Modified Sequentially Rejective Multiple Test Procedures

JULIET POPPER SHAFFER*

Suppose that n hypotheses $H_1, H_2, \ldots, H_n$ with associated test statistics $T_1, T_2, \ldots, T_n$ are to be tested by a procedure with *experimentwise significance level* (the probability of rejecting one or more true hypotheses) smaller than or equal to some specified value α. A commonly used procedure satisfying this condition is the Bonferroni (B) procedure, which consists of rejecting H_i, for any i, iff the associated test statistic T_i is significant at the level $\alpha' = \alpha/n$. Holm (1979) introduced a modified Bonferroni procedure with greater power than the B procedure. Under Holm's sequentially rejective Bonferroni (SRB) procedure, if any hypothesis is rejected at the level $\alpha' = \alpha/n$, the denominator of α' for the next test is $n - 1$, and the criterion continues to be modified in a stagewise manner, with the denominator of α' reduced by 1 each time a hypothesis is rejected, so that tests can be conducted at successively higher significance levels. Holm proved that the experimentwise significance level of the SRB procedure is $\leq\alpha$, as is that of the original B procedure. Often, the hypotheses being tested are logically interrelated so that not all combinations of true and false hypotheses are possible. As a simple example of such a situation suppose, given samples from three distributions, we want to test the three hypotheses of pairwise equality: $\mu_i = \mu_{i'}$ $(i < i' = 1, 2, 3)$, where μ_i is the mean of distribution i. It is easily seen from the relations among the hypotheses that if any one of them is false, at least one other must be false. Thus there cannot be one false and two true hypotheses among these three. If we are testing all hypotheses of pairwise equality with more than three distributions, there are many such constraints. As another example, consider the hypotheses of independence of rows and columns of all 2×2 subtables of a $K \times L$ contingency table. It is shown that if one such hypothesis is false, then at least $(K - 1)(L - 1)$ must be false. When there are logical implications among the hypotheses and alternatives, as in the preceding examples, Holm's SRB procedure can be improved to obtain a further increase in power. This article considers methods for achieving such improvement. One way of modifying the SRB method is as follows: Given that $j - 1$ hypotheses have been rejected, the denominator of α', instead of being set at $n - j + 1$ for the next test as in the SRB procedure, can be set at t_j, where t_j equals the maximum number of hypotheses that could be true, given that at least $j - 1$ hypotheses are false. Obviously, t_j is never greater than $n - j + 1$, and for some values of j it may be strictly smaller, as for $j = 2$ in the first example. Then this modified sequentially rejective Bonferroni (MSRB) procedure will never be less powerful (and typically will be more powerful) than the SRB procedure while (as is proved in the article) maintaining an experimentwise significance level $\leq\alpha$. The MSRB procedure is readily applicable to a wide variety of standard and nonstandard problems. A number of examples are given, and extensions and generalizations are discussed. It is pointed out that the methods may be adapted in some circumstances to the use of non-Bonferroni multiple test procedures.

KEY WORDS: Multiple comparisons; Simultaneous inference; Bonferroni tests; Stagewise multiple tests; Power; Experimentwise error rate.

1. INTRODUCTION

Suppose that n hypotheses $H_1, H_2, \ldots, H_n$ with associated test statistics $T_1, T_2, \ldots, T_n$ are to be tested by a procedure with an experimentwise significance level smaller than or equal to some specified value α, where the experimentwise significance level is defined as the supremum (over all joint distributions F of the T_i that are possible under the assumed model) of the probability of rejecting one or more true hypotheses. A commonly used procedure satisfying this condition is the Bonferroni (B) procedure, based on the simple Bonferroni inequality. The B procedure consists of rejecting H_i, for any i, if and only if the significance probability of T_i—that is, $\Pr_{H_i}(T_i \geq t_i)$—is $\leq \alpha/n$, where t_i is the observed value of T_i and the T_i are defined so that large values lead to rejection.

