ST3241 Categorical Data Analysis I
Three-way Contingency Tables

An Introduction: Conditional Associations

Example: Death Penalty Data

		Death Penalty	
Victims' Race	Defendant's Race	Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

Objectives

- To find association between X and Y by controlling other covariates that can influence the association.
- We study the effect of X on Y by fixing such covariates constant.
- In other words, study the association between X and Y given the levels of Z.

Partial Tables

- Two-way tables between X and Y at separate levels of Z.
- The two-way contingency table obtained by combining the partial tables is called the $X - Y$ marginal table.
- Each cell count in the marginal table is a sum of counts from the same cell location in the partial tables.
- The marginal table, rather than controlling Z, ignores it and does not contain any information about Z.

Example: Death Penalty Data

		Death Penalty	
Victims' Race	Defendant's Race	Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

Example: Death Penalty Data

Defendant's Race	Death Penalty Yes	Death Penalty No
White	53	430
Black	15	176

Notes

- The associations in partial tables are called conditional associations.
- Conditional associations in partial tables can be quite different from associations in marginal tables.

Example: Death Penalty Data

		Death Penalty		
Victims' Race	Defendant's Race	Yes	No	Percentage Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Simpson's Paradox

- This death penalty data is an example of Simpson's paradox.
- The result that a marginal association can have different direction from the conditional associations is called *Simpson's paradox*.
- This result applies to quantitative as well as categorical variables.

Odds Ratios

- Consider $2 \times 2 \times K$ tables, where K denotes the number of levels of a control variable Z.
- Let $\{n_{ijk}\}$ denote the observed frequencies and let $\{\mu_{ijk}\}$ denote their expected frequencies.
- Within a fixed level k of Z,

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

describes conditional $X - Y$ association.

- We refer to them as the $X - Y$ *conditional odds ratios*.

Marginal Odds Ratio

- Expected frequencies in the $X - Y$ marginal table is:

$$\mu_{ij+} = \sum_{k=1}^{K} \mu_{ijk}$$

- The $X - Y$ marginal odds ratio is defined as

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

- Similar formulas with μ_{ijk} substituted by n_{ijk}s provide sample estimates of $\theta_{XY(k)}$ and θ_{XY}.

Sample Odds Ratios

- Sample Conditional Odds Ratio:

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

- Sample $X - Y$ Marginal Odds Ratio:

$$\hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$$

Example: Death Penalty Data

- Sample Conditional Odds Ratio:
  - For Victim's race: White, $\hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43$
  - For Victim's race: Black, $\hat{\theta}_{XY(2)} = 0.0$
- Sample marginal odds ratio:

$$\hat{\theta}_{XY} = 1.45$$

Marginal vs. Conditional Independence

- If X and Y are independent in each partial table, then X and Y are said to be conditionally independent given Z.
- All conditional odds ratios between X and Y are then equal to 1.
- Conditional independence of X and Y, given Z, does not imply marginal independence of X and Y.
- That is, odds ratios between X and Y equal to 1 at each level of Z, the marginal odds ratio may differ from 1.

A Hypothetical Example

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

- Here $\theta_{XY(1)} = 1.0$, $\theta_{XY(2)} = 1.0$ but $\theta_{XY} = 2.0$.
- It is misleading to study only the marginal tables, concluding that successes are more likely with treatment A than with treatment B.

Homogeneous Association

- There is *homogeneous $X - Y$ association* in a $2 \times 2 \times K$ if the conditional odds ratios between X and Y are identical at all levels of Z.
- That is,$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$.
- In such a situation, a single number describes the conditional association.

Notes

- *Conditional independence* is a special case of homogeneous association, where each conditional odds ratio equals 1.0.
- *Homogeneous association* is a symmetric property, applying to any pair of variables viewed across the levels of the third.
- If it occurs, there is said to be *no interaction* between two variables in their effects on the third variable.

