ELSEVIER

Computer Methods and Programs in Biomedicine 43 (1994) 213—216

computer methods and programs in biomedicine

A rapid algorithm and a computer program for multiple test procedures using logical structures of hypotheses

Gerhard Hommel*, Gudrun Bernhard

Institut für Medizinische Statistik und Dokumentation, Universität Mainz, Langenbeckstrasse 1, D-55101 Mainz, Germany

Abstract

It is demonstrated how improvements of general multiple test procedures can be obtained using information about the logical structures among the hypotheses. Based on a procedure of Bergmann and Hommel (B. Bergmann and G. Hommel, Improvements of general multiple test procedures for redundant systems of hypotheses, in Multiple Hypothesenprüfung — Multiple Hypotheses Testing, Eds. P. Bauer, G. Hommel and E. Sonnemann, pp. 100–115 (Springer-Verlag, Berlin, 1988)), a computer program was written by Bernhard (G. Bernhard, Computerunterstützte Durchführung von multiplen Testprozeduren — Algorithmen und Powervergleich, Doctoral thesis (Mainz, 1992)) using this information. It is applicable for a general class of systems of hypotheses which can be expressed in a linear way. By means of a simulation study it is shown that the proposed procedure is often substantially more powerful than other usual multiple test procedures.

Key words: Multiple test procedure; Bonferroni inequality; Logical dependence among hypotheses; Exhaustive index set; Linear hypothesis

1. Introduction

The problem of multiple hypotheses testing arises, if, within the same study, $n > 1$ statistical (null) hypotheses $H_1, \ldots, H_n$ have to be tested. Then it is often necessary to control the multiple level α, i.e. the probability of committing a type I error has to be less than or equal to a predetermined bound α. A multiple test procedure is said to be general if it is independent of specific univariate or multivariate test statistics $(T_1, \ldots, T_n)$, and only based on the P values $P_1, \ldots, P_n$ for the n individual tests. A simple well-known general multiple test procedure controlling the multiple level α is the Bonferroni procedure which rejects a hypothesis H_i whenever $P_i \leq \alpha/n$. An improvement of this procedure is the sequentially rejective Bonferroni procedure by Holm [1]: If $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(n)}$ are the ordered P values and $H_{(1)}, \ldots, H_{(n)}$ the corresponding hypotheses, then $H_{(i)}$ is rejected whenever $P_{(j)} \leq \alpha/(n-j+1)$ for $j = 1, \ldots, i$. Both procedures need no information about the correlation structure of the test statistics; therefore

* Corresponding author.

SSDI 0169-2607(93)1491-W

improvements leading to more powerful procedures are possible when this correlation structure (stochastic dependence) is used explicitly. Another way of improvement is to use logical dependences among the hypotheses if not all combinations of hypotheses are free [1] (Example: For the three hypotheses H_1: $\mu_1 = \mu_2$, H_2: $\mu_1 = \mu_3$, H_3: $\mu_2 = \mu_3$, it cannot occur that only H_1 and H_2 are true). An improvement of Holm's procedure using information about logical dependences is the 'modified sequentially rejective Bonferroni' (MSRB) procedure by Shaffer [2]: Instead of Holm's bounds $\alpha/(n-j+1)$, $j = 1, \ldots, n$, it uses stepwise bounds α/t_j where $t_j = \min\{k\colon k \leq n-j+1$, exactly k hypotheses can be true$\}$.

The Bonferroni (B), Holm (H), and Shaffer (S) procedures are in ascending sequence strictly more powerful (i.e. if a hypothesis H_i is rejected by (B) then it is also rejected by (H), and if it is rejected by (H) then it is rejected by (S)), but they are, in the same sequence, also more complicated to perform.

2. An improved procedure

A further improvement called 'Procedure 4' (P4) was given by Bergmann and Hommel [3], although not as a fully elaborated algorithm. It is still based on the Bonferroni inequality and works as follows: An index set $I \subseteq \{1, \ldots, n\}$ is called exhaustive if and only if it is possible that all null hypotheses H_i, $i \in I$, are true, and all hypotheses H_j, $j \notin I$, are not. Then one determines an 'acceptance set'

$$A = \cup\{I\colon I \text{ exhaustive}, \min\{P_i\colon i \in I\} > \alpha/|I|\}$$

and rejects all H_j with $j \in A$. This procedure is strictly more powerful than the (B), (H), and (S) procedures and still controls the multiple level α, but it is also essentially more complicated than the other three procedures.