Holm (1977, 1979) introduced a class of sequentially rejective multiple test methods that includes a modified Bonferroni procedure with greater power than the B procedure. Holm's sequentially rejective Bonferroni (SRB) procedure modifies the criterion in a stagewise manner, as follows: Let $Y_i = \Pr_{H_i}(T_i \geq t_i)$, let $\{Y_{(i)}\}$ be the order statistics of the Y_i, $Y_{(1)} \leq \cdots \leq Y_{(n)}$, and let $H_{(i)}$ be the hypothesis with test statistic $Y_{(i)}$, $i = 1, \ldots, n$. Then $H_{(1)}$ is rejected iff $Y_{(1)} \leq \alpha/n$; given that $H_{(1)}$ is rejected, $H_{(2)}$ is rejected iff $Y_{(2)} \leq \alpha/(n - 1)$; . . . ; given that $H_{(j-1)}$ is rejected, $H_{(j)}$ is rejected iff $Y_{(j)} \leq \alpha/(n - j + 1)$; and so forth. Acceptance of $H_{(k)}$ implies acceptance of $H_{(l)}$ for all $l > k$, $1 \leq j, k, l \leq n$. Holm proved that the experimentwise significance level of the SRB procedure is α, the same as that of the original B procedure.

It will be assumed that no hypothesis in the set is equivalent to the intersection of any of the others—that is, the hypotheses are *minimal* (Gabriel 1969). A decision on any intersection hypothesis of interest is made by rejecting it iff at least one of the hypotheses $H_1, H_2, \ldots, H_n$ included in the intersection is rejected; clearly, these decisions can be added to the decisions with respect to $H_1, H_2, \ldots, H_n$

* Juliet Popper Shaffer is Senior Lecturer, Department of Statistics, University of California, Berkeley, CA 94720.

Journal of the American Statistical Association
September 1986, Vol. 81, No. 395, Theory and Methods

without changing the experimentwise significance level of the total procedure.

A procedure δ will be called uniformly more powerful than another procedure δ^* for testing a specific set S of hypotheses if the probability of rejecting each false hypothesis in S under δ is greater than or equal to the probability of rejecting it under δ^*, for all joint distributions of the T_i that are possible under the assumed model, with strict inequality for at least one false hypothesis in S under some distribution. The SRB procedure is obviously uniformly more powerful than the B procedure for $H_1, H_2, \ldots, H_n$ and their intersections; in fact it has the stronger property of always rejecting hypotheses that are rejected under B and sometimes rejecting additional ones.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be the set of indexes of the hypotheses that are true in any particular application. In Holm's terminology, the situation is one of "free combinations" if the set $\{H_i: i \in I\}$ can be any subset of the n hypotheses. If these conditions are not satisfied, Holm's procedure remains valid, but it is possible to improve it to obtain a further increase in power. The purpose of this article is to show how the improvement can be achieved and to illustrate its extent in a number of different applications.

2. A MODIFIED SEQUENTIALLY REJECTIVE BONFERRONI PROCEDURE

The SRB procedure described in Section 1 can be modified in the following way: At stage j, instead of rejecting $H_{(j)}$ if $Y_{(j)} \leq \alpha/(n - j + 1)$, reject $H_{(j)}$ if $Y_{(j)} \leq \alpha/t_j$, where t_j equals the maximum number of possibly true hypotheses, given that at least $j - 1$ hypotheses are false. When there are relationships of logical implication among the hypotheses, usually the number m of true hypotheses cannot take on certain values between 0 and n, since the falsity of $j - 1$ hypotheses implies the falsity of some additional hypotheses for some values of j, as will be illustrated in Section 3. For those values of j, t_j will be strictly less than $n - j + 1$, and since t_j is obviously never greater than $n - j + 1$, the modified SRB (MSRB) procedure will be at least as powerful as the SRB procedure, and in most applications with restricted combinations it will be uniformly more powerful.

Given some specific application, let $A = \{a_i: i = 1, \ldots, r\}$ be the set of possible numbers of true hypotheses, $0 \leq a_1 < a_2 < \cdots < a_r \leq n$, and let J be the associated set of possible values of t_j. Then either $J = A$ or, more typically, J is the set of all nonzero values of A, since $t_j = \max_{\{a_i \leq n-j+1\}} a_i$ for all stages j.

That the MSRB procedure has experimentwise significance level $\leq \alpha$ follows directly from Holm's (1979) proof for the SRB procedure. The basic idea behind Holm's proof is that if m hypotheses are true, *an error must occur at or before stage* $n - m + 1$. Therefore, $\Pr(\text{no errors}) \geq \Pr(Y_i > \alpha/m \text{ for all } i \in I) = 1 - \Pr(Y_i \leq \alpha/m \text{ for some } i \in I) \geq 1 - \sum_{i \in I} \alpha/m = 1 - \alpha$. In Section 3 some applications in which the MSRB procedure may be considerably more powerful than the SRB procedure are presented, in Section 4 possible further modifications to achieve still greater power are given, and in Section 5 specific illustrations of the use of the MSRB are provided.

