



ABSTRACT

When using chi-square statistics to evaluate the relationship between two dichotomous variables, the test statistics approach the chi-square distribution as the sample size increases. In practice, the validity of the chi-square approximation depends on the expected cell counts exceeding five. For examining three-way tables using the Cochran-Mantel-Haenszel (CMH) statistic, the same rule does not apply. Mantel and Fleiss (1980) proposed an extension to the ‘minimum of five’ rule for $2 \times 2 \times H$ tables. We propose a user-friendly macro to supplement an analysis that uses the CMH one-degree-of-freedom test.

MANTEL-FLEISS CRITERION

Mantel and Fleiss showed that it is appropriate to compute the CMH p-value from a chi-square distribution with one degree of freedom if

$$\min \left\{ \left[\sum_{k=1}^q m_{h1k} - \sum_{k=1}^q (n_{h1k})L \right], \left[\sum_{k=1}^q (n_{h1k})U - \sum_{k=1}^q m_{h1k} \right] \right\}$$

exceeds 5, where $(n_{h1k})L = \max(0, n_{h1k} - n_{h+2})$ and $(n_{h1k})U = \min(n_{h+1}, n_{h1k})$ (Stokes, Davis and Koch, 2000).

When performing an analysis using the CMH test, PROC FREQ produces asymptotic p-values for the general association (CMH) test statistic.

INTRODUCTION

It is common in many areas of statistical practice to examine the relationship between a binary response variable and a binary predictor variable in the presence of a potentially confounding categorical factor. Data are typically presented in a series of partial 2×2 tables segmented by a stratification factor. Cochran (1954) first proposed a test of conditional independence in $2 \times 2 \times H$ tables, treating the rows in each 2×2 table as independent binomials. Mantel and Haenszel (1959) proposed a similar, but more generalized test based on the hypergeometric distribution (Agresti, 1996).

TABLE STRUCTURE

Visually, the table structure follows where $X=2$, $Y=2$ and h runs from 1 to q .

For h=1			For h=2			For qth table					
X	Y		X	Y		X	Y				
	n_{111}	n_{112}	n_{11+}		n_{211}	n_{212}	n_{21+}		n_{q11}	n_{q12}	n_{q1+}
	n_{121}	n_{122}	n_{12+}		n_{221}	n_{222}	n_{22+}		n_{q21}	n_{q22}	n_{q2+}
	n_{1+1}	n_{1+2}	n_1		n_{2+1}	n_{2+2}	n_2		n_{q+1}	n_{q+2}	n_q

CMH

The CMH method tests whether the conditional odds ratio between the response and predictor variables equals one in each partial 2×2 table. The approach conditions on the row and column totals in each partial 2×2 table and, without loss of generality, utilizes the cell in the first row and column (n_{h11}) (Agresti, 1996).

The CMH test statistic $\frac{(\sum_{h=1}^q n_{h11} - \sum_{h=1}^q m_{h11})^2}{\sum_{h=1}^q n_{h11}}$ is valid with sparse data but requires a large sample size.

CONCLUSIONS

Our macro determines the validity of the chi-square approximation for the CMH one-degree-of-freedom test. If the Mantel-Fleiss criterion is satisfied, the chi-square approximation is valid. However, if the criterion is not satisfied, other techniques are necessary.

Techniques for producing exact p-values are illustrated by Agresti (1996). Both PROC LOGISTIC and PROC MULTTEST provide tools for obtaining exact CMH p-values. The LOGISTIC approach uses the STRATA and EXACT statements.

EXAMPLE DATA AND RESULTS

The following are two hypothetical scenarios comparing treatments A and B between three sites.

In Example 1 the criterion is not satisfied. Therefore the chi-square approximation is not valid. In Example 2 the criterion is satisfied.

For illustration the macro is divided into four steps.

