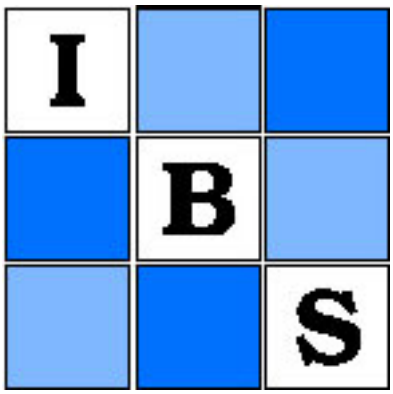

BIOMETRICS 43, 417–423
June 1987

An Improved Sequentially Rejective Bonferroni Test Procedure

Burt S. Holland
Department of Statistics, Temple University, Philadelphia, Pennsylvania 19122, U.S.A.
and
Margaret DiPonzio Copenhaver
Wyeth Laboratories, P.O. Box 8299, Philadelphia, Pennsylvania 19101, U.S.A.

SUMMARY

Holm (1979, *Scandinavian Journal of Statistics* **6**, 65–70) presented a simple and easily implemented multiple comparisons test procedure that is uniformly more powerful than the familiar Bonferroni method. Recently, Shaffer (1986, *Journal of the American Statistical Association* **81**, 826–831) proposed a modification of Holm's procedure that may achieve even more power at the cost of greater complexity. These procedures can be used whenever the observed levels of significance are available for all individual tests. It is shown that both the Holm and Shaffer procedures can be improved under the assumption of positive orthant dependence for the test statistics. It is noted that this assumption is met in many important practical situations and recommended that in these cases the new procedure be used in place of its predecessors whenever the required observed significance levels are available. The methodology is illustrated with a numerical example.

1. Introduction

When faced with the need to test simultaneously several hypotheses, many statisticians believe it is desirable to control not only the individual Type I errors but also the probability of rejecting at least one of those hypotheses being tested that are in fact true. One of the most widely applicable and easily implemented procedures for handling this problem is what has come to be known as the Bonferroni method. Based on the elementary probability inequality

$$\Pr\left(\bigcup_{i=1}^{k} E_i\right) \leqslant \sum_{i=1}^{k} \Pr(E_i),$$

this procedure undertakes to control the probability of rejecting at least one true hypothesis at some specified level α by testing each of the k hypotheses of interest at level of significance α/k.

The Bonferroni method has few competitors in a variety of settings. A prominent example is where one has m normal populations with common variance σ^2 and unknown means $\{\mu_j\}, j \in A_m$ ($A_m = \{1, \ldots, m\}$), a random sample of size n_j from the jth population, and interest in testing a preselected subset of size k of all $m(m-1)/2$ hypotheses of the form $\mu_j = \mu_{j'}, j \neq j'$, where k is quite a bit less than $m(m-1)/2$. But where the n_j are homogeneous, $k = m(m-1)/2$, test statistics are independent, or where interest lies only in comparing all active treatments with a control, other methods are usually more powerful. Another example is the situation where one wishes to conduct tests of independence on

Key words: Generalized Type I error probability; Multiple comparisons; Observed level of significance; Positive orthant dependence; *P*-value; Simultaneous tests of hypotheses.

some or all of the $RC(R-1)(C-1)/4$ 2×2 subtables of an $R \times C$ contingency table. See Miller (1981) for a review of Bonferroni and other multiple comparisons procedures.

We denote the k hypotheses $H_1, \ldots, H_k$ and assume they are a minimal set of hypotheses in the sense that no hypothesis can be expressed as the intersection of other hypotheses in the set. Let $I \subseteq A_k$ be the set of those t indices for which the hypotheses are actually true. The generalized Type I error probability, α, is the probability of rejecting at least one H_i for $i \in I$. (This coincides with what is known as experimentwise, or familywise, Type I error probability when all k hypotheses are true.) For each $i \in A_k$, let X_i be the test statistic and P_i be the P-value, or observed level of significance, of the test of H_i. That is, P_i is the probability, in repeated sampling, of obtaining a test statistic more extreme than that attained by the observed random sample(s) used to construct the statistic to test H_i. The marginal distribution of P_i is uniform on (0, 1) (see Cox and Hinkley, 1979, p. 66).

