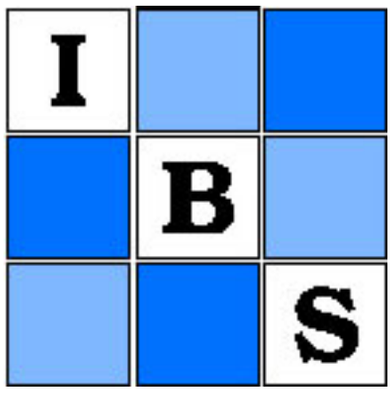



BIOMETRICS 48, 1005–1013
December 1992

# Adjusted *P*-Values for Simultaneous Inference

**S. Paul Wright**
Statistics Department, University of Tennessee, Knoxville, Tennessee 37996, U.S.A.

SUMMARY

This paper proposes that results from simultaneous tests be reported as adjusted *P*-values such that, if the adjusted *P*-value for an individual hypothesis is less than the chosen significance level of $\alpha$, then the hypothesis is rejected with an experimentwise error rate of no more than $\alpha$. Examples are given of adjusted *P*-values for multiple comparisons in the analysis of variance and of adjusted *P*-values based on the Bonferroni procedure and modifications of that procedure by Holm (1979, *Scandinavian Journal of Statistics* **6,** 65–70), Hochberg (1988, *Biometrika* **75,** 800–802), and Hommel (1988, *Biometrika* **75,** 383–386). The modified Bonferroni-based procedures are much more powerful than the original Bonferroni procedure, and they deserve wider use. In addition to the above, a procedure is outlined for obtaining adjusted *P*-values for any closed test procedure.

## 1. Introduction

Without a doubt, the *P*-value has become the "bottom line" for many consumers of statistical analyses. This is not without reason. A *P*-value provides information about whether a statistical hypothesis test is significant or not, and it also indicates something about "how significant" the result is: The smaller the *P*-value, the stronger the evidence against the null hypothesis. Most important, it does this without committing to a particular level of significance as traditional hypothesis tests and confidence intervals do. The dilemma for the consulting statistician is how to convince the client that a *P*-value is not necessarily "significant" just because it is less than .05 (or any other chosen level), when that *P*-value is for one of a possibly large collection of tests conducted during the course of a study. One way out of the dilemma is to report adjusted *P*-values which take into account that multiple tests are being conducted. It is the purpose of this paper to encourage the use of adjusted *P*-values and to extend their use to settings where they have not been used previously.

The idea of adjusted *P*-values is not new. In one sense, the *P*-value produced by a multivariate test statistic can be thought of as being adjusted for simultaneous testing. This is the approach taken by O'Brien (1984) as a means of handling multiple endpoints in clinical trials. This approach, however, does not solve the problem of reaching a conclusion about a particular test; for this, the *P*-value of the individual test needs to be adjusted. The idea of adjusting individual *P*-values is also not new (see Rosenthal and Rubin, 1983), but it does not seem to have gained much favor. One reason is that most adjusted *P*-values have been based on the Bonferroni inequality, leading to very conservative results in most instances. The objective of this paper is to demonstrate how adjusted *P*-values can be obtained using other, less conservative procedures. Section 2 presents methods that can be used for multiple comparisons in an analysis of variance setting. Sections 3–5 address the more general setting for which the Bonferroni adjustment is so often used. Special emphasis is given to recent improvements on the Bonferroni procedure by Holm (1979), Simes

*Key words:* Multiple comparisons; *P*-values; Simultaneous inference.

(1986), Hochberg (1988), and Hommel (1988). In Section 6, further improvements and applications are considered.

First, before specific techniques for adjustment are given, a definition of "adjusted *P*-value" is in order. The ordinary, unadjusted *P*-value for a single hypothesis test can be defined in two equivalent ways: **(1)** The *P*-value is the probability, under the sampling distribution of the test statistic when the null hypothesis is true, of obtaining a result as extreme as or more extreme than the one observed in the sample; **(2)** The *P*-value is the smallest level of significance that results in rejection of the null hypothesis. In applying the *P*-value concept in simultaneous inference, the second definition is more useful since it can be applied to any collection of tests. The adjusted *P*-value for a particular hypothesis within a collection of hypotheses, then, is the smallest overall (i.e., "experimentwise") significance level at which the particular hypothesis would be rejected. An adjusted *P*-value can be compared directly with any chosen significance level $\alpha$: If the adjusted *P*-value is less than or equal to $\alpha$, the hypothesis is rejected.

