



Multiple Tests with Discrete Distributions

Author(s): Peter H. Westfall and Russell D. Wolfinger

Source: *The American Statistician*, Vol. 51, No. 1 (Feb., 1997), pp. 3-8

Published by: American Statistical Association

Stable URL: <http://www.jstor.org/stable/2684683>

Accessed: 20/07/2009 14:31

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=astata>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.



American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*.

<http://www.jstor.org>

Multiple Tests with Discrete Distributions

Peter H. WESTFALL and Russell D. WOLFINGER

We review special issues in multiplicity adjustment where the sampling distributions are discrete. These include (1) incorporating discreteness into the multiplicity adjustments, (2) incorporating correlations versus using Bonferroni or independence-based approximations, and (3) using discrete tails in two-sided tests. Incorporating discrete characteristics can greatly improve the power of the tests that maintain a given familywise error rate. Use of correlations also can improve the power, but it is shown that independence-based multiplicity adjustment is not necessarily a conservative procedure. Exact methods that incorporate discreteness and correlations are generally recommended.

KEY WORDS: Adjusted p value; Bonferroni inequality; Familywise error rate; Multinomial data; Permutation distribution; Stepwise methods.

1. INTRODUCTION

Multiple testing involves a family of hypotheses H_1, \dots, H_k (in null form; alternatives are H'_j), test statistics t_1, \dots, t_k , and a decision rule, such as $t_j \geq t_j(\alpha)$, for accepting or rejecting the H_j based on the values of the t_j . The problem with multiple tests is the large probability of rejecting null hypotheses incorrectly when many or all the nulls in the family are true. To control the familywise error rate (FWE; also called experimentwise error rate) the α levels for individual tests can be adjusted. Hochberg and Tamhane (1987) provide a good general reference for the multiple testing problem. There is controversy over whether multiple testing should be done (e.g., Saville 1990); nevertheless, multiple testing methods continue to be used by practitioners to avoid the pitfalls of "data snooping." Although the issue of whether to adjust is controversial, there is little argument that multiple testing methods should be as powerful as possible when familywise error rates are controlled. The purpose of this paper is to highlight power improvements that are possible when testing with discrete data.

Often, the sampling distributions of the t_j are continuous, having Student's t , F , approximately normal, or related continuous distributions (such as that of $|T|$) when t_j is used to

test two-sided alternatives. Assuming the cumulative distribution function of t_j is F_j , the p values (one- or two-sided) may be defined as $p_j = 1 - F_j(t_j)$, and are uniformly distributed under H_j . In such a case the Bonferroni decision rule is to reject H_j when $p_j \leq \alpha/k$ or, using adjusted p values, when $p'_j \leq \alpha$, where $p'_j = kp_j$ defines the adjusted p value.

It is well known that the Bonferroni method is conservative when the p values are uniformly distributed. There are two reasons for this: (1) the Bonferroni method ignores correlation structure among the p_j , which can be exploited to provide less conservative critical values, and (2) the Bonferroni method makes no allowance for situations where some H_j are clearly false (based on very large t_j). Tukey's method for all pairwise comparisons in the one-way analysis of variance (ANOVA), and more general approaches given by Edwards and Berry (1987) and Naiman and Wynn (1992), are examples of methods that overcome (1). Stepwise testing methods (e.g., Hochberg 1988; Holm 1979; Shaffer 1986) are designed to lessen problem (2), although all remain conservative to a degree.

It is less well known that the Bonferroni method can be conservative due to discreteness of the sampling distributions. This problem potentially is much worse than either of problems (1) or (2). Fortunately, it is easy to incorporate discreteness into the multiplicity adjustments so that the FWE remains protected, yet the tests are potentially much more powerful.

In Section 2 we give the basic adjustment procedures, and compare the continuous and discrete adjustments for the case of independent tests. In Section 3 special considerations needed for two-sided tests are presented. Section 4 presents the effect of incorporating correlations, and Section 5 reviews stepwise testing methods that can improve power further. An application to multiple tests with multinomial data is given in Section 6, and Section 7 presents a simulation study to compare power.

