2.7 Association In Three-Way Tables

2.7.1 Partial Tables and Marginal Tables
2.7.2 Conditional Versus Marginal Associations
2.7.3 Simpson's Paradox
2.7.4 Conditional and Marginal Odds Ratios
2.7.5 Conditional Independence v.s. Marginal Independence
2.7.6 Homogeneous Association
4.3.4 The Cochran-Mantel-Haenszel Test for Conditional Indep.
Mantel-Haenszel Estimate for the Common Odds Ratio

Example — Kidney Stone Treatments

Result for a study comparing 2 treatments for reducing or eliminating kidney stones is shown below.

	Outcome (*Y*)		
Treatment (*X*)	Success	Failure	% Success
A	273	77	78%
B	289	61	83%

Here, "Success" means no stone of size > 2 mm.

- An observational study needs to control for *confounders*.
- 3-way contingency tables can be used to control for a *single* confounder. In later chapters, we can control for more confounders by models.

Example — Kidney Stone Treatments (Cont'd)

a $2 \times 2 \times 2$ table — 2 rows, 2 columns, 2 layers:

$$Y = \text{Outcome (response variable)}$$
$$X = \text{Treatment (explanatory variable)}$$
$$Z = \text{Initial size of kidney stone (control variable)}$$

Initial Size of		*Outcome* (*Y*)		
Kidney Stone (*Z*)	*Treatment* (*X*)	Success	Failure	% Success
Small	A	81	6	93.1%
	B	234	36	86.7%
Large	A	192	71	73.0%
	B	55	25	68.8%
Total	A	273	77	78.0%
	B	289	61	82.6%

Partial Tables

The tables

Z = Small

	Outcome	
Trt	Success	Failure
A	81	6
B	234	36

Z = Large

	Outcome	
Trt	Success	Failure
A	192	71
B	55	25

are called *partial tables*. They control for Z (hold it constant).

The (estimated) *conditional odds ratios* are:

$$Z = \text{Small}: \quad \widehat{\theta}_{XY(1)} = \frac{81 \times 36}{6 \times 234} \approx 2.08$$
$$Z = \text{Large}: \quad \widehat{\theta}_{XY(2)} = \frac{192 \times 25}{71 \times 55} \approx 1.23$$

Controlling for the initial size of kidney stone, odds of success were higher for treatment A than for treatment B.

Marginal Table

Adding the partial tables gives the XY *marginal table*, which ignores the effect of Z.

	Outcome (Y)	
Trt (X)	Success	Failure
A	273	77
B	289	61

(estimated) *marginal odds ratio*

$$= \widehat{\theta}_{XY} = \frac{273 \times 61}{77 \times 289} \approx 0.75$$

Ignoring the initial size of kidney stones, odds of success were lower for treatment A than for treatment B.

Definition: Simpson's Paradox

All partial tables show reverse association from that in marginal table.

- Cause?
- Moral: can be dangerous to "collapse" contingency tables.

Conditional Independence

Definition

X and Y are conditionally independent given Z if they are independent in each partial table.

In a $2 \times 2 \times K$ table this means

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.0$$

Conditional Independence $\not\Rightarrow$ Marginal Independence

Conditional independence of X and Y, given Z, does NOT imply marginal independence of X and Y.

Example.

Clinic	*Treatment*	*Outcome* (Y)			
(Z)	(X)	Success	Failure	% Success	$\widehat{\theta}$
1	A	18	12	60%	1.0
	B	12	8	60%	
2	A	2	8	25%	1.0
	B	8	32	25%	
Total	A	20	20	50%	2.0
	B	20	40	33%	

Homogeneous Association

- Associations of X and Y are the same at each level of Z.
- In a $2 \times 2 \times K$ table this means all partial tables share a common odds ratio:

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)}$$

- Conditional independence is a special case of homogeneous association.

Understanding Homogeneous Association

Example. Suppose we want to compare the effectiveness ($Y = S$ or F) of two treatments ($X = A$ or B) using patients from several hospitals ($Z = 1, 2, \ldots, k$). Let π_{Ai} and π_{Bi} be the prob. of success for the two treatments in Hospital i.

