Categorical Data Analysis
Lab material #4

The respiratory data were collected in the study that included patients at two medical centers and produced the complete data shown in Table 2. These data comprise a set of two 2×2 tables.

Table 1: Respiratory Improvement

Center	Treatment	Yes	No	Total
1	Test	29	16	45
1	Placebo	14	31	45
Total		43	47	90
2	Test	37	8	45
2	Placebo	24	21	45
Total		61	29	90

Investigators were interested in whether there were overall differences in rates of improvement; however, they were concerned that the patient populations at the two centers were sufficiently different that center needed to be accounted for in the analysis. One strategy for examining the association between two variables while adjusting for the effects of others is *stratified analysis*.

In general, the strata may represent explanatory variables, or they may represent research sites or hospitals in a multicenter study. Each table corresponds to one stratum; the strata are determined by the levels of the explanatory variables (one for each unique combination of the levels of the explanatory variables). The idea is to evaluate the association between the row variable and the response variable, while *adjusting*, or *controlling*, for the effects of the stratification variables. In some cases, the stratification results from the study design, such as in the case of a multicenter clinical trial; in other cases, it may arise from a prespecified poststudy stratification performed to control for the effects of certain explanatory variables that are thought to be related to the response variable.

The analysis of sets of tables addresses the same questions as the analysis of a single table: is there an association between the row and column variables in the tables and what is the strength of that association? These questions are investigated with similar strategies involving chi-square statistics and measures of association such as the odds ratios; the key difference is that you are investigating overall association instead of the association in just one table.

For the data in Table 2, there is interest in the association between treatment and respiratory outcome, after adjusting for the effects of the centers. The following DATA step puts all the respiratory data into the SAS data set RESPIRE. Producing a Cochran-Mantel-Haenszel analysis from PROC FREQ requires the specification of multi-way tables. The triple crossing CENTER*TREATMENT*RESPONSE specifies that the data consists of sets of two-way tables. The two rightmost variables TREATMENT and RESPONSE determine the rows and columns of the tables, respectively, and the variables to the left (CENTER) determine the stratification scheme. There will be one table for each value of CENTER.

If there are more variables to the left of the variables determining the rows and columns of the tables, there will be strata for each unique combination of values for those variables.

The CHISQ option specifies that chi-square statistics be printed for each table. The CMH option requests the Cochran-Mantel-Haenszel statistics for the stratified analysis; these are also called summary statistics. The ORDER=DATA option specified that PROC FREQ order the rows and columns according to the order in which the variable values are encountered in the input data.

```
* example4_1.sas;
data respire;
  input center treatmnt $ response $ count @@;
  cards;
  1 test y 29 1 test n 16
  1 placebo y 14 1 placebo n 31
  2 test y 37 2 test n 8
  2 placebo y 24 2 placebo n 21
  ;
proc freq order=data;
  weight count;
  tables center*treatmnt*response / nocol nopct chisq cmh;
run;
```

Output 1 displays the frequency tables and chi-square statistics for each center. For Center 1, the favorable rate for test treatment is 64 percent, versus 31 percent for placebo. For Center 2, the favorable rate for test treatment is 82 percent, versus 53 percent for the placebo. X^2 for Center 1 is 10.0198; X^2 for Center 2 is 8.5981. With 1 df, both of these statistics are strongly significant.

Following the information for the individual tables, PROC FREQ prints out a section titled "SUMMARY STATISTICS FOR TREATMNT BY RESPONSE, CONTROLLING FOR CENTER." This includes tables containing Cochran-Mantel-Haenszel (CMH) statistics, estimates of the common relative risk, and the Breslow-Day test for homogeneity of the odds ratio.