2. THE DISCRETE METHOD

When tests are discrete the statistics t_j often refer to counts or rank sums, and are in different scales. The distribution of $\max t_i$, a staple for typical multiple testing applications, is inappropriate in this case. Instead, it is convenient to define multiplicity adjustments using the distribution of the minimum of the p values because the p values are on the same scale. Hence we define the adjusted p values

$$p'_j = \Pr(\min_{i=1, \dots, k} P_i \leq p_j), \quad (1)$$

Peter H. Westfall is Professor, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409. Russell D. Wolfinger is Senior Research Statistician, SAS Institute Inc., Cary, NC 27513.

Table 1. Data and Upper Tail p Values for $k = 4$ Fisher Exact Tests

Observed data				
Group	Variable 1	Variable 2	Variable 3	Variable 4
Control	0/50	4/50	0/50	6/50
Treated	5/48	3/48	4/48	4/48
Permutation distributions				
Test statistic, t	Variable 1: Total = 5	Variable 2: Total = 7	Variable 3: Total = 4	Variable 4: Total = 10
0	1.00000	1.00000	1.00000	1.00000
1	.96880	.99278	.93625	.99927
2	.80602	.93765	.67580	.99068
3	.48047	.76489*	.29327	.94744
4	.16848	.47697	.05387*	.82409*
5	.02521*	.20129		.60332
6		.04967		.34428
7		.00532		.14250
8				.03946
9				.00645
10				.00047

* Denotes observed p value.

where the P_i refer to the random p values considered under their respective null hypotheses. An additional justification for the use of $\min P_i$ for multiplicity adjustment is that it measures the degree of “surprise” that the analyst should experience after isolating the smallest p value from a long list of p values calculated from a given data set.

Such adjustment using the $\min P_i$ distribution is widely used in the analysis of toxicology data; see Mantel (1980), Farrar and Crump (1988), Heyse and Rom (1988), and Rom (1992). However, this method has much broader scope. The Tukey and Dunnett methods for pairwise comparisons follow from (1) under the homoscedastic normal ANOVA model; the Bonferroni method also is obtained from (1) and the Bonferroni inequality. Additionally, (1) can be modified easily to incorporate stepwise procedures; an example is given in Section 6. There is a simple interpretation of the adjusted p value in (1): p'_j is the probability that a p value as small as p_j will be observed in the entire study when all null hypotheses are true. Westfall and Young (1993) describe various adjustment procedures that follow from (1).

When the distributions of the t_i are discrete, then the observable values of the random p value P_i are $\{p_{it}; t = 1, \dots, m_i\}$, with $\Pr(P_i \leq p_{it}) = p_{it}$. Now, assuming the p values are independent, (1) yields

$$p'_j = 1 - \prod_{i=1}^k (1 - p_{it(j)}), \quad (2)$$

where

$$p_{it(j)} = \begin{cases} \max_t \{p_{it}; p_{it} \leq p_j\}, & \text{if } \min_t \{p_{it}\} \leq p_j \\ 0, & \text{otherwise.} \end{cases}$$

To see the effects of incorporating discreteness in (2), consider the data in Table 1, where $k = 4$ 2×2 contingency tables are considered. There are $n_C = 50$ and $n_T = 48$ observations in the “control” and “treated” groups. The test statistics t_j are the number of occurrences in the treated group, and the upper tail of the permutation distribution

(using the Fisher exact test, conditioning on the total number of occurrences in treated and control groups) of each t_j is given.

The smallest observed upper tail p value is $p_1 = .02521$. Under independence, and incorporating the discrete characteristics, this p value is multiplicity-adjusted to $p'_1 = 1 - (1 - .02521)(1 - .00532)(1 - 0)(1 - .00645) = .03665$, significant at the familywise $\alpha = .05$ level. On the other hand, if the p values are assumed to be uniformly distributed, then (1) yields $p'_1 = 1 - (1 - .02521)^4 = .09709$, barely significant at the familywise $\alpha = .10$ level.

