Paper 264

Advances in Multiple Comparisons and Multiple Tests using the SAS® System

Peter H. Westfall, Mail Stop 2101, Texas Tech University, Lubbock TX 79409
Randall D. Tobias, SAS Institute, Inc.

1. Introduction

This article provides a summary of the upcoming book *Multiple Comparisons and Multiple Tests using the SAS® System*, by Westfall et al. (1999) and introduces some of the new methods and capabilities it presents.

Whenever you want to make more than one decision in a statistically valid way, multiple inferences are involved. There are numerous alternative solutions for multiple inference problems—some are very good, some perform reasonably well, and some are of questionable value. The wide variety of methods that are available can make the choice of technique difficult. In Westfall et al. (1999) we explain the various methods, along with their pitfalls and advantages. The material in the book is self-contained, requiring only a moderate statistical background. We apply the methods to real data, giving examples from many areas, including business, medicine, sociology, and engineering.

One goal we had in writing this book was to unify the presentation of the diverse multiple inference methods, and to simplify the use of software. The proper choice of a multiple inference procedure depends upon your inference objectives and data structure. While there are procedures in SAS/STAT® software for such inferences, such as the GLM, MIXED, and MULTTEST procedures, we realized through the course of writing the book that several types of problems "fall through the cracks," and are not particularly well accommodated by any of the existing procedures. To fill this gap, we developed a set of SAS® macro language programs that implement more general and more recently developed multiple inference techniques than are available in the procedures mentioned above. In addition to handling the usual pairwise comparison applications, these macros can be used for problems as diverse as

- confidence bands for regression functions (linear, logistic, survival analysis,…),
- simultaneous intervals for log-odds ratios in logistic regression, and
- closed testing for covariate-adjusted linear contrasts in multivariate analysis of covariance.

These macros, in conjunction with the existing facilities for multiple comparisons in the SAS system, allow users to carry out multiple inferences in most applications of practical interest. All macros will be made available in the book and on the SAS web site.

Most programs in the book are designed to run in Version 7 of the SAS System. We take liberal advantage of such features as variable and data set names with more than eight characters, as well as the Output Delivery System. Therefore, many of our programs will not run correctly in releases prior to Version 7 without suitable modifications. In particular, two of the more important macros, %SimIntervals and %SimTests, would require extensive modifications to run in previous releases.

Multiple Comparisons versus Multiple Tests

Why is there a distinction between multiple comparisons and multiple tests shown in our title? "Multiple comparisons" usually refers to the comparison of mean values from an experiment with multiple groups. For

example, you might compare consumer perceptions of three different advertising displays, labeled A, B, and C, using the data to compare display A with display B, A with C, and B with C. This is the classic "multiple comparisons" application, and SAS software has long offered a variety of methods for such analyses (e.g., Tukey's method for comparing means in PROC GLM). Multiple testing, on the other hand, concerns a broader class of applications. For example, a clinical trial designed to assess "efficacy" of a pharmaceutical compound might be considered "efficacious" if it reduces fever, *or* if it speeds recovery time, *or* if it reduces headaches. Here, there are three tests—a comparison of active compound with placebo for each of the three outcomes. This is an example of "multiple testing." The distinction between multiple comparisons and multiple tests is that, with multiple comparisons, you typically compare three of more mean values of the *same measurement*, while with multiple testing, you consider *multiple measurements*.

One aim of our book is to balance the presentation of multiple comparisons with multiple testing, thereby filling a gap in previous expositions. There have been fewer methods proposed and thus little software developed for analyzing multiple test data, due to difficulties relating to the covariances among the variables. For example, Jason Hsu's excellent book on multiple comparisons (Hsu, 1996) does not treat this problem at all. One aim of our book is to address this lack. We give numerous examples of multiple testing data, and use the SAS system to solve the problems. We also give numerous examples from the multiple comparisons side, as well as examples that are a combination of both.

Outline

The fields of multiple comparisons and multiple testing have tended to consist of something of a hodge-podge of methods. Instead of organizing the book around these methods, we've organized it by the different multiple comparison/testing *problems* you might want to solve. This makes the interrelationships between the methods clearer and makes it easier for you to decide which analysis is appropriate. The topics covered in the book as well as the software tools that are used are shown in Table 1.

