Preparing Data Using PROC SQL and SAS Macro for Paired Sample Comparison on Intervention Study

Wei Zhao, University of Miami, Miami, FL
Hua Li, University of Miami, Miami, FL

ABSTRACT

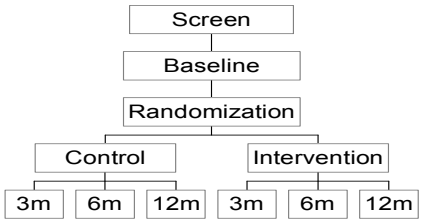
Intervention study is one of the study methodologies in epidemiology research. Appropriate data management and statistical analysis play an important role when evaluating the efficacy of an intervention and converting data into valuable information. SAS is the most powerful data management and statistical analysis tool to better serve this need. This paper demonstrates an example of preparing a SAS data set using PROC SQL and SAS macro for a paired sample comparison on an intervention study to test the changes before and after intervention. The study used the same questionnaire to assess HIV patient risk behaviors for HIV infection before and after a risk reduction intervention. The statistical analysis plan entailed a paired sample proportion comparison using the McNemar test for binary variables and paired sample t test for numeric variables which requires two point data to be one to one match merged. Therefore, all the variable names on one of the data points had to be renamed. We decided to add B after each variable name on the baseline data, since most of the variable names had 8 characters already, SPSS has a limitation to add post-fix in this case, PROC SQL and SAS macro can considerably facilitate SAS data manipulation in both DATA and PROC steps, and other users with little knowledge of SAS can utilize the program to conduct similar tasks. For the beginner/novice SAS programmer this paper is not an attempt to teach PROC SQL and SAS Macro, but rather to inform the users to make an educated choice of available techniques.

Key words: PROC SQL, SAS macro, paired sample, McNemar's Test, Paired T Test.

INTRODUCTION

Intervention study is one of the study methodologies in epidemiology research. Quite often it involves with a questionnaire with more than one thousand variables and most of the variables are categorical. Appropriate data management and statistical analysis play an important role in evaluating the efficacy of the intervention and converting data into valuable information.

Process of Intervention Study: Figure 1.

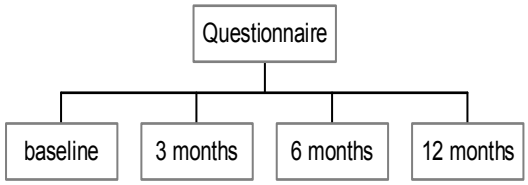


Goals of the Statistical Analysis:

To compare the difference between baseline and over time follow-up. To compare the difference between intervention and control.

Raw data structure:

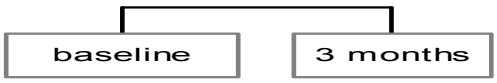
The same questionnaire was used to assess patient behaviors at baseline and over time follow up. Data had been stored separately on different assessments as illustrated in Figure 2.



Therefore data are available to both parallel and concatenated structure depending on the statistical analysis plan. This paper focuses on preparing the SAS data set for the former case.

Required data structure for paired sample analysis: Figure 3.

one to one match merged



It was the researchers' decision which two assessments to compare at the time.

Things needed to be done before match merge:

Match merging of data is done for the purpose of combining records from two or more source tables into a new data file with a combined record layout and is based on a key variable(s). Except for the key variable(s), the

data to be merged should usually contain different fields. If the data sets being merged have common variables, the left-most value(s) is (are) overwritten with the right-most value. It is very important to understand the data in term of content and structure.
Two major things needed to be done before merge:
1) Make sure the key variables below are in each assessment database:
   id: to distinguish each case.
   time: to distinguish the assessment types.
   interven: to distinguish the intervention arms.
2) Rename baseline variables: add B after variables.

Limitation if use other methodologies:

SPSS does not take the variables more than 8 characters. Trimming the variables is going to create problems later on to keep tracking them.

SAS rename for more than one thousand variables might make you feel tedious.

This paper demonstrate an example of preparing a SAS data set using PROC SQL and SAS macro for paired sample comparison on an intervention study to test the changes before and after intervention.

METHODOLOGIES

Structured Query Language (SQL) is a standardized, widely used language for relational database management systems defined by the American National Standard Institute (ANSI).

PROC SQL, a component of base SAS, is the SQL implementation within the SAS system which enables you to retrieve and manipulate data that are stored in tables or views, creates SAS macro variables that contain values from rows in query's results, creates tables, views, and indexes on columns in table.

SAS Macro Facility is a component of base SAS.
Two main tools of SAS Macro Facility are SAS macro variables: typically used to repeatedly insert a piece of text throughout a SAS program and SAS macro programs: uses macro variables and macro programming statements to build SAS programs. Advantages of the SAS Macro Facility are 1) the program you write can become reusable, shorter, and easier to follow, 2) accomplishes repetitive tasks quickly and efficiently, 3) provides a more modular structure to your programs., and 4) make the main program becomes easier to read.