The magnitude of the improvement depends upon the specific characteristics of the discrete distributions. Larger gains are possible when k is large, and where many variables are sparse, like Variable 3 in the previous example. On the other hand, if all distributions are nearly continuous, implying uniformly distributed p values, then $p_{it(j)} \approx p_j$. In this case from (2) we have $p'_j \approx 1 - (1 - p_j)^k$, and there is little difference between the continuous and discrete case.

Tarone (1990) recognized this problem, and proposed a modified Bonferroni method that incorporates only those variables that can possibly contribute to the multiplicity adjustment. In animal carcinogenicity studies where this method sometimes is used, it is common to observe $k = 30$ or more tumor types, each of which must be tested individually for a dose-response relationship. However, it is also common in such studies that only 15 or fewer tumor types are frequent enough to contribute to multiplicity adjustments; the remaining types, like Variable 3 in Table 1, effectively are removed from the list. In this case the Bonferroni multiplier is reduced greatly, from approximately 30 to approximately 15. Incorporating the precise discrete characteristics can bring the multiplier down even further, as seen with the example data of Table 1.

3. TWO-SIDED DISCRETE MULTIPLE TESTS

When tests are two-sided, discrete characteristics can be exploited further by using “reflected” tail probabilities. Where the null expectation of the test statistic T_i is μ_i , the reflected value $R_i = 2\mu_i - T_i$ locates the point corresponding to T_i in the opposite tail of the distribution. The two-tailed p value is then defined as

$$p_i = \begin{cases} \Pr(T_i \leq t_i) + \Pr(T_i \geq r_i), & \text{for } t_i < \mu_i \\ \Pr(T_i \geq t_i) + \Pr(T_i \leq r_i), & \text{for } t_i > \mu_i \\ 1.0, & \text{for } t_i = \mu_i \end{cases} \quad (3)$$

where r_i denotes the observed value of the random R_i . There is some controversy regarding use of such two-tailed p values; Yates (1984) recommends simply doubling the single-tail probability. However, the reflected p values have the nice property that $\Pr(P_i \leq p_i) = p_i$, unlike the doubled p values. Additionally, the reflected values are commonly reported by statistical software packages, and thus have become standard in usage.

Because the reflected p values are typically smaller than the doubled p values, they provide higher power. This is particularly important in multiple testing applications where power is needed.

Table 2. Two-Tail p Values for $k = 4$ Fisher Exact Tests

Test statistic, t	Variable 1: $\mu_1 = 2.45$	Variable 2: $\mu_2 = 3.43$	Variable 3: $\mu_3 = 1.96$	Variable 4: $\mu_4 = 4.90$
0	.05641	.01254	.11762	.00120
1	.36246	.11202	.61747	.01577
2	1.00000	.43640	1.00000	.09202
3	.67445	1.00000*	.35702	.31841
4	.19968	.71208	.05387*	.74096*
5	.02521*	.26364		1.00000
6		.05689		.52019
7		.00532		.19506
8				.04878
9				.00718
10				.00047

* Denotes observed p value.

Table 2 shows the reflected two-sided p values for the data of Table 1. For each 2×2 contingency table the null expected value of the test statistic is $\mu_i = 48 \times \text{Total}_i / 98$, where Total_i is the total number of occurrences in treated and control groups for variable i . The two-tailed p values are calculated using (3) and the distributions shown in Table 1 (from which the needed lower tailed probabilities in (3) also can be obtained).

Under independence the smallest p value .02521 is multiplicity-adjusted to $p'_1 = 1 - (1 - .02521)(1 - .01254)(1 - 0)(1 - .01577) = .05261$, almost significant at the family-wise $\alpha = .05$ level. If the two-tailed p values are defined by doubling the single-tail area, and then are multiplicity-adjusted as if they were uniformly distributed, then $p'_1 = 1 - (1 - 2 \times .02521)^4 = .18693$, substantially larger than .05261. These results suggest that large improvements in power may be obtained by incorporating discrete characteristics in the unadjusted p values, as well as in the multiplicity adjusted p values.

