Paper 410

Using PROC COMPARE for Data Reliability Measurement

Sharon W. Stroud, Medical Review of North Carolina, Cary, NC

ABSTRACT

Data validation and reliability is a critical step in any research project. Reliability refers to the repeatability or reproducibility of data collection. PROC COMPARE can be used to quantitatively measure reliability. Using Excel as a data conversion tool for SAS output we can measure variable specific reliability as well.

INTRODUCTION

At Medical Review of North Carolina, reliability is measured by having abstractors re-abstract a portion of medical records they previously abstracted. The extent to which the abstractors agree with themselves at two different points in time is called the intra-rater reliability. Inter-rater reliability refers to the degree to which two different abstractors agree with each other. Reliability can also be calculated for each variable being abstracted. Our standards require that all reliability measures must be 90 percent or higher.

PROCEDURE

We have two files – one containing the original abstracted data and one with re-abstracted data. First, remove variables for which you would not expect agreement, such as date of abstraction and comment fields. You only need to remove the variables from one datasheet, because only variables that are in common are compared. Additional records in either file are not a problem, because only observations in common, determined by the id variables, are compared. The Compare procedure compares the two datasets and identifies observations that differ. Here is sample code for running PROC COMPARE:

```
proc sort data=original; by hic; run;
proc sort data=valfile; by hic; run;

proc compare data=original compare=valfile;
id hic adm;
run;
```

From this output we calculate the proportion of agreement (reliability) by:

$$1 - \frac{\text{\# of non-matches}}{(\text{\# of cases}) \times (\text{\# of variables})}$$

If we want to do a variable specific analysis, we look at the Values Summary section of the output and the proportion of agreement for each variable is 1 – (# nonmatches / # cases). If several abstractors are working on a project, variable nonmatches must be summed for all abstractors to get the total variable reliability. Doing this by hand from the SAS output would not be difficult if you had only a few variables, but with many variables there is a lot of room for error. PROC COMPARE does produce output datasets, but none that contain the information needed for the variable reliability.

To get the variable summary information into a SAS dataset, save copies of the output, open in Excel, then use DDE (dynamic data exchange) to import the Excel files into SAS datasets.

First, run PROC COMPARE for each abstractor (5 in this case). Use the novalues option to suppress the variable comparison section of the output. Save each output as a text file, then open the file in the SAS program editor and delete everything that comes before the value comparison section. Save the file again so that the only data in the text file is the value comparison.

Then open each file in Excel using the text import wizard. Edit the Excel files so that the variable names are in column 1 and the number missed is in column 2. Delete the columns containing type, len, maxdif and missdif. The excel spreadsheets should look like this:

RACE	2
CHF	2
MI	4
AFIB	2
BUN	3
CREAT	3
CHOL	1
BP	2
DISACE	2
SODIUM	2

Save each Excel file, leave the files open, minimize Excel and return to SAS. Then run the following program.

```
**********************************************************************
* Prior to running this program, run proc compare                    *
* (5 times – one  per abstractor) with the novalues                  *
* option.  Save the output and  open in Excel.                       *
* Create excel spreadsheets with the variable name                   *
* (var) and number missed (n1-n5).  Do not put variable              *
* names in the spreadsheets. Put data in columns 1                   *
* and 2. The spreadsheets must be open when running                  *
* this program.                                                      *
**********************************************************************

filename xlfile DDE
'excel|s:\sas\sharon\anginafu\cwcw.xls!c1:c2';
filename xlfile2 DDE
'excel|s:\sas\sharon\anginafu\sgsg.xls!c1:c2';
filename xlfile3 DDE
'excel|s:\sas\sharon\anginafu\vmvm.xls!c1:c2';
filename xlfile4 DDE
'excel|s:\sas\sharon\anginafu\vbvb.xls!c1:c2';
filename xlfile5 DDE
'excel|s:\sas\sharon\anginafu\\kbkb.xls!c1:c2';

data x1;
 infile xlfile;
 input var $ n1;
run;

data x2;
 infile xlfile2;
 input var $ n2;
run;

data x3;
 infile xlfile3;
 input var $ n3;
run;

data x4;
 infile xlfile4;
```

```
 input var $ n4;
run;

data x5;
 infile xlfile5;
 input var $ n5;
run;

proc sort data=x1; by var;
proc sort data=x2; by var;
proc sort data=x3; by var;
proc sort data=x4; by var;
proc sort data=x5; by var;

data new;
 merge x1 x2 x3 x4 x5;
 by var;
 if n1=' ' then n1=0;
 if n2=' ' then n2=0;
 if n3=' ' then n3=0;
 if n4=' ' then n4=0;
 if n5=' ' then n5=0;
 total=n1+n2+n3+n4+n5;
run;

proc sort data=new;
by descending total;
run;

proc print data=new;
run;

****************** end of program**********************;
```

The results are the number of mismatches per variable, summed across all abstractors. Below is a sample output for 10 variables and 5 abstractors:

The SAS System

OBS	VAR	N1	N2	N3	N4	N5	TOTAL
1	HYPERT	4	5	1	1	3	14
2	SODIUM	2	3	3	1	4	13
3	AFIB	2	2	0	2	3	9
4	BP	2	2	0	0	3	7
5	CHF	2	3	1	0	1	7
6	DISACE	2	1	0	0	3	6
7	MI	4	1	0	0	1	6
8	BUN	3	1	1	0	1	6
9	CREAT	1	1	0	2	1	5
10	RACE	2	1	0	0	1	4

Variable reliability can easily be calculated from here by dividing the total missed for each variable by the number of observations and subtracting from one.

CONCLUSION

PROC COMPARE is an excellent tool for quantitatively measuring reliability. While I have not found any of the output datasets for this procedure useful, using Excel to create SAS datasets from the SAS output works well.

CONTACT INFORMATION

Sharon W. Stroud
Medical Review of North Carolina
5625 Dillard Drive
Cary, NC 27511
(919) 851-2955
sstroud@mrnc.org