NONSTANDARD USES OF PROC MULTTEST: PERMUTATIONAL PETO TESTS; PERMUTATIONAL AND UNCONDITIONAL T AND BINOMIAL TESTS

Peter H. Westfall, Texas Tech University, Lubbock, TX
Keith A. Soper, Merck Research Laboratories, West Point, PA

ABSTRACT

The main feature of PROC MULTTEST is its calculation and tabulation of raw and multiplicity-adjusted p-values, when a user-defined collection of tests is specified. The adjusted p-values are calculated primarily using resampling methods, allowing correlations and distributional characteristics to be incorporated into the multiplicity adjustments.

We discuss lesser-known applications of PROC MULTTEST involving tests for a single variable. The software computes exact permutation tests for stratified 2XC tables; we apply this method to analyze rodent carcinogenicity data with time and cause of death information. The test procedure is the life table method described by Peto et al. (1980), but their large-sample normal approximation to the critical value for this test can be quite inaccurate when the number of observed tumors is small. We also describe methods to adjust permutational p-values for the multiplicity of tests from separate analyses of each tumor type.

Further resampling applications are given, including comparison of bootstrap and permutation (exact) binomial tests, bootstrapped and rerandomized t-tests, and calculation of nonparametric p-values for rank-transformed tests.

1. PROC MULTTEST OVERVIEW

1.a. History

PROC MULTTEST is the successor to PROC MTEST, which in turn was preceded by PROC MBIN. Development of the two precursors was funded entirely by a grant from the Pharmaceutical Manufacturing Association (PMA), in a contract between the PMA and Texas Tech University (TTU). Money for the project was donated by a consortium of participating companies. The programming was done by Dr. Lin Youling at TTU, under the direction of Drs. Peter H. Westfall and S. Stanley Young of Glaxo Inc., with substantial intellectual support from statisticians at participating pharmaceutical companies. The grant stipulated that the software be donated to SAS Institute, Inc., in hopes that it would become a regular SAS® procedure. PROC MULTTEST is currently available as a SAS/STAT® procedure as of the Version 6.07 release. It is essentially identical to PROC MTEST, but contains syntactical modifications.

1.b. Main Use of PROC MULTTEST

The name "MULTTEST" in PROC MULTTEST refers to its main application, multiple testing. It will accept data having multiple-group multivariate structure as input. Use of a blocking variable is also allowed. Each variable may be designated as continuous or binary, and the software calculates test statistics appropriate to the given designation; e.g., t-tests for continuous variables, and a variety of options including permutation tests for binary variables.

To perform multiplicity adjustment, vectors are resampled, and all test statistics are recomputed. The resampled test statistics incorporate dependence structures and possible nonnormal distributional characteristics, frequently making the multiple testing method more powerful, without sacrificing level robustness. This is an attractive feature since multiple testing methods are often thought to be excessively conservative.

Further details concerning the capabilities of the software may be found in *SAS® Technical Report P-229, SAS/STAT® Software: Changes and Enhancements, Release 6.07* (1992), and Westfall and Young (1993).

2. STRATIFIED 2XC TABLES

While the software is primarily used for multiple tests, there are useful applications of this software involving single tests for a single variable. A noteworthy feature of PROC MULTTEST is its ability to calculate exact permutation distributions for score statistics in stratified 2XC tables.

2.1 The Peto Test

The Peto procedure requires distinction between "Fatal" and "Incidental" tumors. The following data set displays reported incidences of a particular tumor type; e.g., tumors of the liver, in a carcinogenicity study.

Fatal Tumors (Count/{at risk})

Day	Control	Low	Medium	High
426	0/59	1/56	0/59	0/57
505	1/58	0/50	0/57	0/49
514	0/57	0/48	1/56	0/47
532	0/57	0/47	0/55	1/45
556	0/55	1/45	0/51	0/40
581	0/48	1/41	0/42	0/34
591	0/47	1/38	0/40	0/33
596	0/46	0/36	1/40	0/31
620	1/42	0/34	0/35	1/25
644	0/37	0/31	1/33	0/19
680	1/35	0/23	0/26	0/11
682	1/34	0/23	0/26	0/11
727	0/26	0/17	1/16	0/6

Incidental Tumors (Count/{# Dead})

Days	Control	Low	Medium	High
0–365	0/1	0/2	0/0	0/1
366–546	1/3	0/12	0/7	0/10
547–644	2/17	1/12	1/18	2/22
645–728	2/9	3/13	5/16	1/13
729+	2/26	3/17	2/15	0/6

To test for increasing tumor incidence with dose level in stratum i, the score statistic $T_i = 0 \times X_{i1} + 1 \times X_{i2} + 2 \times X_{i3} + 3 \times X_{i4}$ is frequently used, where X_{ij} =(number of observed tumors in stratum i, treatment group j). It has been suggested (Mantel, 1980) that all strata be treated as independent when determining the significance level of the test statistic $T = \sum T_i$. The significance level is determined in three steps: (i) determine the exact distribution of T_i within each strata, (ii) convolve the distributions of $T_1, T_2 \ldots$, assuming independence, to arrive at the distrihution of T, and (iii) calculate the upper-tailed p-value for $T = t$ as $P(T \geq t)$.

