Multiple Testing of General Contrasts Using Logical Constraints and Correlations

Peter H. Westfall

Journal of the American Statistical Association, Vol. 92, No. 437. (Mar., 1997), pp. 299-306.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199703%2992%3A437%3C299%3AMTOGCU%3E2.0.CO%3B2-R

Journal of the American Statistical Association is currently published by American Statistical Association.

Multiple Testing of General Contrasts Using Logical Constraints and Correlations

Peter H. WESTFALL

Use of logical constraints among hypotheses and correlations among test statistics can greatly improve the power of step-down tests. An algorithm for uncovering these logically constrained subsets in a given dataset is described. The multiple testing results are summarized using adjusted p values, which incorporate the relevant dependence structures and logical constraints. These adjusted p values are computed consistently and efficiently using a generalized least squares hybrid of simple and control-variate Monte Carlo methods, and the results are compared to alternative stepwise testing procedures.

KEY WORDS: Adjusted p value; Control variate; Linear model; Monte Carlo; Multiple comparisons; Simultaneous inference.

1. INTRODUCTION

Multiple testing means testing more than one hypothesis in a particular study. The well-known problem with such procedures is the inflated probability of erroneous rejections when there is no allowance for multiplicity. This is an important problem, because practicing statisticians routinely deal with large and complex datasets. It becomes even more important when such datasets are used for major decisions, as happens, for example, in the drug approval process.

Multiple testing methods are usually designed to control the *familywise error rate* (FWE; also called experimentwise error rate), which is the probability of incorrectly rejecting at least one hypothesis in a given family of tests (Hochberg and Tamhane 1987). FWE control can be achieved by adjusting the significance levels for individual tests; for example, if there are k tests, then testing individual nulls at the α/k level ensures that FWE $\leq \alpha$. This is the Bonferroni method. Equivalently, testing may be done using *adjusted p* values; in this case the adjusted p value for test j is $\tilde{p}_j = kp_j$, where p_j is the ordinary p value. Rejection when $\tilde{p}_j \leq \alpha$ ensures that FWE $\leq \alpha$. The adjusted p values are not probabilities per se; rather, they are simply convenient statistics to use for identifying significances. They have been defined by various authors, including Dunnett and Tamhane (1991) and Wright (1992).

The Bonferroni method is unnecessarily conservative because (a) it fails to account for dependencies among tests, and (b) the actual FWE of the procedure is potentially much less than the nominal α level when only a few null hypotheses are true. Methods such as Tukey's method for all pairwise comparisons in the analysis of variance (ANOVA) alleviate problem (a), achieving greater power by incorporating dependencies. Sequentially rejective testing methods alleviate problem (b) by reducing the family size sequentially, depending on the ordering of the p values. Holm (1979) used a simple step-down method, and Shaffer (1986) extended Holm's method by exploiting logical constraints among the hypotheses. Using these constraints, the implicit family of tests used at each stage of the step-down algorithm can be restricted to a subset of all tests, as determined by the ordering of the observed p values, and by the constraints. Holland and Copenhaver (1987) and Rom and Holland (1995) offered further refinements by using Šidák's (1967) inequality rather than the Bonferroni inequality, and applied the method to pairwise comparisons in the one-way ANOVA model. These step-down methods are great improvements over the simple Bonferroni method; however, they still use probability inequalities and thus remain conservative. The purpose of this article is to exploit logical constraints *as well as* dependencies, obtaining further improvements in the power of multiple testing procedures applied to a general set of linear contrasts.

The resulting method is superior to competing methods that control the FWE, because it accounts for specific logical constraints and for specific dependence structures, as is demonstrated via theory and examples. This method is applicable to any collection of general contrasts in the linear model. Primarily I consider the homoscedastic normal model, but I also discuss approximate resampling-based solutions in the nonnormal model with possible heteroscedasticity.

Specific contributions of this article include (a) comparing the proposed method to alternative multiple testing methods, (b) proving theoretical results (in the Appendix) showing existence of the logically constrained sets and giving a computationally feasible condition for identifying them, and (c) developing a generalized least squares hybrid of simple and control variate resampling for estimating the multiplicity adjustments.

In Section 2 I review existing multiple testing procedures, redefining all in terms of adjusted p values. In Section 3 I describe an algorithm for finding the logically constrained subsets in a multiple testing problem involving contrasts. In Section 4, I develop a simple and efficient Monte Carlo method for obtaining the multiplicity-adjusted p values and give an example. I provide properties and extensions of the proposed method in Section 5, and a summary in Section 6.

Peter H. Westfall is Professor of Statistics, Department of Information Systems and Quantitative Sciences, College of Business Administration, Texas Tech University, Lubbock, TX 79409. The author thanks an anonymous operator at the sci.math.research usenet news group for suggesting the proof to the lemma in the appendix, and the referees and associate editor for their constructive comments.

Journal of the American Statistical Association
March 1997, Vol. 92, No. 437, Theory and Methods

2. OVERVIEW OF MULTIPLE TESTING METHODS

Usually, multiple testing methods are defined in terms of critical regions, but here I restate them using adjusted p values for ease of use and uniformity of comparison. Unless stated otherwise, I assume that the marginal null distribution of each unadjusted p value is uniform on [0, 1].

A desirable property of a simultaneous testing procedure (STP) is that it control the FWE in the *strong* sense, as defined by Hochberg and Tamhane (1987). For any given STP, if

$$\Pr(\text{reject at least one } H_i, i = j_1, \ldots, j_t | H_{j_1}, \ldots, H_{j_t} \text{ are true}) \leq \alpha,$$

for *any* configuration of true nulls $H_{j_1}, \ldots, H_{j_t}$, then the STP controls the FWE in the strong sense.