4. INCORPORATING CORRELATION STRUCTURES

Using correlation information makes Bonferroni-style multiplicity adjustments less conservative. Although less conservative than the Bonferroni method, the independence-assuming method also is conservative, provided certain positive correlation conditions are met (e.g., Jogdeo 1977; Slepian 1962; Šidák 1967). However, with multivariate discrete data the independence-based multiplicity adjustments (2) can easily be anticonservative. Consider, for example, a multinomial two-group experiment with data as shown in Table 3, where upper tail tests are desired. The comparison of Control versus Treated in category "A" using the Fisher exact test yields $p_A = 4/56 = .071429$. Only the test for category "C" can achieve this low a p value in its permutation distribution; therefore, the independence-assuming adjusted p value (2) is $p'_A = 1 - (1 - 4/56)^2 = .137755$. Note that $\Pr(P_A \leq 4/56 \text{ and } P_C \leq 4/56) = 0$ because it is impossible for both categories to have three occurrences in the Treated group. Hence using (1) we have $p'_A = \Pr(\min(P_A, P_B, P_C) \leq 4/56) = \Pr(P_A \leq 4/56) + \Pr(P_C \leq 4/56) - \Pr(P_A \leq 4/56 \text{ and } P_C \leq 4/56) =$

Table 3. Multinomial Response Example

Group	Binary response vectors		
	A	B	C
Control	0	1	0
Control	0	0	1
Control	0	0	1
Control	0	0	1
Treated	1	0	0
Treated	1	0	0
Treated	1	0	0
Treated	0	1	0

Group	Number in response category			Total
	A	B	C	
Control	0	1	3	4
Treated	3	1	0	4

$2(4/56) = .142857$, and the independence assumption is seen to be anticonservative in this case.

It is simple to bound the true value of (1) using the Bonferroni inequality to avoid anticonservative adjustments. The discrete Bonferroni adjusted p values are

$$p'_j = \sum_{i=1}^k p_{it(j)}. \tag{4}$$

Rather than use the independence-assuming adjustments (2) or the Bonferroni adjustments (4), it is preferable to incorporate dependence structures exactly. In some cases (such as the multinomial example above) the multiple tests refer to multiple variables measured on the same experimental unit, with just one test per variable. In such a case the adjusted p values (1) can be calculated from the permutation distribution of the observation vectors (Table 3 shows the vectors that are permuted in that example). Although it is feasible with small sample sizes and/or small totals to directly enumerate the set of outcomes leading to $\min P_i \leq p_j$, it is difficult to do so with larger sample sizes. Nevertheless, the probabilities may be estimated easily by simulating from this multivariate permutation distribution, drawing enough samples to estimate the adjusted p values with accuracy sufficient for the purposes of the experiment. "Sufficient accuracy" is determined by the simulation standard error of the adjusted p value, which is given by $\{p'_N(1 - p'_N)/N\}^{1/2}$, where p'_N is the adjusted p value obtained from N samples.

The condition under which it is valid to use the global randomization distribution to adjust the p values for all tests is called subset pivotality by Westfall and Young (1993, p. 42). This condition requires that the joint distribution of

Table 4. Multinomial Responses Measuring Respiratory Health

Group	Number in response category					Total
	Very poor	Poor	Fair	Good	Excellent	
Placebo	12	3	17	9	16	57
Active	1	8	12	9	24	54

Table 5. Two-Sided Unadjusted and Step-Down Adjusted p Values Comparing Frequencies in Active and Placebo Groups

Category	Adjusted					
	Unadjusted		Independence			
	Exact	Doubled	Uniform	Doubled	Permutation	Exact*
Very Poor	.00204	.00270	.0102	.0135	.0066	.0065 (.00003)
Poor	.11787	.17070	.3136	.4297	.2525	.2457 (.00014)
Fair	.39446	.48772	.6333	.7376	.5803	.5707 (.00016)
Good	1.00000	1.00000	1.0000	1.0000	1.0000	1.0000 (.00000)
Excellent	.07931	.10964	.2815	.3716	.2322	.2148 (.00013)

* 99% margin of Monte Carlo error in parentheses.

the subvector $\{P_i; i \in K\}$ is identical under the restrictions $\{H_i; i \in K \text{ are true}\}$ (a partial null hypothesis) and $\{H_i; i = 1, \dots, k \text{ are true}\}$ (the complete null hypothesis), for all $K \subset \{1, \dots, k\}$. When this condition is true it is permitted to compute the adjustments (1) under the computationally convenient complete null hypothesis.