PROC MULTTEST performs such an analysis, and will print the entire permutation distribution if requested. The following code is used for the analysis.

```
data;
 input strat$ @ ;
  do tgroup = 1 to 4;
   do tumor = 0 to 1;
    input count @ ;
   output;
  end;
 end;
cards;
F426 59 0 55 1 59 0 57 0
F505 57 1 50 0 57 0 49 0
F514 57 0 48 0 55 1 47 0
F532 57 0 47 0 55 0 44 1
F556 55 0 44 1 51 0 40 0
F581 48 0 40 1 42 0 34 0
F591 47 0 37 1 40 0 33 0
F596 46 0 36 0 39 1 31 0
F620 41 1 34 0 35 0 24 1
F644 37 0 31 0 32 1 19 0
F680 34 1 23 0 26 0 11 0
F682 33 1 23 0 26 0 11 0
F727 26 0 17 0 15 1 6 0
I365 1 0 2 0 0 0 1 0
I546 2 1 12 0 7 0 10 0
I644 15 2 11 1 17 1 20 2
I728 7 2 10 3 11 5 12 1
I729 24 2 14 3 13 2 6 0
;
proc multtest outperm=p;
  class tgroup;
  strata strat;
  freq count;
  test ca(tumor/ upper perm=15);
  contrast "C-A Trend" 0 1 2 3;
proc print data=p;
```

Presence or absence of tumor is denoted by 0 or 1, respectively. The "count" variable records the numbers of 1's and 0's in each stratum/group combination. (The output tables should be examined carefully to check accuracy of the input data.)

The program above reports the upper-tailed permutational p-value= 0.8228, and the permutation distribution of T. The actual computed value of T is $T = 50$, and the upper-tailed p-value 0.8228 is readily found in the "outperm=p" data set, which contains upper-tail probabilities $P(T \geq t)$ by default. The "outperm=p" data set shows that a trends $T \geq 67$ produce p-values less than 0.05, and that trends $T \geq 71$ produce p-values less than 0.01. The distribution is exact under the assumption of independence, since the total number of observed tumors within each strata is less than or equal to 15. Had there been more than 15 tumors in a particular strata, PROC MULTTEST would have computed asymptotic approximations to such distributions, then convolved the approximations with the exact distributions for the strata with totals ≤ 15 to arrive at an approximation for the distribution of T.

Despite the computation of an "exact" probability value in this case, it should be noted that the method is inexact if any tumors are coded "Fatal" because tables are dependent. Mehta et al. (1992) and Soper and Tonkonoh (1993) provide a permutation distribution that is computationally intensive but exact under the assumption of equal censoring hazards. No exact permutation test is known given unequal censoring.

2.2 Multiplicity Adjustment

When all tumors are designated as incidental, PROC MULTTEST performs resampling-based multiplicity adjustments that account for discreteness as well as correlation structure. When some tumors are designated as fatal, PROC MULTTEST performs only large-sample adjustments, incorporating neither correlation nor discreteness.

To adjust for multiplicity effects of several discrete approximation Peto tests, while accounting for the very discrete nature of the data, one may compute separate permutation distributions for each tumor type examined. Heyse and Rom (1988) and Westfall and Young (1993, p. 165) show how these permutation distributions may be used to adjust p-values for multiplicity of tests under independence of tumor types. Such adjustments are usually only slightly more conservative than those which incorporate dependence structures (Soper and Westfall, 1990; Heyse and Rom, 1988; Westfall and Young, 1993 , p. 163–165). A macro program is currently being developed to perform this analysis.

3. UNCONDITIONAL BINOMIAL TESTS

Permutation tests are conditional, in that the totals in the margins of the table are considered fixed. The resulting p-values are then calculated using the (possibly multivariate) hypergeometric distribution.

From the standpoint of the product-binomial model, it is well-known that such tests are conservative, since the probability of observing a p-value less than α is at most α (and often much less the α) in any table with fixed margins (Upton, 1982).

To avoid this problem, some have advocated use of the unconditional p-value, which is the probability of observing a result as extreme as the given result calculated from the product-binomial distribution, with probabilities estimated from the data assuming the null hypothesis of no group differences is true. PROC MULTTEST computes such p-values when one specifies a single test and variable, with bootstrap sampling.

Consider the following data, which compare frequencies of adverse events for control and treated patients in a multicenter clinical trial.