3. APPLICATIONS

3.1 Comparisons Among k Distributions

Consider a class of distributions $G \in \mathcal{G}$ and a function f defined over $\mathcal{G}$ and taking on at least k distinct values. Let $G_1, G_2, \ldots, G_k$ be k unknown distributions in $\mathcal{G}$, and consider the $k(k - 1)/2$ hypotheses

$$f(G_i) = f(G_{i'}), \qquad i < i'. \tag{3.1}$$

The family may or may not restrict the distributions to some specified form, such as normal; the function may be real-valued, such as the mean or variance, or, at the other extreme, $f(G)$ may equal G. Given any set of distributions, they will be said to be homogeneous or different according to whether or not their values of f are equal. The possible numbers of true hypotheses can be determined from the properties of equivalence relationships, as illustrated in Table 1 for $k = 4$, in which case the number of hypotheses $n = 6$. By considering all possible configurations of true and false hypotheses, as in Table 1, we see, for example, that all six hypotheses may be true, but that if any hypothesis is false, at least three must be false, since if any two distributions differ, at least one of these must differ from the remaining ones. As shown, it is also possible to have 2, 1, or 0 true hypotheses, so $A = \{0, 1, 2, 3, 6\}$ in this case.

The possible numbers of true hypotheses, and thus the values of t_j, for $3 \leq k \leq 10$ are given in Table 2. Values for $k > 10$ can be obtained from the recursion formula

$$S(k) = \bigcup_{j=1}^{k} \{\binom{j}{2} + x: x \in S(k - j)\}, \tag{3.2}$$

where $S(k)$ is the set of possible numbers of true hy-

Table 1. Determining Possible Numbers of True Hypotheses for the Application in Section 3.1, Illustrated for $k = 4$

Partitions of 4 populations	*Representation*	*Number of true hypotheses*	*Number of false hypotheses*	*Maximum number of true hypotheses*
1. [(1, 2, 3, 4)]	(4)	$\binom{4}{2} = 6$	0	6 (partition 1)
2. [(1)(234)], [(2)(134)], etc.	(3)(1)	$\binom{3}{2} = 3$	1–3	3 (partition 2)
3. [(12)(34)], [(13)(24)], etc.	(2)(2)	$\binom{2}{2} + \binom{2}{2} = 2$	4	2 (partition 3)
4. [(12)(3)(4)], etc.	(2)(1)(1)	$\binom{2}{2} = 1$	5	1 (partition 4)
5. [(1)(2)(3)(4)]	(1)(1)(1)(1)	0	6	0 (partition 5)
General	$(k_1)(k_2) \cdots (k_t)$	$\sum_{k_i \geq 2} \binom{k_i}{2}$ or 0 if all $k_i = 1$		

Table 2. Possible Numbers of True Hypotheses for the Application in Section 3.1, With k Distributions ($3 \le k \le 10$)

Number of distributions (k)	Number of hypotheses (n)	Possible numbers of true hypotheses
3	3	0, 1, 3
4	6	0–3, 6
5	10	0–4, 6, 10
6	15	0–4, 6, 7, 10, 15
7	21	0–7, 9, 10, 11, 15, 21
8	28	0–13, 15, 16, 21, 28
9	36	0–13, 15, 16, 18, 21, 22, 28, 36
10	45	0–18, 20, 21, 22, 24, 28, 29, 36, 45

NOTE: To use the MSRB procedure for the application in Section 3.4, determine the set $A = \{a_i\}$ of possible numbers of true hypotheses corresponding to the relevant values of K and L. Then, at stage j, for $j = 1, \ldots, \binom{K}{2} \cdot \binom{L}{2}$ test the hypothesis $H_{(j)}$ at significance level α/t_j, where $t_j = \max_{\{a_i \le n-j+1\}} a_i$.

potheses with k distributions, $k \ge 2$, and $S(0) = S(1) = \{0\}$. By testing intersections of these pairwise hypotheses as described in Section 1, tests of all of the $2^k - k - 1$ hypotheses of subset homogeneity of the G_i can be obtained.

Formula (3.2) can be proved by induction. It obviously holds for $k = 2$. Assuming it holds for $k - 1$ distributions, when a new distribution is added to those $k - 1$, it will be one of a set of j homogeneous distributions for some $j \in \{1, 2, \ldots, k\}$ and the other $k - j$ distributions will be different from those. Therefore, the set of possible numbers of true hypotheses (3.1), given j, is $\{\binom{j}{2} + x: x \in S(k - j)\}$, and $S(k)$ = the union of these sets over $j \in \{1, 2, \ldots, k\}$.