Table 1: Example: Chinese Smoking Study

City	Smoking	Lung Cancer Yes	Lung Cancer No	Odds Ratio	μ_{11k}	$\mathrm{Var}(n_{11k})$
Beijing	Smokers	126	100	2.20	113.0	16.9
	Non-Smokers	35	61			
Shanghai	Smokers	908	688	2.14	773.2	179.3
	Non-Smokers	497	807			
Shenyang	Smokers	913	747	2.18	799.3	149.3
	Non-Smokers	336	598			
Nanjing	Smokers	235	172	2.85	203.5	31.1
	Non-Smokers	58	121			
Harbin	Smokers	402	308	2.32	355.0	57.1
	Non-Smokers	121	215			
Zhengzhou	Smokers	182	156	1.59	169.0	28.3
	Non-Smokers	72	98			
Taiyuan	Smokers	60	99	2.37	53.0	9.0
	Non-Smokers	11	43			
Nanchang	Smokers	104	89	2.00	96.5	11.0
	Non-Smokers	21	36			

Cochran-Mantel-Haenszel Test

- To Test: X and Y are conditionally independent given Z.
- So, $H_0 : \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1.0$.
- In the k-th partial table, the row totals are n_{1+k}, n_{2+k} and column totals are n_{+1k}, n_{+2k}.
- Given both these totals, n_{11k} has a hypergeometric distribution and that determines all other cell counts in the k-th partial table.

Cochran-Mantel-Haenszel Test C Continued

- Under the null hypothesis of independence,

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n},$$

$$Var(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{{n_{++k}}^2(n_{++k}-1)}, k = 1, \cdots, K$$

- The test statistic is given by

$$CMH = \frac{[\sum_{k=1}^{K}(n_{11k} - \mu_{11k})]^2}{\sum_{k=1}^{K} Var(n_{11k})}$$

- This is called the $Cochran - Mantel - Haenszel$ (CMH) statistic.
- It has a large sample chi-squared distribution with *df* $= 1$.

Notes

- CMH takes larger values when $(n_{11k} - \mu_{11k})$ is consistently positive or consistently negative.
- This test is inappropriate when the association varies widely among the partial tables.

Example

- In the Chinese Smoking Study,

$$\sum_k n_{11k} = 2930, \sum_k \mu_{11k} = 2562.5, \sum_k Var(n_{11k}) = 482.1$$

- So, $CMH = (2930 - 2562.5)^2/482.1 = 280.1$ with d.f. $= 1$.
- There is extremely strong evidence against conditional independence.

Estimation of Common Odds Ratio

- Assume, homogeneous association, that is, $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$.
- The Mantel-Haenszel estimator of that common value equals

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} (n_{11k} n_{22k} / n_{++k})}{\sum_{k=1}^{K} (n_{12k} n_{21k} / n_{++k})}$$

Standard Error

- The squared standard error for log of MH estimator is:

$$\hat{\sigma}^2(log(\hat{\theta}_{MH})) = \frac{\sum_{k=1}^{K}(n_{11k}+n_{22k})(n_{11k}n_{22k})/n_{++k}^2}{2(\sum_{k=1}^{K}n_{11k}n_{22k}/n_{++k})^2}$$

$$+\frac{\sum_{k=1}^{K}[(n_{11k}+n_{22k})(n_{12k}n_{21k})+(n_{12k}+n_{21k})(n_{11k}n_{22k})]/n_{++k}^2}{2(\sum_{k=1}^{K}n_{11k}n_{22k}/n_{++k})(\sum_{k=1}^{K}n_{12k}n_{21k}/n_{++k})}$$

$$+\frac{\sum_{k=1}^{K}(n_{12k}+n_{21k})(n_{12k}n_{21k})/n_{++k}^2}{2(\sum_{k=1}^{K}n_{12k}n_{21k}/n_{++k})^2}$$

Example: Chinese Smoking Studies

- The MH estimator: $\hat{\theta}_{MH} = 2.17$
- The estimated standard error:

$$\hat{\sigma}(log\hat{\theta}_{MH}) = 0.046$$

- A 95% C.I. for common log odds ratio

$$0.777 \pm 1.96 \times 0.046 = (0.686, 0.868)$$

- A 95% C.I. for common odds ratio

$$(e^{0.686}, e^{0.868}) = (1.98, 2.38)$$

Notes

- If the true odds ratios are not identical but do not vary drastically, $\hat{\theta}_{MH}$ still provides a useful summary of the K conditional associations.