Output 1 Results

```
                    The FREQ Procedure

      Table 1 of treatmnt by response
          Controlling for center=1

treatmnt     response

Frequency|
Row Pct  |y       |n       |  Total
---------+--------+--------+
test     |     29 |     16 |     45
         |  64.44 |  35.56 |
---------+--------+--------+
placebo  |     14 |     31 |     45
         |  31.11 |  68.89 |
---------+--------+--------+
Total          43       47       90


    Statistics for Table 1 of treatmnt by response
              Controlling for center=1

Statistic                    DF       Value      Prob
```

Statistic	DF	Value	Prob
Chi-Square	1	10.0198	0.0015
Likelihood Ratio Chi-Square	1	10.2162	0.0014
Continuity Adj. Chi-Square	1	8.7284	0.0031
Mantel-Haenszel Chi-Square	1	9.9085	0.0016
Phi Coefficient		0.3337	
Contingency Coefficient		0.3165	
Cramer's V		0.3337	

Fisher's Exact Test

Cell (1,1) Frequency (F)	29
Left-sided Pr <= F	0.9997
Right-sided Pr >= F	0.0015
Table Probability (P)	0.0011
Two-sided Pr <= P	0.0029

Sample Size = 90

Table 2 of treatmnt by response
Controlling for center=2

treatmnt response

Frequency| Row Pct	y	n	Total
test	37	8	45
	82.22	17.78	
placebo	24	21	45
	53.33	46.67	
Total	61	29	90

Statistics for Table 2 of treatmnt by response
Controlling for center=2

Statistic	DF	Value	Prob
Chi-Square	1	8.5981	0.0034
Likelihood Ratio Chi-Square	1	8.8322	0.0030
Continuity Adj. Chi-Square	1	7.3262	0.0068
Mantel-Haenszel Chi-Square	1	8.5025	0.0035
Phi Coefficient		0.3091	
Contingency Coefficient		0.2953	
Cramer's V		0.3091	

Fisher's Exact Test

Cell (1,1) Frequency (F)	37
Left-sided Pr <= F	0.9993
Right-sided Pr >= F	0.0031
Table Probability (P)	0.0025
Two-sided Pr <= P	0.0063

Sample Size = 90

Summary Statistics for treatmnt by response
Controlling for center

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	18.4106	<.0001
2	Row Mean Scores Differ	1	18.4106	<.0001
3	General Association	1	18.4106	<.0001

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	4.0288	2.1057	7.7084
(Odds Ratio)	Logit	4.0286	2.1057	7.7072
Cohort	Mantel-Haenszel	1.7368	1.3301	2.2680
(Col1 Risk)	Logit	1.6760	1.2943	2.1703
Cohort	Mantel-Haenszel	0.4615	0.3162	0.6737
(Col2 Risk)	Logit	0.4738	0.3264	0.6877

Breslow-Day Test for
Homogeneity of the Odds Ratios

Chi-Square	0.0002
DF	1
Pr > ChiSq	0.9900

Total Sample Size = 180

To find the value of Q_{CMH}, read the value for any of the statistics in the table labeled "Cochran-Mantel-Haenszel Statistics": "Nonzero Correlation," "Row Mean Scores Differ," or "General Association." These statistics pertain to the situation where you have sets of tables with two or more rows or columns. However, they all reduce to the CMH statistics when you have 2×2 tables and use the CMH option in its default mode (that is, no SCORE= option specified).

Q_{CMH} for these data is $Q_{CMH} = 18.4106$, with 1 df. This is clearly significant. The associations in the individual tables reinforce each other so that the overall association is stronger than that seen in the individual tables. There is a strong association between treatment and response, adjusting for center. The test treatment had a significantly higher favorable response rate than placebo.

The following data are based on a study on coronary artery disease. The sample is one of convenience since the patients studied were people who came to a clinic and requested an evaluation.

Table 2: Coronary Artery Disease Data

Sex	ECG	Disease	No Disease	Total
Female	< 0.1 ST segment depression	4	11	15
Female	≥ 0.1 ST segment depression	8	10	18
Male	< 0.1 ST segment depression	9	9	18
Male	≥ 0.1 ST segment depression	21	6	27

Investigators were interested in whether electrocardiogram (ECG) measurement was associated with disease status. Gender was thought to be associated with disease status, so investigators post-stratified the data into male and female groups. In addition, there was interest in examining the odds ratios.