Of course, many other methods have been proposed for this situation [see, e.g., Einot and Gabriel (1975), and note the modifications described in Sec. 4 here and other possibilities indicated in Sec. 5]. Some detailed comparisons with other approaches can be found in Shaffer (1984), where it is shown that the method described here is competitive with other methods in general use.

3.2 Comparisons Within Several Sets of Distributions

Let f_i be a function, defined over a class of distributions $G_i \in \mathcal{G}_i$, which takes on at least K' distinct values. Suppose there are p sets of unknown distributions $G_{i1}, G_{i2}, \ldots, G_{ik_i}$ $(i = 1, \ldots, p)$, where $\sum_{i=1}^{p} k_i = K'$, and consider the $\sum_{i=1}^{p} k_i(k_i - 1)/2$ within-set hypotheses

$$f_i(G_{ij}) = f_i(G_{ij'}), \qquad j < j'. \tag{3.3}$$

From (3.2) we obtain the recursion formula

$$W(k_1, k_2, \ldots, k_p) = \{x_1 + x_2 + \cdots + x_p: x_i \in S(k_i), \quad i = 1, 2, \ldots, p\}, \tag{3.4}$$

where $W(k_1, k_2, \ldots, k_p)$ is the set of possible numbers of true within-set hypotheses (3.3) with k_i distributions in set i $(i = 1, 2, \ldots, p)$ and $S(k_i)$ is defined as in (3.2). By testing intersections as in the application in Section 3.1, all within-set hypotheses of subset homogeneity of the G_{ij} may be included. A specific application would be the tests usually recommended when there is interaction between two factors in a factorial design: The levels of one of the factors are compared separately within each level of the other factor. If i represents one of the p levels of a factor A, j represents one of the k_i levels of a factor B, with $k_i = k$ for all i, and the $f_i(G_{ij})$ are the means of normal distributions G_{ij} with common variance, then the hypotheses (3.3) are the standard normal-theory analysis-of-variance hypotheses that the effects of B within each level of A equal zero (see also Sec. 5, Illustration 2).

3.3 Comparisons Between Several Sets of Distributions

Given the same situation as in the application in Section 3.2, consider the $\sum\sum_{1 \le i < i' \le p} k_i k_{i'}$ pairwise equality hypotheses

$$f_i(G_{ij}) = f_{i'}(G_{i'j'}), \qquad i < i'. \tag{3.5}$$

[This comparison would generally make sense only when $f_i(G_{ij}) = f(G_{ij})$ for all i.] The possible numbers of true hypotheses can be obtained from the recursion formula

$$B(k_1, k_2, \ldots, k_p) = \bigcup_{(c_1, \ldots, c_p) \in C} \left\{ \sum\sum_{1 \le i < i' \le p} c_i c_{i'} + x: \right.$$
$$\left. x \in B(k_1 - c_1, k_2 - c_2, \ldots, k_p - c_p) \right\}, \tag{3.6}$$

where $B(0, 0, \ldots, 0) = \{0\}$, $B(k_1, k_2, \ldots, k_p)$ = the set of possible numbers of true between-set hypotheses (3.5) with k_i distributions in set i $(i = 1, 2, \ldots, p)$ and $C = \{(c_1, c_2, \ldots, c_p): 0 \le c_i \le k_i$ for $i = 1, 2, \ldots, p$ and $\sum_{i=1}^{p} c_i > 0\}$. The proof of (3.6) is somewhat similar to that of (3.2) and is omitted. By adding consideration of intersections, one obtains tests of the $2^{K'} - K' - 1 - \sum_{i=1}^{p} (2^{k_i} - k_i - 1)$ hypotheses of equality of all subsets containing distributions from more than one set.

An important application is to studies comparing treatments with control groups. As pointed out by Cochran (1983), in many observational studies an ideal control group is not available, in which case it is desirable to compare each treatment group with more than one control group, where each control may be vulnerable to different sources of bias.