2. Examples

One feature of the book is that all of the methods discussed are illustrated by using specific SAS software for some of the most powerful methods currently available on real data from a broad range of applications. The following examples give a good idea of the range of examples in the book. The first employs just the GLM procedure to perform a traditional multiple comparisons analysis, while the other two demonstrate two of the macros that were developed to implement more modern methods.

2.1 - A Simple Balanced ANOVA Example

The following data from Ott (1988) compare weight losses for patients on 5 different treatment groups. The study is balanced, with 10 subjects per group. Display 1 shows how to analyze the data, using Tukey's method (Tukey, 1953).

Table 1 - Outline

Topic/Chapter	Software Tools
Introductory Material	
1: Multiple Comparisons/Testing Applications	
2: General Concepts: Adjusted p-values, p-value plots	PROC MULTTEST, %Rom, %HochBen macros
Comparing treatment means	
3: Balanced one-way	PROC GLM/MEANS statement
4: Unbalanced one-way	PROC GLM/LSMEANS statement
5: Designs with covariates	%SimIntervals macro
6: General functions of means, Confidence Bands	%SimIntervals macro
7: Power and sample size	%SimPower, %PlotSimPower macros
8: Step-down and closure-based testing	REGWQ, PROC MULTTEST %Beggab macro, %Simtests macro
More complicated designs	
9: Generalizations of one-way methods for standard linear models	%SimIntervals, %SimTests macros
10: Heteroscedastic, mixed, and multivariate models	PROC MIXED/LSMEANS statement; %SimIntervals, %SimTests macros
11: Non-normal error distributions	PROC MULTTEST
12: Binary Data	PROC MULTTEST, %MultComp macro
13: Bayesian Methods	%BayesIntervals, %BayesTests macros
14: Further topics	
Logistic regression	%SimIntervals, %SimTests macros
Survival analysis	PROC MULTTEST, %SimIntervals, %SimTests macros
Multiple comparisons with the best	%MCB macro

Display 1. Simple Balanced ANOVA: Program and Output

```
data wloss;
   do diet = 'A','B','C','D','E';
      do i = 1 to 10;
         input wloss @@;
         output;
         end;
      end;
datalines;
12.4 10.7 11.9 11.0 12.4 12.3 13.0
12.5 11.2 13.1  9.1 11.5 11.3  9.7
13.2 10.7 10.6 11.3 11.1 11.7  8.5
11.6 10.2 10.9  9.0  9.6  9.9 11.3
10.5 11.2  8.7  9.3  8.2  8.3  9.0
 9.4  9.2 12.2  8.5  9.9 12.7 13.2
11.8 11.9 12.2 11.2 13.7 11.8 11.5
11.7
;
proc glm data=wloss;
   class diet;
   model wloss=diet;
   lsmeans diet / pdiff cl
                  adjust=tukey;
   ods select DiffMat LSMeanDiffCl;
run;
```

Adjustment for Multiple Comparisons: Tukey
Least Squares Means for effect diet
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: wloss

i/j	1	2	3	4	5
1		0.1604	0.0021	<.0001	0.9988
2	0.1604		0.4547	0.0026	0.0914
3	0.0021	0.4547		0.1828	0.0009
4	<.0001	0.0026	0.1828		<.0001
5	0.9988	0.0914	0.0009	<.0001	

Least Squares Means for Effect diet

i	j	Difference Between Means	Simultaneous 95% Confidence Interval for LSMean(i)-LSMean(j)	
1	2	1.030000	-0.236550	2.296550
1	3	1.780000	0.513450	3.046550
1	4	2.780000	1.513450	4.046550
1	5	-0.120000	-1.386550	1.146550
2	3	0.750000	-0.516550	2.016550
2	4	1.750000	0.483450	3.016550
2	5	-1.150000	-2.416550	0.116550
3	4	1.000000	-0.266550	2.266550
3	5	-1.900000	-3.166550	-0.633450
4	5	-2.900000	-4.166550	-1.633450

Note the following:

- Because the data are balanced, the Tukey method is exact: the simultaneous coverage level of the pairwise intervals is exactly 95% under the usual assumptions.
- The pairwise p-values matrix shows which simultaneous intervals exclude zero, and at what significance levels, thereby allowing simple determinations of significance.
- The program uses a new Version 7 ODS feature to specify that only the p-values and confidence limits for the pairwise comparisons should be displayed.

