Paper 197-29

A SAS® Program Computes the Crude, Stratified and Mantel-Haenszel Odds Ratio for Case-Control Study Analysis of Nx2xK Tables

Ilene Brill, University of Alabama at Birmingham, Birmingham, AL
Fabio Barbone, Universita' di Udine, Udine, Italy

ABSTRACT

For the analysis of epidemiologic case-control studies, a SAS macro program was constructed to compute estimates of the exposure odds ratio (OR) (ratio of exposed to unexposed cases divided by the ratio of exposed to unexposed noncases) beyond the standard 2x2 table. For each exposure level compared to the referent zero exposure level, the program calculates crude ORs, stratified ORs based on all possible combinations of confounders and the Mantel-Haenszel estimator of the OR adjusted for confounding. In epidemiology confounding, defined as the mixing of the effect of another factor (the confounder) associated with both the outcome and the exposure, is a major threat to validity. The program accommodates varying number of confounders and levels for each confounder, one response variable with two levels (case/control) and one exposure with varying number of levels. The user designates as arguments to the macro: temporary input data set name, name prefix for the confounders PREFIX1-PREFIXn, outcome and exposure variable names, number of exposure levels and the variable name for the weights assigned to the frequency values. The epidemiology instructor preferred using these methods of computing ORs over logistic regression to lead students to a better understanding of the relation among the contributing variables.

INTRODUCTION

Epidemiology research studies disease occurrence and its association with various risk factors in human populations. Population studies are the research tool employed to help describe and explain patterns and trends between risk factors and disease outcomes. A selected group of people is followed prospectively or retrospectively over time and information on disease outcomes and related factors are collected in these studies. Alternatively, cross-sectional studies can be done which measure prevalence of disease at a point in time and associated exposures of interest. Case-control studies are one type of study design used for this purpose and can be carried out as cross-sectional or retrospective study designs. Cases are defined as those with the disease and the comparison group consists of the controls who are without disease. The investigator selects the cases and controls from separate populations of available cases and noncases (Kleinbaum, et al., 1982). Data from case-control studies are usually displayed in a standard 2x2 table where each of the four cells represents the frequency of either cases or controls classified according to the presence or absence of an exposure. The table is illustrated as follows:

	Exposed	Unexposed
Cases	a	b
Controls	c	d

The effect measure of interest in a case control study is the exposure odds ratio (EOR), which is the ratio of exposed to unexposed cases (a/b) divided by the ratio of exposed to unexposed noncases (c/d). Alternatively this can be stated as the odds of being exposed among cases divided by the odds of being exposed among noncases (Kleinbaum, et al., 1982). Usually, the EOR can be interpreted as an estimate of the relative risk, the most common measure of association in epidemiology. This crude estimate is simplified to the following equation:

$$EOR = ad/bc$$

From BASE SAS, PROC FREQ (SAS Procedures Guide, Version 8, Reference) you can request to assess the association of the row and column variables in the 2×2 table. In addition, the Cochran-Mantel-Haenszel (CMH) statistics test for association in sets of tables. These statistics assess association between the row and column variables after adjusting for the stratification variable(s). CMH methods are targeted at detecting average effects across strata, and sample size requirements are based on total frequencies, not individual cell sizes. Accordingly, also Mantel-Haenszel (MH) (Mantel N. and Haenszel W., 1959) adjusted ORs can be computed using PROC FREQ option CMH for Nx2x2 tables. What BASE SAS does not do is to compute stratified, crude and adjusted ORs for Nx2xK, i.e., analyses of the effects of an exposure that is described not as dichotomous but as categorized in ≥ 3 levels. For this reason, logistic regression analysis is the common method chosen for computing ORs for exposure variables with ≥ 3 levels with or without adjustment for confounding. In epidemiology confounding, defined as the

mixing of the effect of another factor (the confounder) associated with both the outcome and the exposure, is a major threat to validity. The epidemiology instructor preferred a different method of computation that would enable his students to better understand the relation among the contributing variables to the OR calculations. As requested by the epidemiologist, the programmer designed a macro program in SAS to fulfill the students' needs. For each exposure level compared to the referent zero exposure level, the program calculates crude ORs, stratified ORs based on all possible combinations of confounders and the MH estimator of the OR adjusted for confounding.