EXAMPLE 1					EXAMPLE 2				
Site	Treat.	Yes	No	Total	Site	Treat.	Yes	No	Total
1	A	12	5	17	1	A	12	5	17
	B	7	3	10		B	7	3	10
	TOT	19	8	27		TOT	19	8	27
2	A	1	2	3	2	A	1	2	3
	B	5	2	7		B	5	2	7
	TOT	6	4	10		TOT	6	4	10
3	A	0	9	9	3	A	5	9	14
	B	5	6	11		B	5	6	11
	TOT	5	15	20		TOT	10	15	25

STEPS DESCRIPTION

EXAMPLE 1 LOG

EXAMPLE2 LOG

Step	Description	EXAMPLE 1 LOG	EXAMPLE2 LOG
1	Sum expected values in each cell n_{h11}	SUM OF EXPECTED VALUES IN CELLS NH11: SUM(MH11) = 20.5130	SUM OF EXPECTED VALUES IN CELLS NH11: SUM(MH11) = 19.3630
2	Sum maximum difference between row one total and column two total for each strata	SUM OF LOWER BOUND (SUM[NH11]L) = 9	SUM OF LOWER BOUND (SUM[NH11]L) = 9
3	Sum minimum column one total and row one total across strata	SUM OF UPPER BOUND (SUM[NH11]U) = 25	SUM OF UPPER BOUND (SUM[NH11]U) = 30
4	Compare step one to steps two and three, then find overall minimum	Left end-point: sum[mh11] - sum[(nh11)L] = 11.513 Right endpoint: sum[(nh11)U] - sum[mh11] = 4.487 Mantel-Fleiss R = min[11.513, 4.487] = 4.487 Mantel-Fleiss criterion <u>NOT</u> satisfied (since R <= 5)	Left end-point: sum[mh11] - sum[(nh11)L] = 10.363 Right endpoint: sum[(nh11)U] - sum[mh11] = 10.637 Mantel-Fleiss R = min[10.363, 10.637] = 10.363 Mantel-Fleiss criterion satisfied (since R > 5)

References

- Agresti, A. 1996. *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.
- Cochran, W. G. 1954. "Some Methods for Strengthening the Common χ^2 tests". *Biometrics*. 10: 417-451.
- Dimitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. 2005. *Analysis of clinical trials using SAS: a practical guide*. Cary, NC: SAS Institute Inc.
- Kahn, H. and Sempos, C. 1989. *Statistical Methods in Epidemiology*. New York, NY: Oxford University Press.
- Mantel, N. and Joseph L Fleiss. 1980. "Minimum Expected Cell Size Requirements For The Mantel-Haenszel One-Degree-Of-Freedom Chi-Square Test And A Related Rapid Procedure". *American Journal of Epidemiology*. 112: 129-134.
- Mantel, N. and W Haenszel. 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, 2009. SAS usage note. "Usage Note 32711: How can I get an exact CMH test?" <http://support.sas.com/kb/32/711.html>.
- Stokes, Maura E., Davis, Charles S., Koch, Gary G. 2000. *Categorical Data Analysis Using The SAS System*. Cary, NC: SAS Institute Inc.

Authors

Brandon Welch Email: Brandon_Welch@rhoworld.com
 Rob Woolson Email: Rob_Woolson@rhoworld.com
 Jane Eslinger Email: Jane_Eslinger@rhoworld.com
 Rho Building • 6330 Quadrangle Drive • Chapel Hill, North Carolina 27517 • Phone: (919) 408-8000 • Fax: (919) 408-0999

MACRO CALL: %MantelFleiss(In,Strat,Var1,Var2,Count)

The macro consists of five parameters. Parameters &IN, &STRAT, &VAR1 and &VAR2 are required, whereas &COUNT is optional. The &IN parameter denotes the input data set and requires unique observations. &STRAT, &VAR1 and &VAR2 represent the stratification, row, and column variables respectively. &COUNT issues the WEIGHT statement in PROC FREQ. If the &COUNT parameter is omitted, all observations in the input data set are assumed to have a weight of one.

```
%macro MantelFleiss(In,Strat,Var1,Var2,Count);
```

```

%*-----;
%* &In - input data set ;
%* &Strat - stratification variable (e.g., site, race, etc.) ;
%* &Var1 - row variable ;
%* &Var2 - column variable ;
%* &Count - optional weight variable ;
%*-----;
%*reset output data sets;
PROC DATASETS nodetails nolist;
delete _counts _expect _mantel;
RUN;
QUIT;
```

```

%*determine number of strata levels;
PROC SQL noprint;
select count(distinct &strat) into: stratflag from &In;
QUIT;
%put **Number of strata levels &stratflag;
```

```

%*get expected counts and totals in each partial table;
PROC FREQ data = &In;
tables &Strat* &Var1* &Var2 / expected outexpect out = _expect;
%if %nrquote(&Count.) ne %then weight &Count.;;
ODS output CrossTabFreqs = _counts;
RUN;
```

