Closed Multiple Testing Procedures and PROC MULTTEST

Peter H. Westfall and Russell D. Wolfinger

Peter H. Westfall, Ph.D., Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, has worked as a consultant for several large pharmaceutical companies, is an author or co-author of numerous articles and books on multiple comparisons and multiple tests, including Multiple Comparisons and Multiple Tests Using the SAS System.

Russell D. Wolfinger, Ph.D., Senior Research Statistician at SAS Institute Inc., is the developer for the MIXED and MULTTEST procedures, and is also a coauthor of Multiple Comparisons and Multiple Tests Using the SAS System

Abstract

Multiple comparisons and multiple testing problems arise frequently in statistical data analysis, and it is important to address them appropriately. *Closed testing methods* are among the most powerful multiple inference methods available, and are therefore gaining rapidly in popularity. The purpose of this article is to explain what a closed testing procedure is, why such methods are desirable, and explicitly identify situations for which the MULTTEST procedure provides a closed testing procedure.

Introduction

The area of multiple comparisons and multiple testing has been seemingly taken over with "closed testing" applications in recent years (Grechanovsky and Hochberg, 1999; Koch and Gansky, 1996; Zhang et al., 1997, to name just a few). These methods typically result in "step-wise"-type methods, which are usually more powerful than their "single-step" counterparts. An example of a "single-step" method is the simple Bonferroni method, wherein the multiple inferences must use significance levels α/k, where k denotes the number of distinct inferences and where α denotes the maximum allowable Type I error rate over the set of k inferences. A step-wise method, on the other hand, typically uses critical levels larger than α/k, allowing significant differences more often, meaning that the methods are more powerful.

The goal of multiple testing procedures is to control the "maximum overall Type I error rate," which is the maximum probability that one or more null hypotheses is rejected incorrectly. This quantity also goes by the name "Maximum Experimentwise Error Rate" (MEER) in the documentation for PROC GLM, (See SAS/STAT User's Guide, Version 8, Volumes 1, 2, and 3), and it is called "maximum familywise error rate" (often abbreviated as FWE) by many authors including Hochberg and Tamhane (1987, p. 3). In this article we refer to it as MEER to be consistent with SAS/STAT documentation.

The terms "multiple comparisons" and "multiple tests" are often used somewhat interchangeably. In this article, "multiple comparisons" typically refers to comparisons among mean values of different groups (for example, A vs B, A vs. C, B vs. C). By "multiple tests" we mean multiple tests of a more general nature, but often in the context of multivariate data.

A major disadvantage of closed testing procedures is that there is usually no confidence interval correspondence. However, if you are willing to give up the confidence intervals, then you can gain a lot of power in your multiple comparisons procedure with closed testing methods.

The MULTTEST procedure in SAS/STAT software has used step-wise methods since its inception, roughly in 1993 with Release 6.06. At the time, closed testing methods were not as prevalent as they are today, and the methods of PROC MULTTEST were described in its documentation and in the supporting book by Westfall and Young (1993) without reference to the concept of closure. However, it turns out that the MULTTEST methodology does in fact provide closed tests. In this article we illustrate why, how, and when PROC MULTTEST gives you closed tests.

Closed Testing Methods for Multiple Tests

The idea behind closed testing is wonderfully simple, but requires some notation. Suppose you want to test hypotheses H_1, H_2, and H_3. These might be comparisons of three treatment groups with a common control group, or comparisons of a single treatment against a single control using three distinct measurements. The closed method works as follows.

1. Test each hypothesis H_1, H_2, H_3 using an appropriate α-level test.
2. Create the "closure" of the set, which is the set of all possible intersections among H_1, H_2, H_3, in this case the hypotheses H_{12}, H_{13}, H_{23}, and H_{123}.
3. Test each intersection using an appropriate α-level test. These tests could be F-tests, MANOVA tests, or in general any test that is valid for the given intersection. (There are many possibilities for testing these intersection hypotheses, and each method for testing intersections results in a different closed testing procedure. We present and compare seven such procedures below.)
4. You may reject any hypothesis H_i, with control of the MEER, when the following conditions both hold

- The test of H_i itself yields a statistically significant result, and
- The test of every intersection hypothesis that includes H_i is statistically significant.

We illustrate the method using the following real data set.

```
 data mult;
   input G Y1 Y2 Y3;
   datalines;
0 14.4 7.00 4.30
0 14.6 7.09 3.88
0 13.8 7.06 5.34
0 10.1 4.26 4.26
0 11.1 5.49 4.52
0 12.4 6.13 5.69
0 12.7 6.69 4.45
1 11.8 5.44 3.94
1 18.3 1.28 0.67
1 18.0 1.50 0.67
1 20.8 1.51 0.72
1 18.3 1.14 0.67
1 14.8 2.74 0.67
1 13.8 7.08 3.43
1 11.5 6.37 5.64
```

```
1 10.9 6.26 3.47
;
```

We now discuss seven different types of tests that might be used to test for differences between the means of the G=0 and G=1 groups for each of the three variables Y1, Y2, and Y3. For each test we explain in detail how the closed testing procedure works.

1. Hotelling's T^2

Our first closed testing method uses basic t-tests for the component hypotheses, and Hotelling's T^2 test (refer, for example, to Johnson and Wichern, 1998, p. 302-306) for the composites, computed as follows using PROC REG:

```
proc reg data=mult;
   model Y1 Y2 Y3 = G;
   H1: mtest Y1;
   H2: mtest Y2;
   H3: mtest Y3;
   H12: mtest Y1, Y2;
   H13: mtest Y1, Y3;
   H23: mtest Y2, Y3;
   H123: mtest Y1, Y2, Y3;
run;
```

Each MTEST statement produces a test statistic and p-value. The following diagram lists the p-values for the hypotheses, arranged in a hierarchical fashion to better illustrate the closed testing method.

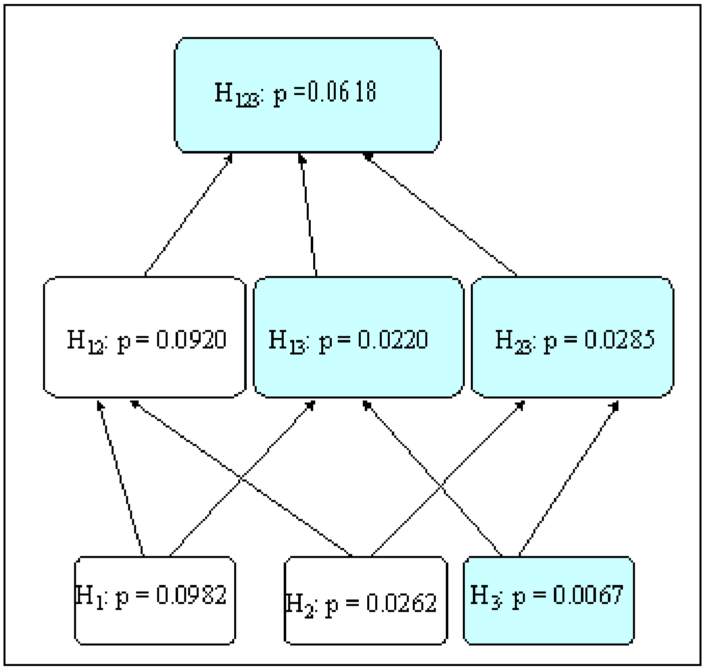


Figure 1: Illustration of the closed testing method using Hotelling's T^2 tests

The shaded areas show how to test hypothesis H_3 using the closed method. You must obtain a statistically significant result for the H_3 test itself (at the bottom of the tree), as well as a significant result for all hypotheses that include H_3, in this case, H_{13}, H_{23}, and H_{123}. Since the p-value for one of the including tests, the H_{123} test in this case, is greater than 0.05, you may not reject the H_3 test at the MEER=0.05 level.