The macro program accommodates varying number of confounders and levels for each confounder, one response variable with two levels (case/control) and one exposure with varying number of levels. The user designates as arguments to the macro: the temporary input data set name, a name prefix for the confounders such as PREFIX1-PREFIXn, the outcome and exposure variable names, the number of exposure levels and the variable name for the weights assigned to the frequency values.

Output produced includes frequency tables of crosstabulations of disease by exposure for all possible combinations of confounding variables and their values; crude ORs, stratified ORs and the MH adjusted estimate of the OR; and the respective upper and lower confidence intervals for each of these OR measurements.

INPUT DATA SET

The input data set provided to the program is a temporary SAS data set that must adhere to a specific structure. It allows some flexibility when naming variables. The temporary SAS data set name is also one of the macro parameter values listed when invoking the macro. The temporary SAS data set can be created by inputting your values using a DATA step with a CARDS statement followed by your data values as shown below. Alternatively, one can read in a permanent data set in a SET statement and output it as a temporary data set in the DATA statement. You can create this data set either in the beginning of the program by replacing the current example data step or alternatively you can insert the programming statements after the end of the macro and before the line invoking the macro. Remember to change your temporary data set name if necessary in the macro invocation (ONE is the default example name). Below is the example data that was initially provided prior to designing this SAS program.

```
    DATA one;
    INPUT c2 w d e c1;
    CARDS;
    1 944 1 1 1
    0 600 1 1 1
    1 709 1 1 0
    0 400 1 1 0
    1 140 1 0 1
    0 200 1 0 1
    1 114 1 0 0
    0  50 1 0 0
    1 344 1 2 1

    0 200 1 2 1
    1 209 1 2 0
    0 100 1 2 0
    1 628 0 1 1
    0 828 0 1 1
    1 300 0 1 0
    0 400 0 1 0
    1 186 0 0 1
    0 286 0 0 1
    1  98 0 0 0
    0 198 0 0 0
    1 086 0 2 1
    0 086 0 2 1
    1  38 0 2 0
    0  98 0 2 0
;
run;
```

In this example data set the third variable D is the outcome variable name. The outcome variable can be any valid SAS name and must be coded with only 2 levels: 0=control or nondiseased, 1=case or diseased. The fourth variable input is E, the exposure variable. The variable name can be any valid SAS name and can have multiple (n) levels coded as (0,1, ... n -1) where 0 is the reference category. This example has two confounder variables C1 and C2 (fifth and first variables input respectively). The confounder variables can have any prefix though the prefix must be

the same for all the confounder variables. There can be an infinite number of confounder variables (we hope this works, limited checking of the program was possible due to time constraints). For example they could be named PREFIX1—PREFIXn where n is the total number of confounding variables. For each confounder there can be any number of levels (n) where this n can be a different value for each confounder. Confounder variables must be coded as (0,1,…n -1) where 0 is the lowest level. It is not necessary to specify in the macro the total number of levels for each confounder. The second variable input is W or the weight variable. In this case W represents the frequency of observations with that particular combination of outcome, exposure and confounding variable values.

MACRO STATEMENT AND PARAMETERS

The main macro within this program is named ORNx2xK. There are 7 macro parameters defined in the macro statement:

```
%macro ORNx2xK(dname,c,n,d,e,en,w);
```

The values for each of these parameters are assigned in the macro invocation statement as follows:

DNAME is the input SAS temporary data set name
C is the prefix for confounder variables PREFIX1-PREFIXn
N is the number of confounder variables
D is the outcome variable name
E is the exposure variable name
EN is the number of levels for the exposure variable
W is the weight variable name

MACRO INVOCATION

For the data example cited above, the macro is invoked with the following statement:

```
%ORNx2xK(one,c,2,d,e,3,w)
```

ONE is the input SAS temporary data set name.
C is the prefix for the confounder variables C1-C2
2 is the number of confounder variables
D is the outcome variable name
E is the exposure variable name
3 is the number of levels for the exposure variable E
W is the weight variable name

DESCRIPTION OF THE OUTPUT

The first part of the output produced is frequency tables of crosstabulations of disease by all exposure levels for all possible combinations of the values of the confounding variables. For the remainder of the output it is necessary to understand a few naming conventions used. When numbers are included in the variable names they refer to values of either the exposure variable or values of the confounding variables. It is easy to understand how variables would be named if you add more exposure or confounder levels or more confounder variables after viewing the examples below. Variables names are listed under each of the explanations below.