2. The "Sequentially Rejective Multiple Test Procedures" of Shaffer and Holm

Holm (1979) presented a simple modification of the Bonferroni method that provides a nonnegligible increase in power of the tests while retaining all of the properties of Bonferroni. However, this modification cannot shorten Bonferroni's confidence intervals for individual (scalar) parameters.

For situations where there are logical implications among the k hypotheses (e.g., the truth of some of the hypotheses necessarily implies the truth of some others), Shaffer (1986) shows that it is possible to modify Holm's procedure to provide a further increase in power at the cost of greater complexity.

Denote by $P_{(1)} \leq \cdots \leq P_{(k)}$ the ordered P-values and let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding hypotheses. Shaffer's procedure is as follows. For $i \in A_k$, define t_i to be the maximum number of possibly true hypotheses given that the specific hypotheses $H_{(1)}, \ldots, H_{(i-1)}$ are false, and let i^* be the smallest index in A_k such that

$$P_{(i^*)} > \alpha/t_{i^*}. \tag{2.1}$$

Then reject $H_{(1)}, \ldots, H_{(i^*-1)}$ and accept $H_{(i^*)}, \ldots, H_{(k)}$. If (2.1) is true for no index i^*, then all k hypotheses are to be rejected.

If $i-1$ hypotheses are false, then at most all of the remaining $k-i+1$ hypotheses are true. But if the hypotheses are logically interrelated, one has $t_i < k-i+1$ for some values of i. For example, consider the situation where one wishes to test all $k = K(K-1)/2$ possible hypotheses $\theta_j = \theta_{j'}, j \neq j'$, involving K populations, where θ_j is some parameter of interest from population j. Suppose $K = 4$ and that one of the six hypotheses is false. Then necessarily at most three of the remaining five hypotheses can be true and thus $t_2 = 3$. In Table 1 we present for $K = 3, 4, \ldots, 10$ the required t_i values for use in this situation. Note that in this particular example the t_i do not depend on which specific hypotheses are true or false or on the order in which the hypotheses are tested.

Holm's (1979) procedure does not consider the logical interrelationships among the k hypotheses, and differs from Shaffer's in that the right-hand side of (2.1) is replaced by $\alpha/(k-i^*+1)$, a quantity at least as small. The Bonferroni method consists of using a still smaller quantity not depending on i in (2.1): α/k. Therefore, any hypothesis rejected by Bonferroni is also rejected by Holm's procedure and any hypothesis rejected by Holm's procedure is also rejected by Shaffer's. Hence, Shaffer's procedure is at least as powerful as Holm's and Holm's is at least as powerful as Bonferroni. Both authors prove that their procedures guarantee control of generalized Type I error probability to be at most α. Holm claims that in actual practice the gain in power with his procedure, as compared to Bonferroni, is nonnegligible because $\alpha/(k-i+1)$ is much larger than α/k for many values of i. When there are no logical implications among the hypotheses, the Shaffer and Holm procedures coincide.

Table 1
t_i values for testing hypotheses $\theta_j = \theta_{j'}$ for all (j, j'), $1 \leq j < j' \leq K$

	K							
i	3	4	5	6	7	8	9	10
1	3	6	10	15	21	28	36	45
2	1	3	6	10	15	21	28	36
3	1	3	6	10	15	21	28	36
4		3	6	10	15	21	28	36
5		2	6	10	15	21	28	36
6		1	4	10	15	21	28	36
7			4	7	15	21	28	36
8			3	7	11	21	28	36
9			2	7	11	16	28	36
10			1	6	11	16	22	36
11				4	11	16	22	29
12				4	10	16	22	29
13				3	9	16	22	29
14				2	7	15	22	29
15				1	7	13	22	29
16					6	13	21	29
17					5	12	18	29
18					4	11	18	28
19					3	10	18	24
20					2	9	16	24
21					1	8	16	24
22						7	15	24
23						6	13	22
24						5	13	22
25						4	12	21
26						3	11	20
27						2	10	18
28						1	9	18
29							8	17
30							7	16
31							6	15
32							5	14
33							4	13
34							3	12
35							2	11
36							1	10
37								9
38								8
39								7
40								6
41								5
42								4
43								3
44								2
45								1