2.1 Bonferroni Method

Given a collection of p values $p_1, \ldots, p_k$, from k tests of hypotheses, $H_1, \ldots, H_k$ (all in null form), one rejects H_j if $p_j \leq \alpha/k$, where α is the preset familywise significance level. Or, as stated in Section 1, one rejects H_j if $\tilde{p}_j \leq \alpha$, where $\tilde{p}_j = kp_j$. It is well known that this method controls the FWE strongly.

2.2 Holm Method

Holm's (1979) step-down testing method is as follows: Order the observed p values as $p_{(1)} \leq \cdots \leq p_{(k)}$, corresponding to hypotheses $H_{(1)}, \ldots, H_{(k)}$. The adjusted p values are initially taken to be $\tilde{p}_{(j)} = (k - j + 1)p_{(j)}$and subsequently are monotonicity enforced as follows to ensure that $\tilde{p}_{(j)} \leq \tilde{p}_{(j+1)}$:

$$\tilde{p}_{(j)} \leftarrow \max(\tilde{p}_{(j)}, \tilde{p}_{(j-1)}), \quad \text{for} \quad j = 2, \ldots, k, \quad \text{sequentially.} \qquad (1)$$

The Holm procedure rejects $H_{(j)}$ at FWE $= \alpha$ if $\tilde{p}_{(j)} \leq \alpha$, where the $\tilde{p}_{(j)}$ are the monotonicity-enforced adjusted p values. Although Holm defined his procedure in terms of critical values, Equation (1) gives the equivalent procedure using adjusted p values. Holm demonstrated that his method controls the FWE in the strong sense.

2.3 Shaffer Method

Shaffer (1986) extended the Holm procedure by incorporating logical constraints among hypotheses. The method starts like the Holm and the simple Bonferroni methods: $\tilde{p}_{(1)} = kp_{(1)}$. Having rejected $H_{(1)}$, one then considers the largest collection of hypotheses that can possibly be true, given that $H_{(1)}$ is false. Often, it is impossible for $H_{(2)}, \ldots, H_{(k)}$ to be all true if $H_{(1)}$ is false; rather, only some subsets can possibly be true, and the Bonferroni multiplier at the second step is the cardinality of an appropriate subset. Examples given by Shaffer include pairwise tests among means in a one-way ANOVA and tests of 2×2 subtables of an $r \times c$ contingency table. Shaffer noted that this procedure (defined formally in the next paragraph) also controls the FWE strongly and is uniformly more powerful than the Holm procedure.

Following Westfall and Young (1993, pp. 68–69), let $p_{(j)} = p_{r_j}$. The hypotheses are rejected sequentially based on the magnitudes of the original p values. If $H_{(j)}$ is rejected, then $H_{(1)}, \ldots, H_{(j-1)}$ must have been previously rejected. Given a particular observed sequence $r_1, \ldots, r_k$, define sets $\bar{S}_j$ of hypotheses *that include* $H_{(j)}$ that can be true at stage j, given that all previously rejected hypotheses are false. Letting $S = \{r_1, \ldots, r_k\} = \{1, \ldots, k\}$, and letting H'_j denote the alternative hypothesis, define

$$\bar{S}_1 = \{S\},$$

$$\bar{S}_2 = \left\{ K \subset S | r_2 \in K \quad \text{and} \quad \left(\bigcap_{l \in K} H_l \right) \cap (H'_{(1)}) \neq \emptyset \right\},$$

$$\bar{S}_3 = \left\{ K \subset S | r_3 \in K \quad \text{and} \quad \left(\bigcap_{l \in K} H_l \right) \cap \ (H'_{(1)} \cap H'_{(2)}) \neq \emptyset \right\},$$

$$\vdots$$

$$\bar{S}_j = \left\{ K \subset S | r_j \in K, \quad \text{and} \quad \left(\bigcap_{l \in K} H_l \right) \cap \left(\bigcap_{l=1}^{j-1} H'_{(l)} \right) \neq \emptyset \right\},$$

$$\vdots$$

$$\bar{S}_k = \{\{r_k\}\}. \qquad (2)$$

The sets may be simplified by eliminating subsets. If $K_1 \in \bar{S}_j$ and $K_2 \in \bar{S}_j$, with $K_1 \subset K_2$, then K_1 is redundant and may be eliminated. Let S_j denote the subset of $\bar{S}_j$ without such redundant elements.

Defining

$$M_j = \max_{K \in S_j} |K|,$$

where $|K|$ denotes cardinality, the step-down adjusted p values implied by Shaffer's method are initially $\tilde{p}_j = M_j p_j$, with monotonicity enforcement as defined in (1).

2.4 Step-Up Methods

Hochberg (1988) noted that under independence of tests, the FWE can be strongly controlled using a *step-up* procedure. Like the Holm procedure, here the adjusted p values are initially taken to be $\tilde{p}_{(j)} = (k - j + 1)p_{(j)}$; however, unlike the step-down procedure, the monotonicity enforcement of the adjusted p values is performed in the *reverse* direction:

$$\tilde{p}_{(j)} \leftarrow \min(\tilde{p}_{(j)}, \tilde{p}_{(j+1)}), \quad \text{for} \quad j = k - 1, \ldots, 1, \quad \text{sequentially.} \qquad (3)$$

Hochberg's adjusted p values are uniformly as small as Holm's; therefore, the Hochberg procedure is uniformly at least as powerful as Holm's.

Dunnett and Tamhane (1992, 1995) developed step-up methods that incorporate correlations among test statistics. These methods are frequently more powerful than the corresponding step-down methods. It would be useful to have such methods available for the present setup, which involves logical constraints, but I leave this as an open problem.