The first number after the OR in the variable name is the exposure level of interest, and the number following it is 0, which is the reference category. This applies to all variable names.

Crude Odds Ratio and its lower and upper confidence limits:

For exposure level 1 compared to exposure level 0:
crudeOR10, LcrudeOR10, UcrudeOR10

For exposure level 2 compared to exposure level 0:
crudeOR20, LcrudeOR20, UcrudeOR20

For exposure level 3 compared to exposure level 0:
crudeOR30, LcrudeOR30, UcrudeOR30

Mantel-Haenszel Odds Ratio and its lower and upper confidence limits:

For exposure level 1 compared to exposure level 0:
adjOR10, LadjOR10, UadjOR10

For exposure level 2 compared to exposure level 0:
adjOR20, LadjOR20, UadjOR20

For exposure level 3 compared to exposure level 0:
adjOR30, LcrudeOR30, UcrudeOR30

Stratified Odds Ratio and its lower and upper confidence limits:

The first number after the C in the variable name refers to Confounder 1 and the second number after the C refers to Confounder 2. If there were three confounders you would see a third number representing the value of the third confounder which is the third character after the letter C in the variable name.

For exposure level 1 compared to exposure level 0:

Confounder 1 equals 0, Confounder 2 equals 0
OR10C00, LOR10C00, UOR10C00

Confounder 1 equals 1, Confounder 2 equals 0
OR10C10, LOR10C10, UOR10C10

Confounder 1 equals 0, Confounder 2 equals 1
OR10C01, LOR10C01, UOR10C01

Confounder 1 equals 1, Confounder 2 equals 1
OR10C11, LOR10C11, UOR10C11

For exposure level 2 compared to exposure level 0:

Confounder 1 equals 0, Confounder 2 equals 0
OR20C00, LOR20C00, UOR20C00

The rest of Exposure level 2 names are similar to the naming convention described above for Exposure level 1 compared to exposure level 0.

PRINTING NOTES

The following OPTIONS statement is included in the program.

```
options pagesize=66 linesize=122
   pageno=1 missing=' ' date
   FORMCHAR="|----|+|---+=|-/\<>*";
```

In the Print Setup you should select a laser jet printer that prints to letter size paper, portrait orientation, with a SAS Monospace, regular font and a font size 8. These specifications are important settings so that the paging of the output looks correct. If you choose the above settings then you can ignore the following WARNING message printed in the SAS log:

The current page size is too small to hold the procedure output and all of the titles and footnotes, so some may have been dropped. Increase the page size to see all of the titles and footnotes.

The program generates the usual SAS output listing and an RTF file using the SAS Output Delivery System. Both the RTF file name and the footnote with the program name are named using the same first level name as provided in the first macro program statement below which defines the macro variable CMT. All title statements have been predefined in this program and should not be altered as this may change the appearance of the output. The title below appears only on the pages with frequency tables. You may change the name of the program in the first line presented below (this appears in the beginning of the program presented in the appendix).

```
%let cmt=OR_nx2xk;
%let pgm=&cmt..sas;
ods rtf file="c:\&cmt..rtf";
footnote "&pgm.";
title1 'Frequencies of Outcome by Exposure for all combinations of confounding
variables';
```

CONCLUSION

It is best that epidemiology students before reporting overall OR effect measures in case-control studies evaluate the distribution of the cases and controls by category of exposure across levels of potential confounders. This SAS macro program with its inherent flexibility easily allows the students to observe the relation between outcome and exposure stratified by various combinations of confounder variable values and to compare the results derived from crude versus adjusted MH methods of computation. According to SAS technical support such a procedure did not exist; therefore, it was necessary to develop our own program for this purpose. The macro facility in SAS enabled an easy method for specifying the distinct features of an input data set through the use of macro parameters. Further testing of this program is still necessary with greater number of confounders and exposure levels. Feedback from others with regard to their success using this program would be greatly appreciated. Meanwhile the epidemiology students appeared grateful for this program and used it for class assignments. We have received continued communication with the students regarding this program even post-class semester; apparently, they found it to be a useful tool.