3.4 Tests of Independence of All 2 × 2 Subtables of a $K \times L$ Contingency Table or Tests of Additivity in All 2 × 2 Subparts of a $K \times L$ Factorial Design

The sets A of possible numbers of true hypotheses are the same in (a) tests of independence of all 2×2 subtables of a $K \times L$ contingency table and (b) tests of additivity in all 2×2 subparts of a $K \times L$ factorial design. If $L = 2$, they reduce to those in the application in Section 3.1: In (a), the hypotheses are then equivalent to the hypotheses $\pi_{i1}/\pi_{i2} = \pi_{i'1}/\pi_{i'2}$, for $i, i' = 1, 2, \ldots, K$, where π_{ij} is the probability of an observation falling in row i and column j; in (b), they are equivalent to the hypotheses $\mu_{i1} - \mu_{i2} = \mu_{i'1} - \mu_{i'2}$ for $i, i' = 1, 2, \ldots, K$, where μ_{ij} is the mean of the distribution at level i of factor A and j of

Table 3. Possible Numbers of True Hypotheses for the Application in Section 3.4, for Selected Values of K × L

K and L	Number of hypotheses	Possible numbers of true hypotheses
$L = 2$		Obtain from Table 2
All K	$\binom{K}{2}$	by setting $K = k$.
$L = 3$		
$K = 3$	9	0–3, 5, 9
$K = 4$	18	0–10, 12, 18
$K = 5$	30	0–16, 18, 22, 30
$L = 4$		
$K = 4$	36	0–21, 24, 27, 36

NOTE: To use the MSRB procedure for the application in Section 3.4, determine the set $A = \{a_i\}$ of possible numbers of true hypotheses corresponding to the relevant values of K and L. Then, at stage j, for $j = 1, \ldots, \binom{K}{2} \cdot \binom{L}{2}$, test the hypothesis $H_{(j)}$ at significance level α/t_j, where $t_j = \max_{\{a_i \le n-j+1\}} a_i$.

factor B. Results for some representative values of $K \times L$ are given in Table 3. Adding intersections permits tests of the hypotheses of independence in all subtables of a contingency table or of the hypotheses of additivity under all subsets of factor level combinations in a factorial design.

If $L > 2$, there is no obvious algorithm for computing the possible numbers of true hypotheses. It can be seen, however, from the sets of possible numbers in Sections 3.1–3.3 and from Table 3 that the main advantage of the MSRB procedure over the SRB procedure appears at the second stage, where the relative difference in criterion significance probabilities is greatest. An explicit expression for t_2 in Section 3.4, proved in the Appendix, is

$$t_2 = [K(K-1)/2][L(L-1)/2] - (K-1)(L-1). \quad (3.7)$$

A compromise procedure, possibly applicable also in other situations, would be to set $t_j = t_2$ for all $2 \le j \le n - t_2 + 1$, and to use the SRB values for all stages $j > n - t_2 + 1$. This approach could also be combined effectively with the modified procedure described in Section 4.1.

4. MODIFICATIONS AND GENERALIZATIONS OF THE MSRB PROCEDURE

4.1 A Modified MSRB Procedure Following Initial Rejection of a More Comprehensive Hypothesis

Often the n hypotheses are not tested separately unless a more comprehensive hypothesis has initially been rejected at significance level α, where such rejection implies that at least some number r of the n hypotheses (but not which ones) are false, $r = 1, 2, \ldots, n-1$. It follows directly from the proof in Section 2 that a further improvement in the MSRB is then possible; the critical values α/t_j for testing $H_{(1)}, H_{(2)}, \ldots, H_{(r)}$ can be replaced by $\alpha/t_{(n-r)}$ without increasing the overall significance level above α. A typical opportunity to apply this modified procedure would arise in the use of a sequentially rejective procedure in the application in Section 3.1 following rejection of the hypothesis $f(G_1) = f(G_2) = \cdots = f(G_k)$ by a composite test based on a statistic other than $Y_{(1)}$ (e.g., rejection of equality of means with an F test in analysis of variance). In the first stage of the MSRB following rejection of this composite hypothesis, α/n would be replaced by α/t_2. For an application of this idea in a somewhat different context, see Shaffer (1979).

4.2 A Modified MSRB Procedure Taking Into Account the Particular Hypotheses Rejected

The power of the MSRB procedure can be increased, at the cost of greater complexity, by substituting for α/t_j at stage j the value α/t_j^*, where t_j^* is the maximum number of hypotheses that could be true, given that the specific hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(j-1)}$ are false. (The dependence of t_j^* on the first $j-1$ rejected hypotheses is suppressed for convenience in the notation.) To prove that this procedure has an experimentwise significance level $\le \alpha$, let $t_{j,L}^*$ be the minimum t_j^* (for $1 \le j \le n - m + 1$) over all subsets of size $j-1$ of false hypotheses. Note that $t_{j,L}^* \ge m$ for all j, where m is the number of true hypotheses. Then $\Pr(\text{one or more errors}) = \Pr(Y_i \le \alpha/t_{j,L}^* \text{ for some } j \le n - m + 1 \text{ and some } i \in I) \le \Pr(Y_i \le \alpha/m \text{ for some } i \in I) \le \alpha$.