## 2. Multiple Comparisons in the Analysis of Variance

One area in which simultaneous inference has a long history is in the use of multiple comparisons in the analysis of variance (AOV). This is a good example of a situation in which *P*-values have been little used, but would be valuable. Although *P*-values are routinely reported for overall *F* tests of effects, they are almost never reported for individual comparisons. Instead, confidence intervals may be constructed; or in the case of pairwise comparisons, pairs of means that are significantly different may be reported verbally or displayed graphically. In either case, a fixed $\alpha$ level must be chosen in advance; it is this commitment to a specific $\alpha$ that can be avoided by using *P*-values.

Analysis of variance multiple comparisons can be divided into two categories: simultaneous test procedures (STPs) and multiple-stage tests (MSTs). In the case of STPs, in which all comparisons are referred to a single sampling distribution, the calculation of a *P*-value is straightforward. For example, in the case of the Scheffé procedure used in a one-way AOV with $g$ groups or treatments, an individual, unadjusted *P*-value is obtained by finding the sum of squares for the comparison (call it $SS_c$), converting this to an *F* statistic ($F_c = SS_c/MSE$, where MSE is the error mean square), and finding the upper-tail area under an *F* distribution with 1 and $(N - g)$ degrees of freedom. To obtain the adjusted *P*-value, the same $SS_c$ is used but is treated as if it had $(g - 1)$ degrees of freedom. That is, $F_c = (SS_c/(g - 1))/MSE$, and the *P*-value is obtained from an *F* distribution with $(g - 1)$ and $(N - g)$ degrees of freedom.

For example, consider a one-way AOV with four groups ($g = 4$), each with 5 observations (so $N = 20$). Suppose that the four cell means are 50.0, 51.0, 55.0, and 59.0, and the error mean square (MSE) is 16.5. The overall *F* statistic is then 5.126 with 3 and 16 degrees of freedom, resulting in a *P*-value of .01127. Table 1 shows, for this example, one convenient way to display the results of pairwise comparisons using *P*-values. The adjusted *P*-values are shown as the upper-right off-diagonal elements, cell means are given along the diagonal, and the unadjusted *P*-values are given below the diagonal. (These unadjusted and Scheffé adjusted *P*-values also appear in Table 2 in a different format to facilitate their comparison with adjusted *P*-values obtained by other procedures discussed below.)

At present, Tukey's procedure using the Studentized range distribution is the most powerful of the STPs for pairwise comparisons in a balanced design. In this procedure all comparisons are referred to the same critical value. The *P*-value for any comparison, which is by its nature adjusted for simultaneous inference, is simply the tail area under the sampling distribution of the Studentized range. Table 2 shows these *P*-values for the one-way AOV example.

**Table 1**
*A P-value matrix for pairwise comparisons in a one-way* AOV *using Scheffé's procedure*

| Group | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 50.0 | .98445 | .32058 | .02472 |
| 2 | .70222 | 51.0 | .50771 | .05028 |
| 3 | .06940 | .13903 | 55.0 | .50771 |
| 4 | .00294 | .00668 | .13903 | 59.0 |

*On the main diagonal:* The mean for each group. *Below the diagonal:* Unadjusted *P*-values for pairwise comparisons. *Above the diagonal:* Adjusted *P*-values.

