Paper ST-06

Comparison of Correlated Proportions using SAS® PROC GLM: a Simulation Study

Mark S. Litaker, Ph.D., and Daron G. Ferris, M.D.
Office of Biostatistics and Bioinformatics and Department of Family Medicine
Medical College of Georgia
Augusta, GA 30912

ABSTRACT

The performance of an analysis of variance (ANOVA) model to compare correlated proportions was evaluated using simulated data and SAS® PROC GLM. The simulation study is based on a clinical study of three diagnostic techniques, which were evaluated for agreement with biopsy results. 145 study subjects received two independent evaluations using each of the three methods. The data consists of 1s, representing agreement with the biopsy diagnosis, and 0s, representing disagreement. Blocks in the ANOVA represent subjects. Least-squares means and 95% confidence intervals were calculated for the means of the three methods. Correlated observations were generated using observed subject means and hypothesized treatment effects as probabilities of agreement. Validity and power of the procedure were estimated using data simulated under the null and two alternative situations. Coverage probabilities of the confidence intervals were estimated as the proportion of replicate analyses in which the true, or simulated, probability was included in the confidence interval. Alpha for the main effect test was estimated as .051. Power to detect a difference in mean agreement among the methods was 0.67 for a difference of .10 between the extreme methods, and 0.96 for a difference of .15 between the extreme methods.

Funded by a grant from the NCI and AHCPR.

INTRODUCTION

Correlated observations are common in biomedical research. Using a subject as his or her own control can provide added power to compare treatments, compared to evaluating the treatment effects on different individuals. However, the correlations that exist between measurements made on the same subject must be incorporated into the statistical tests and into the calculations of confidence intervals. Ignoring these correlations can result in invalid tests, and tends to yield confidence intervals that do not have the nominal coverage probability. If the observations consist of measurements on a continuous scale, repeated measures analysis of variance (RmANOVA) is a well-established procedure for comparing treatment effects with correlated observations. However, if the observations are categorical, the appropriate analysis becomes less obvious. For two dichotomous treatments, proportions may be compared between two treatments by McNemar's chi-square test. For multiple dichotomous responses, such as comparing more than two correlated proportions, a technique analogous to RmANOVA may be implemented using PROC GLM. The goal of this simulation study was to evaluate the performance of ANOVA using subjects as block in comparing correlated proportions resulting from two observations on each of three treatments per subject, and in calculating confidence intervals on the proportions of agreement.

The Telemedicine Colposcopy Study

A study was conducted to evaluate three methods of cervical examination by colposcope: in-person examination, distant examination using computer imaging, and distant examination using a telemedicine system. Each of 145 patients who previously had abnormal Pap smears and who had received biopsy-based diagnoses, were examined twice using each of the three methods, and a diagnosis was assigned at each of the six examinations. Thus, the experimental design is a 3 x 2 doubly repeated measures layout. The method of examination defines the treatment groups for the ANOVA. Subjects are included as blocks in the analysis. The outcome measure that is of interest for this report is the proportion of examinations in which the diagnosis is in agreement with the biopsy result. The goal is to compare the proportions of examinations which agree with the biopsy result among the three methods, and to calculate 95% confidence intervals on these proportions that are suitably adjusted for the correlation due to the repeated measurements.

The Simulation Study

The goal of this simulation study was to evaluate the performance of a blocked ANOVA, implemented with PROC GLM, in comparing correlated proportions. Correlated data were simulated for six examinations for each of 145 patients for the null situation of no difference among mean agreement levels and for two alternative situations, representing differences of 10% and of 15% between the extreme groups, with the third group having a mean agreement halfway between the other two. Each observation is either a 1, representing agreement with the biopsy diagnosis, or a 0, representing disagreement. Correlations among the observations were induced by incorporating a subject term, consisting of the deviation of the mean agreement observed for each subject from the overall mean agreement, into the probability of agreement. The probability of agreement for each observation was the sum of the overall mean, the subject term and a fixed treatment effect. Individual observations were then generated as a "0" or "1", with the probability of a "1" being the value determined by the sum of the subject and treatment effects. The observation was generated by generating a random number from the uniform(0,1) distribution. If the random number was less than the probability defined by overall mean + subject effect + treatment effect, then the observation was coded as 1, that is, an agreement. Otherwise, the simulated observation was coded as 0, a disagreement. Using this coding, the point estimate of the agreement with biopsy for each treatment is the mean of the observations within that treatment. Comparison of these treatment means must account for within-subject correlations.