A substantial problem is to get all exhaustive index sets without passing through the whole power set of $\{1, \ldots, n\}$. Therefore a rapid algorithm has been described by Bernhard [4] (see also Hommel and Bernhard [5]): Denote SE(I) = the smallest exhaustive index set containing I. Then the algorithm is founded upon the lexicographical order for the power set of $\{1, \ldots, n\}$ and checks in each step whether the calculation of SE(I) (which requires intensive computing time) is really necessary to obtain a new exhaustive index set.

Example: Consider all pairwise comparisons of k parameters, then $n = \binom{k}{2}$ hypotheses exist, and one has therefore 2^n index sets I. In Table 1 it is shown that, especially for larger k, only a small part of all index sets has to be checked. (The given numbers of checks ($\bar{n}_c$) are average values, since they depend, although only slightly, on the numbering of the hypotheses.)

If the hypotheses can be represented as linear combinations of some parameters, SE(I) can be determined considering the ranks of matrices corresponding to the combination of certain hypotheses ([3,4]). For this case, the algorithm was written as a part of a FORTRAN program. When using the program, one needs as an input only the multiple level α, the P values $P_1, \ldots, P_n$, and the n matrices $A_1, \ldots, A_n$ corresponding to the hypotheses $H_1, \ldots, H_n$; the output contains the decisions on each hypothesis for different procedures, in particular (P4). As an example of how the matrices may be chosen, assume that, for any three parameters μ_1, μ_2, μ_3, one wants to test the $n = 3$ hypotheses H_1: $\mu_1 = \mu_2$, H_2: $\mu_2 = \mu_3 = 0$, H_3: $\mu_1 = \mu_2 = \mu_3$. Then one can choose the matrices A_i as the following ($i \times 3$)-matrices:

Table 1
All pairwise comparisons of k treatments:

k	$n = \binom{k}{2}$	2^n	n_e	$\bar{n}_c$
4	6	64	14	31
5	10	1024	51	161
6	15	32768	202	884
7	21	2097152	876	5036
8	28	2.7×10^8	4139	30257
9	36	6.9×10^{10}	21146	~ 100000

The table gives the number of exhaustive index sets (n_e) and average number of checks ($\bar{n}_c$) needed to determine whether an index set is exhaustive.

$$A_1 = (1 \quad -1 \quad 0),$$

$$A_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

(H_3 can also be represented by omitting any one of the three lines of A_3).

In case of all pairwise comparisons of k parameters, one has a standard set of $(1 \times k)$-matrices for which a direct option is provided in the program.

3. Simulation results

In case of all pairwise comparisons of k parameters ($k = 3, 4, 5, 6$), a simulation study was performed in order to compare the powers of several procedures with respect to (P4). We considered two power concepts (see Maurer and Mellein [6]):

- Total power = the probability of rejecting all false null hypotheses.
- Mean power = the proportion of all (correctly) rejected null hypotheses among all false null hypotheses.

The sample size per group was 8 throughout; the data were generated (using the SAS function RANNOR) as normally distributed with expectations μ_j, $j = 1, \ldots, k$, and variances 1. We used two-sided t-tests, a multiple level $\alpha = 0.05$ was chosen, and 10000 simulations were performed for each situation.

First of all, we were interested in the amount of gain in power, if (P4) is compared with (H) and (S). In Tables 2 and 3 a comparison of the total and the mean power for different parameter constellations is given. We used (H) as a 'standard', i.e. μ was chosen so that (H) achieved a power of 50%. Typically, (S) showed only slight gain in power with respect to (H), whereas a substantial improvement (up to 19% for the total power and up to 8% for the mean power) was often obtained by (P4). (For $k = 3$, the procedures (S) and (P4) are identical).

Table 2
Total power (percent) of (H), (S), and (P4) for several parameter constellations

Expectations	(H)	(S)	(P4)
$(0,0,\mu)$	50	60	60
$(0,0,0,\mu)$	50	56	63
$(0,0,0,0,\mu)$	50	53	62
$(0,0,0,\mu,\mu,\mu)$	50	50	69
$(0,\mu,2\mu,3\mu)$	50	50	52
$(0,\mu,2\mu,3\mu,4\mu)$	50	50	52
$(0,\mu,2\mu,3\mu,4\mu,5\mu)$	50	50	51

Moreover, (P4) was also compared with the SAS procedure 'anova' using the option 'regwq' [7]. For this comparison, we used two different versions of (P4). The first one ((P4) = (P4/B)) is based on the Bonferroni inequality, as described in Section 2; the second one (P4/SR) is an improvement using Studentized range tests, for which a very similar algorithm is applicable [4]. Here we found that (P4/SR) was sometimes slightly less powerful, but in the majority of the situations (often considerably) more powerful than regwq. Even the simpler and more general procedure (P4/B) was in some cases (for the total power) superior to regwq. In Tables 4 and 5, the results for some parameter constellations are given, using regwq as the 'standard'.