As an illustration, consider the application in Section 3.2 with $p = 2$ and $k_1 = k_2 = 4$. By referring to Table 1, we see that if the two hypotheses $f_1(G_{11}) = f_1(G_{14})$ and $f_1(G_{11}) = f_1(G_{13})$ are false, the number of possibly true hypotheses is 9; if the two hypotheses $f_1(G_{11}) = f_1(G_{14})$ and $f_2(G_{21}) = f_2(G_{24})$ are false, the number of possibly true hypotheses is 6 (see also Sec. 5, Illustration 2).

5. ILLUSTRATIONS

When the number of hypotheses is large, the analysis of relationships among them may be complicated. In many situations that arise in practice, however, the number of hypotheses is small and their logical interrelations are transparent. In such cases, the MSRB procedure and its extensions can be easily applied. Illustration 1 is an example of this kind. In Illustration 2, the MSRB is compared with a more familiar approach to the problem described in Section 3.2.

Illustration 1. Information was available on the proportions of (i) passes, (ii) failures, and (iii) incompletes or withdrawals, in a number of mathematics classes, each of which had been taught by one of two different methods. All classes were of approximately the same size and were taught by different instructors. The experimenter was interested in comparing the proportions (i), (ii), and (iii) for the two methods.

Let p_{ijk} be the proportion of students in the kth class taught by method i who fall in category j, for $i = 1, 2$; $j = 1, 2, 3$; $k = 1, 2, \ldots, n_i$. Assuming the vectors $(p_{i1k}, p_{i2k}, p_{i3k})$ are independent observations from trivariate distributions F_i with mean vectors $(\mu_{i1}, \mu_{i2}, \mu_{i3})$, the three hypotheses to be tested are $\{H_j: \mu_{1j} - \mu_{2j} = 0, j = 1, 2, 3\}$.

Since the sum of the three observations for each class equals 1, it follows that if any of the three hypotheses is false, at most one can be true. Thus the following MSRB methods may be considered.

1. Choose an appropriate test statistic for each hypothesis. Order the hypotheses as in Section 1, and reject $H_{(1)}$ (the hypothesis corresponding to the smallest significance probability) if $Y_{(1)} < \alpha/3$. If $H_{(1)}$ is rejected, reject $H_{(i)}$ if $Y_{(i)} < \alpha$, for $i = 2, 3$.

2. Carry out a level-α test of the hypothesis H_o: $\mu_{1j} - \mu_{2j} = 0$ for all j. Under appropriate conditions on the proportions, a repeated measures analysis of variance or a multivariate analysis of variance would be a reasonable approximate test in this situation (see Shaffer 1981). In view of the result in Section 4.1, if H_o is rejected, test each of H_1, H_2, and H_3 at level α.

Illustration 2. Assume a 2×3 balanced factorial design to be analyzed by a fixed-effects analysis of variance. As pointed out in Section 3.2, if the test for interaction is significant, it is often recommended that the effects of any factor of interest be examined separately within each level of the other factor. Suppose the interaction is significant, and assume that we are interested in all pairwise contrasts among the three levels of factor B for each of the two levels of factor A. Letting μ_{ij} be the mean of the cell for level i of factor A and j of factor B, the six hypotheses to be tested are $\{H_{i(jk)}: \mu_{ij} - \mu_{ik} = 0; i = 1, 2; j < k = 1, 2, 3\}$.

Assume that we want the experimentwise significance level to be α. A typical way of accomplishing this aim is to use a multiple range test for each value of i, with significance level $\alpha/2$ for each. More specifically, given the value of i, the three means are ordered, and the difference between the largest and smallest is considered significant (i.e., the corresponding hypothesis is rejected) if the difference, divided by its estimated standard deviation based on the within-groups mean square, is greater than the $\alpha/2$ critical value of the studentized range distribution for three means. If the difference is significant, the tests of the remaining two differences are based on the studentized range of two means, with the levels depending on the particular multiple range procedure adopted. The optimal levels, consistent with a maximum Type I error probability of $\alpha/2$, are $\alpha/2$ for each of the remaining two differences (see, e.g., Lehmann and Shaffer 1979).

