding elementary hypotheses are detected,

is thus very high.
The smallest possible adjusted level of the BHM I procedure is also $\frac{\alpha/3}{KL(KL-1)/2}$. The adjusted significance levels the two smallest p–values of factor A and B have to be compared with are, however, greater using the BHM II procedure than 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 approach.
Another aspect of multiple test procedures besides that of 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.
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 (2.1) assuming normality for the error terms, homogeneity of variances, and a balanced design. For each factor we allow for 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 single tests are performed as F–tests for the intersection hypotheses and as multiple t–tests for

the pairwise comparisons.

The multiple level α is fixed as 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 greater 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, Pohl, Grycko & Steinert, 1995) normally distributed random numbers are generated. The sample size N is fixed as 100 and the grand mean μ as 0 without loss of generality. Regarding the variance, it has to be taken into account that another parameter may be important to judge the power of the different multiple tests, given as the smallest difference of two means and denoted as δ. Allowing for different values of δ gives us the possibility to get an idea of the capability of the various procedures to detect even 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 absolute 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 stepwidth of 0.03. The obtained Monte–Carlo results are not reported for all choices of δ/σ but only for some selected values yielding 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. Here, power I and II are given only for the two 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 as in the two situations described here in detail. For $|I_{A \times B}| \geq 18$,

however, the results 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 $|I_{A\times B}| \leq 12$. The simulation results are summarized in Tables 1–7. It should be mentioned, that we have chosen only some typical examples out of all possible tables for illustrating the results.

3.2 Results

Let us begin with some further general characteristics of the multiple test procedures. The simultaneous power of the BHM II procedure is exactly zero whenever at least the two factors or a factor and the interaction imply partially true as well as false hypotheses (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 always have to be retained. Thus, the simultaneous power is exactly zero. In addition, power II can never reach 1 in this situation. In fact it always remains below 0.5, since for the reasons given above the BHM II procedure can reject all false elementary hypotheses within one factor, but not those within the other one.
The BHM I procedure comes up with the same simultaneous power I and II as the BHM II procedure, if the two factors or a factor and the interaction imply only true hypotheses (Tables 3, 4).
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. Here, we observe a significance level of 3.8% for the BHM I and II procedure and a value of 0.5% for the BH procedure. Thus, the problem already addressed above, that the nominal level can be far below α, clearly occurs. While the first two procedures are slightly conservative, this effect is extreme for the BH procedure.
For the nominal per–comparison error rate we get a value of 0.08% using the BHM I and II procedure and a value of 0.01% using the BH method. Let us also mention, that the nominal multiple level and the nominal per–comparison error rate are also kept with designs different from the one chosen here.
As a first result w.r.t. power it can be noticed that the simultaneous power depends substantially more on the size of the differences in the means than the power II. To achieve a simultaneous power greater than zero, δ/σ has to be at least – with a few

exceptions – 0.3 if all elementary hypotheses concerning the interaction terms are true. Otherwise δ/σ must be even larger than 0.42. Power II, however, is already greater than zero if the differences in the means are 0.03 times the standard deviation.

Summarizing the remaining simulation results, one should first state that there is no simple answer to the question which of the tests introduced in this paper is best in regard to its power. One should be aware of the fact that the properties of the test procedures are data–dependent. But additional information for instance due to subject–matter knowledge may help to reach a decision.

The BHM I and II procedure are both equally good w.r.t. power I and II if there are either no interactions but main effects concerning one factor (Table 3) or if all null hypotheses related to the interaction are false (Table 5). This is because in the first case the adjusted significance level of $\frac{\alpha/3}{3-i+1}$ of the BHM I and II procedure is much greater for $1 \leq i \leq 3$ than the one of the BH procedure with $\frac{\alpha}{42-i+1}$. In the second case it is due to the fact that the adjusted levels of significance of the BH procedure $\frac{\alpha}{42-i+1}$ are greater than $\frac{\alpha/3}{36-i+1}$ for $1 \leq i \leq 34$, but smaller for $i \geq 35$. Thus, the two greatest p–values are to be compared with a value which is smaller using the BH procedure.

Regarding power I, the BHM II procedure is the best test if there are no interactions but main effects concerning both factors (Table 6) – with the exception of the case described above when the power is exactly zero. If power II is considered, the BHM I procedure turns out to be the best for these situations.

The BH procedure is the most powerful test among the three procedures presented here if there are interactions and more than half of the elementary hypotheses related to the interaction are false (Tables 2, 7, 8). A few choices of δ/σ lead to a nearly similar power I of the BH and BHM II procedure. In regard to power II, the BH procedure again outperforms the other procedures (Table 8).

If less than half of the elementary hypotheses related to the interaction are false then the results are close to those obtained when there is no interaction effect. The only exception is the situation that there are no main effects of both factors. Here, the BH procedure is the most powerful test among the three procedures investigated here.