In this example, we could reject the H_3 hypothesis for MEER levels as low as, but no lower than 0.0618, since this is the largest p-value among all containing hypotheses. This suggests an informative way of reporting the results of a closed testing procedure.

Definition: When using a closed testing procedure, the *adjusted p-value* for a given hypothesis H_i is the maximum of all p-values for tests that include H_i as a special case (including the p-value for the H_i test itself).

The adjusted p-value for testing H_3 is, therefore, formally computed as max(0.0067, 0.0220, 0.0285, 0.0618) = 0.0618.

For this example the joint hypotheses are tested using Hotelling's T^2 tests. There are many other ways you can test these composite hypotheses. Every different method for testing the composite hypotheses leads to a new and different closed testing method, and closed testing methods are best when the tests for the composite hypotheses are as powerful as possible. Sometimes one method is more powerful, sometimes another method is more powerful. Often, there is no unique method that is best for all situations, and the choice of which test to use depends upon the nature of the alternative hypothesis that is anticipated for the given testing situation. While the T^2 test is generally accepted as a good test with high "average power, " it is not always best. In the following sections we define a test based on the minimum p-value, which has higher power when there is a pronounced difference for one of the alternatives.

2. Bonferroni-Holm minP

Another test you might use for the composite hypotheses is the Bonferroni minP test. To use this test, you need only compare the minimum p-value (minp) of the individual component tests to α/k^*, where k* is the number of components in the composite and α is the desired MEER level, and reject the composite when minp$\leq \alpha/k^*$. Equivalently, you can reject the composite when $(k^*)\times$minp$\leq\alpha$, so that $(k^*)\times$minp is the p-value for the composite test, in the same way that the Hotelling's T^2 test produces a p-value for the composites shown in Figure 1. Figure 2 displays the p-values for this method.

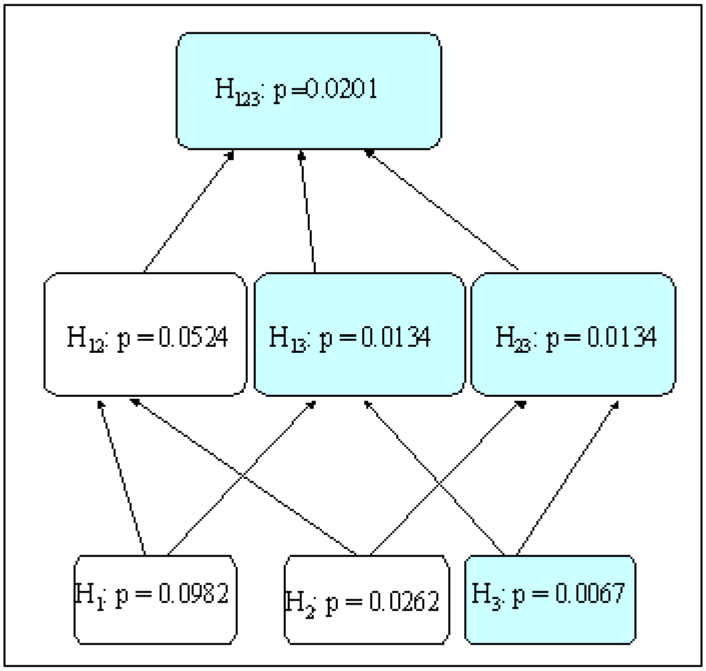


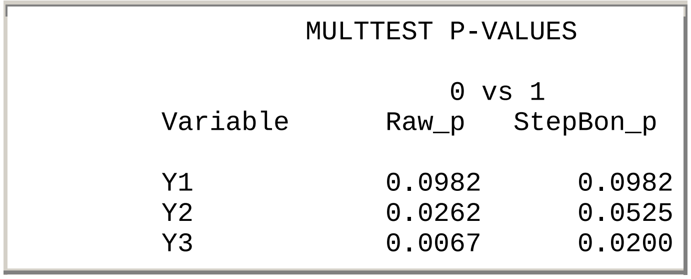
Figure 2: Illustration of the closed testing method using the Bonferroni minP test (Bonferroni-Holm Method)

The p-value for the Bonferroni minP test of H_{123} is p = k*×minp = 3×>0.0067 = 0.0201; for H_{12} it is p = 2×0.0262 = 0.0524; for H_{13} it is p = 2×0.0067 = 0.0134; and for H_{23} it is also p = 2×0.0067 = 0.0134. The shaded areas in Figure 2 show how to test hypothesis H_3 using the closed method with the Bonferroni minP test. You must obtain a statistically significant result for the H_3 test itself (at the bottom of the tree), as well as a significant result for all hypotheses that include H_3, in this case, H_{13}, H_{23}, and H_{123}. Since the p-value for all of the including tests are less than 0.05, you may reject the H_3 test at the MEER=0.05 level. Further, you may reject the H_3 hypothesis for MEER values as low as 0.0201 so that 0.0201 is the adjusted p-value for the test of H_3. Similar reasoning shows that the adjusted p-value for H_2 is 0.0524 and that of H_1 is 0.0982, so that H_1 and H_2 may be rejected at the MEER=0.10 level, but not at the MEER=0.05 level when using the Bonferroni minP closed testing procedure.

Closed testing using the Bonferroni minP test is known as "Holm's Method," (Holm, 1979). Holm showed that you do not need to calculate p-values for the entire tree; you only need to calculate p-values for the nodes of the tree corresponding to the ordered p-values. Refer to Westfall et al. (1999) for further details. Holm's closed minP-based method can be obtained using PROC MULTTEST as follows:

```
proc multtest data=mult holm pvals;
   class g;
   test mean(Y1 Y2 Y3);
   contrast "0 vs 1" -1 1;
run;
```

The output is as follows. Apart from rounding, it agrees exactly with the "by hand" calculations shown in Figure 2.



```
               MULTTEST P-VALUES

                        0 vs 1
         Variable      Raw_p   StepBon_p

         Y1            0.0982      0.0982
         Y2            0.0262      0.0525
         Y3            0.0067      0.0200
```

Since the Holm method does not require actual data, only the p-values, you can also use the "p-value input" mode of PROC MULTTEST to produce the same result:

```
data pvals;
   input test$ raw_p @@;
   datalines;
Y1 .0982 Y2 .0262 Y3 .0067
proc multtest pdata=pvals holm out=results;
proc print data=results;
run;
```

3. Westfall-Young Bootstrap minP

The Bonferroni-Holm minP test is conservative (less likely to reject) because it does not account for correlations among the variables. Westfall and Young (1993) suggest that, rather than comparing the observed minp for a given composite to α/k^*, you can compare it to the actual α-quantile of the MinP null distribution. Formally, this is equivalent to calculating the p-values p = P(MinP $\le$ minp), where MinP denotes the random value of the minimum p-value for the given composite, and minp denotes the value of minp that was actually observed. Then, you simply compare p to the MEER level α to decide whether to reject the composite. Usually, the distribution of MinP is unknown, but can be easily approximated via bootstrap resampling of the centered data vectors, as shown in Westfall and Young (1993). Westfall and Young also show that the p-values for all composite hypotheses need not be computed. Instead, you can use a trick like the Bonferroni-Holm method, and consider only particular subsets corresponding to the ordered p-values. The following chart shows how these composite p-values based on the minP test look when using bootstrap resampling-based tests.