Implementing the Simulation Study

Subject effects were obtained by calculating mean agreement for each subject and across all subjects. A variable, dxbxagree, was coded as 0 or 1 representing disagreement or agreement with biopsy, respectively. This variable was averaged across each subject's six observations, and across all subjects, using PROC SUMMARY.

```
* calculate overall and subject
means from observed data;
proc summary data=dims.agree;
class id method;
var dxbxagree;
output out=avgagree mean=mean;
run;
```

Means were calculated for each method, in order to allow calculation of power based on the observed means. A data set with subject id numbers and subject means each

replicated on six records was created:

```
data subj;
set avgagree;
if _type_ eq 2;
subj=mean;
keep id subj;
run;
data subj;
set subj;
  do i = 1 to 6;
    id = id;
   subj=subj;
   output;
    end;
keep id subj;
run;
```

SPECIFYING THE SUBJECT AND TREATMENT EFFECTS

Subject effects are calculated as the difference between the overall mean agreement and the mean agreement for each subject. Treatment effects are defined as the difference between the overall mean and the treatment means. The following code illustrates the alternative situation with a difference of 10% in agreement between the extreme treatments, with the third treatment halfway between them.

```
overall=.55287;  * overall mean
                   agreement;
subject = subj - overall;
           * subject effect;
method1=0;
           * method 1 treatment
            effect;
method2=.05;
           * method 2 treatment
             effect;
method3=-.05;
           * method 3 treatment
             effect;
```

SIMULATING THE AGREEMENT DATA

For each observation, the probability of agreement, that is, the probability that the observed value is 1, is the sum of the overall mean, the subject effect and the treatment effect. A random number is generated from the uniform(0,1) distribution. If the random number is less than or equal to the calculated probability, then the observation is assigned a value of 1. Otherwise, the observation is assigned a value of 0, representing a disagreement.

```
   *  simulate data for power
         analysis;
*options nonotes;
  %macro simpower;
    %do i = 1 %to 10000;
data rep;
set dims.sim10;
if rep = 1 then sim = ranuni(-1);
if rep = 2 then sim = ranuni(-1);
if (method eq 1) then yhat =
overall + subject + method1 ;
if (method eq 2) then yhat =
overall + subject + method2;
if (method eq 3) then yhat =
overall + subject + method3;

if sim le yhat then agree = 1;
if sim gt yhat then agree = 0;
run;
```

ANALYSIS OF THE GENERATED DATA

PROC GLM is used to perform the analysis of variance, using subject ID as the blocking variable. The p-value for the F-test for method of examination is written to a text file. The least squares means are written to an output data set.

```
proc glm data=rep noprint
outstat=pvals;
class id method ;
model agree = id method;
lsmeans method /out=ls;
run;
data pvals;
set pvals;
if _SOURCE_ eq 'method' and
_TYPE_ eq 'SS3';
file 'p_values.txt' mod;
put PROB;
run;
```

Least squares means and standard errors are written to a text file.

```
data ls;
set ls;
file 'lsmeansci.txt' mod;
put method lsmean stderr;
run;
```

Confidence intervals are calculated for the means of each method, as if there were no repeated measures, and are written to a text file.

```
proc summary data=rep noprint;
class method;
var agree;
output out=mcl mean=mean
lclm=lclm uclm=uclm;
run;
data mcl;
set mcl;
if _type_ eq 1;
file 'meansci.txt' mod;
put method mean lclm uclm;
run;

%end;

%mend ;

%simpower;
run;
```

TABULATION OF RESULTS OF THE SIMULATIONS

p-values for the F-test comparing treatment means are read

from the text file. Power is calculated as the proportion of the total number of analyses in which the null hypothesis is rejected. This is obtained by coding a variable, power, as 1 if the null is rejected, that is, if the p-value is less than or equal to .05, and as 0 otherwise. Then the mean of this variable is the estimated power to detect a difference in treatment means of the size specified in the simulation program. For the null situation, with all treatment effects equal to zero, alpha is calculated in the same manner.

```
data probs;
infile 'p_values.txt';
input pvalue;
if pvalue le .05 then power = 1;
if pvalue gt .05 then power = 0;
run;
options pageno=1;
proc means n mean data=probs;
title1
'simulations for power analysis';
title2
'treatment effects: 0, +5, -5';
title3 '55.2%, 60.2%, 50.2%
agreement';
var power;
run;
```

Adjusted onfidence intervals for the treatment means are calculated using the least squares means and their standard errors. A variable indicating whether the calculated confidence interval includes the true value of the treatment mean, that is, the value that was specified in the simulation, is coded, similar to the procedure for calculating power. The mean of this indicator variable is the estimated coverage probability for the confidence interval.

```
data ci;
infile 'lsmeansci.txt';
input method lsmean stderr;
lower = lsmean - 1.96*stderr;
upper = lsmean + 1.96*stderr;

* identify if ci includes true
mean agreement;
include  = 0;
if (method eq 1) and (lower  le
.55287) and (upper  ge .55287)
then include   = 1;
if (method eq 2) and (lower  le
.60287) and (upper  ge .60287)
then include   = 1;
if (method eq 3) and (lower  le
.50287) and (upper  ge .50287)
then include   = 1;
run;

proc sort data=ci;
by method;
options pageno=1;
proc means data=ci n mean;
title 'performance of adjusted
confidence intervals';
var lsmean include;
by method;
run;
```