The increased complexity of Shaffer's procedure as compared with Holm's arises from the need to determine the series of t_i values. This requires some effort in nonstandard settings and/or when k is large. For instance, suppose in the above situation one does not wish to test some proper subset of size k_- of the k hypotheses. Then the $\{t_i\}$ cannot be read from Table 1. Nor are they determined merely by K and k_-; the configuration of the deleted tests and the ordering of the H_i matter as well.

3. An Improvement on the Shaffer and Holm Procedures for Positively Orthant Dependent Test Statistics

Under conditions that hold in a variety of multiple testing situations, it is possible to make a small but uniform improvement on the Shaffer and Holm procedures.

Definition 3.1 (Dykstra, Hewett, and Thompson, 1973). Random variables $X_1, \ldots, X_k$ are positively orthant dependent if

$$\Pr(X_1 \leq x_1, \ldots, X_k \leq x_k) \geq \prod_{i=1}^{k} \Pr(X_i \leq x_i) \quad \text{for all } x_1, \ldots, x_k.$$

The following theorem gives our improved procedure.

Theorem 3.1. Suppose that the form of the critical region for testing each of the hypotheses $H_1, \ldots, H_k$ is such that H_i is rejected if the ith test statistic is sufficiently large, and that the k test statistics exhibit positive orthant dependence. Let i^* be the smallest index in A_k such that

$$P_{(i^*)} > C(t_{i^*}), \tag{3.1}$$

where $C(x) = 1 - (1 - \alpha)^{1/x}$. Reject $H_{(1)}, \ldots, H_{(i^*-1)}$ and accept $H_{(i^*)}, \ldots, H_{(k)}$. If no index i^* satisfies (3.1), then reject all k hypotheses. This procedure constrains the generalized Type I error probability to be at most α, and is always at least as powerful (and possibly more powerful) than the original Shaffer and Holm procedures.

Proof. The new procedure differs from that of Shaffer in that the right-hand side of (2.1) is replaced in (3.1) by a number at least as large: $1 - (1 - \alpha)^{1/x} \geq \alpha/x$ for any $x \geq 1$ with equality only for $x = 1$. Hence, the new procedure rejects all hypotheses rejected by Shaffer and Holm, and possibly some additional ones as well.

Let t be the number of indices in I and suppose the event $[P_i > C(t_{i^*})$ for all $i \in I]$ occurs. Then all true hypotheses are accepted, and $t \leq k - i^* + 1$. Also, $t_{i^*} \geq t$ since the maximum number of possibly true null hypotheses given that the specific hypotheses $H_{(1)}, \ldots, H_{(i^*-1)}$ are false is at least as large as the actual number of true hypotheses. Hence, $C(t_{i^*}) \leq C(t)$. Invoking the positive orthant dependence yields

$$\begin{aligned}
\Pr(\text{all true hypotheses are accepted}) &= \Pr[P_i > C(t_{i^*}) \text{ for all } i \in I] \\
&\geq \prod_{i \in I} \Pr[P_i > C(t_{i^*})] \\
&\geq \prod_{i \in I} \Pr[P_i > C(t)] \\
&= [1 - C(t)]^t \\
&= 1 - \alpha.
\end{aligned} \tag{3.2}$$

Inequality (3.2) states that the probability that at least one true hypothesis is rejected is at most α.

If one wishes to avoid the task of assembling the list of $\{t_i\}$, (3.1) can be replaced by the rule $P_{(i^*)} > C(k - i^* + 1)$, giving a slightly less powerful test that is a uniform improvement on Holm's procedure. Holm (1979) recognized that his rule $P_{(i^*)} > \alpha/(k - i^* + 1)$ could be improved with $P_{(i^*)} > C(k - i^* + 1)$ for the case of independent test statistics, but he did not indicate that this improvement applies in certain dependent circumstances as well.