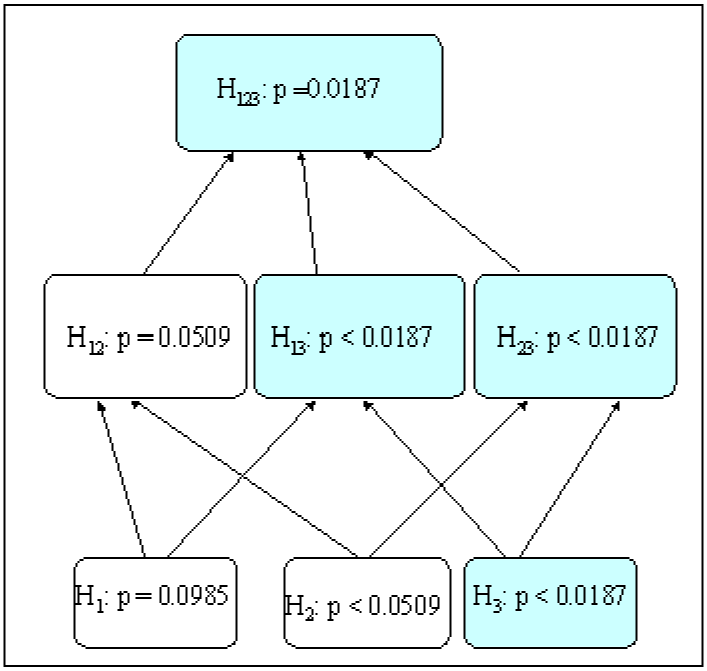


Figure 3: Illustration of the closed testing method using the Westfall-Young minP tests with Bootstrap Resampling

The shaded areas in Figure 3 show how to test hypothesis H_3 using the closed method with the Westfall-Young minP method: you must obtain a statistically significant result for the H_3 test itself (at the bottom of the tree), as well as a significant result for all hypotheses that include H_3, in this case, H_{13}, H_{23}, and H_{123}. Since the p-value for all of the including tests are less than 0.05, you may reject the H_3 test at the MEER=0.05 level. Further, you may reject the H_3 hypothesis for MEER values as low as 0.0187 so that 0.0187 is the adjusted p-value for the test of H_3. Similar reasoning shows that the adjusted p-value for H_2 is 0.0509, and that of H_1 is 0.0985, so that H_1 and H_2 may be rejected at the MEER=0.10 level, but not at the MEER=0.05 level when using the Westfall-Young minP closed testing procedure.

Note also the following about this method:

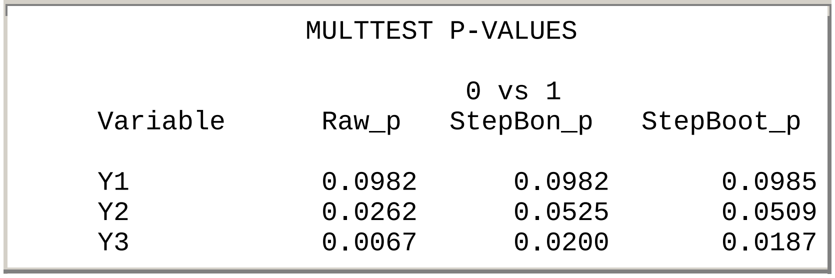
- The adjusted p-values for the Westfall-Young method are smaller than those of the Bonferroni-Holm method, since their method incorporates correlations.

- All tests are approximate, since they are based on bootstrap sampling. The accuracy improves as the sample size of the original data set increases.

- As in the case of the Bonferroni-Holm minP method, not all nodes need to be tested using the Westfall-Young method. Here, the H_{13} node need not be tested, since it is guaranteed that its p-value will be smaller than that for the H_{123} node. Those nodes in Figure 3 shown as "p<" rather than "p=" are calculated by implication, not directly. For example, it is known that $P(\min(P_2,P_3) \le 0.0067) < P(\min(P_1,P_2,P_3) \le 0.0067) = 0.0187$, so the p-value for the "H_{23}" hypothesis need not be calculated.

The Westfall-Young closed minP-based method can be obtained using the STEPBOOT option in PROC MULTTEST as follows:

```
proc multtest data=mult holm stepboot pvals n=1000000;
   class g;
   test mean(Y1 Y2 Y3);
   contrast "0 vs 1" -1 1;
run;
```



```
                   MULTTEST P-VALUES

                              0 vs 1
     Variable      Raw_p   StepBon_p   StepBoot_p

     Y1            0.0982      0.0982       0.0985
     Y2            0.0262      0.0525       0.0509
     Y3            0.0067      0.0200       0.0187
```

The N=1000000 option specifies that 1000000 resampled data sets are computed. The Monte Carlo standard error of the StepBoot_p estimates is sqrt{p(1-p)/1000000}; for example, the Monte Carlo standard error for the estimate of the true bootstrap p-value corresponding to 0.0187 is sqrt{0.0187(1-0.0187)/1000000} = 0.000135, which is reasonably small. The StepBoot_p values account for correlations and are usually smaller than the StepBon (or Bonferroni-Holm) p-values, thus the Westfall-Young method allows potentially more significances than does Bonferroni-Holm. Note, however, that for this example the Y1 StepBoot_p value is slightly larger than the StepBon_p value; this discrepancy can be attributed to the fact that the StepBon_p p-value assumes normality and the StepBoot_p p-value does not.

4. Exact Permutational minP Method

Westfall and Young (1993) show that the bootstrap method is asymptotically valid (meaning correct with large sample sizes) and quite accurate even in small sizes. Nevertheless, you might criticize this method because of its approximate nature. You can modify the method to produce an exact, permutation-based minP test for each component test (Westfall and Young, 1993, p. 133-121; see also Chung and Fraser, 1958). Here, you simply compare the observed minp value to the α quantile of the distribution of MinP obtained via permuting the (Y1,Y2,Y3) vectors between treatment and control groups. Formally, this is equivalent to calculating the p-value p = P(MinP $\leq$ minp), where MinP is the minimum p-value for the given composite calculated from a random permutation of the vectors; and where minp denotes the fixed, observed minimum p-value from the given composite. The criterion p$\leq$ 0.05 provides an exact test of the given composite hypothesis. Westfall and Young show that, as in the case of bootstrap p-values, the permutation p-values for all composite hypotheses need not be computed. The following chart shows how these p-values look in the case of the permutational resampling-based tests using the MinP statistic.