- X and Y are conditionally indep. if $\pi_{Ai} = \pi_{Bi}$ for all i. In this case, the two treatments are equally effective, but different hospitals can have different probability of success (due to difference in the demographics of patients or in the quality of the hospitals, etc).
- XY have homogeneous association if

$$\frac{\pi_{Ai}}{1-\pi_{Ai}} = \theta \frac{\pi_{Bi}}{1-\pi_{Bi}} \quad \text{for some constant } \theta \text{ for all } i$$

In this case, different hospitals can have different probabilities of success, and changing the treatment from B to A just change the odds of success by a constant θ.

Homogeneous Association

In a 3-way table, if XY has homogeneous association given Z, then so do YZ given X and XZ given Y.

	$Z = 1$		$Z = 2$	
	$X = 1$	$X = 2$	$X = 1$	$X = 2$
$Y = 1$	a	b	A	B
$Y = 2$	c	d	C	D

Homogeneous XY association given Z means

$$\theta_{XY(1)} = \frac{ad}{cb} = \frac{AD}{CB} = \theta_{XY(2)}$$
$$\iff \theta_{YZ(1)} = \frac{aC}{cA} = \frac{bD}{dB} = \theta_{YZ(2)}$$

which means homogeneous YZ association given X.

	$X = 1$		$X = 2$	
	$Z = 1$	$Z = 2$	$Z = 1$	$Z = 2$
$Y = 1$	a	A	b	B
$Y = 2$	c	C	d	D

- The "Kidney Stone Treatments" examples have illustrated
  - it is not appropriate to use marginal odds ratio to examine the association of two variables X and Y when there is a confounding variable Z,
  - the need to use conditional odds ratios
- Therefore, the population parameters of interest are those conditional odds ratios rather than the marginal odds ratio.
- If XY associations (odds ratios) change with Z, in this case, we should discuss the XY relations at each level of Z by analyzing the partial tables at each level of Z.
- If XY associations (odds ratios) do not change too much across different levels of Z, we may
  - estimate the common odds ratio using the Mantel-Haenszel estimate of the common odds ratio
  - test the conditional independence using the Cochran-Mantel-Haenszel test

Example: Effect of Smoking

The following table shows the result of a survey conducted in 1972-74 cross-classifying 1314 women in the United Kingdom by their smoking status and age in 1972-1974 and their survival status 20 years later (determined by a follow-up survey 20 years later)[1].

Age	18-34		35-54		55-64		65+	
	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Smoker	5	174	41	198	51	64	42	7
Nonsmoker	6	213	19	180	40	81	165	28
Odds ratio	1.02		1.96		1.61		1.02	

In this example, the effect of smoking appears to depend on age.

- Association between smoking and survival was much stronger for middle age women (35-54 and 44-64) than young (18-34) or old (65+) women.

[1] Source: Example 1 in p.90 and Example 16 on p.134 of the book *Statistics – The Art and Science of Learning from Data* (3rd) by Agresti and Franklin

Cochran-Mantel-Haenszel (CMH) Test of Conditional Independence

Suppose the XY partial table for $Z = k$ is

	$Z = k$		
	$Y = 1$	$Y = 2$	row total
$X = 1$	$\boxed{n_{11k}}$	n_{12k}	R_{1k}
$X = 2$	n_{21k}	n_{22k}	R_{2k}
column total	C_{1k}	C_{2k}	T_k

Recall that in Fisher's exact test, under the H_0 of (conditional) independence, n_{11k} has a hypergeometric distribution. It can be show that

$$\mathbb{E}[n_{11k}] = \frac{R_{1k}C_{1k}}{T_k}$$

$$\mathrm{Var}(n_{11k}) = \frac{R_{1k}R_{2k}C_{1k}C_{2k}}{T_k^2(T_k - 1)}$$

Cochran-Mantel-Haenszel (CMH) Test of Conditional Independence

For testing

H_0: XY are conditionally independent across all levels of Z,

H_a: XY are not independent in at least one level of Z,

the Cochran-Mantel-Haenszel (CMH) statistic is

$$\text{CMH} = \frac{\text{sum of } (n_{11k} - \mathbb{E}[n_{11k}]) \text{ over all partial tables}}{\sqrt{\text{sum of } \mathrm{Var}(n_{11k}) \text{ over all partial tables}}}.$$

Under H_0, the CMH statistic is approximately $N(0, 1)$.
(Or equivalently $(\text{CMH})^2$ is approximately χ^2_1.)

Example: Lung Cancer and Passive Smoking

To study the effect of passive smoking and lung cancer, the 3 tables below summarize results of case-control studies from 3 countries among nonsmoking women married to smokers.