REFERENCES

Kleinbaum, D.G., Kupper L.L. and Morgenstern, H. (1982), *Epidemiologic Research, Principles and Quantitative Methods*, Belmont, CA: Lifetime Learning Publications.

Mantel, N. and Haenszel, W. (1959), “Statistical aspects of the analysis of data from retrospective studies of disease,” *Journal of the National Cancer Institute*, 22(4), 719-748.

SAS Institute Inc. (1999), *SAS Procedures Guide, Version 8, Reference, Volume 1*, Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

Many thanks to Robert Matthews and Dhong-Jin Kim for their network, hardware, software support and technical advice.

CONTACT INFORMATION

Ilene Brill
Department of Epidemiology
University of Alabama at Birmingham
Ryals School of Public Health, Room 533B
1665 University Boulevard
Birmingham, AL 35294-0022
Work Phone: (205) 934-7160
Fax: (205) 975-7058
Email: ibrill@ms.soph.uab.edu

APPENDIX 1

```
Note:  Add programming statements in PRINTING NOTES here.
Note:  You may add data step next (see example under INPUT DATA SET)
Note:  You may choose to remove or comment out the OPTIONS statement below.
options mprint symbolgen mtrace;
```

```
**** Macro to list all confounders in PROC FREQ TABLES statement;

 %macro names(number,prefix);
%local i;
%do i=&number %to 1 %by -1;
 &prefix&i.*
%end;
%mend names;

**** Lists macro variables CN1-CNn which will have values for the # of levels
     in each confounder macro variable C1-Cn (used in statement assigning these
     variables as global macro variables);

%macro globname(number,prefix=cn);
%local i;
%do i=1 %to &number;
 &prefix&i
%end;
%mend globname;

%macro ornx2xk(dname,c,n,d,e,en,w);

options pagesize=66 linesize=122 pageno=1 missing=' ' date
        FORMCHAR="|----|+|---+=|-/\<>*";

*** Puts in the first argument for macro NAMES the # of confounding variables - i.e.
    c1 - cn where n=2 then the argument will be 2, and into argument 2 the prefix
    for confounding variable names.  Creates frequency tables for all combinations
    of confounder variables by outcome by exposure variable and weighted by the
    given weight variable;
proc freq data=&dname; tables %names(&n,&c) &d*&e/sparse out=aaa; weight &w; run;

*** Variables for number of levels (CN1 .. CNn) for each confounding variable
    C1.. Cn are assigned as global macro variables;
%global %globname(&n);

data aaa;
  set aaa nobs=totn;

*** Creates variable TOTTABLE and macro variable TOTT for total number of tables
    created by the PROC FREQ (this is really just the # of possible confounder level
    combinations);
tottable=totn/(&en*2);
call symput('tott',trim(left(put(tottable,4.))));
run;

data aaa1(drop=x i ndim j s);
 set aaa(drop=percent) end=last;
length id1-id&tott $ 50;

array tableid {*} $ id1-id&tott;
array cs {*} &c.1-&c&n;


**** Creates variables TABLEID1-TABLEIDn and macro variables S1-Sn in CALL SYMPUT
     statement below with values which are just the numerical values of confounders
     1-n which characterize each table produced as output from the PROC FREQ