A recent addition to the step-up literature concerns methods that control the *false discovery rate* (FDR). The FDR criterion allows more rejections, but procedures that control the FDR usually do not control the FWE. Benjamini and Hochberg (1995) devised a method for controlling the FDR using p values under the assumption of independence. The p values are initially adjusted to $\tilde{p}_{(j)} = (k/j)p_{(j)}$, then monotonicity enforced in step-up fashion, as defined by (3).

3. FINDING THE LOGICALLY CONSTRAINED SUBSETS

In this section I describe computational aspects of finding the sets $K \in S_j$ of (2).

3.1 An Example

To fix ideas, consider an example of an analysis of covariance (ANCOVA) involving four treatments and two covariates. The treatments are dosage levels (in 0, 5, 50, and 500 units) of a compound administered to rodent dams, the response is the average nonbirth weight of the litter, and the covariates are gestation time and litter size. The data were provided and described in more detail by Westfall and Young (1993, pp. 99–101). Suppose that all $\binom{4}{2} = 6$ pairwise comparisons of the adjusted means are of interest, as are the trend contrasts

$$\mathbf{c}'_1 = (-1.5, -.5, .5, 1.5),$$

$$\mathbf{c}'_2 = (-138.75, -133.75, -88.75, 361.25),$$

and

$$\mathbf{c}'_3 = (-.795, -.105, .305, .595).$$

Note that $\mathbf{c}_1$ is a simple ordinal trend contrast, $\mathbf{c}_2$ is a trend contrast suggested by the arithmetic dosage levels, and $\mathbf{c}_3$ is a contrast suggested by a log-ordinal dose–response relationship. The reason for selecting several trend contrasts is that the actual form of the dose–response relationship is unknown.

In this example there are nine contrasts, $\mathbf{c}_1, \mathbf{c}_2$, and $\mathbf{c}_3$, as well as $\mathbf{c}_4 = (-1, 1, 0, 0), \ldots, \mathbf{c}_9 = (0, 0, -1, 1)$. If $\boldsymbol{\beta}_1$ is the adjusted mean vector with elements $\beta_{1l}, l = 1, 2, 3, 4$, then the hypotheses considered are H_j: $\mathbf{c}'_j\boldsymbol{\beta}_1 = 0$. Because the compound is expected to *lower* weights, all alternative hypotheses are lower-tailed. Table 1 displays the contrasts in order of most significant to least significance, or in order of t ratios from most negative to most positive, and the corresponding sets S_j at each step. The sets in the table refer to the contrasts ordered by significance; that is, the set $K = \{2, 3\}$ refers to the contrasts $\{(-.795, -.105, .305, .595), (-1, 0, 0, 1)\}$. Note that $\{2, 3\} \in S_2$ because $2 \in \{2, 3\}$ and because the equations

$$-.795\beta_{11} - .105\beta_{12} + .305\beta_{13} + .595\beta_{14} = 0$$

and

$$-\beta_{11} + \beta_{14} = 0$$

do not contradict $-\beta_{11} + \beta_{12} \neq 0$. However, inclusion of any $j > 3$ in the set $\{2, 3\}$ would yield a system of three equations that force $-\beta_{11} + \beta_{12} = 0$; hence $\{2, 3\}$ is a maximal contrast set in S_2.

3.2 A General Method

With the foregoing example in mind, return to the general problem of finding the sets $K \in S_j$ of (2). Suppose that k contrasts are to be tested. As demonstrated in the corollary in the Appendix, the sets S_j are nonempty, provided that no pair of contrasts is collinear. Let $\mathbf{C}_K$ denote the matrix whose columns are the contrast vectors $\mathbf{c}_i, i \in K$ (taken in any order), and let $\mathcal{C}(\mathbf{A})$ denote the real vector space spanned by the columns of a real matrix $\mathbf{A}$. Then a necessary and sufficient condition (proven in the proposition in the Appendix) for $K \in S_j$ is that $r_j \in K$ and $\mathbf{c}_{(i)} \notin \mathcal{C}(\mathbf{C}_K)$ for all $i = 1, \ldots, j-1$.

To check this condition numerically, let $\mathbf{C}_{(j)}$ denote the matrix whose columns are the vectors $\mathbf{c}_{(1)}, \ldots, \mathbf{c}_{(j-1)}$. Construct the matrix $\mathbf{C}_{(j)}^{\perp K} = (\mathbf{I} - \mathbf{C}_K(\mathbf{C}'_K\mathbf{C}_K)^{-}\mathbf{C}'_K)\mathbf{C}_{(j)}$; if all columns of $\mathbf{C}_{(j)}^{\perp K}$ are nonzero, then $K \in S_j$. (To verify that a column vector is nonzero, one can check that its norm is not within machine error of zero.) The set S_j then can be determined by searching all subsets of $\{r_j, \ldots, r_k\}$ that include r_j to determine which ones meet the criterion for inclusion in S_j.

Table 1. Maximal Sets of Contrasts S_j for Litter Weight Data

j	*r_j*	*Contrast coefficients*				*Estimated contrast*	*Standard error*	*t value*	*Maximal subset, S_j*
1	4	−1	1	0	0	−3.35	1.29	−2.60	{{1, 2, 3, 4, 5, 6, 7, 8, 9}}
2	3	−.795	−.105	.305	.595	−1.94	.96	−2.02	{{2, 3}, {2, 4}, {2, 5}, {2, 6}, {2, 7}, {2, 8}, {2, 9}}
3	6	−1	0	0	1	−2.68	1.33	−2.01	{{3, 4, 7}, {3, 5, 9}, {3, 6}}
4	5	−1	0	1	0	−2.29	1.33	−1.72	{{4, 5}, {4, 6}, {4, 8}}
5	1	−1.5	−.5	.5	1.5	−3.49	2.08	−1.68	{{5, 6}, {5, 7}, {5, 8}}
6	2	−138.75	−133.75	−88.75	361.25	−315.32	408.15	−.77	{{6, 7}, {6, 8}, {6, 9}}
7	9	0	0	−1	1	−.39	1.45	−.27	{{7, 8, 9}}
8	8	0	−1	0	1	.68	1.33	.51	{{8}}
9	7	0	−1	1	0	1.06	1.39	.76	{{9}}