Confidence intervals that do not take into account the repeated nature of the data were also calculated, and the coverage probabilities calculated. These intervals were expected to be too wide. That is, they were expected to have coverage probabilities in excess of .95, if a nominal confidence level of 95% is specified.

```
data mci;
infile 'meansci.txt';
input method mean lclm uclm;
* identify if ci includes true
mean agreement;
include  = 0;
if (method eq 1) and (lclm  le
.55287) and (uclm  ge .55287)
then include   = 1;
if (method eq 2) and (lclm  le
.60287) and (uclm  ge .60287)
then include   = 1;

if (method eq 3) and (lclm  le
.50287) and (uclm  ge .50287)
then include   = 1;
run;
proc sort data=mci;
by method;
options pageno=1;
proc means data=mci n mean;
title  'performance of unadjusted
confidence intervals';
var mean include ;
by method;
run;
```

METHODS

Simulations were run for four scenarios. The null situation, with all treatment effects equal to zero, was simulated in order to evaluate test validity. That is, whether the test achieves the specified alpha level. Two alternative scenarios were simulated with differences of sizes that would be considered clinically important. In the first of these, the two extreme treatment groups differed by 10% in mean agreement, and in the second, the two extreme groups differed by 15% . In each case, the third treatment group was defined with a mean halfway between the extreme groups. Additionally, simulations were run utilizing the observed treatment means, in order to calculate power for the effect size that was actually found in the study. For each scenario, 10,000 replicate data sets were generated. Power, or alpha for the null scenario, was calculated as the proportion of the total number of analyses resulting in rejection of the null hypothesis. The nominal confidence level was set at 95% for all tests and confidence intervals.

RESULTS

Overall mean agreement between examination and biopsy diagnoses in this study was 55.3%. For the null situation, with all treatment effects set to zero, 507 of the 10,000 analyses resulted in rejection of the null hypothesis. Thus, the alpha level for this set of analyses is 0.0507. For the scenario with a 10% difference in agreements, reflecting treatment mean agreements of 55.3%,

60.3%, and 50.3%, 6693 of 10,000 simulations resulted in rejection of the null hypothesis, or power of 66.9%. A difference of 15% between the treatment groups, or treatment means of 55.3%, 62.8%, and 47.8%, resulted in a power of 95.8%. The observed treatment means of 56.9%, 53.4%, and 55.5% correspond to treatment effects of 1.6%, -1.9%, and 0.2%. Simulations using these observed treatment effects yielded a power of 12.1%.

Coverage probabilities for 95% confidence intervals on the true agreement, based on least squares means, were 94.7%, 95.0%, and 94.9% for the three treatments in the null situation. The overall rate for inclusion of the true means was 94.8% across all three treatments. The corresponding unadjusted confidence intervals showed actual confidence levels of 98.9%, 99.1%, and 99.0% for the three groups. In the 10% difference scenario, the adjusted confidence intervals showed inclusion rates of 94.6%, 93.8%, and 93.8% for the three treatment groups, and was 94.1% overall. Inclusion rates for the unadjusted confidence intervals were 98.8%, 98.1%, and 98.5%, respectively. The 15% difference scenario showed inclusion rates for the adjusted confidence intervals of 95.2%, 90.8%, and 93.0%, with 93.0% overall, and for the unadjusted confidence intervals, 98.9%, 97.3%, and 98.1%.

CONCLUSION

The implementation of a blocked ANOVA using PROC GLM to compare correlated proportions resulting from a 3 x 2 doubly repeated measures experimental design resulted in a valid test, in that the observed alpha level was very close to the specified alpha. The test also demonstrated moderate to high power to detect clinically important treatment differences. Power to detect the much smaller differences in agreement with biopsy which were actually observed in the study was much smaller, at only 12%. The low power in this situation is not an indictment of the procedure, as the observed treatment effects reflect a very small effect size, and in this sense were very close to the null situation. As expected, the ordinary confidence intervals, not taking into account the repeated nature of the data, were unacceptably wide. In all the scenarios, these nominal 95% confidence intervals showed inclusion probabilities of 98 - 99%. Performance of the adjusted confidence intervals is somewhat less clear. In the scenario with a true null hypothesis, and thus no differences among treatments, the adjusted confidence intervals performed well, with inclusion probabilities very close to 95%. However, in the scenarios with true treatment differences, the inclusion probabilities were near 95% for the “middle” treatment, that is, the treatment with a mean equal to the overall mean, but the intervals were shorter than required for the treatment groups with the extreme means. Inclusion probabilities for these intervals ranged from 93.8% for effects of + or - 5%, to 93.0% and 90.8% for treatment effects of 7.5% and -7.5%, respectively. The use of ANOVA with subjects as blocks and 0's and 1's as observations may be useful as an easily-implemented method for comparison of correlated proportions. Basing adjusted confidence intervals on least squares means gives more conservative confidence intervals, compared to unadjusted ones. However, this preliminary investigation suggests that the resulting confidence limits may be over-adjusted, giving intervals that are shorter than would be required to achieve the nominal confidence level. The performance of this adjustment should be investigated further before recommending the procedure as establishing the desired level of coverage.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mark S. Litaker, Ph.D.
Medical College of Georgia
Office of Biostatistics and Bioinformatics
Augusta, GA 30912-4900
Work Phone: (706) 721-4846
Fax: (706) 721-6294
Email:mlitaker@mail.mcg.edu