Note that the improvement of (3.1) over (2.1) is analogous to Šidák's improvement on the original Bonferroni procedure for simultaneously testing equality of means with multivariate t statistics (see Miller, 1981, pp. 254–255).

4. Discussion

The assumption of positive orthant dependence of the k test statistics holds in many common testing situations. Examples are the following:

(i) Absolute statistics from various multivariate t distributions, including those applicable in tests for k pairs of cell means in analysis of variance (see Karlin and Rinott, 1981).
(ii) Certain multivariate statistics that are asymptotically chi-square, including ones obtained by the weighted least squares or the maximum likelihood methods for categorical data analysis (see Dykstra, 1980).
(iii) F statistics with common denominators and possibly dependent numerators, as occur in mean square ratios in analysis of variance tables (see Dykstra, 1980).

Table 2
Implementation of Theorem 3.1 *procedure for Rhizobium data assuming generalized Type I error probability* .05

j:	1	2	3	4	5	6	$s_{\bar{x}_j - \bar{x}_{j'}} = 2.17$
Ordered sample mean $\bar{x}_j$:	13.26	14.64	18.70	19.92	23.98	28.82	df = 24

Conclusion[a]: 13.26 14.64 — 18.70 19.92 — 23.98 — 28.82

i	t_i	$.05/t_i$	$1 - .95^{1/t_i}$	Ordered P-Values[b]	j, j'	
1	15	.0033	.0034	.0000	1, 6	Reject H_0: $\mu_j = \mu_{j'}$ vs H_1: $\mu_j \neq \mu_{j'}$
2	10	.0050	.0051	.0000	2, 6	
3	10	.0050	.0051	.0000	1, 5	
4	10	.0050	.0051	.0001	3, 6	
5	10	.0050	.0051	.0002	2, 5	
6	10	.0050	.0051	.0004	4, 6	
7	7	.0071	.0073	.0053	1, 4	
8	7	.0071	.0073	.0194	1, 3	Accept H_0: $\mu_j = \mu_{j'}$ vs H_1: $\mu_j \neq \mu_{j'}$
9	7	.0071	.0073	.0229	2, 4	
10	6	.0083	.0085	.0229	3, 5	
11	4	.0125	.0127	.0354	5, 6	
12	4	.0125	.0127	.0738	4, 5	
13	3	.0167	.0170	.0738	2, 3	
14	2	.0250	.0253	.5311	1, 2	
15	1	.0500	.0500	.5794	3, 4	

[a] If two sample means are underlined by the same line, the corresponding population means are declared "not significantly different"; otherwise, they are declared "significantly different."

[b] $2\Pr(t_{24} > |\bar{x}_j - \bar{x}_{j'}|/s_{\bar{x}_j - \bar{x}_{j'}})$, where t_ν has a central Student's t distribution with ν degrees of freedom.

(iv) Absolute multivariate normal statistics with zero mean and arbitrary dispersion (see Šidák, 1967).

(v) The characteristic roots of a Wishart matrix having identity dispersion matrix. These roots are used to test whether a dispersion matrix $\mathbf{V}$ is a designated value $\mathbf{V}_0$ (see Dykstra and Hewett, 1978).

We have successfully used the new procedure in the following situation. Let Y_{iju} be independent $N(\mu_{ij}, \sigma^2)$ random variables, $u = 1, \ldots, n_{ij}$; $i = 1, \ldots, r$; $j = 1, \ldots, c$. Interest is in simultaneously testing only all hypotheses of the form $\mu_{ij} = \mu_{ij'}$, $j \neq j'$. Here $k = rc(c-1)/2$ is often substantially less than all $rc(rc-1)/2$ possible pairs of means and so most competitors to the Bonferroni method are inefficient because they protect against errors in a far larger family of hypotheses than is desired. A detailed comparative investigation of the new procedure vs competitors in this setting is underway.

It may be noted that the gain in power of the new procedure over the Shaffer and Holm originals is slight because $C(x) - \alpha/x$ is small for most α and $x > 1$. However, the improved procedure detailed herein requires only one or two additional statements of program code to implement and so seems worthwhile.