Noting that r_j must be included in every subset, $2^{(k-j)}$ subsets are to be checked at step j. Summing over $j = 2, \ldots, k$, there are $2^{(k-1)} - 1$ evaluations of subsets for inclusion. Obviously, this large number of evaluations precludes using the method for large k. On the other hand, the computations are as feasible as all-subsets regression analysis with $k-1$ predictor variables. In the previous example with $k = 9$, the calculations used to find the subsets listed in Table 1 took less than 7 seconds using a 486DX-33 IBM-compatible personal computer, running a program written in SAS/IML® (SAS Institute, Inc. 1989).

Once the sets S_j are found, the maximum dimensions M_j are easily determined. Using the M_j, adjusted p values are easily computed using the Shaffer method with the Bonferroni inequality. In the next section I give a method for incorporating specific dependence structures.

4. INCORPORATING CORRELATIONS

4.1 The Adjusted *p* Values

Using the Bonferroni inequality simplifies the problem of evaluating the adjusted p values precisely. Following Westfall and Young (1993, pp. 71–72), the adjusted p values corresponding to Shaffer's method that incorporate dependencies are defined sequentially as follows:

$$\tilde{p}_{(1)} = \Pr(\min_{l \in \{r_1, r_2, \ldots, r_k\}} P_l \le p_{(1)} | H_0),$$

$$\tilde{p}_{(2)} \leftarrow \max\left[\tilde{p}_{(1)}, \max_{K \in S_2} \Pr\left(\min_{l \in K} P_l \le p_{(2)} \middle| \bigcap_{i \in K} H_i\right)\right],$$

$$\vdots$$

$$\tilde{p}_{(j)} \leftarrow \max\left[\tilde{p}_{(j-1)}, \max_{K \in S_j} \Pr\left(\min_{l \in K} P_l \le p_{(j)} \middle| \bigcap_{i \in K} H_i\right)\right],$$

$$\vdots$$

$$\tilde{p}_{(k)} \leftarrow \max[\tilde{p}_{(k-1)}, \Pr(P_{r_k} \le p_{(k)} | H_{r_k})]. \tag{4}$$

Here P refers to the random (preobserved) p value, and the p refer to the observed p values. Note also that the $\{r_j\}$ are fixed in all probability calculations.

Often, the distributions considered are those of maxima of t statistics rather than minima of p values, but these are equivalent formulations. Using p values is convenient, because one-sided and two-sided testing can be considered under a unifying notation. In addition, if some tests in the family are two-sided and some are one-sided, then using the $\min P$ distribution provides a more "balanced" multiplicity adjustment (Westfall and Young 1993, pp. 50–51).

In the case of t tests of contrasts among location parameters, the partial null conditions $\cap_{i \in K} H_i$ and H_{r_k} may be replaced by the complete null condition $H_0 = \cap_{i \in S} H_i$, simplifying calculations. This is true because the distribution of any subset K of test statistics is identical under $\cap_{i \in K} H_i$ and H_0, a circumstance called the *subset pivotality* condition by Westfall and Young (1993, pp. 42–43).

4.2 Calculating the Adjusted *p* Values

The calculations in (4) require a probability model. Assume the normal homoscedastic linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is a fixed $(n \times p)$ design matrix assumed to be full rank without loss of generality, $\boldsymbol{\beta}$ is a fixed and unknown $(p \times 1)$ parameter vector, and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of iid $N(0, \sigma^2)$ random variables. Where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $s^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$, the t statistic for testing H_i: $\mathbf{c}_i'\boldsymbol{\beta} = 0$ is

$$t_i = \frac{\mathbf{c}_i'\hat{\boldsymbol{\beta}}}{s\{\mathbf{c}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_i\}^{1/2}},$$

which is distributed as Student's t with $n-p$ df under H_i.

Consider the adjusted p value $\tilde{p} = \Pr(\min_{i \in K} P_i \le p)$, where the condition H_0 in (4) is implicit. If the tests are two-tailed, then an equivalent form is $\tilde{p} = \Pr(\max_{i \in K} |T_i| \ge |t|)$, where the T_i and t denote the t statistics corresponding to the random (preobserved) and observed p values.

I propose estimating the adjusted p values via simulation for every K in every index set S_j to estimate (4). To do this, I require an accurate Monte Carlo method, yet one that takes relatively few basic calculations for each adjusted p value. It is important that the algorithm use few basic calculations, as the calculations must be performed for many sets K and many Monte Carlo samples.

The algorithm that I propose is a generalized least squares hybrid of *simple* and *control variate* (CV) Monte Carlo methods. Although less efficient than the method of Naiman and Wynn (1992), this proposed method requires less storage and fewer function calls, an important advantage because the probabilities are to be evaluated for a very large collection of subsets for many Monte Carlo samples.

In the following description, let s denote the simulation counter, with $s = 1, \ldots, N$. Suppose also that the simulated values $T_s^* = (T_{sr_1}^* \quad T_{sr_2}^* \quad \ldots)$ are generated as iid vectors from the multivariate Student's t distribution with degrees of freedom $n-p$ and dispersion matrix $\operatorname{corr}(\mathbf{C}_K'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_K)$. ($\operatorname{Corr}(\cdot)$ maps covariance matrices to correlation matrices.)

