Paper 171-29

Two Approaches to the Summarization of Data

Huijuan Xu, BiogenIdec MA, Inc., Cambridge MA
John Hotaling, BiogenIdec MA, Inc., Cambridge MA

ABSTRACT

Clinical study reports revolve around the analysis of data relevant to a drug or therapy that is captured during a clinical trial. The data can be presented in an informative manner that is esthetically pleasing through the use of summary tables.

In order to create summary tables for a clinical study report, the data collected for the study may need to be manipulated for analysis. Once the data has been analyzed, it may require some manipulation to fit into the format of the summary table.

The manipulation of data for pre-analysis, the summary of data, and the post-analysis manipulation of data can all be accomplished through the use of either data steps or procedures.

Examples of both approaches are provided in this paper.

INTRODUCTION

For a clinical study report, raw data will be manipulated, analyzed, and presented in summary tables. In general, two kinds of approaches can be used to accomplish this work.

In the first approach, data steps can be used to create derived variables and summary variables and to transpose the data structure into a format that will fit the table.

In the second approach, using procedures rather than data steps, the SQL procedure can be used to create derived variables and summary variables. Procedures such as PROC MEANS, PROC SUMMARY or PROC UNIVARIATE can also be used to create summary variables. The TRANSPOSE procedure can be used to transpose the data structure.

USING DATA STEPS

The use of DATA steps when summarizing data allows the user complete control of the manipulation and analysis of the data. Through use of the RETAIN statement, the user can create flags which track the minimum and maximum values of a variable within a certain by-group. At the end of the group, the mean value for the variable can be created as well.

Once the statistics have been calculated, the data needs to be “flipped” so that the parameters will appear as rows with treatment courses as the columns. This can be accomplished as part of a data step through the use of IF…THEN statements.

SAMPLE CODE FOR DATA STEP

An example of data manipulation required prior to analysis would be the formatting of treatment group or visit information.

```
data blchem;
  set bchem;
  length course $1.;
  course=substr(left(reverse(tpt_name)),1,1);
  visnum=input(compress(upcase(tpt_name),
     'VISITABC'), best3.);
run;
```

Data steps can be used to calculate the n, min, and max through the use of the FIRST, LAST, and RETAIN statements.

```
proc sort data=blchem;
  by visnum course pid;
run;

data summ(keep=visnum course mcalt lcalt scalt count);
  set blchem;
  by visnum course pid;
  retain count tcalt lcalt scalt;
  if first.course then do;
    count=0;
    tcalt=0;
```

```
      lcalt=.;
      scalt=.;
    end;
    count+1;
    tcalt+bcalt;
    lcalt=max(of bcalt lcalt);
    scalt=min(of bcalt scalt);
    if last.course then do;
      mcalt=tcalt/count;
      output;
    end;
  run;

  data bcalt(keep=visnum course dis1-dis3);
    set summ;
    by visnum course;
    length dis1-dis3 $8.;
    dis1=put(count,3.);
    dis2=put(mcalt,4.1);
    dis3=right(put(scalt,3.))||','||
      left(put(lcalt,3.));
  run;
```

Many times, the treatment groups will be presented in a table as columns and the summary statistics as rows. This necessitates a transposition of the data. One way to accomplish this in a data step is through the use of IF... THEN statements.

```
  %macro trans(outdat=, var=, ord=);
      data &outdat(keep=visnum ord colA
               colB colC);
          set bcalt(keep=visnum course &var);
          by visnum;
          length colA colB colC $8.;
          retain colA colB colC;
          ord=&ord;
          if course='A' then colA=&var;
          else if course='B' then colB=&var;
          else if course='C' then colC=&var;
          if last.visnum;
      run;
  %mend trans;

  %trans(outdat=t1, var=dis1, ord=1);
  %trans(outdat=t2, var=dis2, ord=2);
  %trans(outdat=t3, var=dis3, ord=3);

  data all;
    set t1 t2 t3;
  run;
  proc sort data=all;
    by ord visnum;
  run;
```

USING PROCEDURES

The use of procedures when summarizing data takes full advantage of the power of SAS®. The SQL procedure can be used to manipulate data prior to analysis and can also be used to summarize the data.