When the multiple tests refer to multiple comparisons of the same variable over different combinations of treatment groups, then the subset pivotality condition fails, and the probabilities in (1) must be computed from the appropriately restricted randomization distributions. If the adjustments are calculated from the global randomization distribution, then there is a risk of excess Type I errors, as discussed by Petrondas and Gabriel (1983).

5. STEPWISE METHODS

Stepwise methods provide a further enhancement to improve the power of multiple testing methods. These techniques are not unique to discrete distributions, but must be mentioned because they improve the power of the tests. The p values in (1) can be adjusted in step-down fashion by adjusting the smallest p value according the $\min P$ distribution, while the second-smallest is adjusted according to the distribution of $\min P$ over all variables excluding the variable whose unadjusted p value was smallest, and so on. At each step the $\min P$ distribution incorporates fewer and fewer variables. The probabilities can be calculated by sampling from the multivariate permutation distribution, as described by Westfall and Young (1993).

Step-up methods are more powerful in some cases (although not uniformly so) than step-down methods. References based on Bonferroni inequalities and/or independence assumptions include Simes (1986) and Hochberg (1988). Step-up methods for use with discrete distributions are under development; Troendle (1996) makes progress in this direction.

6. AN APPLICATION

Data from an experiment on the efficacy of a respiratory therapy given by Koch, Carr, Amara, Stokes, and Uryniak (1990) are partially summarized in Table 4. We analyze a rating of respiratory health as a multinomial response, and compare the rating categories for the active and placebo groups.

Table 5 compares the results of step-down testing of the Koch et al. data using two-sided tests for each variable.

Among the adjusted p values the "Exact" method is preferred. These values are calculated in step-down fashion by sampling 10,000,000 times from the multivariate permutation distribution using PROC MULTTEST of SAS Institute Inc. (1995). (The execution time for this analysis was approximately 4.0 hours using an IBM-compatible with the Intel Pentium 60 MHz processor chip.) Margins of error (99%) for these Monte Carlo estimates are given in parentheses.

For comparison, step-down adjustments using (2) (the "Independence" columns) also are given in several forms. The "Uniform" adjustments assume that the unadjusted p values are from a uniform distribution. The "Doubled" adjustments use the "doubled" unadjusted two-sided p values, and are multiplicity-adjusted as if these values were uniformly distributed. The "Permutation" method uses (2) and the permutation distributions.

The value of incorporating discreteness appears again in this example. Assuming a uniform distribution for the p values, the "Uniform" and "Doubled" tests are not significant at the familywise $\alpha = .01$ level. By using the discrete nature of the distributions, significance is attained at this level, both by the "Independence Permutation" method and the "Exact" method. We see also by comparing the "Independence Permutation" and "Exact" methods that the effect of incorporating correlations is small in this example.

As a final note the independence adjustment is seen to be conservative in this example, not anticonservative as shown in the multinomial example of Section 4. The reason is that the tests are two-sided. With multinomial observations two-sided p values are positively correlated: a large value of t_i tends to be associated with a small value of t_j for some $j \neq i$, implying that the two-sided p values p_i and p_j are both small. With positively correlated p values incorporating dependence structure typically makes the multiplicity adjustments less conservative than independence-assuming multiplicity adjustments.

7. POWER INVESTIGATION

The discrete multiple testing method improves the power of the test without sacrificing FWE control. To investigate the degree of FWE control and the power improvements we performed a simulation analysis using the data set of Section 6 to select parameters.