The following algorithms are defined in terms of two-sided tests. For one-sided tests, remove the absolute value signs, insert the inequality appropriate for the given alternative, and replace "max" with "min" for lower-tailed tests.

Simple Monte Carlo. Let $x_s = I(\max_j |T_{sr_j}^*| > |t|)$, where $I(\cdot)$ denotes the indicator function. Noting that the x_s are iid with $E(x_s) = \tilde{p}$, a consistent estimate of $\tilde{p}$ is

$$\hat{\tilde{p}}_{\mathrm{S}} = \frac{1}{N}\sum_s x_s,$$

with simulation standard error

$$\mathrm{SE}(\hat{\tilde{p}}_{\mathrm{S}}) = \{\hat{\tilde{p}}_{\mathrm{S}}(1 - \hat{\tilde{p}}_{\mathrm{S}})/N\}^{1/2} = \{\hat{\tau}_{11}/N\}^{1/2}.$$

Control-Variate Monte Carlo. Let $y_s = \sum_j I(|T_{sr_j}^*| > |t|)$ and $z_s = y_s - x_s$. Note that $\tilde{p} = E(y_s) - E(z_s) =$

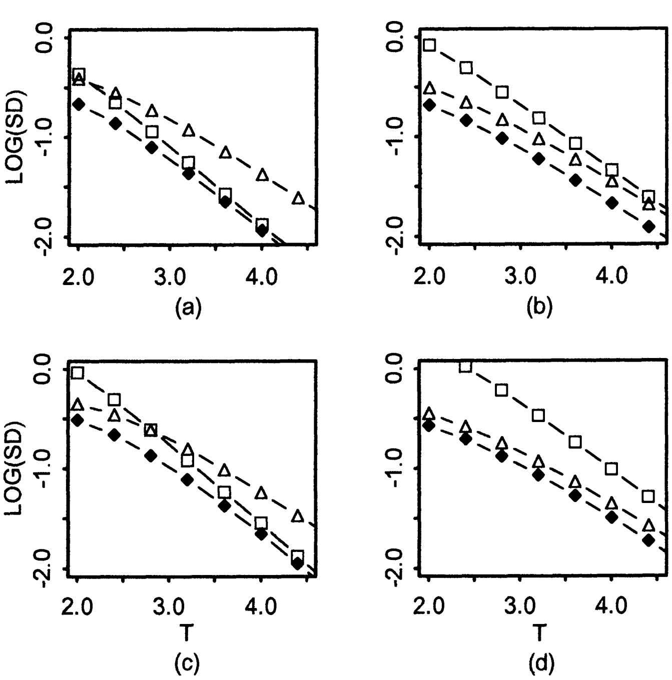

Figure 1. Base 10 Logarithms of Monte Carlo Standard Deviations for Simple (Triangle), Control Variate (Square), and Generalized Least Squares (GLS) Hybrid (Solid Diamond) Estimates of $\tilde p$ = Pr(max T_i > t). The vector $\{T_1, \ldots, T_k\}$ has the equicorrelated multivariate Student T distribution with 30 df; (a) k = 10, ρ = .3; (b) k = 10, ρ = .7; (c) k = 20, ρ = .3; (d) k = 20, ρ = .7. Actual standard errors are obtained by dividing the standard deviation by $N^{1/2}$, where N is the number of Monte Carlo samples. The GLS method is the most efficient.

$|K|p_0 - E(z_s)$, where $p_0 = P(|T_{n-p}| > |t|)$ can be readily evaluated using available software. Because the z_s are iid with finite variance, a consistent estimate of $\tilde p$ is

$$\hat{\tilde p}_{\mathrm{C}} = |K|p_0 - \frac{1}{N}\sum_s z_s,$$

with simulation standard error

$$\mathrm{SE}(\hat{\tilde p}_{\mathrm{C}}) = \left\{\sum (z_s - \bar z)^2\right\}^{1/2} \Big/ N = \{\hat\tau_{22}/N\}^{1/2}.$$

The CV estimate will tend to be more accurate when the variance of z_s is small. Note that z_s measures how many times $|T^*_{sr_j}| \geq |t|$ occurs, less one, when there is at least one occurrence. The variance of this quantity can be large when the correlations among the $T^*_{sr_j}$ are large; in this case the tendency is for either zero occurrences of $|T^*_{sr_j}| \geq |t|$ or $|K|$ occurrences, and the control variate estimate will tend to be inferior to the simple estimate. However, experience has shown that in a wide variety of applications with moderate correlation or large $|t|$, the CV estimate frequently outperforms the simple estimate.

It is difficult to predict which estimate will be more accurate in general. I thus propose a simple combination of the two estimates, which has uniformly smaller variance than either one individually.

Generalized Least Squares Hybrid. Let

$$\operatorname{cov}\begin{pmatrix} X_s \\ Z_s \end{pmatrix} = \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{12} & \tau_{22} \end{pmatrix}.$$

Noting that

$$\begin{pmatrix} \hat{\tilde p}_{\mathrm{S}} \\ \hat{\tilde p}_{\mathrm{C}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tilde p + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix},$$

where

$$\operatorname{cov}\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} = \frac{1}{N}\begin{pmatrix} \tau_{11} & -\tau_{12} \\ -\tau_{12} & \tau_{22} \end{pmatrix},$$

the generalized least squares estimate is obtained:

$$\tilde p_{\mathrm{G}} = \frac{(1 \quad 1)\begin{pmatrix} \tau_{11} & -\tau_{12} \\ -\tau_{12} & \tau_{22} \end{pmatrix}^{-1}\begin{pmatrix} \hat{\tilde p}_{\mathrm{S}} \\ \hat{\tilde p}_{\mathrm{C}} \end{pmatrix}}{(1 \quad 1)\begin{pmatrix} \tau_{11} & -\tau_{12} \\ -\tau_{12} & \tau_{22} \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix}}.$$

