Biometrika (1988), **75**, 2, *pp.* 383-6
Printed in Great Britain

A stagewise rejective multiple test procedure based on a modified Bonferroni test

BY G. HOMMEL

Institut für Medizinische Statistik und Dokumentation, University of Mainz, D-6500 Mainz, Federal Republic of Germany

SUMMARY

Simes (1986) has proposed a modified Bonferroni procedure for the test of an overall hypothesis which is the combination of n individual hypotheses. In contrast to the classical Bonferroni procedure, it is not obvious how statements about individual hypotheses are to be made for this procedure. In the present paper a multiple test procedure allowing statements on individual hypotheses is proposed. It is based on the principle of closed test procedures (Marcus, Peritz & Gabriel, 1976) and controls the multiple level α.

Some key words: Closed test procedure; Control of multiple level; Modified Bonferroni procedure; Multiple test procedure; Quasi-coherence.

1. INTRODUCTION

When n hypotheses $H_1, \ldots, H_n$ with associated test statistics $T_1, \ldots, T_n$ are to be tested, one can make use of the corresponding p-values $P_1, \ldots, P_n$. A first step to aim for an overall statement can be based on a test of the overall hypothesis $H_0 = \cap\{H_i: i = 1, \ldots, n\}$. Application of the Bonferroni inequality leads to a very simple level α test of H_0: reject H_0, if $P_{(1)} \leqslant \alpha/n$, where $P_{(1)}$ is the smallest one of the p-values.

A disadvantage of this procedure is that it may be very conservative, in particular, if the test statistics are highly correlated; moreover, it is often inappropriate to use only the smallest p-value. Another level α test which might avoid this disadvantage is based on Rüger's (1978) inequality: reject H_0, if $P_{(k)} \leqslant k\alpha/n$, where $P_{(k)}$ is the kth smallest of the p-values; here k $(2 \leqslant k \leqslant n)$ has to be determined before performing the n tests.

If one wishes to avoid the problem of choosing k in advance, one can combine the Bonferroni test and all $(n-1)$ possible Rüger tests and obtain the following level α test of H_0 (Hommel, 1983): reject H_0, if $P_{(k)} \leqslant k\alpha/(nC_n)$ for at least one k $(1 \leqslant k \leqslant n)$, where $C_n = 1 + \frac{1}{2} + \ldots + 1/n$. A very similar test of H_0 which is less conservative because of omitting the constant C_n has been proposed by Simes (1986): reject H_0, if $P_{(k)} \leqslant k\alpha/n$ for at least one k $(1 \leqslant k \leqslant n)$.

Since the inequalities of Bonferroni, Rüger and Hommel are all strict, there will be constellations of dependencies among the test statistics where the test of H_0 has exactly the level α; it seems, however, that these situations are rather pathological. In practical applications, the corresponding tests of H_0 can be expected to be conservative. As Simes pointed out, his procedure does not always lead to a level α test of H_0; nevertheless, he suggested by a simulation study that the level of his procedure is less than or equal to α for a large family of multivariate distributions of $(T_1, \ldots, T_n)$, and he proved that the level is exactly equal to α if the test statistics are independent. Therefore, in such cases application of Simes's procedure is recommended since it is strictly more powerful than each of the other three procedures.

When, by any of these procedures, H_0 has been rejected, the question remains which of the individual hypotheses H_i $(i = 1, \ldots, n)$ should be rejected. An answer is easy for the Bonferroni procedure, where one can reject all H_i with $P_i \leqslant \alpha/n$. For the other procedures, however, it is not quite clear which of the H_i should be rejected. Simes has proposed for his procedure to reject

in an exploratory sense the individual hypotheses $H_{(1)}, \ldots, H_{(j)}$, where $j = \max\{k: P_{(k)} \leq k\alpha/n\}$, $H_{(i)}$ being the hypotheses corresponding to $P_{(i)}$ for $i = 1, \ldots, j$. However, this procedure is not always satisfactory. Suppose that the test statistics T_i $(i = 1, \ldots, n)$ are independent, that m individual hypotheses are true, and that the other $(n-m)$ hypotheses H_i are false to such an extent that $\text{pr}(P_i \leq \alpha/n)$ is nearly equal to 1. Then the probability of rejecting at least one of the m true H_i is nearly equal to $1-\{1-(n-m+1)\alpha/n\}^m$. If, for example, $\alpha = 0{\cdot}05$, $n = 100$ and $m = 50$, then the probability of committing a type I error is 0·725, and it tends to 1 for $m = \frac{1}{2}n$ and $n \to \infty$.

