Scand J Statist 6: 65–70, 1979

A Simple Sequentially Rejective Multiple Test Procedure

STURE HOLM

Chalmers University of Technology, Göteborg

Received December 1977, revised September 1978

ABSTRACT. This paper presents a simple and widely applicable multiple test procedure of the sequentially rejective type, i.e. hypotheses are rejected one at a time until no further rejections can be done. It is shown that the test has a prescribed level of significance protection against error of the first kind for any combination of true hypotheses. The power properties of the test and a number of possible applications are also discussed.

Key words: multiple test, simultaneous test

1. Introduction

The statistical problems arising in applications often involve a number of detail problems, i.e. there are often a number of interesting parameters to be estimated and/or a number of interesting hypotheses to be tested. In some cases these detail problems may be treated separately without any connection to each other. But in most cases the detail problems are connected to each other and the totality of solutions to the detail problems are used to get a general picture. In this latter case the statistician is faced with a multiple statistical inference problem, where he has to take into consideration that the different detail problems should be treated simultaneously.

Multiple statistical inference has been a vital research area within statistical inference theory the past 50 years, and methods have been proposed for several situations of practical interest. A good presentation of the earlier main results is given by Miller (1966). The multiple statistical inference methods are separated into two main types, multiple confidence interval methods and multiple test methods.

For multiple test procedures there has been suggested several types of properties, which the tests should have in order to give satisfactory protection against wrong decisions. Some of those are based on decision theoretic conceptions, while others are based on probabilities of making wrong decisions. In this paper we will study multiple test procedures and we will use the most common type of protection against error of the first kind by requiring the tests to have a small probability of rejecting any true hypotheses. The methodological motivation and exact definition is the following.

Let the (detail) hypotheses in a multiple test problem be denoted by $H_1, H_2, \ldots, H_n$ and the alternatives to those by $K_1, K_2, \ldots, K_n$. A (non-randomized) multiple test procedure is a rule assigning to each outcome a set of rejected hypotheses (which might be empty). This means that there are also n critical regions $C_1, C_2, \ldots, C_n$ consisting of those outcomes for which the corresponding hypotheses are rejected.

In a test of a single null hypothesis H_1 against an alternative K_1 the size of the test is defined as the supremum of the probability of the critical region C_1 when the hypothesis H_1 is true. This probability of error of the first kind is always kept at (or below) a small predetermined level α. The philosophical reason for this is that when we have made a 'discovery' by rejecting the null hypothesis we can quite safely claim that the null hypothesis is not true, because if it was true, we should have accepted it with a probability of at least $1-\alpha$. This also implies that we do not make any 'discovery' by accepting the null hypotheses, because we do not have such a protection against errors of the second kind.

In a multiple test of a number of hypotheses $H_1, H_2, \ldots, H_n$ there are a lot of possible combinations of null hypotheses. If we want to make our 'discoveries' in form of rejected null hypotheses to be safely claimed, we must keep the probability of rejecting any true null hypotheses small, how many and which the true hypotheses may be. Thus we are led to the following definition

Definition. A multiple test procedure with critical regions $C_1, C_2, \ldots, C_n$ for testing hypotheses $H_1, H_2, \ldots, H_n$ is said to have a *multiple level of significance* α (for free combinations) if for any non-empty index set $I \subseteq \{1, 2, 3, \ldots, n\}$ the supremum of the probability $P(\bigcup_{i \in I} C_i)$ when H_i are true for all $i \in I$ is smaller than or equal to α.

The words 'for free combinations' are put into the definition in order to underline that all subsets of

null hypotheses could appear as the set of true hypotheses. There might be situations in which all subsets are not allowed for some reason, for instance situations where the truth of two hypotheses implies the truth or falseness of a third hypothesis. It is to be observed that a multiple level of significance α for some restricted combinations imposes fewer conditions on the test procedure than a multiple level of significance α for free combinations, i.e. a test procedure with multiple level of significance α for free combinations has a multiple level of significance α for any type of restricted combinations.

In our setting the basic hypotheses $H_1, H_2, ..., H_n$ are minimal in the sense of Gabriel (1969). This means that if $\omega_1, \omega_2, ..., \omega_n$ are the parameter sets where the hypotheses $H_1, H_2, ..., H_n$ are true than the only (secondary) hypotheses to be tested are the hypotheses that the parameter belongs to intersections $\bigcap_{i\in I} \omega_i$ of sets ω_i for different index sets $I \subseteq \{1, 2, ..., n\}$.