While the ODS SELECT statement is new, the application of Tukey's method for balanced ANOVA has been available in the SAS System for more than 15 years. Next, we consider an application that is considerably more complex and shows off some of the newer features.

2.2 - Multiple Tests with Multivariate Analysis of Covariance

Suppose you have multivariate multiple-group response data with covariates, where you want to perform the multiple tests associated with a multivariate analysis of covariance, or MANCOVA. A MANCOVA example like this is discussed by Morrison

(1990, pp. 234-236), with response variables Creatinine = amount of the pigment creatinine and Chloride = amount of chloride in urine samples. These are to be compared for subjects in four different obesity groups ("lighter underweight," "heavier underweight," "lighter obese," "heavier obese"), adjusting for a single covariate, Volume. There are six pairwise comparisons of the amount of creatinine and six of the amount of chloride. The family of inferences will include all 12 confidence intervals for the differences in means, all covariate-adjusted. The data, code, and output are given in Display 2:

Display 2. Multiple Comparisons of Means in MANCOVA: Program and Output

```
data Obesity;
   input Group $ Creatinine Chloride
         Volume @@;
   Subject = _n_;
   datalines;
LU 17.6  5.15 205  LU 13.4  5.75 160
LU 20.3  4.35 480  LU 22.3  7.55 230
LU 20.5  8.50 235  LU 18.5 10.25 215
LU 12.1  5.95 215  LU 12.0  6.30 190
LU 10.1  5.45 190  LU 14.7  3.75 175
LU 14.8  5.10 145  LU 14.4  4.05 155
HU 18.1  9.00 220  HU 19.7  5.30 300
HU 16.9  9.85 305  HU 23.7  3.60 275
HU 19.2  4.05 405  HU 18.0  4.40 210
HU 14.8  7.15 170  HU 15.6  7.25 235
HU 16.2  5.30 185  HU 14.1  3.10 255
HU 17.5  2.40 265  HU 14.1  4.25 305
HU 19.1  5.80 440  HU 22.5  1.55 430
LO 17.0  4.55 350  LO 12.5  2.65 475
LO 21.5  6.50 195  LO 22.2  4.85 375
LO 13.0  8.75 160  LO 13.0  5.20 240
LO 10.9  4.75 205  LO 12.0  5.85 270
LO 22.8  2.85 475  LO 16.5  6.55 430
LO 18.4  6.60 490  HO 12.5  2.90 105
HO  8.7  3.00 115  HO  9.4  3.40  97
HO 15.0  5.40 325  HO 12.9  4.45 310
HO 12.1  4.30 245  HO 13.2  5.00 170
HO 11.5  3.40 220
;
data ObesityU;/*Change multivariate
                data format to
                MIXED data format */
   set Obesity;
   Compound = 'Creatinine';
   Amount   =  Creatinine;  output;
   Compound = 'Chloride'  ;
   Amount   =  Chloride;    output;
   keep Subject Group Compound
        Amount Volume;
run;

proc mixed data=ObesityU order=data;
  class Group Compound Subject;
  model Amount = Group*Compound
                 Volume*Compound
     / ddfm=satterth s;
  repeated/type=un subject=Subject;
  lsmeans Group*Compound / cov;
  contrast 'F test'
    Group*Compound 1 0 -1 0 0 0 0 0,
    Group*Compound 1 0 0 0 -1 0 0 0,
    Group*Compound 1 0 0 0 0 0 -1 0,
    Group*Compound 0 1 0 -1 0 0 0 0,
    Group*Compound 0 1 0 0 0 -1 0 0,
    Group*Compound 0 1 0 0 0 0 0 -1;
   ods output LSmeans   = LSmeans;
   ods output Contrasts = Contrasts;
run;

%macro Contrasts;
   C = { 1  0 -1  0  0  0  0  0 ,
         1  0  0  0 -1  0  0  0 ,
         1  0  0  0  0  0 -1  0 ,
         0  0  1  0 -1  0  0  0 ,
         0  0  1  0  0  0 -1  0 ,
         0  0  0  0  1  0 -1  0 ,
         0  1  0 -1  0  0  0  0 ,
         0  1  0  0  0 -1  0  0 ,
         0  1  0  0  0  0  0 -1 ,
         0  0  0  1  0 -1  0  0 ,
         0  0  0  1  0  0  0 -1 ,
         0  0  0  0  0  1  0 -1 }';
   Clab = {"Creatine,LU-HU",
           "Creatine,LU-LO",
           "Creatine,LU-HO",
           "Creatine,HU-LO",
           "Creatine,HU-HO",
           "Creatine,LO-HO",
           "Chloride,LU-HU",
           "Chloride,LU-LO",
           "Chloride,LU-HO",
           "Chloride,HU-LO",
           "Chloride,HU-HO",
           "Chloride,LO-HO"};
%mend;

%macro Estimates;
use Contrasts;
read all var {DenDf} into df;
use LSmeans;
read all var {Cov1 Cov2 Cov3 Cov4
              Cov5 Cov6 Cov7 Cov8}
     into cov;
read all var {Estimate} into EstPar;
%mend;

%SimIntervals(seed=121211,
              nsamp=50000)
```