**Table 2**
*P-values for pairwise comparisons in a one-way* AOV

| $i$ | Means | $p_i$ | $p_{\text{Scheffé}}$ | $p_{\text{Tukey}}$ | $p_{\text{REGWF}}$ | $p_{\text{REGWQ}}$ | $p_{\text{CTPF}}$ |
|---|---|---|---|---|---|---|---|
| 1 | 1-vs-4 | .00294 | .02472 | .01405 | .01127 | .01405 | .01127 |
| 2 | 2-vs-4 | .00668 | .05028 | .03057 | .02259 | .01736 | .02259 |
| 3 | 1-vs-3 | .06940 | .32058 | .24873 | .15236 | .15815 | .15236 |
| 4 | 2-vs-3 | .13903 | .50771 | .42907 | .25873 | .25873 | .15236 |
| 5 | 3-vs-4 | .13903 | .50771 | .42907 | .25873 | .25873 | .30296 |
| 6 | 1-vs-2 | .70222 | .98445 | .97927 | .91133 | .91133 | .70222 |

$p_i$: Unadjusted *P*-value for the *t*-test that the two means are equal; $p_{\text{Scheffé}}$: Adjusted *P*-value using Scheffé's procedure; $p_{\text{Tukey}}$: Adjusted *P*-value using Tukey's procedure; $p_{\text{REGWF}}$ and $p_{\text{REGWQ}}$: Adjusted *P*-values using REGWF and REGWQ procedures; $p_{\text{CTPF}}$: Adjusted *P*-values from closed test procedure using *F* tests.

Obtaining adjusted *P*-values for MSTs is a greater problem. The problem is to determine, for each comparison of interest (e.g., each pairwise comparison), what is the smallest experimentwise level of significance that would result in the comparison being declared significant. With unbalanced data, this might involve computations similar to those described in Sections 4 and 5. For the balanced one-way AOV example, it is fairly easy to obtain adjusted *P*-values. The method will be demonstrated using the MST of Ryan (1959, 1950), Einot and Gabriel (1975), and Welsch (1977) based on the *F* distribution. The method has been implemented in SAS (1990), where it is referred to as the REGWF procedure. Similar calculations using the Studentized range distribution give the REGWQ procedure.

The REGWF procedure, like many of the balanced-data procedures, is most easily carried out using the ordered means. For this reason, the means in the AOV example above are intentionally listed in order from smallest (mean 1) to largest (mean 4). REGWF declares a pair of means significantly different by conducting *F* tests on subsets of means. For example, means 1 and 4 are significantly different if the null hypothesis $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ is rejected by an *F* test. Similarly, means 1 and 3 are significantly different if $H_0$: $\mu_1 = \mu_2 = \mu_3$ is rejected. The experimentwise error rate is maintained at the desired level ($\alpha$) by using adjusted significance levels for different subsets of the $g$ means. If $k$ is the number of means in a subset, then the *F* test for equality of those $k$ means is done with significance level $\alpha$ when $k = g$ or $k = g - 1$, and with level $1 - (1 - \alpha)^{k/g}$ when $k < g - 1$ (see Hochberg and Tamhane, 1987, p. 69). For each *F* test there is a corresponding unadjusted *P*-value, call it $p_F$. The hypothesis that a set of $k$ means are equal is rejected if the corresponding $p_F \leqslant \alpha$ when $k \geqslant g - 1$ or when $p_F \leqslant 1 - (1 - \alpha)^{k/g}$ when $k < g - 1$. The adjusted *P*-value is obtained by rearranging these expressions to obtain a quantity that leads to rejection of the hypothesis when the quantity is $\alpha$ or less. Thus when $k \geqslant g - 1$, the adjusted *P*-value (call it $p_{\text{REGWF}}$) is just $p_F$. When $k < g - 1$, rearrangement gives

$p_{\text{REGWF}} = 1 - (1 - p_F)^{g/k}$. Table 2 shows the results for the AOV example for the REGWF procedure and also for the REGWQ procedure.

## 3. The Bonferroni Procedure and Some Simple Modifications

The appealing characteristic of the Bonferroni procedure is that it is applicable in essentially any simultaneous inference situation. The price of this generality is a lack of power, but recently proposed modifications to the Bonferroni procedure have improved the situation.