Adverse Events (Count/{# Patients})

Center	Control	Treatment
A	0/50	3/50
B	1/50	3/45
C	4/50	7/48

The large-sample approximate binomial test and the bootstrapped version (100,000 bootstrap samples) of this large-sample test are determined by the first MULTTEST invocation of the following code. The exact permutation test is given by the second invocation.

```
data one;
do center = 1 to 3;
 input strat$ @ ;
  do tgroup = 1 to 2;
   do event = 0 to 1;
    input count @ ;
   output;
  end;
 end;
end;
cards;
A 50 0 47 3  B 49 1 42 3  C 45 5 41 7
;
proc multtest bootstrap nsample=100000
       seed=98781;
 class tgroup;
 strata strat;
 freq count;
 test ca(event/upper);
 contrast "C vs T" 0 1;
proc multtest data=one outperm=p;
 class tgroup;
 strata strat;
 freq count;
 test ca(event/upper permutation=15);
 contrast "C vs T" 0 1;
```

The large-sample binomial test produces the p-value Raw_p= 0.0364, and the bootstrapped version of this test produces the p-value Adj_p= 0.0384, both from the first MULTTEST invocation. The second invocation produces an exact p-value Raw_p= 0.0600, which is more conservative than the large-sample binomial and bootstrapped binomial tests. It is known that the actual significance bootstrap binomial tests approximates the nominal α better than the permutation test in 2X2 tables (Westfall and Young, 1993, p. 169–173; see also Upton's analysis of "Liddell's Test").

4. CONTINUOUS DATA: BOOTSTRAP AND PERMUTATION TESTS

The following code uses PROC MULTTEST to analyze measurements taken on patients in control (C) and treated (T) groups. The question is whether the measurements are larger in the treated group.

```
data one;
 input y g$ @@;
 cards;
```

```
45 C 40 T  63 C 54 T  65 C 79 T
45 C 53 T  49 C 57 T  59 C 81 T
55 C 45 T  54 C 71 T  20 C 85 T
48 C 80 T  42 C 14 T
;
proc rank;
 var y;
 ranks yrank;
proc multtest bootstrap nsample=100000
        seed=91819;
 class g;
 test mean(y/upper);
 contrast "C vs T" -1 1;
proc multtest permutation nsample=100000
        seed=11721;
 class g;
 test mean(y/upper);
 contrast "C vs T" -1 1;
proc multtest permutation nsample=100000
        seed=87165;
 class g;
 test mean(yrank/upper);
 contrast "C vs T" -1 1;
```

All MULTTEST invocations use 100,000 simulated data sets. The first performs the usual two-sample (pooled variance) t-test, resulting in an upper-tailed p-value Raw_p= 0.0933, and a bootstrapped version Adj_p= 0.0928. The bootstrapped version is computed by resampling the pooled, mean-centered data, as discussed in Westfall and Young (1993, Chapter 3). Despite some outliers in the data, usual t-test performs well in this case, since it agrees well with the bootstrapped p-value. The second invocation of MULTTEST calculates the same Raw_p value, but calculates the significance level by permutation resampling of the noncentered data, rather than by bootstrap sampling, resulting in a permutational p-value Adj_p= 0.0946. In both resampling analyses, the resampling-based p-value is simply the proportion of resampled data sets that produce a recomputed upper-tail p-value less than or equal to the observed value 0.0933. Note also that the usual t-test appears to be an adequate approximation to the permutation test in this example.

The third invocation of MULTTEST uses the rank-transformed data. Use of the two-sample t-test on the rank-transformed data yields essentially the Wilcoxon two-sample test, but with an approximate significance level obtained from the t-distribution (Raw_p $= 0.1098$). Use of permutation resampling allows us to evaluate the accuracy of this approximation: the proportion of the 100,000 permuted data sets yielding a smaller rank-transformed p-value is Adj_p= 0.1096, showing the t-approximation to be reasonable.

5. CONCLUSION

PROC MULTTEST is useful for multiple testing, but also has useful univariate testing capabilities. The permutation test for scores with stratified multiple-group binary data (including the Peto test as a special case) is available. It also performs randomization and bootstrap tests for continuous data, and it can estimate exact significance levels of rank tests.

REFERENCES

Heyse,J.F. and Rom,D.(1988). "Adjusting for Multiplicity of Statistical Tests in the Analysis of Carcinogenicity Studies," *Biom. Journal* **30** 883–896.

Mantel,N.(1980). "Assessing Laboratory Evidence for Neoplastic Activity," *Biometrics* **36** 381–399.

Mehta,C.R.,Patel,N., and Senchaudhuri,P.(1992). "Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data," *Journal of Comp. and Graph. Stat.* **1** 21–40.

Peto,R.,Pike,M.,Day,N.,Gray,R., Lee,P., Parish, S.,Peto,J., Richards,S. and Wahrendorf,J. (1980). "Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-Term Animal Experiments," *Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal* IARC Monographs, Annex to Supplement 2, 311–426. Lyon: International Agency for Research on Cancer.

SAS Institute (1992). *Technical Report P-229, SAS/STAT® Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Institute Inc.

Soper,K.A. and Tonkonoh,N.(1993). "The Discrete Distribution Used for the Log-Rank Test Can Be Inaccurate," *Biom. Journal* **35** 291–298.

Soper,K.A. and Westfall,P.H.(1990). "Monte Carlo Estimation of Significance Levels for Carcinogenicity Tests Using Univariate and Multivariate Models," *Journal of Stat. Comp. and Sim.* **37** 189–209.

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

Upton,G.J.G.(1982). "A Comparison of Alternative Tests for the 2 x 2 Comparative Trial," *Journal Royal Stat. Soc. A* **148** 86–105.

SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.