The following statements produce the SAS data set CA and request a stratified analysis. The first TABLES statement requests chi-square tests for the association of gender and disease status. The second TABLES statement requests the stratified analysis, including the generation of odds ratios with the MEASURES option.

```
 * example4_2.sas;
 data ca;
   input gender $ ECG $ disease $ count;
cards;
female <0.1 yes 4
female <0.1 no 11
female >=0.1 yes 8
female >=0.1 no 10
male <0.1 yes 9
male <0.1 no 9
male >=0.1 yes 21
male >=0.1 no 6
;
proc freq;
weight count;
tables gender*disease / nocol nopct chisq;
tables gender*ECG*disease / nocol nopct cmh chisq measures;
run;
```

Output 1 contains the table of GENDER by DISEASE. X^2 takes value 7.0346 and G^2 takes the value 7.1209. Obviously there is a strong association between gender and disease status. Males are much more likely to have symptoms of coronary artery disease than females. The idea to control for gender in a stratified analysis is a good one.

Output 1 GENDER$\times$DISEASE

```
     Table of gender by disease

gender     disease

Frequency|
Row Pct  |no       |yes      |  Total
---------+---------+---------+
female   |     21  |     12  |     33
```

```
         |  63.64 |  36.36 |
---------+--------+--------+
male     |     15 |     30 |     45
         |  33.33 |  66.67 |
---------+--------+--------+
Total          36       42       78
```

Statistics for Table of gender by disease

Statistic	DF	Value	Prob
Chi-Square	1	7.0346	0.0080
Likelihood Ratio Chi-Square	1	7.1209	0.0076
Continuity Adj. Chi-Square	1	5.8681	0.0154
Mantel-Haenszel Chi-Square	1	6.9444	0.0084
Phi Coefficient		0.3003	
Contingency Coefficient		0.2876	
Cramer's V		0.3003	

Fisher's Exact Test

Cell (1,1) Frequency (F)	21
Left-sided Pr <= F	0.9981
Right-sided Pr >= F	0.0075
Table Probability (P)	0.0056
Two-sided Pr <= P	0.0114

Sample Size = 78

Output 2 and Output 3 display the individual tables results for ECG×disease status; included are the table of chi-square statistics generated by the CHISQ option and only the "Estimates of the Relative Risk" table part of the output generated by the MEASURES option.

Output 2 Results for Females

Statistics for Table 1 of ECG by disease
Controlling for gender=female

Statistic	DF	Value	Prob
Chi-Square	1	1.1175	0.2905
Likelihood Ratio Chi-Square	1	1.1337	0.2870
Continuity Adj. Chi-Square	1	0.4813	0.4879
Mantel-Haenszel Chi-Square	1	1.0836	0.2979
Phi Coefficient		0.1840	
Contingency Coefficient		0.1810	
Cramer's V		0.1840	

Fisher's Exact Test

Cell (1,1) Frequency (F)	11
Left-sided Pr <= F	0.9233
Right-sided Pr >= F	0.2450
Table Probability (P)	0.1683

Two-sided Pr <= P	0.4688

Statistic	Value	ASE
Gamma	0.3750	0.3232
Kendall's Tau-b	0.1840	0.1686
Stuart's Tau-c	0.1763	0.1623
Somers' D C|R	0.1778	0.1636
Somers' D R|C	0.1905	0.1743
Pearson Correlation	0.1840	0.1686
Spearman Correlation	0.1840	0.1686
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0667	0.2951
Lambda Symmetric	0.0370	0.1666
Uncertainty Coefficient C|R	0.0262	0.0485
Uncertainty Coefficient R|C	0.0249	0.0462
Uncertainty Coefficient Symmetric	0.0256	0.0473

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	2.2000	0.5036	9.6107
Cohort (Col1 Risk)	1.3200	0.7897	2.2063
Cohort (Col2 Risk)	0.6000	0.2240	1.6073

Sample Size = 33

X^2 is 1.1175 for female, with a p-value of 0.2905. The odds ratio for the females is OR=2.2000, with a 95 percent confidence interval that includes 1. Those female with higher ST segment depression levels had 2.2 times the odds of CA disease than those with lower levels.