We will exclusively discuss a type of multiple test procedures, which may be called sequentially rejective because basic hypotheses are rejected one at a time according to certain rules. Thus we do not make separate tests of all the (secondary) hypotheses that the parameter belongs to intersections $\bigcap_{i\in I} \omega_i$ for different $I \subseteq \{1, 2, ..., n\}$. We always consider such (secondary) hypotheses to be rejected as soon as any of the included basic hypotheses are rejected.

A test procedure is called coherent if it prevents the contradiction of rejecting a hypothesis without also rejecting all other hypotheses implying it. It is called consonant if it avoids dissonances consisting in rejecting a hypothesis and not rejecting any other hypotheses implied by it. (See Gabriel, 1969, pp. 229 and 231.) The sequentially rejective tests are coherent and consonant by their very definition.

In many applications there are logical implications among the basic hypotheses i.e. some combinations of falseness of different basic hypotheses are not allowed because there are no possible parameter points corresponding to those combinations. Then we do not want the multiple test procedure to end up with a statement that the parameter belongs to such an empty set. This requirement has to be studied separately for each kind of logical implication. We will consider only the type of logical implications arising when we have two-sided alternatives for some parameters, and want to make one-sided statements.

The sequentially rejective multiple tests are not completely new. Tests of the same type are discussed by Naik (1975, p. 522), and the consonant closed procedures discussed by Marcus et al. (1976, p. 656) are equivalent to sequentially rejective tests. Marcus et al. (1976) give one particular example of such a test in an analysis of variance situation, and indicate that others can be constructed. But they do not seem to have thought of the simple and general procedure we present in the next section, because that procedure can easily be used to make one-sided rejections, which they have posed as a difficult problem. Our test is based on the simple Boole inequality and can be applied to any parametric or non-parametric model, but yet it has good power properties. It will be shown by examples that it may have considerably higher power than classical multiple test procedures. It also has surprisingly small loss of power compared to the special sequentially rejective tests (or equivalent consonant closed tests) that can be constructed in different parametric models, for instance analysis of variance models.

2. A simple sequentially rejective test

In this section we will present a simple sequentially rejective test, which is based on the Boole inequality. The use of the Boole inequality within multiple inference theory is usually called the Bonferroni technique, and for this reason we will call our test the sequentially rejective Bonferroni test.

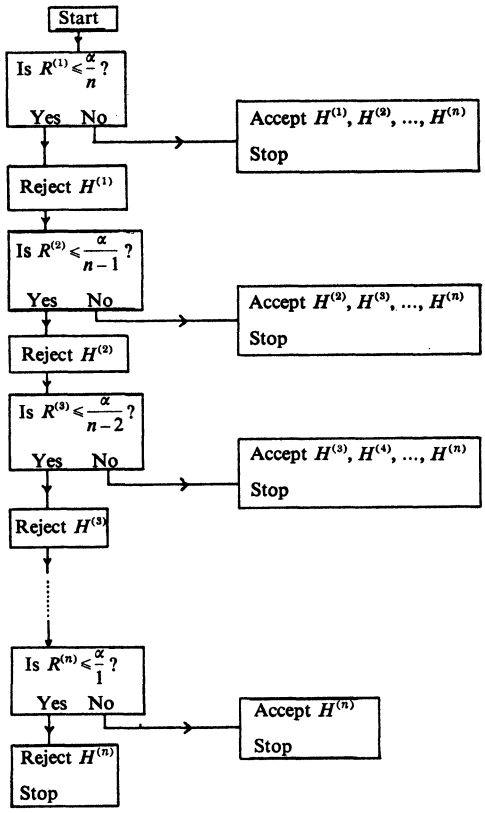
When the n hypotheses $H_1, H_2, ..., H_n$ are tested separately by using tests with the level α/n it follows immediately from the Boole inequality that the probability of rejecting any true hypotheses is smaller then or equal to α. This constitutes then a multiple test procedure with the multiple level of significance α for free combination, the classical Bonferroni multiple test procedure.

The separate tests in the classical Bonferroni multiple test are usually performed by using some test statistics, which we will denote here by $Y_1, Y_2, ..., Y_n$. We suppose now that this is the case, and also that these test statistics have a tendency of obtaining greater values when the corresponding hypothesis is not true. The critical level $\hat{\alpha}_k(y)$ for the outcome y of the test statistic Y_k is then equal to the supremum of the probability $P(Y_k \geq y)$ when the hypothesis H_k is true. Defining now the obtained levels $R_1, R_2, ..., R_n$ by

$$R_k = \hat{\alpha}_k(Y_k)$$

the classical Bonferroni test can be performed by comparing all the obtained levels $R_1, R_2, ..., R_n$ with α/n.