DEMONSTRATION

Baseline data: Table 1.

id	interven	time	v1	v2 …
1	0	1	0	1
2	1	1	1	0
3	0	1	0	1
4	1	1	1	0

Follow up data: 3m (6m / 12m) Table 2.

id	interven	time	v1	v2 …
1	0	2	0	1
2	1	2	1	1
3	0	2	0	1
4	1	2	1	1

Step by step example to merge the data:

/* input baseline sample data set */

```
data bs;
 input id interven time v1 v2;
 cards;
 1 0 1 0 1
 2 1 1 1 0
 3 0 1 0 1
 4 1 1 1 0 ;
```

Same input step as above for 3 month data.

/* Use PROC SQL to put all variable names into macro variables named */

/* NAME1-NAME(n). In this case the old name was concatenated with 'B' */

/* You should assign a number that's at least as high as the number of variables you need to rename. Your library and data set names will go where 'WORK' and 'BS' are now (they must be uppercased) */

```
proc sql noprint;
  select trim(name) || '=' || trim(name)||'B', count(*)
    into :name1 through:name5, :count
  from dictionary.columns
  where libname='WORK' and  memname='BS';
quit;
```

/* Within a macro, use a macro do loop on the RENAME statement to rename the current variables to the new name.*/

```
%macro rename;
  data newbs;
    set bs;
       rename %do i=1 %to &count;
       &&name&i
                %end;;
  run;

  proc print;
  run;
%mend;
```

```
/* invoke the macro   */
  %rename
```

Output of new baseline data: Table 3.

idB	intervenB	timeB	v1B	v2B
1	0	1	0	1
2	1	1	1	0
3	0	1	0	1
4	1	1	1	0

```
/* sort renamed baseline data */

data newbssort;
  set work.newbs(rename=(idB=id));
    id1=id;
proc sort;
  by id;
run;

/* sort 3months data */

data m3sort;
  set work.m3;
    id2=id;
proc sort;
   by id;
run;

/* one to one merge baseline and 3m data */

bs3mPaired(drop=id1 id2);
  merge newbssort (in =in1)  m3sort (in =in2);
    by id;
    if id1=id2;
    if in1;
    if in2;
  run;
```

Output of ready to analyzed data: Table 4.

id	intervenB	timeB	v1B	v2B	interven	time	v1	v2
1	0	1	0	1	0	2	0	1
2	1	1	1	0	1	2	1	1
3	0	1	0	1	0	2	0	1
4	1	1	1	0	1	2	1	1

Statistical Analysis:

For binary variables:
/* test difference by time: McNemar's Test. */

```
proc freq;
  tables v1B * v1 / agree;
  title' paired sample comparison';
run;

/* test changes over time by intervention */

data t;
  set work.bs3mPaired;

/* create new variables for changes due to intervention:
eg. v1c=change of crack use */
   if ((v1B=1 and v1=0)
   or (v1B=0 and v1=0)) then v1c=1;
   else if ((v1B=0 and v1=1)
   or (v1B=1 and v1=1) then v1c=0;

  proc freq;
    tables interven * v1c / chisq;
run;
```

For numeric variables:
/* test difference by time: Paired T Test. */

```
proc ttest;
paired v2 * v2B;
run;

/* test changes over time by intervention */
data t;
  set work.bs3mPaired;

/* create new variables for changes due to intervention:
e.g. knowledge scores change */

     v2C=v2B-v2;

proc ttest;
      class interven;
      var v2C;
run;
```

CONCLUSIONS

SAS is a data management and statistical analysis tool. SAS SQL and SAS macro can considerably facilitate SAS data manipulation in both DATA and PROC steps, especially for a questionnaire with more than one thousand variables. SAS has been giving customers around the world 'The Power to Know.'

REFERENCES

Michele M. Burlew. (2001), SAS MACRO Programming Made Easy. SAS Institute Inc., Cary, NC, USA.

Craig Dickstein, Ray Pass (2002), Data Step vs PROC SQL: What's a neophyte to do?, Proceedings of the 28 Annual SUGI

John Q. Zhang, (2002), More About "INTO: Host-Variable" in PROC SQL: Examples, Proceedings of the 28 Annual SUGI

ACKNOWLEDGEMENTS

Thanks to Dr Lisa Metsch*, Dr. Gayle Dakof * for supporting the presentation of the paper at the meeting.

Thanks to Dr. Shari Messinger* and Dr. Kris Arheart* for comments.

Thanks to Amanda Coltes* for proof reading.

*Department of Epidemiology and Public Health, University of Miami

CONTACT INFORMATION

You comments and questions are valued and encouraged. Contact the authors at:

Wei Zhao
Department of Epidemiology and Public Health, University of Miami
1801 NW 9th Ave, STE300
Miami, FL 33136
Tel: 305-243-6317 (w)
Email: weizhao@med.miami.edu

Hua Li
Department of Epidemiology and Public Health, University of Miami
1400 NW 10th Ave, Suite 1109A
Miami, Florida 33136.
Tel: 305-243-6828 (w)
Email: hli2@med.miami.edu