Discussion

From the above simulation results it becomes obvious that no simple and generally valid rule can be given for one of the procedures being the best test. Such a rule does even not exist if it is restricted to particular situations since the performance of the tests heavily depends on the true, but unknown model. Thus, it would of course be helpful to have some further knowledge of the empirical situation when choosing the best test. Typically, such an information is, however, not known in advance. Without going into details, one possible way–out might be to perform special tests in order to reach a decision for the final test procedure. Such an approach can be regarded as an adaptive procedure where the finally selected multiple test depends on the given data. However, when using such an adaptive procedure, it has, however, again to be checked whether the multiple level is still kept and how the simultaneous power or power II behave. To summarize, the results of Section 3 may be understood as rough hints only when being 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%. This is especially true for the original Bonferroni–Holm procedure. 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).

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 most improbable. Here, other techniques based on modelling the correlation structure and thus avoiding any adjustments may be more appropriate (cf. Bretz, 1999), 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.

References

Bretz, F. (1999), *Powerful modifications of Williams' test on trends.* PhD–Thesis, University of Hannover.

Gabriel, K.R. (1969), Simultaneous test procedures – some theory of multiple comparisons. *Annals of Mathematical Statistics* **40**, 224-250.

Hartley, H. O. (1955), Some recent developments in analysis of variance. *Communications on Pure and Applied Mathematics* **8**, 47-72.

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

Holm, S. (1979), A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70.

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. P. Bauer, G. Hommel & E. Sonnemann), 48-66. Springer Verlag, Berlin.

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

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.

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

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

Shaffer, J. P. (1986), Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* **81**, 826-831.

Tables

Table 1: Power I and power II for the situation of main effects for exactly two levels of each factor A and B and 36 true null hypotheses for the interactions, i.e. no interactions.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.39	0.195	0.681	0.000	0.408	0.004	0.203
0.42	0.311	0.791	0.000	0.453	0.010	0.330
0.45	0.516	0.859	0.000	0.477	0.160	0.465
0.48	0.682	0.918	0.000	0.488	0.193	0.582
0.51	0.800	0.958	0.000	0.497	0.330	0.727
0.54	0.937	0.987	0.000	0.498	0.586	0.856
0.57	0.982	0.996	0.000	0.499	0.708	0.905
0.60	0.987	0.997	0.000	0.499	0.850	0.957
0.63	0.996	0.999	0.000	0.500	0.944	0.983

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 null hypotheses for the interactions. The results in brackets are for the same designs but with 5 true 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
0.48	0.041	0.664	0.000	0.577	0.072	0.736
	(0.080)	(0.830)	(0.000)	(0.658)	(0.164)	(0.880)
0.51	0.114	0.763	0.000	0.685	0.172	0.838
	(0.176)	(0.889)	(0.000)	(0.723)	(0.330)	(0.940)
0.54	0.260	0.858	0.000	0.776	0.361	0.917
	(0.382)	(0.935)	(0.000)	(0.783)	(0.611)	(0.973)
0.57	0.353	0.905	0.000	0.830	0.510	0.949
	(0.468)	(0.965)	(0.000)	(0.836)	(0.690)	(0.986)
0.60	0.496	0.946	0.000	0.863	0.694	0.974
	(0.697)	(0.983)	(0.000)	(0.881)	(0.855)	(0.993)
0.63	0.722	0.977	0.000	0.880	0.834	0.989
	(0.832)	(0.992)	(0.000)	(0.907)	(0.921)	(0.996)
0.66	0.801	0.986	0.000	0.891	0.912	0.995
	(0.901)	(0.997)	(0.000)	(0.919)	(0.934)	(0.998)
0.69	0.849	0.993	0.000	0.897	0.958	0.998
	(0.950)	(0.998)	(0.000)	(0.921)	(0.967)	(0.999)
0.72	0.926	0.996	0.000	0.904	0.968	0.999
	(0.962)	(0.999)	(0.000)	(0.922)	(0.991)	(0.999)

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 36 true null hypotheses for the interactions, i.e. 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
0.30	0.052	0.386	0.052	0.386	0.007	0.030
0.33	0.123	0.466	0.123	0.466	0.016	0.062
0.36	0.265	0.591	0.265	0.591	0.030	0.104
0.39	0.382	0.674	0.382	0.674	0.070	0.193
0.42	0.561	0.784	0.561	0.784	0.153	0.291
0.45	0.675	0.846	0.675	0.846	0.274	0.434
0.48	0.822	0.918	0.822	0.918	0.410	0.568
0.51	0.902	0.956	0.902	0.956	0.608	0.729
0.54	0.960	0.983	0.960	0.983	0.742	0.837
0.57	0.981	0.993	0.981	0.993	0.839	0.904
0.60	0.998	0.999	0.998	0.999	0.906	0.951
0.63	0.999	0.999	0.999	0.999	0.964	0.980