The sequentially rejective Bonferroni test will also be defined by the obtained levels. Denoting by $R^{(1)} \leq R^{(2)} \leq ... \leq R^{(n)}$ the ordered obtained levels and by $H^{(1)}, H^{(2)}, ..., H^{(n)}$ the corresponding hypotheses, the procedure can most easily be described by scheme 1, where α, $0 < \alpha < 1$, is a fixed number.



Scheme 1

The test is performed by starting at the top of the scheme and going down step by step until no further rejection can be done. This can happen either by accepting all remaining hypotheses or rejecting the last hypothesis $H^{(n)}$.

Theorem 1. *The sequentially rejective Bonferroni test described by scheme 1 has the multiple level of significance α for free combinations.*

Proof. Let I be the set of indexes of the true hypotheses. By the Boole inequality we then have

$$P\left(R_i > \frac{\alpha}{m} \text{ for all } i \in I\right)$$
$$= 1 - P\left(R_i \leq \frac{\alpha}{m} \text{ for some } i \in I\right)$$
$$\geq 1 - \sum_{i \in I} P\left(R_i \leq \frac{\alpha}{m}\right) \geq 1 - m\frac{\alpha}{m} = 1 - \alpha$$

where m is the number of elements in I. But if the event

$$\{R_i > \frac{\alpha}{m} \text{ for all } i \in I\}$$

occurs then

$$R^{(n+1-m)} > \frac{\alpha}{m},$$

and the sequentially rejective test stops in the step $n+1-m$ or earlier. This implies however that all hypotheses corresponding to obtained levels $R_i > \alpha/m$ will be accepted and this set of hypotheses includes the set of true hypotheses. □

In the sequentially rejective Bonferroni test the obtained levels are compared to the numbers

$$\frac{\alpha}{n}, \frac{\alpha}{n-1}, \cdots, \frac{\alpha}{1}$$

whereas in the classical Bonferroni test they are compared to α/n. This means that the probability of rejecting any set of (false) hypotheses using the classical Bonferroni test is smaller than or equal to the same probability using the sequentially rejective Bonferroni test based on the same test statistics. The classical Bonferroni test has been used mainly in situations where no other (more special) multiple test procedure is available. It can always be replaced by the corresponding sequentially rejective Bonferroni test without loosing any probability of rejecting false hypotheses. Except in trivial non-interesting cases the sequentially rejective Bonferroni test has strictly larger probability of rejecting false hypotheses and thus it ought to replace the classical Bonferroni test at all instants where the latter usually is applied.

The power gain obtained by using a sequentially rejective Bonferroni test instead of a classical Bonferroni test depends very much upon the alternative. It is small if all the hypotheses are 'almost true', but it may be considerable if a number of hypotheses are 'completely wrong'. If m of the n basic hypotheses are 'completely wrong' the corresponding levels attain small values, and these hypotheses are rejected in the first m steps with a big probability. The other levels are then compared to α/k for $k = n-m$, $n-m-1$, $n-m-2$, ..., 2, 1, which is equivalent to performing a sequentially rejective Bonferroni test only on those hypotheses that are not 'completely wrong'.

A very simple example will indicate how big the power gain may be. Suppose that Y_k, $k = 1, 2, \ldots, 10$ are independent and normally distributed with parameters μ_k and 1 for $k = 1, 2, \ldots, 10$ and that we want to test the hypotheses H_k: $\mu_k = 0$ against the alternatives $\mu_k > 0$ for $k = 1, 2, \ldots, 10$ at a multiple level of significance 0.05. If four of the μ_k's are equal to 0.0, four of them are equal to 6.0 and the remaining two are equal to 3.0, the classical Bonferroni test rejects

both the latter hypotheses with probability 0.439, while the sequentially rejective Bonferroni procedure rejects both with probability 0.565.

The great advantage with the sequentially rejective Bonferroni test (as well as with the classical Bonferroni test) is its flexibility. There are no restrictions on the type of tests, the only requirement being that it should be possible to calculate the obtained level for each separate test. Further there are no problems in including in the analysis only the *a priori* interesting hypotheses, while more special multiple tests usually include all hypotheses of a certain kind. But when there exist logical implications among the hypotheses problems arise which we have to take into consideration.