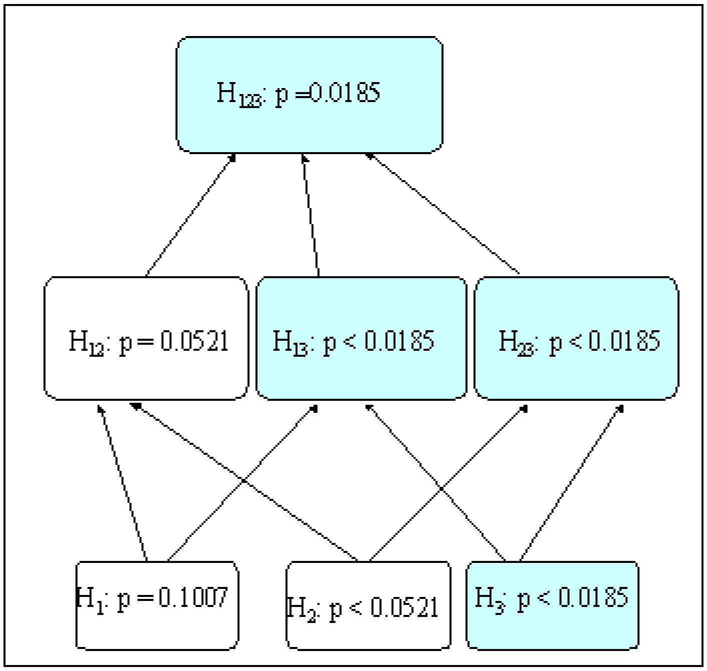
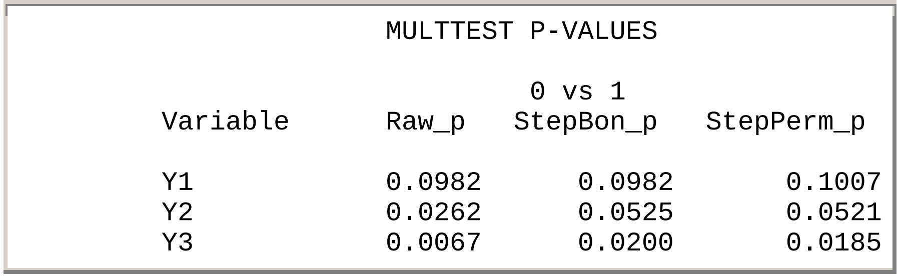


Figure 4: Illustration of the closed testing method using the exact permutational minP tests

The shaded areas show how to test hypothesis H_3 using the closed method with the permutational closed minP method: you must obtain a statistically significant result for the H_3 test itself (at the bottom of the tree), as well as a significant result for all hypotheses that include H_3, in this case, H_{13}, H_{23}, and H_{123}. Since the p-value for all of the including tests are less than 0.05, you may reject the H_3 test at the MEER=0.05 level. Further, you may reject the H_3 hypothesis for MEER values as low as 0.0185, so that 0.0185 is the adjusted p-value for the test of H_3. Similar reasoning shows that the adjusted p-value for H_2 is 0.0521, and that of H_1 is 0.1007, so that H_1 and H_2 may be rejected at the MEER=0.10 level, but not at the MEER=0.05 level when using the permutational closed minP tests.

This analysis can be obtained using PROC MULTTEST as follows.

```
proc multtest data=mult holm stepperm pvals n=1000000;
   class g;
   test mean(Y1 Y2 Y3);
   contrast "0 vs 1" -1 1;
run;
```



```
                        MULTTEST P-VALUES

                                 0 vs 1
         Variable      Raw_p   StepBon_p   StepPerm_p

         Y1            0.0982      0.0982       0.1007
         Y2            0.0262      0.0525       0.0521
         Y3            0.0067      0.0200       0.0185
```

As with StepBoot_p in the previous example, the StepPerm_p value 0.1007 is larger than the StepBon_p 0.0982 due to the fact that different test procedures are used for the base tests (normality-assuming versus permutation).

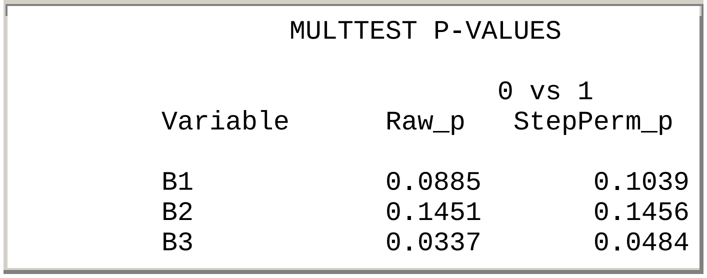
5. Closed Testing with Multivariate Binary Data

The permutational minP method works very well with multivariate binary data. Often, data are coded as binary, such as cases where the observation lived or died, or cases where the result was a success or failure. Sometimes it is useful to recode continuous measures as binary. The following program recodes the variables Y1, Y2, and Y3 as binary, with thresholds 15, 5, and 3 (in practice, choice of particular thresholds requires substantive justification).

```
data mult_bin;
   set mult;
   B1 = (Y1>15);
   B2 = (Y2>5);
   B3 = (Y3>3);
run;
```

Now, suppose we want to test whether the frequencies of B1, B2, and B3 differ between the groups G=0 and G=1. We can test these hypotheses using the two-sided Fisher exact test, with multivariate permutation sampling, as shown in the previous section, using the following program:

```
proc multtest data=mult_bin pvalsum stepperm
   n=200000 seed=121211;
   class g;
    test fisher(B1 B2 B3);
   contrast "0 vs 1" -1 1;
run;
```



```
                 MULLTTEST P-VALUES

                             0 vs 1
         Variable      Raw_p   StepPerm_p

         B1            0.0885       0.1039
         B2            0.1451       0.1456
         B3            0.0337       0.0484
```

Note that the minimum p-value is adjusted very little - from 0.0337 to 0.0484. It often happens that there is relatively little multiplicity adjustment when using binary tests; see Westfall and Young (1993), Westfall and Wolfinger (1997), and Westfall et al. (1999).

Note also that the permutation sample size was selected as 200000 to establish that the "B3" comparison is indeed significant MEER=0.05 level. The standard error from the resampling is {0.0484*(1-0.0484)/200000$^{1/2}$ = 0.00048, so the estimate differs significantly from 0.05 (z=(0.0484-0.05)/0.00048 = -3.33).

6. Exact Nonparametric Testing Using Ranks

You can also perform exact nonparametric tests by first ranking the data, then applying the permutational minP tests. Here is the code for doing so and the resulting output.

```
proc rank data=mult out=rankmult;
   var Y1 Y2 Y3;
   run;
```

```
proc multtest data=rankmult holm stepperm pvals n=1000000;
   class g;
   test mean(Y1 Y2 Y3);
   contrast "0 vs 1" -1 1; run;
```

```
                      MULTTEST P-VALUES

                                 0 vs 1
         Variable      Raw_p   StepBon_p   StepPerm_p

         Y1            0.1762      0.1762       0.1821
         Y2            0.0457      0.0914       0.1008
         Y3            0.0042      0.0125       0.0166
```

In this particular example, there is much less significance using the nonparametric procedure; however, when the data are outlier-prone, the rank-based nonparametric method is preferred.