In the following, multiple test procedures are proposed which are based on the described tests of the overall hypothesis and keep the probability of committing a type I error less than or equal to α.

2. Closed test procedures

We apply the following modification of the principle of 'closed test procedures' (Marcus et al., 1976; Sonnemann, 1982).

THEOREM (*Hommel*, 1986). *Let there be given, for $n \geq 1$, n individual hypotheses $H_1, \ldots, H_n$, and define $H_I = \cap\{H_i: i \in I\}$ for all $I \in K$, where K is the set of all nonempty subsets of $\{1, \ldots, n\}$. Assume that there exists for each $I \in K$ a level α test based on a test statistic T_I. Reject H_I if it is rejected by T_I and if all H_J with $J \supseteq I$, $J \in K$, are rejected by T_J, too. Then this multiple test procedure controls the multiple level α; that is the probability of committing any type I error when testing all H_I, $I \in K$, is at most α irrespective of which of the H_I are true.*

Since every H_I is the intersection of the individual hypotheses H_i, $i \in I$, it can be interpreted as a possibly 'small' overall hypothesis. Hence a level α test of H_I which is only based on the p-values P_i, $i \in I$, can be found as described in § 1. If one chooses Bonferroni tests for testing H_I with the decision rule 'reject H_I if there is at least one P_i, $i \in I$, with $P_i \leq \alpha/|I|$', then the application of the Theorem leads directly to Holm's (1979) sequentially rejective procedure which is an improvement of the classical multiple Bonferroni procedure. Test strategies arising when the theorem is applied to overall tests based on the inequalities of Rüger or of Hommel, are described by Hommel (1986). When all overall tests are tests as proposed by Simes, the arising multiple test procedure can be presented by the flow chart of Hommel (1986, Fig. 3) with the choice $\delta_{kj} = k\alpha/j$. The decisions for the individual hypotheses can be performed in the following simpler way: compute $j = \max\{i \in \{1, \ldots, n\}: P_{(n-i+k)} > k\alpha/i \text{ for } k = 1, \ldots, i\}$. If the maximum does not exist, reject all H_i $(i = 1, \ldots, n)$, otherwise reject all H_i with $P_i \leq \alpha/j$. It follows that this procedure controls the multiple level α provided each of Simes's tests for H_I is a level α test. In particular, the multiple level α is kept if the n tests are independent.

3. Use of logical relations among the hypotheses

Shaffer (1986) gives the following improvement of Holm's (1979) general procedure. Let, for a given system of hypotheses $H_1, \ldots, H_n$, S be the set of all $j \in \{1, \ldots, n\}$ such that it can occur that exactly j of the n hypotheses are true and the remaining $(n-j)$ are false. Define $t_i = \max\{j \in S: j \leq n-i+1\}$ for $i = 1, \ldots, n$. Then the stagewise rejective procedure using the stepwise significance bounds α/t_i instead of Holm's bounds $\alpha/(n-i+1)$ controls the multiple level α.

As an example, consider all $n = 10$ pairwise comparisons of 5 distributions. Then $S = \{1, 2, 3, 4, 6, 10\}$ (Shaffer, 1986, Table 2), and $t_1 = 10$, $t_2 = t_3 = t_4 = t_5 = 6$, $t_6 = t_7 = 4$, $t_8 = 3$, $t_9 = 2$, $t_{10} = 1$.

An analogous improvement can be found for the multiple Simes procedure provided each of Simes's tests for H_I is a level α test. For this improved procedure, one has to compute $j = \max\{i \in S: P_{(n-i+k)} > k\alpha/i \text{ for } k = 1, \ldots, i\}$; then the decisions for the individual hypotheses can be taken as described for the general procedure.

4. Numerical example

Assume that in a study with $n = 10$ statistical tests the p-values are ordered as following: $P_1 = 0{\cdot}0021$, $P_2 = 0{\cdot}0074$, $P_3 = 0{\cdot}0093$, $P_4 = 0{\cdot}0106$, $P_5 = 0{\cdot}0121$, $P_6 = 0{\cdot}0218$, $P_7 = 0{\cdot}0238$, $P_8 = 0{\cdot}0352$, $P_9 = 0{\cdot}0466$, $P_{10} = 0{\cdot}0605$. Let $\alpha = 0{\cdot}05$ be chosen as the multiple level.