Let as before $\omega_1, \omega_2, \omega_3, \ldots, \omega_n$ denote the parameter sets where the hypotheses $H_1, H_2, H_3, \ldots, H_n$ are true. Then there exists a logical implication as soon as there is some index set I such that $\bigcap_{i\in I} \complement\, \omega_i = \phi$. In words this means that some combination of falseness of the different hypotheses is not possible, and the natural condition is of course that we should not end up the multiple test with a statement that the true parameter point is in an empty set. Each type of logical implication requires a special analysis of the properties of the test statistics in order to ensure that the test can not end up with such statements. The only type of logical implication we will consider is the one arising in connection with two-sided rejections.

Let γ be a (one-dimensional) parameter and suppose that H_1: $\gamma \leqslant \gamma_0$ and H_2: $\gamma \geqslant \gamma_0$ are basic hypotheses in a multiple test problem. Then $\complement\, \omega_1 \cap \complement\, \omega_2 = \phi$ and both these hypotheses should not be rejected in the multiple test procedure. It is natural to use the same test statistic to test both hypotheses and since we have the convention of rejecting the hypotheses for high values of the test statistics we should have $Y_2 = -Y_1$. Now for the outcomes y_1 of Y_1 and $y_2 = -y_1$ for Y_2 the obtained levels $\hat{\alpha}_1(y_1)$ and $\hat{\alpha}_2(y_2)$ satisfy

$$\begin{aligned}
\hat{\alpha}_2(y_2) &= \sup_{\gamma\geqslant\gamma_0} P(Y_2 \geqslant y_2) \\
&\geqslant \sup_{\gamma=\gamma_0} P(Y_2 \geqslant y_2) = \sup_{\gamma=\gamma_0} (1 - P(Y_2 < y_2)) \\
&= \sup_{\gamma=\gamma_0} (1 - P(Y_1 > y_1)) \\
&= 1 - \inf_{\gamma=\gamma_0} P(Y_1 < y_1) \\
&\geqslant 1 - \inf_{\gamma=\gamma_0} P(Y_1 \geqslant y_1) \\
&\geqslant 1 - \sup_{\gamma=\gamma_0} P(Y_1 \geqslant y_1) \\
&\geqslant 1 - \sup_{\gamma\leqslant\gamma_0} P(Y_1 \geqslant y_1) = 1 - \hat{\alpha}_1(y_1)
\end{aligned}$$

This means that for any outcomes y_1 and $y_2 = -y_1$ of Y_1 and Y_2 at least one of the obtained levels $\hat{\alpha}_1(y_1)$ and $\hat{\alpha}_2(y_2)$ is $\geqslant \frac{1}{2}$, and thus both hypotheses H_1 and H_2 can not be rejected in a sequentially rejective Bonferroni test (or a classical Bonferroni test) for any multiple level of significance $\alpha \leqslant \frac{1}{2}$.

If there are a number of pairs of one-sided hypotheses and no logical implications beside those within the pairs all illogical statements will still be avoided if the same statistics with opposite signs are used within the pairs and the multiple level of significance α is smaller than or equal to $\frac{1}{2}$. These tests are also coherent and consonant.

3. Applications and extensions

The sequentially rejective Bonferroni test can be applied in all situations where the classical Bonferroni test is usually applied. And it ought to replace the classical Bonferroni test in these cases because it gives only slightly more complicated computations and a non-negligable increase of power. It should however be noted that the sequentially rejective Bonferroni test can not be used to construct smaller confidence sets than those constructed by the classical Bonferroni test. This is so because the confidence set consists of the parameter points that would not be rejected as true parameter points in separate single tests. And when a confidence set is constructed from multiple tests it consists of the parameter points for which none of the detail hypotheses are rejected, which is in fact a special construction of a single test from a multiple test. If the sequentially rejective Bonferroni test is used in this way it is equivalent to the classical Bonferroni test.

The great advantage of the sequentially rejective Bonferroni test (as well as the classical Bonferroni test) is its computational simplicity, which arises from the reduction of the distributional problems to one dimension when the Boole inequality is used. The same computational simplicity is obtained when the test statistics are independent. It is easily seen that a sequentially rejective procedure with multiple level of significance α can be constructed by replacing the comparison constants $\alpha/n, \alpha/(n-1), \ldots, \alpha/1$ in the sequentially rejective Bonferroni test by $1-(1-\alpha)^{1/n}$, $1-(1-\alpha)^{1/(n-1)}, \ldots, 1-(1-\alpha)^{1}$, which are greater. This means that we get a more powerful test, but the increase in power is not very big. Among the numerous possible applications of the sequentially rejective Bonferroni test we will next mention a few.