To estimate this quantity, simply replace the given covariance parameters with their sample estimates. The calculations are streamlined considerably by noting that $\tilde p_{\mathrm{G}} = a_1\hat{\tilde p}_{\mathrm{S}} + a_2\hat{\tilde p}_{\mathrm{C}}$, where $a_1 + a_2 = 1$. Thus

$$(a_1 \quad a_2) = \frac{(\tau_{22} + \tau_{12} \quad \tau_{11} + \tau_{12})}{\tau_{11} + 2\tau_{12} + \tau_{22}}.$$

The estimate of τ_{12} requires no new simulation counters, as $\hat\tau_{12} = \sum x_s z_s/N - \bar x\bar z = \bar z(1 - \bar x)$. The estimate of $\tilde p_{\mathrm{G}}$ so obtained is denoted by $\hat{\tilde p}_{\mathrm{G}} = \hat a_1\hat{\tilde p}_{\mathrm{S}} + \hat a_2\hat{\tilde p}_{\mathrm{C}}$, and the simulation standard error of this estimate is

$$\mathrm{SE}(\hat{\tilde p}_{\mathrm{G}}) = \left\{(\hat a_1 \quad \hat a_2)\begin{pmatrix} \hat\tau_{11} & -\hat\tau_{12} \\ -\hat\tau_{12} & \hat\tau_{22} \end{pmatrix}\begin{pmatrix} \hat a_1 \\ \hat a_2 \end{pmatrix} \Big/ N\right\}^{1/2}.$$

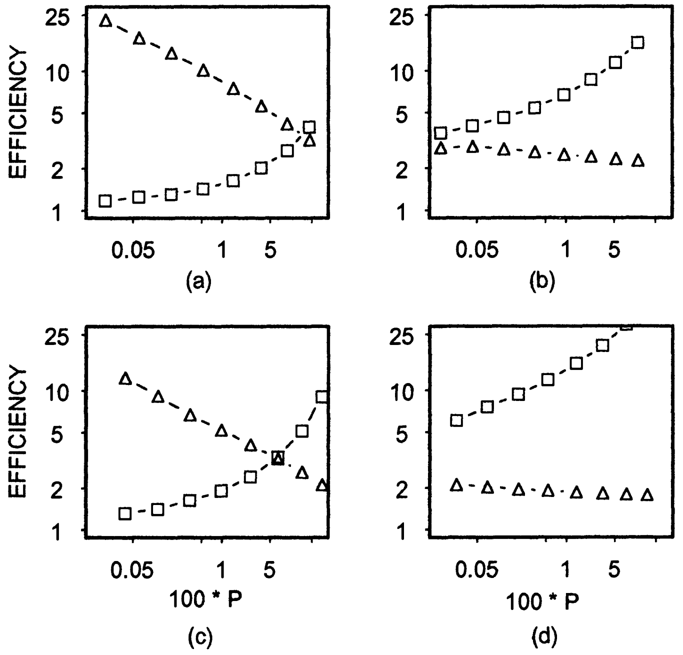

Figure 2. Efficiencies of the GLS Method Relative to the Simple (Triangle), and Control Variate (Square) Methods, Defined as the Ratio of Variances, Calculated Using the Data and Parameter Settings Shown in Figure 1. Both axes are in log scale, and the horizontal axis is 100$\tilde p$, where $\tilde p$ = Pr(max T_i > t). When k = 20 and ρ = .3 in an equicorrelated multivariate Student T distribution with 30 df, the simple method requires approximately 10 times as many Monte Carlo samples to achieve the same margin of error as the GLS method when $\tilde p$ = .0005.

Table 2. p Value Adjustments for Litter Weight Data

Contrast	Raw	Adjusted p values					
j	p value	Bonferroni	Holm	Hochberg	Shaffer	FDR	Proposed[c]
1	.005708	.0514	.0514	.0514	.0514	.0514	.0318
2	.023544	.2119	.1884	.1694[b]	.0514[b]	.0726[b]	.0454
3	.024193	.2177	.1884[b]	.1694	.0726	.0726	.0639
4	.044895	.4041	.2694	.2440[b]	.0898	.0878[b]	.0878
5	.048805	.4392	.2694[b]	.2440	.0976	.0878	.0897
6	.221227	1.0000[a]	.8849	.7758[b]	.4424	.3318	.3946
7	.395867	1.0000[a]	1.0000[a,b]	.7758[b]	1.0000[a]	.5090	.7276
8	.693051	1.0000[a]	1.0000[a]	.7758[b]	1.0000[a,b]	.7276[b]	.7276[b]
9	.775755	1.0000[a]	1.0000[a,b]	.7758	1.0000[a,b]	.7758	.7758

[a] Truncated to 1.0.

[b] Monotonicity enforced.

[c] Maximum 99.7% margin of error is .0003.

Figures 1 and 2 display how the three methods compare. I estimated $\tilde{p} = \Pr(\max T_i > t)$ for various t, where the $\{T_i; i = 1, \ldots, k\}$ are from an equally correlated multivariate t distribution with 30 df. All calculations are based on $N = 10{,}000{,}000$ Monte Carlo samples. When the correlation (ρ) is small and/or t is large, the CV approach outperforms the simple method. In all cases, the GLS method dominates.

As a final note, when $k = 1$, both the CV and GLS methods provide exact answers to the trivial problem, but the simple Monte Carlo method provides only an approximation.