The classic Bonferroni procedure is well known. Given a collection of hypotheses, $H_1, H_2, \ldots, H_n$, and an experimentwise error rate of $\alpha$, each individual hypothesis $H_i$ is tested at a reduced significance level of $\alpha_i$ such that $\sum \alpha_i = \alpha$. Typically, $\alpha_i = \alpha/n$, but unequal allocation can also be used. Let $p_i$ be the unadjusted $P$-value for testing $H_i$; then, using equal allocation ($\alpha_i = \alpha/n$), $H_i$ is rejected when $np_i \leq \alpha$. That is, the Bonferroni adjusted $P$-value (call it $p_{\text{Bonf}}$) is $np_i$. The Bonferroni adjusted $P$-values for the AOV multiple comparisons appear in Table 3. These $P$-values were intentionally not truncated to 1.0 to emphasize their conservatism.

Holm (1979) presented a sequentially rejective Bonferroni procedure that is much less conservative but that still maintains the experimentwise error rate at $\alpha$. In Holm's procedure, the unadjusted $P$-values are ordered (as they have been in Table 3) so that $p_1 \leq p_2 \leq \cdots \leq p_n$, and each $p_i$ is compared to $\alpha/(n - i + 1)$ rather than $\alpha/n$. That is, the smallest $p_i$ is compared to $\alpha/n$, the next smallest to $\alpha/(n - 1)$, etc. In other words, it is $(n - i + 1)p_i$ that is compared to $\alpha$. These values appear in the column labeled $r_i$ in Table 3.

The $r_i$ values are not necessarily the adjusted $P$-values for Holm's procedure, for in this procedure hypotheses are tested sequentially beginning with the smallest $P$-value, $p_1$. Testing is stopped when a nonsignificant result is obtained, and all untested hypotheses are considered nonsignificant. That is, $H_i$ is rejected if $(n - i + 1)p_i \leq \alpha$ provided that $(n - j + 1)p_j \leq \alpha$ for all $j < i$. For example, the smallest significance level at which $H_1$ would be rejected is .01766. This is the Holm adjusted $P$-value (call it $p_{1(\text{Holm})}$) for $H_1$. Similarly, the $r_i$ values are the Holm adjusted $P$-values for $H_2$, $H_3$, and $H_4$. However, $p_{5(\text{Holm})}$ is not .27806. In the first place, the unadjusted $p_4$ and $p_5$ are the same, so it would be illogical for their adjusted $P$-values to differ. Even if $p_4$ and $p_5$ were different, it could still happen that $r_4 > r_5$, as it is here. In Holm's sequentially rejective procedure, $H_5$ cannot be rejected unless $H_4$ is rejected; and the smallest $\alpha$ that permits this is .41709.

Instead of testing sequentially starting with the smallest $P$-value, one might start with the

**Table 3**
*Bonferroni-like adjusted P-values for pairwise comparisons in a one-way* AOV

| $i$ | Means | $p_i$ | $p_{\text{Bonf}}$ | $r_i$ | $p_{\text{Holm}}$ | $p_{\text{Hoch}}$ | $p_{\text{Hommel}}$ | $p_{\text{Šidák}}$ | $r_{i(\text{Šidák})}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1-vs-4 | .00294 | .01766 | .01766 | .01766 | .01766 | .01766 | .01753 | .01753 |
| 2 | 2-vs-4 | .00668 | .04009 | .03341 | .03341 | .03341 | .03341 | .03942 | .03296 |
| 3 | 1-vs-3 | .06940 | .41643 | .27762 | .27762 | .27762 | .20821 | .35052 | .25003 |
| 4 | 2-vs-3 | .13903 | .83417 | .41709 | .41709 | .27806 | .27806 | .59268 | .36179 |
| 5 | 3-vs-4 | .13903 | .83417 | .27806 | .41709 | .27806 | .27806 | .59268 | .25873 |
| 6 | 1-vs-2 | .70222 | 4.2133 | .70222 | .70222 | .70222 | .70222 | .99930 | .70222 |