The problem of comparing several treatments with one control have been studied by several authors. For the case of normally distributed observations the multiple test procedure suggested by Dunnett (1955) is commonly used. It requires the same number of

observations for all treatments, and it is based on the assumption that the variance is the same for the control and all treatments. Marcus et al. (1976) have proposed a closed test procedure, which is a refinement of the Dunnett procedure and which is more powerful. Their procedure is equivalent to a sequentially rejective procedure presented in Holm (1977).

The sequentially rejective Bonferroni test can also be used in this situation although it is of course less powerful than the refined Dunnett test. In most cases the difference is however not very big. In order to illustrate this we consider the case of comparing 9 treatments with one control based on four observations for the control and for each treatment. The refined Dunnett test then consists in successively comparing the ordered individual t-statistics for comparing one treatment with the control with the numbers

1.70, 1.99, 2.15, 2.25, 2.33, 2.40, 2.45, 2.50, 2.54,

while the sequentially rejective Bonferroni test consists in successively comparing the same statistics with the numbers

1.70, 2.04, 2.23, 2.36, 2.46, 2.54, 2.60, 2.66, 2.71.

In the classical Dunnett test they are all compared to 2.54.

There are two different variations of this problem, where the sequentially rejective Bonferroni test can easily be applied, whereas the classical and the refined Dunnett test can not so easily be applied. One is the case of non-equal sample sizes. Then the classical Dunnett type of test requires much computation, because tables are not available. The corresponding closed procedure requires even more computation, since a number of critical values of classical Dunnett test statistics are needed. The sequentially rejective Bonferroni test requires only a number of critical values of ordinary t statistics.

The other variation is the case where one-sided rejections are wanted. This is easily obtained by using a sequentially rejective Bonferroni test and introducing two one-sided hypotheses for each comparison of a treatment with the control.

For comparison of a number of treatments with one control in the case of non-normal distribution a many-one-rank test can be used. Such a test can be refined to a corresponding sequentially rejective test (equivalent to a corresponding closed test) with a higher power. See Holm (1977). If the number of observations are not the same for all treatments and the control, computational problems arise since tables are not available. These difficulties are avoided if a sequentially rejective Bonferroni test is used, since a table for the ordinary two-sample Wilcoxon test statistic is the only table needed for this test.

In the analysis of contingency tables the distributional problems connected with the construction of simultaneous tests are so big that the only possibility in practice is to use the Bonferroni technique. See e.g. Haberman (1974) chapter four. In such cases the power would be higher if the sequentially rejective Bonferroni test was used instead of a classical Bonferroni test. Other fields where big computational problems call for the use of Bonferroni technique are time series analysis and analysis of multidimensional distributions.

In all these cases as well as in others it may happen that some hypotheses are more important than the others, which may imply the use of higher levels of significance for the most important hypotheses and smaller levels of significance for the less important hypotheses when the Bonferroni technique is applied. At a first glance it seems to be impossible to obtain such an arrangement with the sequentially rejective Bonferroni test. But this is not true, since it is possible to generalise Theorem 1 to the case of different weights by slightly changing the procedure.

Let as before $H_1, H_2, \ldots, H_n$ be the hypotheses to be tested and $R_1, R_2, \ldots, R_n$ be the obtained levels of some suitable test statistics for those hypotheses. Further let $c_1, c_2, \ldots, c_n$ be positive constants indicating the importance of the hypotheses in the sense that the constants corresponding to more important hypotheses are greater then those corresponding to less important hypotheses. The precise meaning of these constants will be made clear later. Now introduce the statistics $S_k = R_k/c_k$ for $k = 1, 2, \ldots, n$, let $S^{(1)} \leqslant S^{(2)} \leqslant \ldots \leqslant S^{(n)}$ be the ordered statistics in this series, let $H^{(1)}, H^{(2)}, \ldots, H^{(n)}$ be the corresponding hypotheses and let $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ be the corresponding constants. Then a generalized sequentially rejective Bonferroni test can be described by scheme 2 on the next page.