     statement.  i.e. C1=0 C2=0  then TABLEID1 is '00';

ndim=dim(cs);
if _N_=1 then i=1;
if _N_= i*&en*2 then do;
```

```
 do j=1 to ndim;
  tableid{i}=trim(tableid{i})|| trim(left(put(cs{j},2.)));
 end;
  i=i+1;
end;

retain id1-id&tott i;

if last then do;
 %do k=1 %to &tott;
  s=trim(left(tableid{&k}));
  call symput("s&k",s);
 %end;
end;


*** Creates macro variables CN1 .. CNn for the total number of levels for each
    confounding macro variable C1 .. Cn;
if last then do;
 %do i=1 %to &n;
    x=&c&i+1;
    call symput("cn&i",x);
 %end;
end;
run;

%put _global_;

*** Outputs variable for total # of tables;
data more(keep=tottable);
 set aaa1 end=last;
if last then output;
run;

proc transpose data=aaa1(drop=tottable id1-id&tott) out=taaa; run;

data taaa1;
if _N_=1 then set more;
 set taaa;
where _name_='COUNT';

%global ncol ncell nfrac;

*** Creates variable TOTCOL and macro variable NCOL for the total number of columns
    generated by PROC TRANSPOSE which equals:
the total number of tables X 2(# of outcome levels) X the number of exposure levels;
totcol=tottable*2*&en;
call symput('ncol',trim(left(put(totcol,4.))));

*** Creates variable TOTCELL and macro variable NCELL for the total number of cells
    used in the Crude Odds ratio calculations:
    4 x number of exposure levels (2 x 2 tables will always have 4 cells);
totcell=4*(&en-1);
call symput('ncell',trim(left(put(totcell,4.))));

*** Creates variable TFRAC and macro variable NFRAC for the total # of exposure
    levels minus 1;
tfrac=&en-1;
call symput('nfrac',trim(left(put(tfrac,3.))));
run;

%put _global_ _local_;
```

```
data taaa2(drop=i j);
 set taaa1;
array colnum {*} col1-col&ncol;
array cells {*} cell1-cell&ncell;
array numer {*} numer1-numer&nfrac;
array denom {*} denom1-denom&nfrac;
array denomci {*} denomci1-denomci&nfrac;

do i=1 to totcell;
  cells{i}=0;
end;

do i=1 to tfrac;
 numer{i}=0;
 denom{i}=0;
 denomci{i}=0;
end;

%do i=1 %to &en-1;
 do j=1 to tottable;
  cells{4+((&i-1)*4)}=cells{4+((&i-1)*4)} + colnum{((j-1)*(&en*2))+(&en+1+&i)};
  cells{1+((&i-1)*4)}=cells{1+((&i-1)*4)} + colnum{((j-1)*(&en*2))+1};
  cells{2+((&i-1)*4)}=cells{2+((&i-1)*4)} + colnum{((j-1)*(&en*2))+(1+&i)};
  cells{3+((&i-1)*4)}=cells{3+((&i-1)*4)} + colnum{((j-1)*(&en*2))+(&en+1)};
 end;
%end;


*** Crude OR ***;
%do i=1 %to &en-1;

 crudeOR&i.0=((cells{4+((&i-1)*4)})*(cells{1+((&i-1)*4)}))/((cells{2+((&i-1)*4)})*
                  (cells{3+((&i-1)*4)}));

 LcrudeOR&i.0=exp(log(crudeOR&i.0)-1.96*(sqrt((1/(cells{1+((&i-1)*4)}))+
   (1/(cells{2+((&i-1)*4)}))+(1/(cells{3+((&i-1)*4)}))+(1/(cells{4+((&i-1)*4)})))));

 UcrudeOR&i.0=exp(log(crudeOR&i.0)+1.96*(sqrt((1/(cells{1+((&i-1)*4)}))+
   (1/(cells{2+((&i-1)*4)}))+(1/(cells{3+((&i-1)*4)}))+(1/(cells{4+((&i-1)*4)})))));

%end;


*** Stratified OR for combinations of levels of confounding variables C1..Cn with
    values Cn=0, Cn=1 .... Cn=n, E has EN values ***;
%do i=1 %to &en-1;
  %do j=1 %to &tott;