$p_i$: Unadjusted $P$-value for the $t$-test that the two means are equal; $p_{\text{Bonf}}$: Bonferroni adjusted $P$-value; $r_i$: The sequentially adjusted $P$-value using Bonferroni's procedure; $p_{\text{Holm}}$: Adjusted $P$-value based on Holm's procedure; $p_{\text{Hoch}}$: Adjusted $P$-value based on Hochberg's procedure; $p_{\text{Hommel}}$: Adjusted $P$-value based on Hommel's procedure; $p_{\text{Šidák}}$: Adjusted $P$-value based on Šidák's procedure; $r_{i(\text{Šidák})}$: The sequentially adjusted $P$-value using Šidák's procedure.

largest $P$-value, stopping when a significant result is obtained and declaring all untested results to be significant. That is, $H_i$ is rejected if $(n - j + 1)p_j \leq \alpha$ for any $j \geq i$. This is Hochberg's (1988) procedure. It is clearly more powerful than Holm's procedure, but it still has an experimentwise error rate of $\alpha$ provided that the Simes (1986) test, on which it is based, maintains an error rate of $\alpha$. (Simes' test is discussed below.) Table 3 shows the Hochberg adjusted $P$-values ($p_{\text{Hoch}}$).

An attractive feature of the adjusted $P$-values based on the Holm and Hochberg procedures is that they form a nondecreasing sequence from top to bottom. This means that if the test of $H_j$ is "more significant" than that of $H_i$ as indicated by their unadjusted $P$-values, then the test of $H_j$ cannot be "less significant" than that of $H_i$ based on the adjusted $P$-values. In addition, Hochberg's procedure has the nice characteristic that no adjusted $P$-value can be larger than the largest of the unadjusted $P$-values. Consequently, no adjusted $P$-value can be larger than 1.00.

## 4. Adjusted *P*-Values for Any Closed Test Procedure

Holm's procedure, as well as the even more powerful procedure of Hommel (1988) to be discussed below, is based on the "closed test procedure" principle (Marcus, Peritz, and Gabriel, 1976). The following version of the procedure is given in Hommel (1986, 1988). Let $H_1, H_2, \ldots, H_n$ be a collection of $n$ hypotheses. Define all possible combinations of subsets of these hypotheses: $H_I = \cap\{H_i: i \in I\}$ for all $I \in K$, where $K$ is the set of all nonempty subsets of $\{1, 2, \ldots, n\}$. Let there exist for every $H_I$ a test based on statistic $T_I$. For a given $\alpha$, $H_I$ is rejected if every $H_J$ is rejected at level $\alpha$ by the corresponding $T_J$, where $J \in K$ and $J \supseteq I$ (that is, subset $I$ is included among the subsets $J$). The probability of falsely rejecting one or more hypotheses when testing all $H_I$ is at most $\alpha$.

To use this procedure, one would ordinarily start with the global test $H_I = \cap\{H_i: i = 1, 2, \ldots, n\}$. If this test is rejected at level $\alpha$, one proceeds to test, still at level $\alpha$, each subset of $(n - 1)$ hypotheses. As long as hypotheses continue to be rejected at level $\alpha$, one continues testing, eventually reaching subsets of size 1: the individual hypotheses, $H_i$. It is not always necessary to test every possible combination of hypotheses, however; sometimes shortcuts are available. For example, when the aim is to test the individual hypotheses $H_i$, if the test statistic $T_I$ is a Bonferroni test (i.e., reject $H_I$ for a subset of hypotheses of size $m$ if the smallest unadjusted $P$-value for hypotheses in the subset is less than or equal to $\alpha/m$), Holm's procedure results. If the test of Simes (1986) is used for each subset, the result is Hommel's (1988) procedure, to be discussed in the next section.

The closed test procedure can be restated so that it generates an adjusted $P$-value for each $H_I$ as follows. Let $p_I$ be the unadjusted $P$-value for test $T_I$ of hypothesis $H_I$. $H_I$ is rejected only if $p_J \leq \alpha$ for all $H_J$ where $J \supseteq I$ (again note that the set of $H_J$ includes $H_I$). Therefore, the adjusted $P$-value for $H_I$ (the smallest $\alpha$ at which $H_I$ could be rejected) must be the largest of the $p_J$ values. Unfortunately, in the general case, in order to obtain an adjusted $P$-value for each individual hypothesis $H_i$, one would have to conduct the test and obtain the unadjusted $P$-value for every possible subset of hypotheses. The total number of tests that must be conducted is therefore $\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1$. Fortunately, in certain special cases, shortcuts exist, such as the procedure for Holm adjusted $P$-values given above.

