Calculating a Nonparametric Estimate and Confidence Interval Using SAS® Software

Chris Decker, Glaxo Wellcome Inc., Research Triangle Park, NC

ABSTRACT

In clinical trials nonparametric analysis methods are sometimes used to assess the differences between treatment groups. Most people use either the NPAR1WAY procedure or the FREQ procedure to perform nonparametric analysis. However, no procedures currently exist within SAS software to produce a nonparametric estimate of the difference between treatment groups or a confidence interval to assess the magnitude of that difference. This paper describes the process for calculating the nonparametric estimate of the difference and the confidence interval to assess the magnitude of that difference with the use of the SQL procedure statement and a few DATA steps.

INTRODUCTION

In a clinical trial setting, the data collected are often not normally distributed. Since classical parametric analysis methods require the data to be normally distributed, clinical trials data may require the use of nonparametric analysis methods to assess the differences between treatment groups. One approach commonly used to assess the difference between two treatments is to perform a nonparametric test, such as a Wilcoxon Rank Sum test. This test provides a p-value that may be used to assess the efficacy or safety of a particular treatment.

In addition to the p-value, an estimate (e.g.; Hodges-Lehmann estimate for a Wilcoxon Rank Sum Test) of the difference between two treatment groups, and the confidence interval to assess the magnitude of that difference, is often constructed to assess the difference between two treatments. Though currently several SAS software procedures will calculate the test statistic and associated p-value for a Wilcoxon Rank Sum test, no procedures currently exist within SAS software to produce a nonparametric estimate and confidence interval.

This paper shows how to easily calculate a nonparametric estimate (Hodges-Lehmann) and distribution-free confidence interval (Moses) using PROC SQL and a few data steps. This provides an estimate and confidence interval that are representative of the nonparametric statistical test being performed.

This paper will describe an example using the methodology associated with the Wilcoxon Rank Sum test. However, the SAS code in this example can be modified and applied to any nonparametric method.

BACKGROUND

Data in clinical trials usually consists of two independent random samples, a sample from one treatment group and a sample from another treatment group. When the data is normally distributed the classical parametric analysis methods can be used. However, the data must follow strict assumptions to use these methods. If clinical trial data do not meet these assumptions, nonparametric statistical methods are used. These nonparametric methods only require the data to be on a continuous scale. More often than not, the nonparametric procedures are only slightly less efficient than their parametric counterparts when the underlying populations are normally distributed, and they can be much more efficient than the parametric counterparts when the underlying populations are not normally distributed.

A parametric test, such as the t-test, compares the means of the two samples. A nonparametric method, such as the Wilcoxon Rank Sum Test compares the entire distributions of the two independent samples. The null hypothesis of the Wilcoxon Rank Sum test says the two samples can be viewed as a single sample from one population. The alternative hypothesis is that the first treatment group has a different distribution (or location) than the second treatment group.

The treatment effect, denoted as Δ, is the difference between treatment groups. If parametric methods were used, means could be calculated for each treatment group, and a subtraction of the means can be used to estimate Δ. However, when the data are not normally distributed and the median value of the response variable of interest is calculated for each treatment group, the estimate of the difference in treatment groups is not as straightforward as subtracting one median from the other. Since you are attempting to obtain an estimate based on a difference between distributions, subtracting two medians, or 50 percentiles, is not accurate.

PROCEDURE FOR CALCULATING THE HODGES-LEHMANN ESTIMATE OF THE DIFFERENCE IN TWO MEDIANS

The difference in medians is estimated using the methodology of Hodges-Lehmann. It is a very simple approach. The following steps can be used to estimate Δ:

- form all possible differences between the first treatment group and the second treatment group, in the response variable of interest. For example, if there are 100 patients in each group then 10,000 (100*100) differences would be calculated.
- the estimator Δ is the median of those 10,000 differences.