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
0.45	0.023	0.491	0.023	0.491	0.036	0.591
	(0.032)	(0.714)	(0.032)	(0.714)	(0.060)	(0.790)
0.48	0.089	0.612	0.089	0.612	0.158	0.715
	(0.093)	(0.797)	(0.093)	(0.797)	(0.184)	(0.860)
0.51	0.116	0.713	0.116	0.713	0.239	0.813
	(0.173)	(0.862)	(0.173)	(0.862)	(0.263)	(0.905)
0.54	0.234	0.795	0.234	0.795	0.347	0.864
	(0.352)	(0.919)	(0.352)	(0.919)	(0.502)	(0.951)
0.57	0.360	0.881	0.360	0.881	0.524	0.936
	(0.608)	(0.962)	(0.608)	(0.962)	(0.687)	(0.975)
0.60	0.475	0.917	0.475	0.917	0.593	0.955
	(0.658)	(0.978)	(0.658)	(0.978)	(0.776)	(0.987)
0.63	0.626	0.961	0.626	0.961	0.763	0.978
	(0.856)	(0.991)	(0.856)	(0.991)	(0.870)	(0.994)
0.66	0.788	0.985	0.788	0.985	0.873	0.992
	(0.898)	(0.995)	(0.898)	(0.995)	(0.917)	(0.997)
0.69	0.870	0.992	0.870	0.992	0.925	0.996
	(0.939)	(0.998)	(0.939)	(0.998)	(0.959)	(0.999)
0.72	0.918	0.996	0.918	0.996	0.964	0.998
	(0.976)	(0.999)	(0.976)	(0.999)	(0.990)	(0.999)
0.75	0.962	0.998	0.962	0.998	0.979	0.999
	(0.991)	(0.999)	(0.991)	(0.999)	(1.000)	(1.000)

Table 5: Power I and power II for the situation of no main effects of the factors A and B and no true null hypotheses for the interactions, i.e. all possible interactions present.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.42	0.045	0.892	0.045	0.892	0.043	0.902
0.45	0.217	0.927	0.217	0.927	0.178	0.932
0.48	0.486	0.964	0.486	0.964	0.406	0.963
0.51	0.623	0.975	0.623	0.975	0.524	0.971
0.54	0.796	0.988	0.796	0.988	0.699	0.984
0.57	0.936	0.996	0.936	0.996	0.878	0.994
0.60	0.953	0.998	0.953	0.998	0.923	0.997
0.63	0.972	0.999	0.972	0.999	0.959	0.998
0.66	0.996	1.000	0.996	1.000	0.977	0.999

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
0.36	0.083	0.752	0.138	0.577	0.000	0.407
0.39	0.168	0.878	0.265	0.701	0.000	0.469
0.42	0.398	0.899	0.544	0.852	0.000	0.535
0.45	0.636	0.950	0.759	0.924	0.030	0.621
0.48	0.822	0.975	0.893	0.963	0.080	0.728
0.51	0.939	0.994	0.972	0.992	0.259	0.877
0.54	0.989	0.998	0.995	0.997	0.454	0.889
0.57	0.996	0.999	0.997	0.998	0.742	0.954
0.60	0.998	0.999	0.999	0.999	0.849	0.973

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 of factors A and B.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.48	0.065	0.642	0.078	0.651	0.083	0.722
	(0.077)	(0.825)	(0.005)	(0.718)	(0.152)	(0.878)
0.51	0.168	0.759	0.198	0.780	0.207	0.842
	(0.196)	(0.881)	(0.037)	(0.786)	(0.314)	(0.932)
0.54	0.207	0.826	0.279	0.842	0.344	0.897
	(0.340)	(0.934)	(0.092)	(0.849)	(0.484)	(0.961)
0.57	0.319	0.880	0.379	0.907	0.487	0.935
	(0.535)	(0.968)	(0.262)	(0.908)	(0.668)	(0.981)
0.60	0.494	0.930	0.550	0.948	0.603	0.961
	(0.719)	(0.985)	(0.568)	(0.956)	(0.831)	(0.992)
0.63	0.628	0.965	0.701	0.972	0.788	0.980
	(0.788)	(0.990)	(0.783)	(0.984)	(0.879)	(0.995)
0.66	0.769	0.982	0.831	0.987	0.874	0.993
	(0.918)	(0.996)	(0.915)	(0.995)	(0.927)	(0.998)
0.69	0.871	0.994	0.909	0.996	0.959	0.998
	(0.953)	(0.998)	(0.948)	(0.997)	(0.978)	(0.999)
0.72	0.923	0.996	0.950	0.998	0.983	0.999
	(0.979)	(0.999)	(0.972)	(0.999)	(0.992)	(0.999)

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 of factors A and B.

	BHM I		BHM II		BH	
δ/σ	Power I	Power II	Power I	Power II	Power I	Power II
0.42	0.026	0.889	0.035	0.805	0.074	0.907
0.45	0.198	0.933	0.221	0.864	0.241	0.943
0.48	0.446	0.960	0.458	0.913	0.483	0.968
0.51	0.642	0.979	0.664	0.950	0.672	0.983
0.54	0.786	0.993	0.817	0.980	0.829	0.995
0.57	0.893	0.995	0.907	0.988	0.903	0.996
0.60	0.954	0.997	0.964	0.995	0.959	0.997
0.63	0.971	0.999	0.986	0.999	0.976	0.999