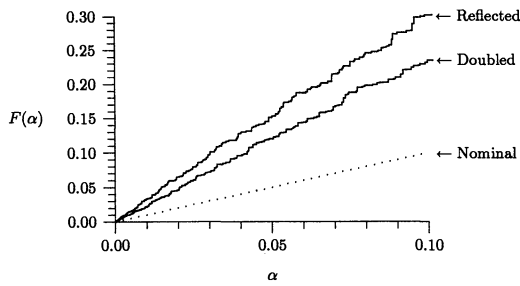


Figure 1. Tail Portions of Estimated Cumulative Distributions of $\min P_j$ for Reflected and Doubled p Values.

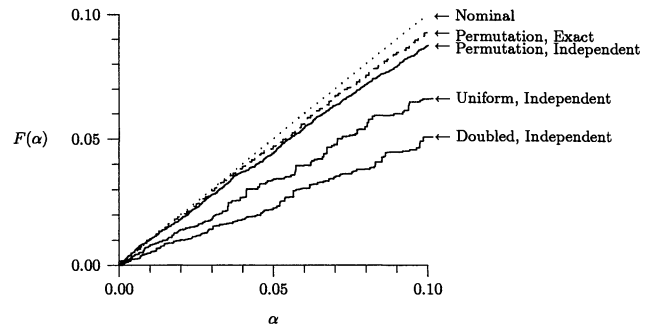


Figure 2. Tail Portions of Estimated Cumulative Distributions of $\min P'_j$ for Multiplicity-Adjusted p Values.

First, we selected a complete null configuration where the multinomial probabilities were chosen as the pooled probabilities from the data shown in Table 4. Then, we investigated the FWE characteristics of the six testing methods (two unadjusted and four adjusted) shown in Table 5. Figure 1 shows the probability (estimated using 10,000 simulated data sets) that the minimum of the five p values is less than α for the exact and doubled p values, not adjusted for multiplicity. It is seen that these methods do not control the FWE because their cumulative distribution functions lie entirely above the $F(\alpha) = \alpha$ line. On the other hand, the adjusted methods all control the FWE, as shown in Figure 2, with the “Exact” method having the closest approximation to the nominal $F(\alpha) = \alpha$ line. Note that all values are below the nominal line. Permutational testing methods are generally conservative because the p values can never attain the α level exactly (e.g., Lancaster 1961; Reid 1995). The “Exact” permutational method was computed using 1,000 resampled data sets for each of the 10,000 simulations.

Second, we evaluated the power of the various methods by selecting alternative parameter configurations. To simulate “small,” “medium,” and “large” effects we changed the probability of a “Very Poor” response in the Placebo group to the pooled probability (13/111) plus .03, .06, and .09, yielding probabilities .147, .177, and .207, respectively.

We changed the corresponding probabilities in the “Active” group to $13/111 - .03 = .087$, $13/111 - .06 = .057$, and $13/111 - .09 = .027$, respectively. The probabilities for the “Excellent” categories were similarly adjusted to make all probabilities add to one. No other probabilities were changed, so that the null hypothesis of no treatment difference is true for three out of the five tests considered.

An additional parameter setting was considered in which the population probabilities for each group were set equal to the observed sample proportions; for example, the population proportion of “Very Poor” responses in the Placebo group was taken to be 12/57. We called this setting the “Post Hoc” alternative.

Table 6 shows the results of comparing power and FWE levels for .05 level tests using the “ $\pm .03$,” “ $\pm .06$,” and “ $\pm .09$ ” configurations. Again, the FWE levels of the unadjusted tests are larger than the nominal .05 level and the levels of the adjusted tests are smaller than the .05 level; further, among the adjusted tests, the “Exact” procedure most closely approximates the nominal level.