Testing Homogeneity of Odds Ratios

- To test for homogeneous association in $2 \times 2 \times K$ tables, $H_0 : \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$.
- The Breslow-Day test statistic has the form:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$$

  where $\hat{\mu}_{ijk}$ is the expected cell frequency.
- The formula for computing $\hat{\mu}_{ijk}$ is complicated.

Notes:

- Under null hypothesis, the Breslow-Day statistic has a large sample chi-squared distribution with degrees of freedom $K - 1$.
- The sample size should be relatively large in each partial table. .. It can be computed using standard statistical software.

SAS Codes: Input Data

```
data cmh;
input center smoke cancer count @@;
datalines;
 1   1   1   126     1   1   2   100     1   2   1    35     1   2   2    61
 2   1   1   908     2   1   2   688     2   2   1   497     2   2   2   807
 3   1   1   913     3   1   2   747     3   2   1   336     3   2   2   598
 4   1   1   235     4   1   2   172     4   2   1    58     4   2   2   121
 5   1   1   402     5   1   2   308     5   2   1   121     5   2   2   215
 6   1   1   182     6   1   2   156     6   2   1    72     6   2   2    98
 7   1   1    60     7   1   2    99     7   2   1    11     7   2   2    43
 8   1   1   104     8   1   2    89     8   2   1    21     8   2   2    36

;
run;
```

SAS Codes: Partial Tables

```
proc freq data=cmh;
  weight count;
  table center*smoke*cancer/ relrisk cmh norow nocol
  nopercent ;
  output out=temp or;
run;
proc print data=temp (rename=(_RROR_=oddsratio))
  noobs;
var center oddsratio;
  title 'Odds Ratio by Center';
run;
title ;
```

Output

```
Table 1 of smoke by cancer
Controlling for center=1
smoke     cancer
     smoke   cancer
 Frequency |      1 |    2 | Total
-----------+--------+------+
         1 |    126 |  100 |   226
-----------+--------+------+
         2 |     35 |   61 |    96
-----------+--------+------+
     Total |    161 |  161 |   322
```

Output

```
Table 8 of smoke by cancer
Controlling for center=8
smoke     cancer
     smoke   cancer
 Frequency |      1 |    2 | Total
-----------+--------+------+
         1 |    104 |   89 |   193
-----------+--------+------+
         2 |     21 |   36 |    57
-----------+--------+------+
     Total |    125 |  125 |   250
```

Output

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

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	280.1375	$<$.0001
2	Row Mean Scores Differ	1	280.1375	$<$.0001
3	General Association	1	280.1375	$<$.0001

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence	Limits
Case-Control	Mantel-Haenszel	2.1745	1.9840	2.3832
(Odds Ratio)	Logit	2.1734	1.9829	2.3823
Cohort	Mantel-Haenszel	1.5192	1.4417	1.6008
(Col1 Risk)	Logit	1.5132	1.4362	1.5942
Cohort	Mantel-Haenszel	0.6999	0.6721	0.7290
(Col2 Risk)	Logit	0.7011	0.6734	0.7300

Output

```
Breslow-Day Test for
Homogeneity of the Odds Ratios
 ------------------------------
 Chi-Square                    5.1997
 DF                                 7
 Pr > ChiSq                    0.6356
Total Sample Size = 8419
```

Output: PROC PRINT

```
Odds Ratio by Center
 center   oddsratio
    1       2.19600
    2       2.14296
    3       2.17526
    4       2.85034
    5       2.31915
    6       1.58796
    7       2.36915
    8       2.00321
```

R Codes: Input Data

```
lung<-read.table(
  "F:/ST3241/lectdata/chinese.txt",
  header=T, sep="\t")
lungtab<-
  xtabs(count Smoking+LungCancer+Center,
  data=lung)
mantelhaen.test(lungtab,correct=F)
```

Output

```
Mantel-Haenszel chi-squared test without continuity
  correction
data:  lungtab
Mantel-Haenszel X-squared = 280.1375, df = 1, p-value
  < 2.2e-16
alternative hypothesis:  true common odds ratio is not
  equal to 1
95 percent confidence interval:
1.984002 2.383249
sample estimates:
common odds ratio
2.174482
```