In this example, because of $P_1 \leq \alpha/10$ and $P_2 > \alpha/9$, Holm's procedure rejects H_1 as the only individual hypothesis. If the general multiple Simes procedure is applied, one obtains $j = 5$, and therefore all H_i with $P_i \leq \alpha/5$; that is H_1, H_2, H_3 are rejected.

If it is known that the 10 tests are all pairwise comparison tests of 5 distributions, because of $5 \notin S$ one obtains $j = 4$. Therefore all H_i with $P_i \leq \alpha/4$, that is $H_1, \ldots, H_5$, are rejected by the improved procedure.

In order to ensure that the multiple level α for this procedure is kept, it is sufficient that Simes's test for each H_I is a level α test. A simulation study was performed for the case that each pairwise comparison test is based on a test statistic $T_i = |X_l - X_m|/2^{\frac{1}{2}}$, where X_l, X_m $(1 \leq l < m \leq 5)$ are independently $N(0, 1)$-distributed, which is fulfilled asymptotically for many types of pairwise comparisons. Simulations were carried out at an IBM/AT personal computer using the SAS function RANNOR for generating random normal variables. From the results in Table 1, the level $\alpha = 0{\cdot}05$ is exceeded for no type of hypothesis, and the type I error rates of Simes's tests are slightly higher than those of the Bonferroni tests.

Table 1. *Type I error rates* for Simes's, S, and Bonferroni, B, overall tests for all types of intersection hypotheses H_I, for* 10 *pairwise comparisons of* 5 *distribution parameters* $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$; *$|I|$ = number of individual hypotheses implied by H_I*

'Typical' hypothesis	$\|I\|$	S	B	'Typical' hypothesis	$\|I\|$	S	B
$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	10	0·044	0·040	$\mu_1 = \mu_2 = \mu_3$	3	0·046	0·044
$\mu_1 = \mu_2 = \mu_3 = \mu_4$	6	0·045	0·041	$\mu_1 = \mu_2$ and $\mu_3 = \mu_4$	2	0·05†	0·049375†
$\mu_1 = \mu_2 = \mu_3$ and $\mu_4 = \mu_5$	4	0·047	0·045	$\mu_1 = \mu_2$	1	0·05†	0·05†

* Based on 20000 simulations each; estimated standard error $\leq 0{\cdot}0015$. † Exact error rates.

5. Discussion

The proposed multiple test procedure is strictly not less powerful than Holm's procedure as well as all other procedures mentioned by Hommel (1986); in many cases, it seems to be considerably more powerful. The computations needed for testing the individual hypotheses are very simple. If decisions for all H_I, $I \in K$, are to be taken, it is recommended to use a computer program based on Hommel (1986, Fig. 3). This can be performed also for a large n, since the computational time is proportional to n^2.

An important logical property of multiple test procedures is coherence (Gabriel, 1969); i.e. if a hypothesis is retained, all its implications also have to be retained. As Hommel (1986) pointed out, general multiple test procedures, as the Bonferroni or Holm's procedure, need not be coherent, but they should be quasi-coherent; i.e. if $H_I = \cap\{H_i: i \in I\}$ is retained, all H_J with $J \subseteq I$ are retained. Since the theorem is applied, the proposed procedure is quasi-coherent.

In § 3 it is shown how logical dependencies in a given system of hypotheses can lead to an improvement of the procedure. Another question is how one can make use of stochastical dependencies between the test statistics. A solution of this problem seems to be more difficult; on the other hand, the overall tests according to Simes are much more flexible against different structures of stochastical dependence than, for example, Bonferroni overall tests.

Acknowledgement

The author would like to thank Peter Bauer for helpful comments and suggestions.

REFERENCES

GABRIEL, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224–50.

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

HOMMEL, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biom. J.* **25**, 423–30.

HOMMEL, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika* **33**, 321–36.

MARCUS, R., PERITZ, E. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–60.

RÜGER, B. (1978). Das maximale Signifikanzniveau des Tests "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika* **25**, 171–8.

SHAFFER, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Am. Statist. Assoc.* **81**, 826–31.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–4.

SONNEMANN, E. (1982). Allgemeine Lösungen multipler Testprobleme. *EDV in Med. u. Biol.* **13**, 120–8.

[*Received July* 1987. *Revised October* 1987]