To use the MSRB, note that the significance of the interaction implies that the six hypotheses are not all true. It is then easily seen intuitively by the kind of argument in Section 3.1, and formally from the results of Section 3.2, that at most four of them are true. Thus, ordering the hypotheses as in Section 1, and using the modification of the MSRB discussed in Section 4.1, hypothesis $H_{(1)}$ would be rejected if the difference between the corresponding means were larger than the $\alpha/4$ critical value of the studentized range of two means. Given a rejection, $H_{(2)}$ would also be tested at $\alpha/4$. Making use of the modification in Section 4.2, $H_{(3)}$ would be tested at $\alpha/4$ or $\alpha/2$, depending on whether $H_{(1)}$ and $H_{(2)}$ referred to the same or different values of i, respectively. At each subsequent stage, the appropriate level for the test would be easily determined.

If the degrees of freedom for error are large, the multiple range and MSRB approaches can be compared by examining critical values of ranges of standard normal random variables. The first test, for example, would be based approximately on the $\alpha/2$ critical value of the range of three means for the multiple range procedure and the $\alpha/4$ critical value of the range of two means for the MSRB procedure. For $\alpha = .05$, the respective values are 3.68 and 3.53. In other words, the probability of finding at least one significant pairwise difference is greater with the modified MSRB procedure than with the range procedure. Some further comparisons are possible by direct consideration of critical values required by the two procedures. For instance, the probability of finding at least one significant difference within each level of i is greater with the modified MSRB than with the multiple range procedure, as is the probability of rejecting all of the hypotheses. Further consideration of the procedures suggests, as a rough approximation, that the multiple range procedure is more powerful when the false hypotheses are all within a single level of factor A, whereas the MSRB procedure has the advantage when true mean differences occur within both levels.

6. DISCUSSION

Note that the improvements in multiple test procedures discussed in this article are based on logical analysis of the relationships among the hypotheses and are independent of the particular test statistics used, except for knowledge of their respective marginal distributions. As in the usual use of the Bonferroni inequality, the methods are, therefore, highly flexible and easily used in nonstandard situations. Other approaches to multiple testing use more powerful methods based on the joint distribution of the test statistics, ranging from the use of improved Bonferroni inequalities that are based on some properties of the joint distribution of subsets of the test statistics (e.g., Worsley 1982), to the full use of the joint distribution, as, for example, when the test statistics are independent or in the comparison of means of normal distributions with equal variance. In many circumstances it may be feasible to combine logical and distributional considerations to obtain multiple testing methods better than those obtainable using either type alone; these would be modifications of the more general class of sequentially rejective methods considered by Holm (1977).

APPENDIX: PROOF OF (3.7)

The proof will be carried out in the contingency table framework. To apply it to factorial designs, substitute *means* for *expected frequencies* and substitute *equivalence if different only by translation* for *equivalence if different only by a scale factor.*

Consider a $K \times L$ contingency table, with entries equal to expected frequencies under the true model, as a row of L column vectors $c_1, c_2, \ldots, c_L$ of length K. Two vectors will be said to be equivalent if they differ only by a scale factor. Then given any L' columns, $2 \leq L' \leq L$, all 2×2 subtables of the $K \times L'$ contingency table consisting of the K rows and those L' columns satisfy the hypotheses (of independence) iff all column vectors j included in the L' columns are equivalent.

It will be shown that a table satisfying the maximum number of true hypotheses of independence, given that not all are true,

is one in which $L - 1$ vectors are equivalent and the Lth vector would be equivalent to these others if a single element were changed. The number of true hypotheses in such a table is readily seen to be (3.7).

Given a specific table that does not have all column vectors equivalent, let $r_{jj'}$ = the number of independent 2×2 subtables in the $K \times 2$ subtable consisting of the K rows and columns j and j', $0 \le r_{jj'} \le K(K-1)/2$. The number of independent 2×2 tables is

$$\sum_{1 \le j < j' \le L} \sum r_{jj'}. \tag{A.1}$$

We want to choose vectors to maximize (A.1) subject to the restriction that not all hypotheses are true.