Output 3 Results for Males

Statistics for Table 2 of ECG by disease
Controlling for gender=male

Statistic	DF	Value	Prob
Chi-Square	1	3.7500	0.0528
Likelihood Ratio Chi-Square	1	3.7288	0.0535
Continuity Adj. Chi-Square	1	2.6042	0.1066
Mantel-Haenszel Chi-Square	1	3.6667	0.0555
Phi Coefficient		0.2887	
Contingency Coefficient		0.2774	
Cramer's V		0.2887	

Fisher's Exact Test

Cell (1,1) Frequency (F)	9
Left-sided Pr <= F	0.9880
Right-sided Pr >= F	0.0538

Table Probability (P)	0.0417
Two-sided Pr <= P	0.1049

Statistics for Table 2 of ECG by disease
Controlling for gender=male

Statistic	Value	ASE
Gamma	0.5556	0.2284
Kendall's Tau-b	0.2887	0.1462
Stuart's Tau-c	0.2667	0.1377
Somers' D C|R	0.2778	0.1424
Somers' D R|C	0.3000	0.1517
Pearson Correlation	0.2887	0.1462
Spearman Correlation	0.2887	0.1462
Lambda Asymmetric C|R	0.0000	0.2828
Lambda Asymmetric R|C	0.1667	0.1964
Lambda Symmetric	0.0909	0.2091
Uncertainty Coefficient C|R	0.0651	0.0662
Uncertainty Coefficient R|C	0.0616	0.0629
Uncertainty Coefficient Symmetric	0.0633	0.0645

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	3.5000	0.9587	12.7775
Cohort (Col1 Risk)	2.2500	0.9680	5.2298
Cohort (Col2 Risk)	0.6429	0.3883	1.0642

Sample Size = 45

X^2 is 3.7500 for males, with a p-value of 0.0528. The odds ratio for the males is OR=3.5000, with a 95 percent confidence interval that barely contains the value 1. Those men with higher ST segment depression levels had 3.5 times the odds of CA disease than those with lower levels.

Output 4 contains the Q_{CMH} statistic, which takes the value 4.5026 with a p-value of 0.0338. By combining the genders, the power has been increased so that the association detected by Q_{CMH} is significant at the $\alpha = 0.05$ level of significance.

Output 4 Stratified Analysis

Summary Statistics for ECG by disease
Controlling for gender

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	4.5026	0.0338
2	Row Mean Scores Differ	1	4.5026	0.0338
3	General Association	1	4.5026	0.0338

Output 5 contains the estimates of the common odds ratios. $\hat{\theta}_{MH} = 2.8467$ and $\hat{\theta}_L = 2.8593$. The confidence intervals for both of estimator do not contain the value 1.

Output 5 Odds Ratios

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	2.8467	1.0765	7.5279
	Logit	2.8593	1.0807	7.5650
Cohort (Col1 Risk)	Mantel-Haenszel	1.6414	1.0410	2.5879
	Logit	1.5249	0.9833	2.3647
Cohort (Col2 Risk)	Mantel-Haenszel	0.6299	0.3980	0.9969
	Logit	0.6337	0.4046	0.9926

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square	0.2155
DF	1
Pr > ChiSq	0.6425

Total Sample Size = 78

Finally, the Breslow-Day test is printed at the bottom and does not contradict the assumption of homogeneous odds ratios for these data. $Q_{BD} = 2.155$ with $p = 0.6425$.