OR&i.0C&&s&j=((colnum{((&j-1)*(&en*2))+(&en+1+&i)})*(colnum{((&j-1)*(&en*2))+1}))/
             ((colnum{((&j-1)*(&en*2))+(1+&i)})*(colnum{((&j-1)*(&en*2))+(&en+1)}));

LOR&i.0C&&s&j=exp(log(OR&i.0C&&s&j)-1.96*(sqrt((1/colnum{((&j-1)*(&en*2))+1})+
              (1/colnum{((&j-1)*(&en*2))+(1+&i)})+
              (1/colnum{((&j-1)*(&en*2))+(&en+1)})+
              (1/colnum{((&j-1)*(&en*2))+(&en+1+&i)}))));

UOR&i.0C&&s&j=exp(log(OR&i.0C&&s&j)+1.96*(sqrt((1/colnum{((&j-1)*(&en*2))+1})+
              (1/colnum{((&j-1)*(&en*2))+(1+&i)})+
              (1/colnum{((&j-1)*(&en*2))+(&en+1)})+
              (1/colnum{((&j-1)*(&en*2))+(&en+1+&i)}))));
  %end;
%end;
```

```
*** MH adjusted OR ***;
%do i=1 %to &en-1;
 do j=1 to tottable;
 *** For numerator and denominator of MH adjusted odds ratios;
 numer{1+((&i-1)*1)}=numer{1+((&i-1)*1)} +
                    ((colnum{((j-1)*(&en*2))+(&en+1+&i)}*colnum{((j-1)*(&en*2))+1})/
                     (colnum{((j-1)*(&en*2))+1} + colnum{((j-1)*(&en*2))+(1+&i)} +
             colnum{((j-1)*(&en*2))+(&en+1)} + colnum{((j-1)*(&en*2))+(&en+1+&i)}));

 denom{1+((&i-1)*1)}=denom{1+((&i-1)*1)} +
                   ((colnum{((j-1)*(&en*2))+(1+&i)}*colnum{((j-1)*(&en*2))+(&en+1)})/
                    (colnum{((j-1)*(&en*2))+1} + colnum{((j-1)*(&en*2))+(1+&i)} +
             colnum{((j-1)*(&en*2))+(&en+1)} + colnum{((j-1)*(&en*2))+(&en+1+&i)}));

 *** For denominator part of MH adjusted odds ratios lower and upper confidence
    intervals;
aa=1/colnum{((j-1)*(&en*2))+1};              if aa=. then aa=0;
bb=1/colnum{((j-1)*(&en*2))+(1+&i)};         if bb=. then bb=0;
cc=1/colnum{((j-1)*(&en*2))+(&en+1)};        if cc=. then cc=0;
dd=1/colnum{((j-1)*(&en*2))+(&en+1+&i)};     if dd=. then dd=0;
denomci{1+((&i-1)*1)}=denomci{1+((&i-1)*1)} + (1/(aa + bb + cc + dd));
 end;
%end;

%do i=1 %to &en-1;
  adjOR&i.0=numer{&i}/denom{&i};
  LadjOR&i.0=exp(log(adjOR&i.0)-1.96*(sqrt(1/denomci{&i})));
  UadjOR&i.0=exp(log(adjOR&i.0)+1.96*(sqrt(1/denomci{&i})));
%end;

**** Print output for Crude, Stratified and Adjusted MH Odds Ratios;
options pagesize=20 linesize=122 date number formdlim=' ';
title1 '--- Crude Odds Ratios and Adjusted MH Odds Ratio Calculations ---';
footnote ' ';

%do i=1 %to &en-1;
 proc print data=taaa2 noobs;
 var crudeOR&i.0 lcrudeOR&i.0 ucrudeOR&i.0 adjOR&i.0 LadjOR&i.0 UadjOR&i.0;
 run;
 options pagesize=15 linesize=122 nodate nonumber formdlim=' ';
 title1 '--- Stratified Odds Ratio Calculations ---';
   %do j=1 %to &tott;
     proc print data=taaa2 noobs;
     var OR&i.0C&&s&j LOR&i.0C&&s&j UOR&i.0C&&s&j;
     run;
     ods rtf startpage=off;
     title1 ' ';
   %end;
ods rtf startpage=on;
title1 '--- Crude Odds Ratios and Adjusted MH Odds Ratio Calculations ---';
options pagesize=20 linesize=122 date number formdlim='';
%end;
run;
options formdlim='';

%mend ornx2xk;

**** You can create a temporary SAS data set here.  Change macro arguments below as
     necessary.;
%ornx2xk(one,c,2,d,e,3,w)
ods rtf close;
```