Let $r_M = \max_{1 \le j < j' \le L} r_{jj'}$ among those that are $< K(K-1)/2$, and let c_{j*} and c_{j**} be any two column vectors for which $r_{j*j**} = r_M$. Replace each column vector c_j by a copy of c_{j*} (if $r_{jj*} \ge r_{jj**}$) or c_{j**} (otherwise). Note that with each of these replacements, (A.1) does not decrease: since $r_{jj'}$ becomes either $K(K-1)/2$ (if c_j and $c_{j'}$ become copies of the same vector c_{j*} or c_{j**}) or r_M (if c_j and $c_{j'}$ become copies of the two different vectors), no $r_{jj'}$ can decrease. After this replacement, the vectors are in two groups of L_1 and $L - L_1$ vectors, respectively, where the vectors within each group are equivalent; the number of true hypotheses is

$$[L_1(L_1-1)/2][K(K-1)/2] + [(L-L_1)(L-L_1 - 1)/2][K(K-1)/2] + L_1(L-L_1)r_M. \tag{A.2}$$

Since the maximum number of true hypotheses must occur in a table of this form, it remains only to maximize (A.2) with respect to L_1 and r_M.

Maximization of (A.2) With Respect to L_1. Since the sum of the coefficients of $K(K-1)/2$ and r_M in (A.2) is fixed, and $r_M < K(K-1)/2$, the maximum is found by maximizing the coefficient of $K(K-1)/2$. Since this coefficient is a quadratic in L_1 with a minimum at $L_1 = L/2$, its integer-valued maximum occurs for $L_1 = 1$ (or $L - 1$), in which case (A.2) becomes

$$[(L-1)(L-2)/2][K(K-1)/2] + (L-1)r_M. \tag{A.3}$$

Maximization of (A.3) With Respect to r_M. We want to maximize r_M, the number of true hypotheses in a $K \times 2$ contingency table, given $r_M < K(K-1)/2$. As noted in Section 3, the set of possible numbers of true hypotheses in tests of independence in $K \times 2$ contingency tables is equivalent to the set of possible numbers of true hypotheses in tests of pairwise equality among k populations (see the application in Sec. 3.1). Considering that application, if at least one of the hypotheses (3.5) is false, there are at least two distinct values of $f(G_i)$; we may assume that there are exactly two, since the number of true hypotheses can never be decreased if two different values are replaced by a single value. If the two distinct values are designated as v_1 and v_2, and d = the number of distributions i such that $f(G_i) = v_1$ $(0 < d < k)$, then the number of true hypotheses is

$$d(d-1)/2 + (k-d)(k-d-1)/2. \tag{A.4}$$

The expression (A.4) is a quadratic in d with a minimum at $d = k/2$ and its integer-valued maximum at $d = 1$ (or $k - 1$). Substituting this value for d in (A.4) gives $(k-1)(k-2)/2$ as the maximum number of true hypotheses in (3.1) with at least one false hypothesis. Therefore, the maximum value of r_M smaller than $K(K-1)/2$ is $(K-1)(K-2)/2$, achieved when the vector that is not equivalent to the $L - 1$ others differs from such equivalence in a single element. Finally, (A.3) with r_M replaced by $(K-1)(K-2)/2$ equals (3.7).

[*Received August 1984. Revised January 1986.*]

REFERENCES

Cochran, W. G. (1983), *Planning and Analysis of Observational Studies*, New York: John Wiley.

Einot, Israel, and Gabriel, K. R. (1975), "A Study of the Powers of Several Methods of Multiple Comparisons," *Journal of the American Statistical Association*, 70, 574–583.

Gabriel, K. R. (1969), "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *Annals of Mathematical Statistics*, 40, 224–250.

Holm, Sture (1977), "Sequentially Rejective Multiple Test Procedures," Statistical Research Report, University of Umea (Sweden), Institute of Mathematics and Statistics.

——— (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Lehmann, E. L., and Shaffer, Juliet Popper (1979), "Optimum Significance Levels for Multistage Comparison Procedures," *The Annals of Statistics*, 7, 27–45.

Shaffer, Juliet Popper (1979), "Comparison of Means: An F Test Followed by a Modified Multiple Range Procedure," *Journal of Educational Statistics*, 4, 14–23.

——— (1981), "The Analysis of Variance Mixed Model With Allocated Observations: Application to Repeated Measurement Designs," *Journal of the American Statistical Association*, 76, 607–611.

——— (1984), "Issues Arising in Multiple Comparisons Among Populations," in *Proceedings of the Seventh Conference on Probability Theory*, ed. M. Iosifescu, Bucharest, Romania: Editura Academiei Republicii Socialiste Romania.

Worsley, K. J. (1982), "An Improved Bonferroni Inequality and Applications," *Biometrika*, 69, 297–302.