Table 2 shows the adjusted $P$-values from the closed test procedure for the one-way AOV example. The $T_I$ used in the procedure was an $F$ test. It is noteworthy that using the closed test procedure results in different adjusted $P$-values for the two hypotheses (2-vs-3 and 3-vs-4) whose unadjusted $P$-values are the same. Thus, while the closed test procedure is a powerful, general-purpose approach to simultaneous testing, it is not without its unappealing aspects, at least for certain choices of the test statistic $T_I$.

## 5. Adjusted *P*-Values from Hommel's Procedure

Simes (1986) introduced the following global test for all hypotheses in a set of $n$ hypotheses: Reject $H_0 = \{H_1, H_2, \ldots, H_n\}$ if $p_i \leq i(\alpha/n)$ for at least one $i$ (where the $p_i$ are, as before, the ordered, unadjusted *P*-values). Simes proved that this test has level $\alpha$ when the tests are independent and provided simulations to indicate that it also has level $\alpha$, except perhaps in unusual circumstances, when tests are dependent. Since $H_0$ is rejected if any $np_i/i \leq \alpha$, the *P*-value for the global Simes test is just the smallest of the $np_i/i$ values. Simes' test, however, does not address the problem of testing the individual $H_i$. Hommel's (1988) procedure does this by using Simes' test as the $T_I$ in the closed test procedure. Hommel (1989) showed that this procedure is more powerful than Hochberg's procedure, but it still has level $\alpha$ (provided that the Simes tests have level $\alpha$). The computations are more involved than for Hochberg's procedure, but the full closed testing procedure need not be done. For some fixed $\alpha$, Hommel (1988) showed that the following shortcut is sufficient. Let $j$ be the number of hypotheses in the largest subset of hypotheses for which the Simes test is not significant. That is,

$$j = \max\{m \in \{1, \ldots, n\}: p_{(n-m+k)} > k(\alpha/m) \quad \text{for all } k = 1, \ldots, m\}.$$

If there are no nonsignificant Simes tests, then all $H_i$ are rejected. Otherwise, reject $H_i$ when $p_i \leq \alpha/j$.

The computation of adjusted *P*-values for Hommel's procedure is perhaps most easily understood by referring directly to the closed test procedure. The full closed test procedure would require that, for each hypothesis $H_i$, the Simes test *P*-value be obtained for every subset of hypotheses containing $H_i$. The Hommel adjusted *P*-value is then the largest of these Simes test *P*-values. However, it is not necessary to test every subset. For subsets containing $m$ of the $n$ hypotheses, it is sufficient to test the single subset containing the largest $(m - 1)$ *P*-values in addition to $p_i$. This could be called the "least significant subset" of size $m$ that contains $H_i$. Clearly, if the Simes test is significant at level $\alpha$ for this subset, it will be significant at level $\alpha$ for all other subsets of size $m$ containing $H_i$. For each $m = 1, \ldots, n$, the Simes test *P*-value is obtained for the least significant subset that contains $H_i$. The largest of these $n$ Simes test *P*-values is the Hommel adjusted *P*-value. Table 4 demonstrates these calculations for hypothesis $H_3$ in the AOV example. While the procedure used in Table 4 is useful for illustrating conceptually how Hommel adjusted *P*-values are obtained, it still involves some duplication of effort. In practice, an algorithm that is more efficient and easily programmed on a computer is desirable. Such an algorithm is given in the Appendix. Table 3 shows the adjusted *P*-values based on Hommel's procedure for the one-way AOV example.