4. Discussion

The proposed procedure (P4) is often by far more powerful than the usual general multiple

Table 3
Mean power (percent) of (H), (S), and (P4) for several parameter constellations

Expectations	(H)	(S)	(P4)
$(0,0,\mu)$	50	53	53
$(0,0,0,\mu)$	50	54	56
$(0,0,0,0,\mu)$	50	53	57
$(0,0,0,\mu,\mu,\mu)$	50	52	58
$(0,\mu,2\mu,3\mu)$	50	52	56
$(0,\mu,2\mu,3\mu,4\mu)$	50	51	55
$(0,\mu,2\mu,3\mu,4\mu,5\mu)$	50	51	55

Table 4
Total power (percent) of regwq, (P4/SR), and (P4/B) for several parameter constellations

Expectations	regwq	(P4/SR)	(P4/B)
$(0,0,0,0,0,\mu)$	50	40	25
$(0,0,\mu,\mu)$	50	60	53
$(0,0,0,\mu,\mu)$	50	57	38
$(0,\mu,2\mu,3\mu)$	50	67	61
$(0,\mu,2\mu,3\mu,4\mu,5\mu)$	50	76	64

test procedures, as the Bonferroni and Holm [1] procedure. By means of the special algorithm a substantial reduction of computational time is possible; a FORTRAN program is available which is very easy to handle.

It should be emphasized that (P4) is not only applicable for the comparison of means under normal assumptions, as the simulation study might suggest, but also for comparisons of parameters of any type (e.g. variances, proportions, distribution functions) under arbitrary distributional assumptions. Moreover, it is not restricted to pairwise comparisons, but can consider arbitrary linear hypotheses, as contrast hypotheses or tests in factorial designs. An important kind of application of this type arises if the system of hypotheses is completely or partially nested, as it is for model search problems. For this case a special test strategy ('arrow test strategy' [8]) seems to be very powerful in some situations.

Table 5
Mean power (percent) of regwq, (P4/SR), and (P4/B) for several parameter constellations

Expectations	regwq	(P4/SR)	(P4/B)
$(0,0,0,0,0,\mu)$	50	49	36
$(0,0,\mu,\mu)$	50	53	44
$(0,0,0,\mu,\mu)$	50	52	39
$(0,\mu,2\mu,3\mu)$	50	52	48
$(0,\mu,2\mu,3\mu,4\mu,5\mu)$	50	52	45

References

[1] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70.
[2] J.P. Shaffer, Modified sequentially rejective multiple test procedures, J. Am. Stat. Assoc. 81 (1986) 626–633.
[3] B. Bergmann and G. Hommel, Improvements of general multiple test procedures for redundant systems of hypotheses, in Multiple Hypothesenprüfung — Multiple Hypotheses Testing, Eds. P. Bauer, G. Hommel and E. Sonnemann, pp. 100–115 (Springer-Verlag, Berlin, 1988).
[4] G. Bernhard, Computerunterstützte Durchführung von multiplen Testprozeduren — Algorithmen und Powervergleich, Doctoral thesis (Mainz, 1992).
[5] G. Hommel and G. Bernhard, Multiple hypotheses testing, in: Computational Aspects of Model Choice, Ed. J. Antoch, pp. 211–235 (Physica, Heidelberg, 1992).
[6] W. Maurer and B. Mellein, On new multiple tests based on independent p-values and the assessment of their power, in Multiple Hypothesenprüfung — Multiple Hypotheses Testing, Eds. P. Bauer, G. Hommel and E. Sonnemann, pp. 48–66 (Springer-Verlag, Berlin, 1988).
[7] SAS/STAT™, Guide for Personal Computers, Version 6 (SAS Institute, Cary NC, 1987).
[8] G. Hommel and G. Bernhard, A multiple test procedure for nested systems of hypotheses, in Computational Statistics, Eds. P. Dirschedl and R. Ostermann (Physica, Heidelberg, in press).