Two-Sample Multivariate Mean Comparisons
Estimated 95% Quantile = 2.941194

Contrast	Estimate	Standard Error	t Value	--- Pr > |t| -- Raw	Adjusted	95% Confidence Interval	
Creatine,LU-HU	-0.7685	1.2700	-0.61	0.5485	0.9944	-4.5037	2.9667
Creatine,LU-LO	1.5011	1.4203	1.06	0.2969	0.9121	-2.6763	5.6785
Creatine,LU-HO	3.6803	1.4213	2.59	0.0133	0.1136	-0.5000	7.8606
Creatine,HU-LO	2.2695	1.2740	1.78	0.0824	0.4889	-1.4776	6.0166
Creatine,HU-HO	4.4488	1.4435	3.08	0.0037	0.0354	0.2031	8.6945
Creatine,LO-HO	2.1792	1.5905	1.37	0.1783	0.7600	-2.4988	6.8573
Chloride,LU-HU	0.5073	0.7805	0.65	0.5194	0.9918	-1.7882	2.8029
Chloride,LU-LO	0.1501	0.8729	0.17	0.8643	1.0000	-2.4172	2.7174
Chloride,LU-HO	2.1061	0.8735	2.41	0.0206	0.1678	-0.4630	4.6752
Chloride,HU-LO	-0.3572	0.7830	-0.46	0.6507	0.9989	-2.6601	1.9456
Chloride,HU-HO	1.5988	0.8872	1.80	0.0791	0.4752	-1.0105	4.2081
Chloride,LO-HO	1.9560	0.9775	2.00	0.0522	0.3535	-0.9190	4.8310

Note the following concerning this program and analysis:

- The covariate-adjusted Creatinine level is significantly larger for the "heavier underweight" group than for the "heavier overweight" group, with no other comparisons statistically significant.
- PROC MIXED was used to estimate the covariance matrix of the creatinine and chloride measures, adjusted for group and covariate, using the unstructured covariance matrix. Thus, the analysis incorporates separate variances (heteroscedasticity) and multivariate correlation.
- The %SimIntervals macro, described in the book, computes simultaneous confidence intervals for any user-specified collection of contrasts, using the simulation method of Edwards and Berry (1987).
- In this example, the distribution of the test statistics is approximated using a multivariate t distribution with covariance structure and degrees of freedom estimated by PROC MIXED.
- The %Estimates and %Contrasts macros are inputs to the %SimIntervals macro, and give the method great flexibility, allowing you to compute confidence intervals for *any* linear functions of *any* estimates for which you can compute a (possibly approximate) covariance matrix.