5. Example

We illustrate the use of the procedure in Theorem 3.1 with the Rhizobium data of Erdman (1946), which was selected by both Steel and Torrie (1980) and the ANOVA procedure of the Statistical Analysis System (1985) to illustrate various multiple comparisons procedures.

The ordered means of samples of size 5 from each of the six populations are shown in Table 2 along with the estimated standard deviation of the difference between any two means. As in the references we will assume that all $k = \binom{6}{2} = 15$ possible comparisons of means are of interest, but we emphasize that our procedure (i) permits the family of hypotheses under consideration to include fewer than all possible comparisons; and (ii) does not require equal sample sizes. We also assume $\alpha = .05$.

Observe in Table 2 that column 4 exceeds column 5 for the first seven rows, while the reverse is true in the last eight rows. Therefore, we reject only those hypotheses corresponding to the first seven rows. It may be noted that the conclusion of the usual Bonferroni multiple comparisons procedure for these data differs from Table 2 in not finding a significant difference between μ_1 and μ_4.

6. Conclusion

The improvement to the Bonferroni multiple comparisons procedure discussed here is very easy to implement now that *P*-values are available or readily obtainable for many popular tests. Since the improvement in terms of power is uniform and appreciable, we recommend use of the new procedure described in Theorem 3.1 in all circumstances where Bonferroni is now preferred provided the positive orthant dependence condition holds. We are investigating the use of the new procedure in some situations where the Bonferroni method is not customarily used.

ACKNOWLEDGEMENTS

The work of the first author was supported in part by a Research and Study Leave from Temple University. We thank Sanat K. Sarkar and anonymous referees for helpful comments, and Paul N. Somerville for introducing us to Holm's work.

RÉSUMÉ

Holm (1979, *Scandinavian Journal of Statistics* **6**, 65–70) a proposé une procédure de comparaisons multiples simple et facile à mettre en oeuvre; cette procédure est uniformément meilleure que la méthode classique dite de Bonferroni. Plus récemment, Shaffer (1986, *Journal of the American Statistical Association* **81**, 826–831) a proposé une modification de la procédure de Holm, qui au prix d'une complexité plus grande, atteint une meilleure puissance. Les deux procédures ne peuvent être utilisées qui si on peut calculer les probabilités limites correspondant à chacun des tests individuels.

Dans cet article, on montre que l'on peut améliorer les deux procédures dans le cas de la dépendance du quadrat positif des statistiques de test. Il faut souligner que ce cas de figure recouvre d'importantes situations pratiques. Lorsque c'est possible, il vaut dont mieux utiliser cette nouvelle procédure, plutôt que les précédentes. Un exemple numérique illustre la méthodologie.

REFERENCES

Cox, D. R. and Hinkley, D. V. (1979). *Theoretical Statistics.* New York: Chapman and Hall.

Dykstra, R. L. (1980). Product inequalities involving the multivariate normal distribution. *Journal of the American Statistical Association* **75**, 646–650.

Dykstra, R. L. and Hewett, J. E. (1978). Positive dependence of the roots of a Wishart matrix. *The Annals of Statistics* **6**, 235–238.

Dykstra, R. L., Hewett, J. E., and Thompson, W. A., Jr. (1973). Events which are almost independent. *The Annals of Statistics* **1**, 674–681.

Erdman, L. W. (1946). Studies to determine if antibiosis occurs among rhizobia. *Journal of the American Society of Agronomy* **38**, 251–258.

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

Karlin, S. and Rinott, Y. (1981). Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. *The Annals of Statistics* **9**, 1035–1049.

Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd edition. New York: Springer-Verlag.

SAS Institute, Inc. (1985). *SAS® User's Guide: Statistics, Version 5 Edition.* Cary, North Carolina: SAS Institute, Inc.

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

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**, 626–633.

Steel, R. G. D. and Torrie, J. H. (1980). *Principles and Procedures of Statistics—A Biometrical Approach*, 2nd edition. New York: McGraw-Hill.

Received June 1986; *revised February* 1987.