7. Hommel's Method Based on Simes' Test

Simes (1986) proposed testing a composite hypothesis using the entire set of ordered p-values, rather than just their minimum. Assuming the p-values are ordered as $p_1 \leq p_2 \leq ... \leq p_k$, you can reject the composite if $p_j \leq j\alpha/k$ for at least one j. Equivalently, the test rejects the composite if $\min(kp_j/j) \leq \alpha$, so p = $\min(kp_j/j)$ is the Simes p-value for the composite. The test is clearly better than the Bonferroni-Holm minP test since the Simes p-value $\min(kp_j/j)$ is always as small or smaller than the Bonferroni-Holm p-value k×minp = $k \times p_1$. Sarkar (1998) and Sarkar and Chang (1997) showed that the Simes test is valid in most cases, and at least approximately valid in others, so the method is generally usable. (When correlations between variables are negative, the test can sometimes allow slightly more Type I errors than the stated MEER level). In the example used throughout this article, the Simes test and the Bonferroni minP test provide exactly the same results: Figure 2 would look identical if the Simes test were used instead of the Bonferroni minP test. Just to illustrate calculations, however, note that the Simes p-value for H_{123} is min(3(0.0067)/1, 3(0.0262)/2, 3(0.0982)/3) = min(0.0201, 0.0524, 0.0982) = 0.0201; and the Simes p-value for H_{12} is min(2(0.0262)/1, 2(0.0982)/2) = min(0.0524, 0.0982) = 0.0524.

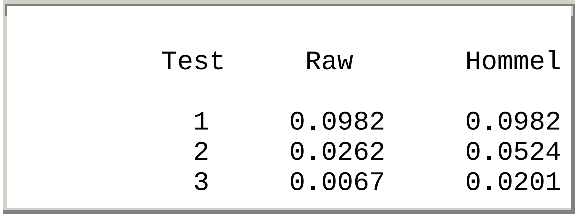
In many cases, you get adjusted p-values that are substantially smaller using the Simes test. Suppose, for example, that all p-values were 0.04. Then the composite p-value would be 0.04 using the Simes test, or statistically significant, but 0.12 using the Bonferroni-Holm test, insignificant even at the 0.10 level.

While the Simes test is uniformly better than the Holm-Bonferroni minP methods, it is not necessarily better than the Westfall-Young minP methods. Power calculations by Dunnett and Tamhane (1993, 1995) and others show that the minP-based methods tend to be preferred in screening applications where there are only a few true alternatives among a large collection of true nulls.

When you perform the Simes test in closed fashion, you can make conclusions about the individual hypotheses. The resulting method is called Hommel's (1988) method, and it will be available in PROC MULTTEST in Release 8.1 of the SAS System. The following is the associated code and its output.

```
proc multtest pdata=pvals hommel;
run;
```

```
                        p-Values
```



```
        Test      Raw        Hommel

           1     0.0982     0.0982
           2     0.0262     0.0524
           3     0.0067     0.0201
```

8. Fisher Combination

The Fisher combination test provides yet another way to calculate a p-value for a composite hypothesis, but it requires independent p-values. The composite statistic is $\chi^2 = -2\Sigma \ln(p_i)$, which is distributed as Chi-Square with 2 × k* degrees of freedom under the composite null, again assuming independent p-values. Figure 5 displays the results of using the procedure in the closed testing tree. NOTE: the method is NOT valid in this case because the variables must be assumed correlated. This graph is for illustrative purposes only.

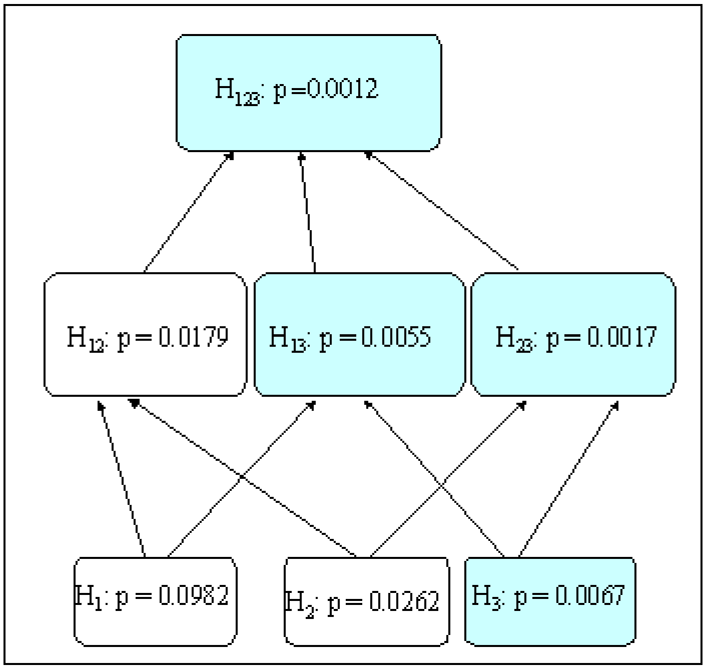


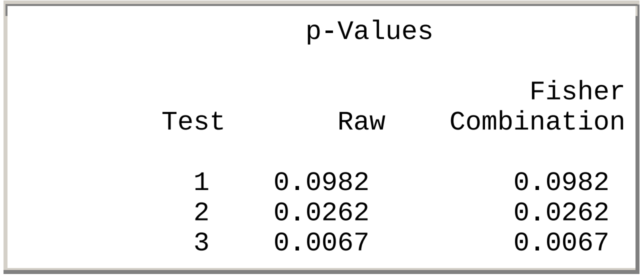
Figure 5: Illustration of the closed testing method using the Fisher combination test

The shaded areas show how to test hypothesis H_3 using the Fisher combination closed method: you must obtain a statistically significant result for the H_3 test itself (at the bottom of the tree), as well as a significant result for all hypotheses that include H_3, in this case, H_{13}, H_{23}, and H_{123}. Since the p-values for all of the including tests are less than 0.05, you may reject the H_3 test at the MEER=0.05 level. Further, you may reject the H_3 hypothesis for MEER values as low as 0.0067, so that 0.0067 is the adjusted p-value for the test of H_3. Similar reasoning shows that the adjusted p-value for H_2 is 0.0262, and that of H_1 is 0.0982, so that no adjustment for multiple comparisons is needed at all!

The result that no adjustments are needed at all tends to occur when the p-values are all reasonably small, and therefore supportive. However, the Fisher closed test can be much worse than the minP-based tests and the Simes test when there are several large p-values. Also, recall that the test is not valid in this application anyway because the variables are correlated.