PROCEDURE FOR CALCULATING THE DISTRIBUTION-FREE CONFIDENCE INTERVAL (MOSES)

The distribution-free confidence interval (Moses), based on the Wilcoxon Rank Sum test, is not quite as straightforward to calculate as the Hodges-Lehmann estimate. The 1-α confidence interval (Δ_L, Δ_U) is given by:

$$\Delta_L = O^{(C\alpha)} \qquad \Delta_U = O^{(XY+1-C\alpha)}$$

where $O^{(1)} \ldots O^{(k)}$ denotes the vector of ordered values of all the possible differences between the two treatment groups (e.g.; the 10,000 differences described above). X is the sample size for the first treatment group and Y is the sample size for the second treatment group. C_α is an integer that approximates the ordered value of the lower confidence interval. For large samples (>30) C_α is a integer approximated by the following:

$$C_\alpha \approx \frac{XY}{2} - Z_{\alpha/2}\left[\frac{XY(X+Y+1)}{12}\right]^{1/2}$$

In general the value of the right-hand side above is not an integer, so round to the closest integer and use that in the confidence interval equation above.

PREPARING YOUR DATA

For this paper it is assumed the data are contained in one data set and are structured as one record per patient. In the code below the data set is called NONPAR and has three variables, PATIENT, TREAT, and RESPONSE, which is the variable being analyzed. The first step is to create two separate data sets, one for each treatment group, and create a separate variable for the RESPONSE within each data set.

```
data treat1 treat2;
  set nonpar;
  if treat='TREAT1' then do;
    resptr1=response;
    output treat1;
  end;
  else if treat='TREAT2' the do;
     resptr2=response;
     output treat2;
  end;
run;
```

USING THE PROC SQL CODE

The next step is to calculate all possible differences between the two treatment groups. In putting all these combinations together the phrase 'many to many merge' comes to mind. One way to calculate this within SAS software is to use PROC SQL.

Below is the PROC SQL code that creates a record for every possible combination.

```
proc sql;
  create table all as
  select treat1.resptr1 , treat2.resptr2,
           (resptr1-resptr2) as diff,
           (1) as merge
  from treat1 treat2
  order by diff;
quit;
```

While SQL code is probably not as intuitive as SAS code most people use, once you use it a few times it's a fairly straightforward and very powerful tool. The code above selects the response variable from each data set and creates a new variable called DIFF, which is the difference between the two values for every possible combination of RESPTR1 and RESPTR2. The data set created from this code is called ALL and contains X*Y records where X is the number of patients in the first treatment group and Y is the number of patients in the second treatment group.

Note that you must calculate DIFF in the order in which you want the difference to be displayed.

CALCULATING THE HODGES-LEHMANN ESTIMATE

Once you have the data set from the PROC SQL, calculating the difference in the medians, the Hodges-Lehmann estimate, is simple. As described earlier, it is the median of the X*Y differences calculated above. Using the data set ALL and the variable DIFF the following code will give you the estimate:

```
proc univariate data=all;
  var diff;
    output out=hlest median=hlest;
run;
```

The above procedure will create a data set called HLEST and contain a variable called HLEST. This is the Hodges-Lehmann estimate of the median difference between the two treatment groups in the response variable of interest.

CALCULATING THE DISTRIBUTION-FREE CONFIDENCE INTERVAL

The following steps will calculate the distribution-free confidence interval (Moses) based on Wilcoxon's Rank Sum Test. The data set ALL, created from the PROC SQL above, contains all the possible X*Y differences between the two treatment groups. An extra code section in the PROC SQL, 'order by diff', orders the differences from low to high. These are the ordered values that will be used in the formulas described earlier.