4.3 The Example and Comparisons

Return to the example of Section 3.1. The data necessary to carry out the method are summarized as follows:

$$\hat{\beta}_1 = \begin{pmatrix} -48.757 \\ -52.109 \\ -51.049 \\ -51.434 \end{pmatrix}, \qquad \text{MSE} = 15.978, \qquad \text{df} = 68,$$

and

$$[(\mathbf{X}'\mathbf{X})^{-1}]_{11} = \begin{pmatrix} 37.586 & 37.759 & 37.248 & 37.690 \\ 37.759 & 38.036 & 37.468 & 37.915 \\ 37.248 & 37.468 & 37.021 & 37.397 \\ 37.690 & 37.915 & 37.397 & 37.905 \end{pmatrix}.$$

(The "1" and "11" subscripts denote that these are the relevant submatrices. Covariate estimates and their covariances are not needed.)

The summary of the tests and the raw and adjusted p values are given in Table 2. For comparison, adjusted p values are also displayed using competing methods. The proposed adjustments (rightmost column) are evaluated using 10,000,000 Monte Carlo samples.

Comparing the "Holm" and "Shaffer" columns, which are Bonferroni adjustments using multipliers $k - j + 1$ and M_j, a clear benefit to incorporating logical constraints is evident: Five tests are significant at the $\alpha = .10$ level using Shaffer, but only one with Holm. The proposed method finds all of Shaffer's $\alpha = .10$ significances and two $\alpha = .05$ significances not found using Shaffer. The reason for the discrepancy between the .0514 and .0318 adjustments is that correlations between the $k = 9$ estimated contrasts are incorporated into the multiplicity adjustment for the proposed method.

It is also noteworthy that in many cases, the Shaffer method is a good approximation to the more precise proposed method. For example, comparing the adjustments for contrasts 4 and 5, there is relatively little difference. In these cases the contrast set used for multiplicity adjustment has only two elements, and apparently the correlations between these elements are not large enough to make a substantial difference in the proposed method.

Note also that Benjamini and Hochberg's step-up method (FDR in the table) produces larger adjusted p values at the low range in this example. Despite the fact the FDR procedure aims to control a much less stringent error rate, this method does not necessarily produce more significances than the proposed method. This is because the FDR procedure does not account for correlations and logical constraints.

Table 3 displays all precise comparisons that can be stated generally between the adjusted p values compared in Table 2. All entries labelled "$\geq$" indicate that the row procedure produces adjusted p values that are uniformly as large as the column procedure; entries labelled "$>, <$" indicate that there is no such uniform ordering. The proposed method produces uniformly smaller p values than do all but the Hochberg and FDR methods, but these latter methods are guaranteed to work properly only in the independence case. Further, the FDR procedure controls a different error rate.

5. PROPERTIES AND EXTENSIONS

The proposed method has a nice consistency property not shared by most competing methods. If hypotheses in

Table 3. Uniform Comparisons of Adjusted p Values

	Holm	Hochberg	Shaffer	FDR	Proposed*
Bonferroni	$\geq$	$\geq$	$\geq$	$\geq$	$\geq$
Holm		$\geq$	$\geq$	$\geq$	$\geq$
Hochberg			$>, <$	$\geq$	$>, <$
Shaffer				$>, <$	$\geq$
FDR					$>, <$

* Assumes exact evaluation.

$K \subseteq \{1, \dots, k\}$ are true and the rest false, for $K \neq \emptyset$, then the actual FWE level converges to the nominal, provided that the noncentrality parameters of the test statistics tend to infinity (in the correct direction, for one-sided tests). This result was proven by Westfall and Young (1993, 213–214) and is not necessarily sample-size driven: Noncentrality parameters will also tend to infinity in a study where one wishes to assess the performance of tests where the effect sizes get large and the sample size remains constant. This consistency property of the proposed method is not shared by any method that uses a conservative probability inequality (such as Bonferroni) for its base.

A useful feature of the proposed method is that it can be readily adapted to the analysis of nonnormal and/or heteroscedastic data using resampling methods. To evaluate $\tilde{p} = \Pr(\max_{i \in K} |T_i| \geq |t|)$ in a nonnormal situation, a model is required. If we assume a location shift linear model whose residuals are distributed as F, then $\tilde{p} = \Pr(\max_{i \in K} |T_i| \geq |t| | F) \approx \Pr(\max_{i \in K} |T_i| \geq |t| | \hat{F})$, where $\hat{F}$ is the empirical distribution of the least squares residuals. The probability may then be simulated by drawing bootstrap samples $\mathbf{Y}^*$ from the residuals, computing the test statistics T_i^* using the $\mathbf{Y}^*$ data, and tabulating the proportion of such Monte Carlo samples yielding $\max_{i \in K} |T_i^*| \geq |t|$. The resulting adjustments are usually more robust to nonnormality than are the normality-based adjustments.

Another possibility in multiple-group ANOVA is to draw bootstrap samples from each individual group's residuals, thereby incorporating possible heteroscedastic effects into the multiplicity adjustments. Details on the location shift and heteroscedastic bootstraps, as they apply to this problem, were given by Westfall and Young (1993).

6. CONCLUSIONS

Stepwise testing methods can be made more powerful by incorporating logical constraints and correlations. The degree of improvement depends on how the constraints reduce the cardinalities of the S_j and on the dependence structure among tests. Those applications with many possible restrictions and large correlations among test statistics will exhibit the most improvement. For a moderate number of tests, it is computationally feasible both to identify the logically constrained subsets and to evaluate the adjusted p values. The SAS/IML (SAS Institute Inc. 1989) code for calculating the proposed normal-based homoscedastic multiplicity adjustments of this article is available from http://lib.stat.cmu.edu/jasasoftware/mtest.