The closed Fisher combination method will be available in PROC MULTTEST in Release 8.1 of the SAS System. The following is the associated code (in input p-value form) and its output.

```
proc multtest pdata=pvals fisher_c;
run;
```



```
                  p-Values

                                Fisher
         Test        Raw    Combination

           1    0.0982         0.0982
           2    0.0262         0.0262
           3    0.0067         0.0067
```

Closed Testing Methods for Multiple Comparisons

As demonstrated by Westfall and Young (1993), the bootstrap resampling methodology of PROC MULTTEST provides valid tests of composites when the data come from a "location shift" model, which does not require normal distributions. The location shift model requires that the distributions of the data in each group are identical, with possibly different locations (medians), but allows that the distributions may not be normal. In such cases, the PROC MULTTEST methodology provides a valid (at least asymptotically) and closed multiple testing procedure.

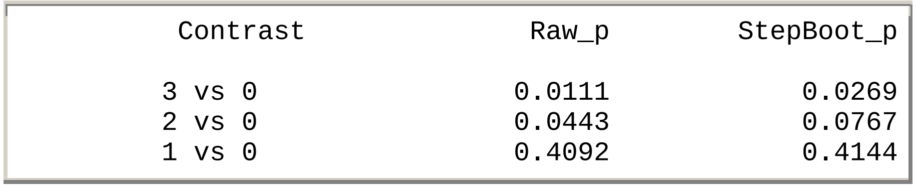
The following example shows a case where PROC MULTTEST is a valid, closed multiple comparisons procedure, when multiple comparisons among groups are considered. In this case we compare several groups against a common control to screen for differences.

Step-down Bootstrap Adjustment for Mean Comparisons

```
data compare;
   do group = 0 to 3;
      do rep = 1 to 4;
         input y @@;
         output;
      end;
   end;
   datalines;
89.8 93.8 88.4 112.6
84.4 116.0 84.0 68.6
64.4 79.8 88.0 69.4
75.2 62.4 62.4 73.8
;

proc multtest data=compare stepboot n=100000;
   class group;
   test mean(y);
   contrast "3 vs 0" -1 0 0 1 ;
   contrast "2 vs 0" -1 0 1 0 ;
   contrast "1 vs 0" -1 1 0 0 ;
run;
```

The output is as follows:



```
          Contrast                    Raw_p         StepBoot_p

         3 vs 0                     0.0111             0.0269
         2 vs 0                     0.0443             0.0767
         1 vs 0                     0.4092             0.4144
```

The values in the "StepBoot_p" column are the result of closed testing as shown in Figure 6.

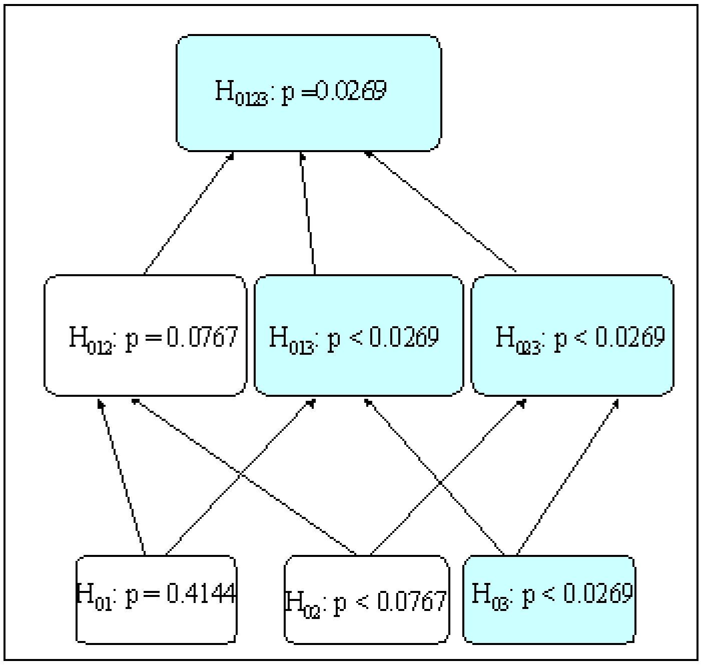


Figure 6: Illustration of closed testing for group comparisons using PROC MULTTEST with bootstrap sampling

In Figure 6, H_{01} denotes the hypothesis that $\mu_0 = \mu_1$, H_{02} denotes the hypothesis that $\mu_0=\mu_2$, etc. H_{012} denotes the intersection hypothesis $H_{01} \cap H_{02}$, which implies $\mu_0=\mu_1=\mu_2$, and so on. As in Figure 3 and Figure 4, PROC MULTTEST does not compute p-values for all nodes: those indicated by "p<" are subsumed by implication. For example, since $P(\min(P_{01},P_{02},P_{03}) \leq 0.0111) = 0.0269$, we know that $P(\min(P_{02},P_{03}) \leq 0.0111) < 0.0269$, so the H_{023} p-value need not be computed.

As before, this is a bootstrap method, and therefore the p-values are approximate. On the other hand, they only require location shift data, not normally distributed data, for their justification.

A subtle but very important point in this analysis is that PROC MULTTEST always calculates probabilities under the global null hypothesis (the one at the top of the closed diagram). When we claim that $P(\min(P_{02},P_{03}) \leq 0.0111) < 0.0269$, we are implicitly assuming that this probability is calculated under the global null case. However, the H_{023} test is not a test of the global null, it is a test of the partial null that μ_0=μ_2=μ_3, and allows that μ_1 might differ. The MULTTEST method, therefore, is valid only when the joint distribution of the p-values is the same, no matter whether calculated assuming the global null, or assuming the partial null. This condition is called the "subset pivotality" condition and is discussed in Westfall and Young (1993, p. 42). In the case of multiple comparisons using t-statistics with location-shift data, the subset pivotality condition is indeed satisfied (Westfall and Young, p. 123), so the MULTTEST procedure provides a valid (albeit approximate because of use of the bootstrap), closed testing procedure in this case.

Comparisons of Restricted Combinations

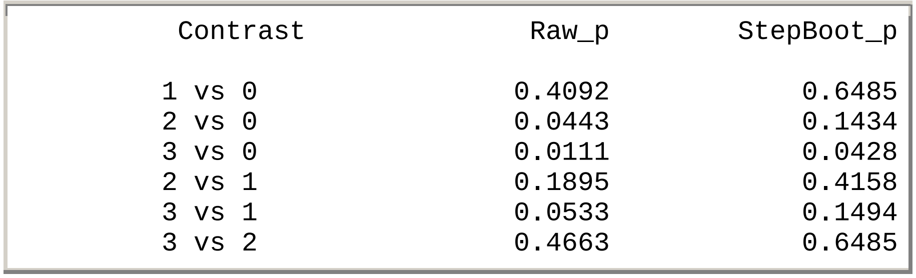
All the preceding examples are cases of "free combinations," since every possible combination of intersections produces a distinct, plausible hypothesis. Hypotheses form "restricted combinations" if some intersections are either nondistinct or nonplausible. The simplest example of this is the case of all pairwise comparisons, where intersections involving different components can produce the same composite. For example, the intersection of H_{12}: μ_1=μ_2 and H_{13}: μ_1=μ_3 produces H_{123}: μ_1=μ_2=μ_3; the intersection of H_{12}: μ_1=μ_2 and H_{23}: μ_2=μ_3 produces the same composite H_{123}: μ_1=μ_2=μ_3. In these cases, the closure tree does not contain nearly as many nodes as in the case of free combination tests, and the adjusted p-values can be reduced considerably. PROC MULTTEST, however, operates in "free combination" mode, meaning that, while the tests are valid, and technically closed, they are not as powerful as they might be.