Theorem 2. *The generalized sequentially rejective Bonferroni test described by scheme 2 has the multiple level of significance α for free combinations.*

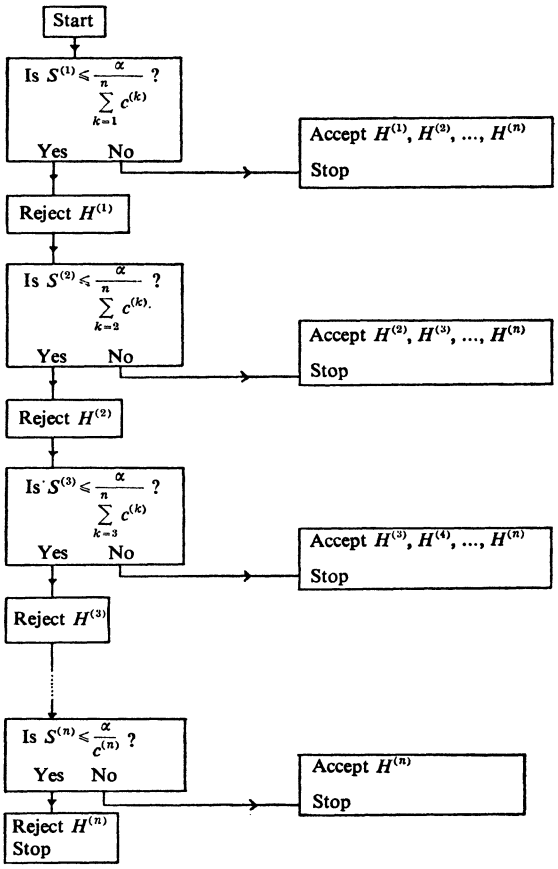
Proof. Let I be the set of indexes for the true hypotheses. By the Boole inequality we have

$$P\left(S_i > \frac{\alpha}{\sum_{j\in I} c_j} \text{ for all } i \in I\right)$$

$$= 1 - P\left(S_i \leqslant \frac{\alpha}{\sum_{j\in I} c_j} \text{ for some } i \in I\right)$$

$$= 1 - P\left(R_i \leqslant \frac{c_i\alpha}{\sum_{j\in I} c_j} \text{ for some } i \in I\right)$$

$$\geqslant 1 - \sum_{i\in I} \frac{c_i\alpha}{\sum_{j\in I} c_j} = 1 - \alpha.$$



Scheme 2

Now suppose that the event

$$\left\{S_i > \frac{\alpha}{\sum_{j \in I} c_j} \quad \text{for all } i \in I\right\}$$

occurs, and let ν be the smallest order number in the series $S^{(1)} \leq S^{(2)} \leq \ldots \leq S^{(n)}$ attained by the variables $\{S_i: i \in I\}$. Then

$$S^{(\nu)} > \frac{\alpha}{\sum_{j \in I} c_j} \geq \frac{\alpha}{\sum_{j=\nu}^{n} c^{(j)}}$$

which implies that the procedure will stop in step ν or earlier and that all true hypotheses will be accepted. □

From the definition of the generalized sequentially rejective Bonferroni test and the proof of Theorem 2 it can easily be seen what role is played by the constants $c_1, c_2, \ldots, c_n$. At each step in the procedure the obtained levels for the not yet rejected hypotheses are compared to parts of α, which are proportional to the corresponding constants. Compared to the 'ordinary' sequentially rejective Bonferroni test this implies an increase of power for alternative to hypotheses with high values of c_k at the cost of decrease of power for alternatives to hypotheses with small values of c_k, which is the reason of introducing the generalized test. When all c_k are equal the generalized test reduces to the ordinary test.

The previous discussion indicates a good way of handling multiple test problems in complicated applications. One can start by choosing a number of relevant hypotheses, then assign to every hypothesis a suitable test statistic, whose one-dimensional distribution is known exactly or approximately, and finally direct the power towards the most important hypotheses by choosing proper constants in a generalized sequentially rejective Bonferroni test. Of course it is also a desire to have the different test statistics exactly or approximately independent not for computational reasons but because a good experimental design requires the different hypotheses to be tested by variables 'not related to each others'.

Acknowledgement

The referee's suggestions for improving the presentation are gratefully acknowledged.

References

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* **50**, 1096–1121.

Gabriel, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224–250.

Haberman, S. J. (1974). *The analysis of frequency data*. The University of Chicago Press.

Holm, S. A. (1977). *Sequentially rejective multiple test procedures*. Statistical research report 1977-1, University of Umeå, Sweden.

Marcus, R., Peritz, E. & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Miller, R. G., Jr. (1966). *Simultaneous statistical inference*. McGraw-Hill, New York.

Naik, U. D. (1975). Some selection rules for comparing p processes with a standard. *Communications in Statistics* **4**, 519–535.

Sture Holm
Department of Mathematics
Chalmers University of Technology
S-412 96 Göteborg
Sweden