2.3 - Frequentist and Bayesian Tests for Multiple Endpoint Data

In clinical trials, one often measures multiple endpoints on each subject, such as (i) a physician's assessment of patient health, (ii) the patient's self-assessment of health, and (iii) an objective measurement such as a chemical analysis of blood sample. You can perform multiple tests for such data easily using PROC MULTTEST, which incorporates the correlations among the variables, as well as possibly non-normal distributional characteristics of the data. This is a frequentist method, and among frequentist methods, the method of PROC MUTLTEST has excellent power and level properties (Reitmeir and Wassmer, 1996).

Lately, Bayesian methods have gained much popularity. Gönen and Westfall (1998) developed a method for analyzing the multiple endpoint data from a Bayesian viewpoint, and calculate posterior

probabilities of multiple null hypotheses which allow both prior correlations and data correlations. The null hypotheses of interest are H_j:θ_j=0, j=1, …, k, which you can test by computing the posterior probabilities p_j = P(H_j is true | Data), as an alternative to the usual frequentist p-values.

To calculate these posterior probabilities, you need priors. The prior used by Gönen and Westfall has the following properties: (i) it allows positive probability on each H_j, (ii) it allows correlation among the binary outcomes (H_j either true or false), and (iii) it allows correlation among the non-zero θ_j realizations. The method is coded in the %BayesTests macro described in the book. Display 3 shows how to use the macro to analyze multiple endpoint data from a real clinical trial of a pharmaceutical compound.

Display 3. Bayesian Multiple Testing: Program and Output.

```
data MultipleEndpoints;
   Treatment = 'Placebo';
   do Subject = 1 to 54;
      input Endpoint1-Endpoint4 @@;
      output;
   end;
   Treatment = 'Drug';
   do Subject = 54+1 to 54+57;
      input Endpoint1-Endpoint4 @@;
      output;
   end;
datalines;
4 3 3 5  5 0 1 7  1 0 1 9  4 0 3 5
3 0 2 9  4 1 2 6  2 0 4 6  2 2 5 5
3 0 1 7  2 0 1 9  4 6 5 5  2 0 2 8
2 7 1 7  1 2 2 9  4 0 3 7  3 0 1 6
3 0 1 6  4 1 4 6  6 0 4 7  3 0 1 8
3 0 1 9  2 1 2 7  6 2 3 5  3 0 4 7
3 0 1 9  2 0 1 9  6 9 6 3  4 9 2 6
2 0 1 7  1 0 1 9  4 0 4 7  3 1 4 6
3 0 3 7  1 0 1 8  6 7 5 4  4 6 2 5
6 19 7 5 6 3 6 6  3 0 5 6  2 4 2 8
1 0 1 8  4 21 5 5 2 0 2 9  4 7 3 5
3 1 2 8  3 3 3 8  4 3 4 6  1 0 1 10
1 0 2 9  3 0 4 5  3 1 1 6  3 4 4 6
5 8 5 5  5 1 5 4  1 0 4 8  1 0 1 10
1 0 1 9  2 1 2 7  4 1 2 5  5 0 5 6
1 4 5 6  5 6 4 6  2 0 2 9  2 2 2 5
1 0 1 10 3 2 3 6  5 4 6 6  2 1 2 8
2 1 2 6  2 1 1 8  3 0 3 9  3 1 2 6
1 0 2 9  1 0 1 9  3 0 3 9  1 0 1 10
1 0 1 9  1 0 1 10 2 0 4 7  5 1 2 6
4 0 5 7  4 0 4 6  2 1 3 6  2 1 1 6
4 0 4 6  1 0 1 8  1 0 2 9  4 1 3 6
4 3 4 5  4 2 5 5  1 0 1 10 3 0 2 8
4 2 2 8  3 0 2 9  1 0 1 10 1 0 1 9
2 0 2 9  2 1 2 8  3 0 3 8  2 4 2 6
2 1 1 9  2 2 2 9  4 0 1 4  3 3 1 8
4 4 3 6  2 0 1 10 4 2 3 6  1 0 1 8
2 0 2 8  5 1 5 5  4 0 4 6
;
data multend1;
   set MultipleEndpoints;
   Endpoint4 = -Endpoint4;
run;

ods listing close;
proc glm data=multend1;
  class Treatment;
  model
    Endpoint1-Endpoint4 = Treatment;
  estimate "Treatment vs Control"
    Treatment -1 1;
  manova h=Treatment / printe;
  ods output Estimates =Estimates
             PartialCorr=PartialCorr;
run;
ods listing;

%macro Estimates;
   use Estimates;
   read all var {tValue}
        into EstPar;
   use PartialCorr;
   read all var
("Endpoint1":"Endpoint4") into cov;
%mend;

%BayesTests(rho=.5,Pi0  =.5);
```