These issues are discussed further by Shaffer (1986), Holland and Copenhaver (1987), Westfall and Young (1993), Rom and Holland (1995), Hochberg and Rom (1995), Westfall (1997), Hommel and Bernhard (1999), and Westfall et al (1999).

The following is an example of this method using the data set from the previous example. It illustrates how PROC MULTTEST works in the case of all pairwise comparisons.

```
proc multtest data=compare stepboot n=100000;
   class group;
   test mean(y);
   contrast "1 vs 0" -1 1 0 0 ;
   contrast "2 vs 0" -1 0 1 0 ;
   contrast "3 vs 0" -1 0 0 1 ;
   contrast "2 vs 1" 0 -1 1 0 ;
   contrast "3 vs 1" 0 -1 0 1 ;
   contrast "3 vs 2" 0 0 -1 1 ;
run;
```

The output is as follows:



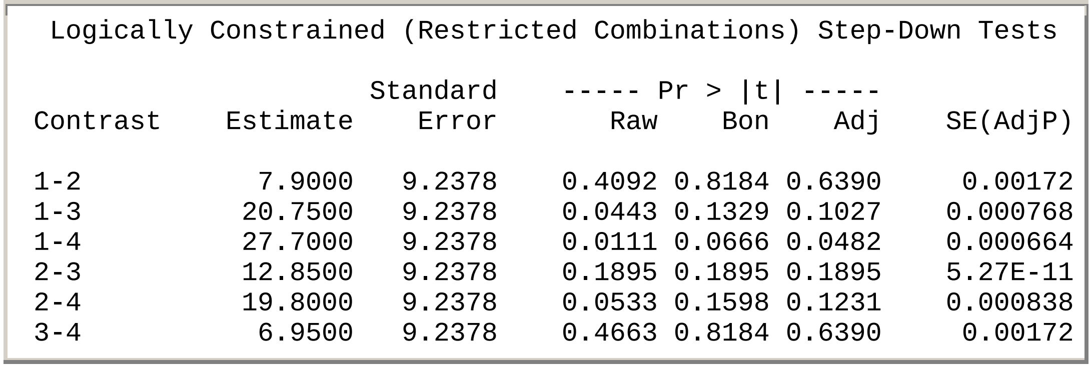
```
          Contrast                Raw_p        StepBoot_p

         1 vs 0                  0.4092            0.6485
         2 vs 0                  0.0443            0.1434
         3 vs 0                  0.0111            0.0428
         2 vs 1                  0.1895            0.4158
         3 vs 1                  0.0533            0.1494
         3 vs 2                  0.4663            0.6485
```

The StepBoot_p values provide valid, and closed multiplicity-adjusted p-values, but they are conservative because they do not account for the fact that the closure tree "collapses" in various places.

Westfall (1997) devised a minP-based method for any set of linear contrasts that respects the collapsing in the closure tree, as well as intercorrelations among the variables. This method is coded using the %SimTests macro in Westfall et al. (1999). The code and output are as follows.

```
%MakeGLMStats(dataset=compare,
   yvar=y,
   classvar=group,
   model=group,
   contrasts=all(group)
);
%SimTests(nsamp=20000,type=LOGICAL);
```



```
  Logically Constrained (Restricted Combinations) Step-Down Tests

                      Standard    ----- Pr > |t| -----
 Contrast    Estimate    Error        Raw    Bon    Adj    SE(AdjP)

 1-2           7.9000   9.2378    0.4092 0.8184 0.6390     0.00172
 1-3          20.7500   9.2378    0.0443 0.1329 0.1027    0.000768
 1-4          27.7000   9.2378    0.0111 0.0666 0.0482    0.000664
 2-3          12.8500   9.2378    0.1895 0.1895 0.1895    5.27E-11
 2-4          19.8000   9.2378    0.0533 0.1598 0.1231    0.000838
 3-4           6.9500   9.2378    0.4663 0.8184 0.6390     0.00172
```

The numbers in the "Adj" column correspond with those in the "StepBoot_p" column of the MULTTEST output, with one major difference and two minor differences. The major difference is that the %SimTests macro incorporates the logical contraints. Because of this, the Adj p-values are smaller than the StepBoot_p values in all cases but one. The minor differences are (1) the labels are slightly different (for example, 1-2 instead of 0-1), and (2) %SimTests is essentially exact under the "usual" assumptions, while MULTTEST is approximate, but does not require normality, only location-shift distributions.

Cases Where PROC MULTTEST Does Not Produce Closed Tests

The PROC MULTTEST methodology is not closed, and sometimes performs badly in the following cases:

1. When there are multiple comparisons of the means of several (more than two) treatment groups using permutation resampling
2. When there are multiple comparisons of the proportions (binary data) of several (more than two) treatment groups using either permutation resampling or bootstrap resampling.

Note that the commonality here is the case where you have multiple (more than two) treatment groups. Note also that the case of bootstrap resampling and multiple comparisons of means is conspicuously absent because, as shown in the previous section, PROC MULTTEST results in a closed testing procedure in that case (again, assuming location shift distributions).

Here is an example (suggested by Teresa Neeman, 1997) to illustrate what can go wrong in the case of multiple comparisons of binary data. Suppose you have the following data, independently sampled from groups A, B, and C.

Group	Count/n	Percent
A	3 / 4	75%
B	1 / 4	25%
C	0 / 2000	0%

You want to compare A with B, and A with C. The following program tells PROC MULTTEST to perform these two comparisons using the upper-tailed Fisher exact test, and to perform multiplicity adjustments using permutation resampling.

```
data binary;
   input b f g$;
   datalines;
0 1 A
1 3 A
0 3 B
1 1 B
0 2000 C
;
proc multtest stepperm order=data;
   class g;
   freq f;
   test fisher(b/upper);
   contrast " b - a " -1 1 0 ;
   contrast " a - c " 1 0 -1 ;
run;
```

The output is as follows:

```
         Contrast                  Raw_p        StepPerm_p

         b - a                    0.9857            0.0076
         a - c                    0.0001            0.0001
```

This seems to suggest that B (25%) is greater than A (75%) after multiplicity adjustment (p=0.0076)! This counterintuitive result is a consequence of failure of the subset pivotality condition. Figure 7 may help to explain.