**Table 4**
*Calculation of Hommel adjusted P-value for* $H_3$ (*group* 1 *vs* 3) *in the one-way* AOV *example*

| | Tests in least significant set | $mp_i/i$ | Smallest $mp_i/i$ |
|---|---|---|---|
| $m = 6$ | 1, 2, 3, 4, 5, 6 | .01766, .02004, .13881, .20854, .41709, .70222 | .01766 |
| $m = 5$ | 2, 3, 4, 5, 6 | .03341, .17351, .23171, .17379, .70222 | .03341 |
| $m = 4$ | 3, 4, 5, 6 | .27762, .27806, .18537, .70222 | .18537 |
| $m = 3$ | 3, 5, 6 | .20821, .20854, .70222 | .20821 |
| $m = 2$ | 3, 6 | .13881, .70222 | .13881 |
| $m = 1$ | 3 | .06941 | .06941 |
| | | $p_{3(\text{Hommel})}$ = Largest value = .20821 | |

$m$: Number of hypotheses in a subset; $p_i$: Unadjusted *P*-value for hypothesis $H_i$.

## 6. Discussion

This paper has made two main points. The primary point is that for any fixed level-$\alpha$ simultaneous inference procedure, it is possible to obtain suitably adjusted $P$-values. These adjusted $P$-values have much to recommend them, and it is to be hoped that they will become as much a part of accepted statistical practice as unadjusted $P$-values are now. A secondary point is that, for any simultaneous inference situation where the Bonferroni procedure is applicable, more powerful procedures exist. In particular, Holm's procedure is just as general in applicability as the Bonferroni procedure, it is nearly as simple to carry out, and it is much more powerful. The procedures of Hochberg and Hommel are even more powerful, but strictly speaking, they are known to have the desired experimentwise error rate only for independent tests. The same is true for the more complex procedure of Rom (1990), which is even more powerful than that of Hochberg (1988).

Despite being more powerful than the Bonferroni procedure, the procedures of Holm, Hochberg, and Hommel still tend to be conservative. Several ways to increase their power are available. One approach would be to base the procedures not on the Bonferroni inequality but on the Šidák (1967) inequality. Holland and Copenhaver (1987) discuss the circumstances (positive-orthant-dependent test statistics) when this is appropriate. In practice, this means that in the classic Bonferroni procedure, $p_{i(\text{Bonf})} = np_i$ becomes $p_{i(\text{Šidák})} = 1 - (1 - p_i)^n$. In the sequentially rejective procedures of Holm and Hochberg, $r_i = (n - i + 1)\,p_i$ becomes $r_{i(\text{Šidák})} = 1 - (1 - p_i)^{(n-i+1)}$. The $p_{i(\text{Šidák})}$ and $r_{i(\text{Šidák})}$ values for the one-way AOV example appear in Table 3. For small $P$-values, which are the ones that receive the most attention, the reduction in magnitude of the adjusted $P$-value (compared to Bonferroni) is quite small.

Another route to improvement is to take advantage of logical relationships among hypotheses, as in multiple comparisons among means, as suggested for Holm's procedure by Shaffer (1986). Hommel (1988) showed how to apply this modification to his procedure.

As Hommel (1988) points out, another route to improvement would be to make use of stochastic dependencies among test statistics. This, of course, is what makes use of the Šidák inequality more powerful than the Bonferroni inequality; it is also what gives the Tukey, REGWF, and REGWQ procedures some of their additional power. Another promising approach to using dependencies among statistics to obtain adjusted $P$-values is the resampling approach of Westfall and Young (1989). In their approach, in exchange for additional computational requirements, one obtains considerably more power. For example, Westfall and Young report the three smallest unadjusted $P$-values, obtained from 24 Fisher exact tests on data from Brown and Fears (1981), as .035, .053, and .125. The corresponding adjusted $P$-values (untruncated) using the Bonferroni procedure are .840, 1.272, 3.000; for Hommel's procedure they are .805, .938, .944. Using the procedure of Westfall and Young, they are .270, .362, and .732, a considerable improvement.

### Acknowledgements

I am grateful to the following colleagues at the University of Tennessee for helpful discussions, comments, and suggestions during the preparation of this paper: Dr Esteban Walker, Dr Robert Mee, and Dr James Schmidhammer. The detailed comments of an associate editor and a referee also resulted in great improvements to the presentation of the material.