The MEANS, SUMMARY, and UNIVARIATE procedures are used to summarize the data quickly and accurately once the data is in an analyzable format.

Once the statistics have been calculated, the data may need to be “flipped”. The TRANSPOSE procedure does exactly what it says, it transposes the data to match the format of the summary table.

SAMPLE CODE FOR PROCEDURES

An example of data manipulation required prior to analysis would be the formatting of treatment group or visit information. This can be done with the SQL procedure.

```
proc sql;
  create table blchem as
  select pid, input(compress(upcase(tpt_name),
    'VISITABC'), best3.) as visnum,
    substr(left(reverse(tpt_name)),1,1) as course length=1, bcalt
  from bchem;
quit;
```

The data can be easily summarized with some standard procedures such as MEANS, SUMMARY, or UNIVARIATE.

```
proc means data=blchem nway noprint;
  var bcalt;
  class visnum course;
  output out=summ(keep=visnum course count
    mcalt lcalt scalt)
  n=count mean=mcalt max=lcalt min=scalt;
run;
```

The SQL procedure can also be used to summarize data.

```
proc sql;
  create table summ as
  select course, visnum, count(pid) as count,
         avg(bcalt) as mcalt, min(bcalt) as scalt, max(bcalt) as lcalt
  from blchem
  group by visnum, course
  order by visnum, course;
quit;
```

Many times, the treatment groups will be presented in a table as columns and the summary statistics as rows. This necessitates a “flip” of the data. The TRANSPOSE procedure handles this efficiently.

```
%macro trans(outdat=, var=, ord=);
    proc transpose data=bcalt
        out=&outdat(drop=_name_) prefix=col ;
        var &var;
        id course;
        by visnum ;
    run;

    data &outdat;
        set &outdat;
        by visnum ;
        if last.visnum;
        ord=&ord;
    run;
%mend trans;

%trans(outdat=t1, var=dis1, ord=1);
%trans(outdat=t2, var=dis2, ord=2);
%trans(outdat=t3, var=dis3, ord=3);

data all;
  set t1 t2 t3;
run;
proc sort data=all;
  by ord visnum;
run;
```

COMPARISONS

There are advantages and disadvantages to using data steps and to using procedures to summarize data.

Data steps allow the programmer more flexibility in the summarization of complex data such as the handling of missing values. However, this can make code a little more complicated. In addition, the data step will need to have sorted data in order to use the FIRST and LAST statements.

Procedures do not require a sort prior to summarization and can be more straightforward than data steps. In addition, it is easier to generate statistics such as medians, quartiles, and standard deviations. However, they do not provide the flexibility that data steps do for more complicated summarizations.

Most often, the most efficient way to summarize data will be a combination of both data steps and procedures. Data steps can be used for pre- and post-processing of data and procedures can be used for the summarization and transposing of data.

CONCLUSION

SAS® provides several methods of summarizing data. Depending upon the data and the comfort level of the programmer, data steps and procedures can be used either exclusively or used together to produce accurate, esthetically pleasing results.

REFERENCES

SAS Institute Inc., SAS Language: Rerence Version 6, First Edition, 1990
SAS Institute Inc., SAS Procedures Guide: Version 6, Third Edition, 1990
SAS Institute Inc., SAS Guide to the SQL Procedure: Usage and Reference, Version 6, First Edition, 1989
SAS Institute Inc., SAS Macro Language: Reference, First Edition, 1997

AUTHOR CONTACT INFORMATION

Huijuan Xu
BiogenIdec MA, Inc.
14 Cambridge Center
Cambridge, MA 02142
617-679-2373
Huijuan.xu@biogenidec.com

John Hotaling
BiogenIdec MA, Inc.
14 Cambridge Center
Cambridge, MA 02142
617-679-6475
john.hotaling@biogenidec.com