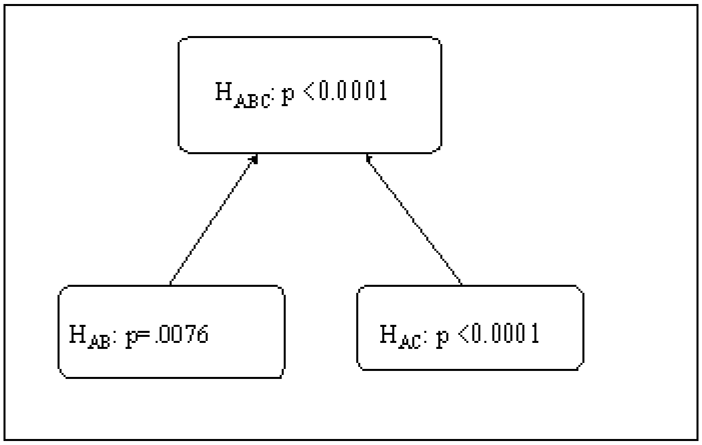


Figure 7: Illustration of problem with PROC MULTTEST with multiple comparisons of frequencies

The problem is that, in the resampling-based calculation of the p-value for the H_{AB} hypothesis, the global permutation distribution is used, rather than just the permutation distribution involving the eight "A" and "B" combined observations. Since there are so many observations in the "C" group, all zeros, most permutation samples will have the four occurrences in the "C" group, and the p-value for the B-A comparisons will be 1.0, since the comparison of 0/4 with 0/4 gives a p-value of 1.0. Thus, in only 0.0076 of the permutation samples do we observe a p-value less than or equal to 0.9857. So in this case, the fact that MULTTEST uses the global distribution for its minP probability calculations makes the procedure not closed, and not valid. This type of problem has been noted also by Petrondas and Gabriel (1983) and Hsu (1996).

It is easy to fix the problem in this case, simply use the valid p-value for the B-A test in the closed testing tree. The adjusted p-value is then max(0.9857, <.0001) = 0.9857. In more complex cases involving more comparisons, the closure tree is larger, and the solution is not quite so simple. See Westfall et al. (1999) for valid, closed methods that use PROC MULTTEST for multiple comparisons of proportions in these more complex applications; see also Westfall and Wolfinger (1997) and Rom (1992).

The example of comparisons of frequencies above is a rather extreme case. Generally, the problem is most pronounced when there is a rather large difference in one of the several groups, which has a large sample size. This is very similar to the case where you perform multiple comparisons of means when one has a very large sample size, and an extremely different variance: pooling the variances across all groups means that the estimated standard errors will be grossly inaccurate for certain contrasts.

Comparison of Freeman-Tukey and Fisher Exact Tests

In the case of binary data, one way to alleviate the problem illustrated in the previous example is to use the "FT," or Freeman-Tukey double-arcsine transformed proportion (for example, Westfall and Young, 1993, p. 153-155), in the multiple tests. This transformation is used to stabilize the variance, and therefore alleviate the type of problem shown in Figure 7. While it does not solve the problem shown for that pathological data set, it does generally perform better than the Fisher exact test for multiple comparisons of proportions using PROC MULTTEST.

To verify that the FT method works better than the Fisher exact method in PROC MULTTEST, we performed a simulation study. We identified a "worst case" scenario involving four treatment groups and a control, where the control and three of the treated groups have background rate 10%, while the fourth group has background rate 1%. We specified 50 observations in all groups except group 4, whose sample size we allowed to vary as 50,100,200,400,800. We performed the pairwise comparisons with control using step-down tests, using either the Fisher exact tests and permutation resampling, or the FT tests with bootstrap resampling. (As a general rule of thumb, bootstrap sampling is more consistent with the use of FT tests, since they are derived using the "population" sampling model, where the observations are assumed to be selected randomly from a larger population. Permutation sampling, on the other hand, is more consistent with the Fisher exact test, the latter being based upon the "randomization" model, where the randomness is assumed to come only from the random assignment of observations to treatment groups.) We tabulated the percentage of simulated data sets (this is a simulation of a simulation-based method) for which one or more of true null comparisons (any but group 4 against control) was found statistically significant, with MULTTEST adjusted p-value less than or equal to either 0.05 or 0.10. The resulting overall Type I error rates are tabulated as follows:

	α=0.05	α=0.10

N	Fisher	FT	Fisher	FT
50	0.049	0.040	0.097	0.076
100	0.037	0.028	0.091	0.073
200	0.060	0.041	0.117	0.070
400	0.094	0.045	0.166	0.092
800	0.155	0.072	0.261	0.134

As expected, there are situations where the overall Type I error rate exceeds the nominal α level. However, the problem is not bad unless the sample sizes are extremely unbalanced, and even then, the FT procedure performs reasonably well.

It should also be noted that these problems occur only for binary tests where there are multiple comparisons between more than two groups. These problems do not occur at all when MULTTEST is used to make a single comparison involving two or more groups for each of several binary variables. In those cases the subset pivotality condition is satisfied, and PROC MULTTEST provides a valid and closed testing procedure, exactly as illustrated in Figures 3 and 4.

Conclusions

PROC MULTTEST provides a valid, closed multiple test procedure in many applications. These include

- Multiple tests of means of multivariate data and/or multiple tests of between-group contrasts, with continuous location-shift data, and bootstrap resampling
- Multiple tests of proportions of multiple binary variables involving multiple groups, but where only one test per variable is used.

PROC MULTTEST does not provide closed tests, and therefore, caution is urged, in the following situations:

- Multiple comparisons of means involving three or more groups, using permutation resampling
- Multiple comparisons of binary variables involving three or more groups.

References

Chung, J.H. and Fraser, D.A.S. (1958). Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association* **53**, 729-735.

Dunnett, C.W., and Tamhane, A.C. (1993). Power comparisons of some step-up multiple test procedures. *Statistics & Probability Letters* **16**, 55-58.

Dunnett, C.W., and Tamhane, A.C. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics* **51**, 217-227.

Grechanovsky, E. and Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference* **76**, 79-91.

Hochberg, Y., and Rom, D. (1995). Extensions of multiple testing procedures based on Simes' test. *Journal of Statistical Planning and Inference* **48**, 141-152.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparisons Procedures*, John Wiley & Sons, Inc., New York.

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

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

Hommel, G. and Bernhard, G. (1999). Bonferroni procedures for logically related hypotheses. *Journal of Statistical Planning and Inference* **82**, 119-128.

Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*, Chapman and Hall, London.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. Fourth Edition, Prentice-Hall, New Jersey.

Koch, G.G. and Gansky, S.A. (1996). Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal* **30**, 523-534.

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.

Neeman, T. (1997). E-mail correspondences with Peter Westfall.

Petrondas, D.A. and Gabriel, K.R. (1983). Multiple comparisons by randomization tests*. Journal of the American Statistical Association* **78**, 949-957.

Rom, D.M. (1992). Strengthening some common multiple test procedures for discrete data. *Statistics in Medicine* **11**, 511-514.

Rom, D.M. and Holland, B. (1995). A new closed multiple testing procedure for hierarchical families of hypotheses. *Journal of Statistical Planning and Inference* **46**, 265-275.

Sarkar, S. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Annals of Statistics* **26**, 494-504.

Sarkar, S., and Chang, C.K. (1997). Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* **92**, 1601-1608.

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

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

Westfall, P.H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association* **92**, 299-306.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*, Cary, NC: SAS Institute Inc.

Westfall, P.H. and Young, S.S (1993). *Resampling-Based Multiple Testing*. John Wiley & Sons, Inc., New York.

Westfall, P.H. and Wolfinger, R.D. (1997). Multiple tests with discrete distributions. *American Statistician* **51**, 3-8.

Zhang, J., Quan, H., Ng, J., and Stepanavage, M.E., (1997). Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials* **18**, 204-221.

Modified code is not supported by the author or SAS Institute Inc.