The first step is to calculate the number of patients with a value in each treatment group. This result is then used to calculate C_α, and the upper and lower ordered values. The following data step uses the original raw data set to perform both these steps.

```
data sampord;
  set nonpar end=last;
   retain nt1 nt2 0;
 /*Count the number of patients in each*/
 /*treatment group*/
   if treat='TREAT1' then nt1+1;
   else if treat='TREAT2' then nt2+1;
 /*Calculate Cα and lower and upper*/
 /*ordered values only on the last*/
 /*record*/
 /*Replace α with actual value*/
 if last then do;

   calpha=round((nt1*nt2/2)-

(probit(α/2)*sqrt((nt1*nt2*(nt1+nt2+1))
   /12)),1);

   loword=calpha;
   uppord=round(nt1*nt2+1-calpha,1);

   merge=1;  /*Dummy variable for merge*/
   output;
 end;
run;
```

Note that you must insert the α-level into the CALPHA equation. This data step will produce lower and upper ordered values of size α.

The final step is to find those ordered values in the ordered data set ALL from the PROC SQL. The following code performs this step.

```
data limits;
  merge all
        sampord
        end=last;
    by merge;
    retain lowcl uppcl;
 /*_n_ is the SAS system variable that*/
 /*contains the observation number*/
  if _n_=loword then lowcl=diff;
  if _n_=uppord then uppcl=diff;
  if last then output;
run;
```

You now have a 1-α distribution-free confidence interval based on the Wilcoxon Rank Sum Test.

SOURCE CODE

Below is the source code collated together.

```
data treat1 treat2;
  set nonpar;
  if treat='TREAT1' then do;
    resptr1=response;
    output treat1;
  end;
  else if treat='TREAT2' the do;
     resptr2=response;
     output treat2;
  end;
run;

proc sql;
  create table all as
  select treat1.resptr1 , treat2.resptr2,
            (resptr1-resptr2) as diff,
            (1) as merge
  from treat1 treat2
  order by diff;
quit;

proc univariate data=all;
  var diff;
    output out=hlest median=hlest;
run;

data sampord;
  set nonpar end=last;
   retain nt1 nt2 0;
 /*Count the number of patients in each*/
 /*treatment group*/
   if treat='TREAT1' then nt1+1;
   else if treat='TREAT2' then nt2+1;
 /*Calculate Cα and lower and upper*/
 /*ordered values only on the last*/
 /*record*/
 /*Replace α with actual value*/
 if last then do;

   calpha=round((nt1*nt2/2)-
   (probit(α/2)*sqrt((nt1*nt2*(nt1+nt2+1))
   /12)),1);

   loword=calpha;
   uppord=round(nt1*nt2+1-calpha,1);

   merge=1;  /*Dummy variable for merge*/
   output;
 end;
run;

data limits;
  merge all
        sampord
        end=last;
    by merge;
    retain lowcl uppcl;
 /*_n_ is the SAS system variable that contains*/
 /*the observation number*/
  if _n_=loword then lowcl=diff;
  if _n_=uppord then uppcl=diff;
  if last then output;
run;
```

OTHER SOFTWARE SOLUTION

Proc-StatXact™ 4 for SAS® users has a procedure available that calculates the Hodges-Lehmann estimate and distribution-free confidence interval. However, since not everyone has this software available to them, SAS is a viable alternative. Please see the reference below for this procedure.

CONCLUSION

With three DATA steps, a PROC UNIVARIATE, and a PROC SQL, one can produce an estimate and distribution-free confidence interval based on the Wilcoxon Rank Sum Test. And until SAS software makes this available in one of it's nonparametric procedures hopefully the steps described in this paper will help you produce results that are more representative of the nonparametric statistical test being performed.

REFERENCES

Hollander, Myles and Douglas A. Wolfe. (1973*), Nonparametric Statistical Methods*, New York: John Wiley & Sons, 75-82.

SAS Institute Inc. (1989), *SAS Guide to the SQL Procedure: Usage and Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

Cytel Software Corporation. (1999), *Proc-StatXact 4 for SAS Users, User Manual*, Cambridge: Cytel Software Corporation.

CONTACT INFORMATION

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

Chris Decker
Glaxo Wellcome, Inc.
5 Moore Drive 17.1463B
Research Triangle Park, NC 27709
Work Phone: 919-483-8989
Fax: 919-483-0272
Email: cd41920@glaxowellcome.com