### Résumé

Cet article propose que les résultats de tests simultanés soient exprimés avec des $P$-valeurs ajustées telles que, si la $P$-valeur ajustée pour une hypothèse individuelle est inférieure au seuil de signification

$\alpha$ choisi, alors l'hypothèse est rejetée avec un taux d'erreur expérimentale inférieur à $\alpha$. Des exemples de *P*-valeurs ajustées sont donnés pour des comparaisons multiples dans une analyse de variance ainsi que des *P*-valeurs fondées sur la procédure de Bonferroni et des modifications de cette procédure par Holm (1979, *Scandinavian Journal of Statistics* **6,** 65–70), Hochberg (1988, *Biometrika* **75,** 800–802) et Hommel (1988, *Biometrika* **75,** 383–386). Les procédures de Bonferroni modifiées sont plus puissantes que la procédure de Bonferroni originale, et méritent un plus large emploi. Par ailleurs, une procédure est proposée pour obtenir des *P*-valeurs ajustées pour tout test de comparaisons multiples.

## References


Brown, C. C. and Fears, T. R. (1981). Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics* **37,** 763–774.

Einot, I. 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.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Holland, B. S. and Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* **43,** 417–423.

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

Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika* **33,** 321–336.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75,** 383–386.

Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika* **76,** 624–625.

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

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40,** 1079–1087.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77,** 663–665.

Rosenthal, R. and Rubin, D. B. (1983). Ensemble adjusted *P*-values. *Psychological Bulletin* **94,** 540–541.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin* **56,** 26–47.

Ryan, T. A. (1960). Significance tests for multiple comparisons of proportions, variances, and other statistics. *Psychological Bulletin* **57,** 318–328.

SAS Institute Inc. (1990). *SAS/STAT™ User's Guide, Version* 6, *Volume* 2, 4th 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.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association* **72,** 566–575.

Westfall, P. H. and Young, S. S. (1989). *P*-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association* **84,** 780–786.


*Received July* 1990; *revised March and July* 1991; *accepted July* 1991.

## Appendix

The following algorithm calculates adjusted *P*-values based on Hommel's procedure. As usual, the $p_i$ are the ordered, unadjusted *P*-values. Let $a_i$ be the final adjusted *P*-values.

**1.** Initially set $a_i = p_i$ for all $i$.
**2.** For each $m = n, (n - 1), \ldots, 2$ (in that order), do the following:

**2a.** For $i > (n - m)$, **(i)** Calculate values $c_i = (mp_i)/(m + i - n)$.

**(ii)** Find the smallest of the above $c_i$ values; call it $c_{\min}$.

**(iii)** If $a_i < c_{\min}$, then let $a_i = c_{\min}$.

**2b.** For $i \leqslant (n - m)$, **(i)** Let $c_i = \min(c_{\min}, mp_i)$.

**(ii)** If $a_i < c_i$, then let $a_i = c_i$.

The basic rationale of the algorithm is to start with unadjusted *P*-values (step 1) and to adjust them upward as necessary while evaluating Simes test *P*-values for various subsets of hypotheses. (Recall that the adjusted *P*-value is the *largest* Simes test *P*-value encountered during the closed test procedure.) The upward revisions occur in steps 2a(iii) and 2b(ii). In step 2a, "$i > (n - m)$" selects the largest $m$ *P*-values, which might be called the "very least significant subset of size $m$" or $\text{VLSS}_m$ for short. $c_{\min}$ is just the Simes test *P*-value for this subset. For the hypotheses within the $\text{VLSS}_m$, $a_i$ can be no smaller than $c_{\min}$, thus the upward revision in step 2a(iii). Step 2b considers hypotheses that are outside the current $\text{VLSS}_m$. The $c_i$ of step 2b(i) is just the Simes test *P*-value for the "least significant subset of size $m$ that contains $H_i$," obtained by substituting $p_i$ for the smallest unadjusted *P*-value in the $\text{VLSS}_m$. Again, $a_i$ can be no smaller than $c_i$, so upward revision occurs in step 2b(ii).