APPENDIX: PROOFS

Lemma. Let $\mathbf{A} = \{a_{ij}\}$ be an $r \times c$ matrix having at least one nonzero element in each row. Then there exists $\mathbf{v} \in \mathcal{C}(\mathbf{A})$ such that all elements of $\mathbf{v}$ are nonzero.

Proof. Let the columns of $\mathbf{A}$ be $\mathbf{A}_j, j = 1, \dots, c$. Now choose $x_i, i = 1, \dots, c$, as iid random variables, uniformly distributed on [0, 1], and consider $\mathbf{v} = \sum x_j \mathbf{A}_j$. Because $v_i = \sum_j x_j a_{ij} = 0$ occurs only on a set of measure zero, we have that $\cup_i \{\{x_1, \dots, x_c\}; v_i = 0\}$ also has measure zero, and thus $\cap_i \{\{x_1, \dots, x_c\}; v_i \neq 0\}$ has measure 1. Thus there exists $\mathbf{v} \in \mathcal{C}(\mathbf{A})$ for which $v_i \neq 0$, all $i = 1, \dots, c$. (In fact, *almost all* elements of $\mathcal{C}(\mathbf{A})$ have this property.)

Proposition. Suppose that the hypotheses tested are linear contrasts of a $p \times 1$ vector $\boldsymbol{\mu}$ whose range is $\Re^p$. A necessary and sufficient condition for

$$\left(\bigcap_{l \in K} H_l\right) \cap \left(\bigcap_{l=1}^{j-1} H'_{(l)}\right) \neq \emptyset$$

is that $\mathbf{c}_{(i)} \notin \mathcal{C}(\mathbf{C}_K)$, for all $i = 1, \dots, j-1$.

Proof. Suppose that $\mathbf{c}_{(i)} \in \mathcal{C}(\mathbf{C}_K)$ for at least one $i = 1, \dots, j-1$. Then $\mathbf{c}_{(i)} = \sum_{l \in K} a_l \mathbf{c}_l$, and $(\cap_{l \in K} H_l)$ implies $\mathbf{c}'_l \boldsymbol{\mu} = 0$, all $l \in K$. In this case $\mathbf{c}'_{(i)} \boldsymbol{\mu} = 0$ and $(\cap_{l \in K} H_l) \cap (\cap_{l=1}^{j-1} H'_{(l)}) = \emptyset$. This contrapositive argument proves the necessity portion of the theorem.

To prove sufficiency, suppose that $\mathbf{c}_{(i)} \notin \mathcal{C}(\mathbf{C}_K)$, all $i = 1, \dots, j-1$. We must show only that there exists $\boldsymbol{\mu} \in \mathcal{C}^{\perp}(\mathbf{C}_K)$ where $\mathbf{C}'_{(j)} \boldsymbol{\mu}$ has all nonzero elements, where, as before, the columns of $\mathbf{C}_{(j)}$ are the vectors $\mathbf{c}_{(i)}, i = 1, \dots, j-1$. Letting $\mathbf{w}_1, \dots, \mathbf{w}_m$ denote an orthonormal basis for $\mathcal{C}^{\perp}(\mathbf{C}_K)$, we have $\mathbf{c}_{(i)} = \mathbf{v}_i + \sum_j a_{ij} \mathbf{w}_j$, where $\mathbf{v}_i \in \mathcal{C}(\mathbf{C}_K)$ and $a_{ij} \neq 0$ for at least one j. Now $\boldsymbol{\mu} = \sum b_i \mathbf{w}_i$, for arbitrary b_i. Thus $\mathbf{C}'_{(j)} \boldsymbol{\mu} = \mathbf{A}\mathbf{b}$, where $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{b} = \{b_i\}$, and each row of $\mathbf{A}$ has at least one nonzero element. Applying the lemma, we conclude that there exists $\mathbf{b}$ such that $\mathbf{A}\mathbf{b}$ has all elements nonzero; hence $(\cap_{l \in K} H_l) \cap (\cap_{l=1}^{j-1} H'_{(l)}) \neq \emptyset$, and the proposition is proved.

Corollary. The sets $\bar{S}_j$ of (2) (and hence S_j) are nonempty, provided that no pairs $\mathbf{c}_i, \mathbf{c}_j$ are collinear for $i \neq j$.

Proof. It will suffice to show that $\{r_j\} \in \bar{S}_j$, all $j = 1, \dots, t$. Provided that there are no pairwise collinearities, this result is implied by the proposition.

[Received August 1995. Revised May 1996.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Dunnett, C. W., and Tamhane, A. C. (1991), "Step-Down Multiple Tests for Comparing Treatments With a Control in Unbalanced One-Way Layouts," *Statistics in Medicine*, 10, 939–947.

——— (1992), "A Step-Up Multiple Test Procedure," *Journal of the American Statistical Association*, 87, 162–170.

——— (1995), "Step-Up Multiple Testing of Parameters With Unequally Correlated Estimates," *Biometrics*, 51, 217–227.

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, 417–424.

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

Naiman, D. Q., and Wynn, H. P. (1992), "Inclusion-Exclusion-Bonferroni Identities and Inequalities for Discrete Tube-Like Problems via Euler Characteristics," *The Annals of Statistics*, 20, 43–76.

Rom, D. M., and Holland, B. (1995), "New Step-Down and Closed Bonferroni Multiple Testing Procedures for Hierarchical Families of Hypotheses," *Journal of Statistical Planning and Inference*, 46, 265–275.

SAS Institute, Inc. (1989), *SAS/IML® Software: Usage and Reference, Version 6*, Cary, NC: Author.

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.

Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, New York: Wiley-Interscience.

Wright, S. P. (1992), "Adjusted P-Values for Simultaneous Inference," *Biometrics*, 48, 1005–1013.