Prior Probability on Individual Nulls is .5
Prior Probability on Joint Null is 0.2000000001
Prior Correlation Between Nulls is .5

Z Statistic	Prior Mean Effect Size	Prior StdDev Effect Size	Posterior Probability	Cov1	Cov2	Cov3	Cov4
2.55256	2.5	1.41421	0.09780	1.00000	0.38262	0.63745	0.69522
2.49145	2.5	1.41421	0.08925	0.38262	1.00000	0.44755	0.42592
1.29349	2.5	1.41421	0.26810	0.63745	0.44755	1.00000	0.63235
2.37971	2.5	1.41421	0.11358	0.69522	0.42592	0.63235	1.00000

You can make decisions concerning whether the nulls are true by using the posterior probabilities. Noting that Bayesian posterior probabilities tend to be much larger than ordinary frequentist p-values (Berger and Sellke, 1987), it is reasonable to consider a probability of 0.10 or less as reasonable evidence against the null hypothesis.

The preceding analysis used rho=.5 and implied a joint prior probability of 0.2 for all null hypotheses. In this study, there was doubt as to whether any of the endpoints were affected by the drug, but there was no doubt that the hypotheses are correlated *a priori*.

3. Conclusions

The analyses presented here offer just a sampling of the modern advances in multiple comparisons that are discussed, with SAS software provided, in Westfall et al. (1999). Not only is the scope of applications extremely large, covering almost all situations of practical interest, but the methods are among the most powerful currently available.

So, what's in this book for you? First of all, we hope to motivate you to take multiple inference problems into account in your data analysis. To meet this need, we present some of the best and most powerful multiple testing/multiple comparisons methods that are currently available. You will see, through our many examples, how to carry out such analyses and how to interpret the results. In some cases, you will find that the improvements obtained using more advanced method over the "usual" multiple comparisons methods (like Bonferroni) are phenomenal, with no cost in terms of increased error rate. In other cases, you will see that there is little gain from using "fancy" multiplicity adjustment procedures. Overall, regardless of the situation, we will emphasize the magnitude of difference between multiplicity adjusted methods versus non-multiplicity adjusted methods, and highlight the benefits of multiplicity adjusted analyses.

References

Berger, J.O., and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence (with comments). *Journal of the American Statistical Association* 82, 112-122.

Edwards, D. and Berry, J.J. (1987). The efficiency of simulation-based multiple comparisons. *Biometrics* 43, 913-928.

Gönen, M. and Westfall, P.H. (1998). Bayesian multiple testing for multiple endpoints in clinical trials. *Proceedings of the American Statistical Association, Biopharmaceutical Subsection*, to appear.

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

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

Morrison, D.F. (1990). *Multivariate Statistical Methods*, 3rd ed, McGraw-Hill, New York.

Ott, L. (1988). *An Introduction to Statistical Methods and Data Analysis*, 3rd Edition, PWS-Kent, Boston.

Reitmeir, P. and Wassmer, G. (1996). One-sided multiple endpoint testing in two-sample comparisons. *Communications in Statistics: Simulation and Computation* 25, 99-117.

Tukey, J.W. (1953). The problem of multiple comparisons. Mimeographed Notes, Princeton University.

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., Johnson,W.O., and Utts,J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84, 419-427.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D. and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests using the SAS® System*, SAS Books by Users, to appear.

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

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