To compare power we calculated the average number of correct rejections per simulated data set. For the $\pm .03$, $\pm .06$, and $\pm .09$ configurations this average is between 0 and 2;

Table 6. FWE's and Average Number of Rejections in 10,000 Simulations for Alternative Parameter Configurations

Familywise error rates						
Alternative	Unadjusted		Adjusted			
	Exact	Doubled	Independence			Exact*
			Uniform	Doubled	Permutation	
$\pm .03$.100	.078	.021	.014	.028	.030
$\pm .06$.103	.078	.025	.017	.034	.036
$\pm .09$.104	.076	.027	.020	.037	.038
Correct rejections						
Alternative	Unadjusted		Adjusted			
	Exact	Doubled	Independence			Exact*
			Uniform	Doubled	Permutation	
$\pm .03$.201	.176	.058	.046	.074	.076
$\pm .06$.653	.611	.309	.259	.360	.369
$\pm .09$	1.294	1.262	.908	.824	.973	.985
Post Hoc	1.812	1.677	1.210	1.075	1.321	1.332

* Based on 1000 resampled data sets each.

however, for the “Post Hoc” configuration all nulls are false, so the average may lie anywhere between 0 and 5.

Of course, the unadjusted procedures have the highest power, at the expense of not controlling the FWE. Among the adjusted procedures the exact method has the highest power, but the power is only slightly higher than the independence-assuming adjustments in this example. Note also that the “Uniform” and “Doubled” adjustments lose substantial power. Thus power of multiple test procedures can be improved substantially by incorporating the discrete characteristics of the distributions of the test statistics.

[Received January 1996. Revised August 1996.]

REFERENCES

- Edwards, D., and Berry, J. J. (1987), “The Efficiency of Simulation-Based Multiple Comparisons,” *Biometrics*, 43, 913–928.
- Farrar, D. B., and Crump, K. S. (1988), “Exact Tests for Any Carcinogenic Effect in Animal Bioassays,” *Fundamental and Applied Toxicology*, 11, 652–663.
- Heyse, J. F., and Rom, D. M. (1988), “Adjusting for Multiplicity of Statistical Tests in the Analysis of Carcinogenicity Studies,” *Biometrical Journal*, 30, 883–896.
- 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: John Wiley.
- Holm, S. (1979), “A Simple Sequentially Rejective Multiple Test Procedure,” *Scandinavian Journal of Statistics*, 6, 65–70.
- Jogdeo, K. (1977), “Association and Probability Inequalities,” *The Annals of Statistics*, 5, 495–504.
- Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak, T. J. (1990), “Categorical Data Analysis,” in *Statistical Methodology in the Pharmaceutical Sciences*, ed. D. A. Berry, New York: Marcel Dekker, chap. 3.
- Lancaster, H. O. (1961), “Significance Tests in Discrete Distributions,” *Journal of the American Statistical Association*, 56, 225–234.
- Mantel, N. (1980), “Assessing Laboratory Evidence for Neoplastic Activity,” *Biometrics*, 36, 381–399.
- 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.
- Petrondas, D. A., and Gabriel, K. R. (1983), “Multiple Comparisons by Randomization Tests,” *Journal of the American Statistical Association*, 78, 949–957.
- Reid, N. (1995), “The Roles of Conditioning in Inference,” *Statistical Science*, 10, 138–157.
- Rom, D. M. (1992), “Strengthening Some Common Multiple Test Procedures for Discrete Data,” *Statistics in Medicine*, 11, 511–514.
- SAS Institute, Inc. (1995), *MULTTEST Procedure: Experimental Upgrade*, Cary, NC: SAS Institute Inc.
- Saville, D. J. (1990), “Multiple Comparison Procedures: The Practical Solution,” *The American Statistician*, 44, 174–180.
- 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.
- Slepian, D. (1962), “The One-Sided Barrier Problem for Gaussian Noise,” *Bell Systems Technical Journal*, 41, 463–501.
- Tarone, R. E. (1990), “A Modified Bonferroni Method for Discrete Data,” *Biometrics*, 46, 515–522.
- Troendle, J. F. (1996), “A Permutational Step-Up Method of Testing Multiple Outcomes,” *Biometrics*, 52, 125–138.
- Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing*, New York: John Wiley.
- Yates, F. (1984), “Tests of Significance for 2×2 Contingency Tables” (with discussion), *Journal of the Royal Statistical Society, Ser. A*, 147, 426–463.