STEP 1

```

%*get sum of expected cell counts in cell n11 in each partial table
(assign to macro var SUMEXP);
DATA _null_;
set _expect end = eof;
by &Strat;
retain sumexp 0;
if first.&Strat then do;
sumexp = sumexp + expected;
end;
if eof then call symput('sumexp',put(sumexp,8.4));
RUN;
%put Sum of expected values in cells nh11: sum(mh11) = %cmpres(&sumexp);
```

```

PROC SORT data = _counts;
by &Strat &Var1 &Var2;
RUN;
```

```

%*Subset to the needed totals output by PROC FREQ
number the &Strat, &Var1, and &Var2 for use below
0 - totals
1 - first &Strat/&Var1/&Var2
2 - second &Strat/&Var1/&Var2;
```

```

DATA _mantelf1;
%*only include marginal totals from ODS;
set _counts(where = (_type_ not in ('100' '111')));
by &Strat &Var1 &Var2;
retain &Strat._0 &Var1._ &Var2._;
```

```

%*reset for change in strata;
if first.&Strat then do;
&Var1._ = 0; &Var2._ = 0;
end;
```

```

%*get numeric version of strata variable;
if first.&Strat then &Strat._ = &Strat._ + 1;
```

```

%*define zero as total for any &Var1/&Var2;
if missing(&Var1) then &Var1._ = 0;
else &Var1._ + 1;
```

```

if missing(&Var2) then &Var2._ = 0;
else &Var2._ + 1;
```

```

keep &Strat._ &Var1._ &Var2._ &Strat. &Var1. &Var2. frequency;
RUN;
```

```

DATA _mantelf2;
set _mantelf1;
by &Strat._;
```

STEP 2

```

retain n11_ n1_1 n1_2;
if first.&Strat._ then do;
n11_ = 0;
n1_1 = 0;
n1_2 = 0;
end;
```

```

%*get each total;
if &Var1._ = 1 then n11_ = frequency; %*row one in the hth strata;
if &Var2._ = 1 then n1_1 = frequency; %*column one in the hth strata;
if &Var2._ = 2 then n1_2 = frequency; %*column two in the hth strata;
```

STEP 3

```

%*define (nh11)L and (nh11)U for final computation;
max_ = max(0,n11_ - n1_2);
min_ = min(n1_1, n11_);
```

```
if last.&Strat._;
```

```
RUN;
```

```

%*sum across min and max to get lower/upper bounds of criterion;
DATA _mantelf3(keep = mf_suml mf_sumu mflendpt mfrendpt mf_r);
set _mantelf2 end = eof;
```

```

retain mf_suml mf_sumu;
if _n_ = 1 then do;
mf_suml = 0; mf_sumu = 0;
end;
mf_suml = mf_suml + max_;
mf_sumu = mf_sumu + min_;
```

STEP 4

```

if eof then do;
mflendpt = &sumexp - mf_suml;
mfrendpt = mf_sumu - &sumexp;
mf_r = min(mflendpt, mfrendpt);
```

```

put "Sum of lower bound (sum[(nh11)L]) = " mf_suml;
put "Sum of upper bound (sum[(nh11)U]) = " mf_sumu;
put "Left end-point: sum[mh11] - sum[(nh11)L] = " mflendpt;
put "Right endpoint: sum[(nh11)U] - sum[mh11] = " mfrendpt;
put "Mantel-Fleiss R = min[" mflendpt ", " mfrendpt "] = " mf_r;
```

```

if mf_r > 5 then put "Mantel-Fleiss criterion satisfied (since R > 5)";
else put "Mantel-Fleiss criterion NOT satisfied (since R <= 5)";
output;
```

```
end;
```

```

label
mf_suml = "Summation of Maximum (sum[(nh11)L])"
mf_sumu = "Summation of Minimum (sum[(nh11)U])"
mflendpt= "Left End-Point: sum[mh11] - sum[(nh11)L]"
mfrendpt= "Right End-Point: sum[(nh11)U] - sum[mh11]"
mf_r = "R Value: min[sum[mh11]-sum[(nh11)L],sum[(nh11)U]-sum[mh11]]";
RUN;
%mend;
```