Spouse	Japan		UK		US	
Smoked	Case	Control	Case	Control	Case	Control
Yes	73	188	19	38	137	363
No	21	82	5	16	71	249
Odds ratio	1.52		1.60		1.32	

Though the estimated odds ratios of the 3 partial tables are all > 1, none of them are significant by Pearson's X^2 test or Fisher's exact test. (The two-sided P-values of the two tests for each of the 3 tables are as follows).

P-value	Japan	UK	US
Pearson X^2	0.14	0.42	0.09
Fisher Exact	0.15	0.58	0.10

Example: Lung Cancer and Passive Smoking

- The associations in the 3 partial tables are not significant might be due to the **small sample sizes** of the 3 studies
- As the 3 partial tables indicate association in the same direction ($\theta > 1$), can we combine evidence from the 3 tables and make a test on all 3 tables simultaneously?
- Simply combining 3 tables and doing Pearson's X^2 or Fisher's exact test on the combined table (marginal table) will ignore the country effect, which might be associated with both passive smoking and lung cancer
  - There might be a higher percentage of people suffering from passive smoking in one country than in another.
  - The prevalence of lung cancer may also change from country to country.

  Ignoring the country effect might result in Simpson's paradox.
- CMH test allows us to combine evidence from the 3 tables while taking the country effect into account.

Example: Lung Cancer and Passive Smoking (CMH-test)

Spouse	Japan			UK			US		
Smoked	Case	Control	total	Case	Control	total	Case	Control	total
Yes	73	188	261	19	38	57	137	363	500
No	21	82	103	5	16	21	71	249	320
total	94	270	364	24	54	78	208	612	820
$\mathbb{E}(n_{11})$	$\frac{261\cdot 94}{364} \approx 67.4$			$\frac{57\cdot 24}{78} \approx 17.5$			$\frac{500\cdot 208}{820} \approx 126.8$		
$\mathrm{Var}(n_{11})$	$\frac{261\cdot 103\cdot 94\cdot 270}{364^2(364-1)} \approx 14.2$			$\frac{57\cdot 21\cdot 24\cdot 54}{78^2(78-1)} \approx 3.3$			$\frac{500\cdot 320\cdot 208\cdot 612}{820^2(820-1)} \approx 37.0$		

To test conditional independence of passive smoking and lung cancer, the CMH statistic

$$CMH = \frac{(73 - 67.4) + (19 - 17.5) + (137 - 126.8)}{\sqrt{14.2 + 3.3 + 37.0}} \approx 2.34$$

The two-sided P-value is $2P(Z > 2.34) \approx 2\%$, which shows that the association between passive smoking and lung cancer is significant.

Three Way Tables in R

To enter 3-way contingency table data (X, Y, Z) into R as a 3-dimensional array, we first write the cell counts as a vector in the order that

$$XY \text{ table for } Z = 1,\ XY \text{ table for } Z = 2,\ \ldots$$

and within each XY table, the counts are entered **by column**.
For the "lung cancer and passive smoking" study, we can enter the data as follows.

```
PSM =
array(c( 73, 21, 188,  82,      # table for Japan
         19,  5,  38,  16,      # table for UK
        137, 71, 363, 249),     # table for US
      dim = c(2, 2, 3),
      dimnames = list(
          SpouseSmoking = c("Yes", "No"),
          LungCancer = c("Case", "Control"),
          Country = c("Japan", "UK", "US")))
```

Three Way Tables in R

```
> PSM
, , Country = Japan
             LungCancer
SpouseSmoking Case Control
          Yes   73     188
          No    21      82

, , Country = UK
             LungCancer
SpouseSmoking Case Control
          Yes   19      38
          No     5      16

, , Country = US
             LungCancer
SpouseSmoking Case Control
          Yes  137     363
          No    71     249
```

CMH Test in R

The R command for CHM test is `mantelhaen.test()`

```
> mantelhaen.test(PSM, correct = F)

        Mantel-Haenszel chi-squared test without continuity correction

data:  PSM
Mantel-Haenszel X-squared = 5.4497, df = 1, p-value = 0.01957
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.053554 1.821709
sample estimates:
common odds ratio
         1.385377
```

By default, R perform CMH test with a continuity correction. To go without the correction, we need to add `correct=F`.
R use $(\mathrm{CHM})^2 = (2.34)^2 = 5.4756$ as the test statistic, which has a χ^2 distribution.

CMH Test in R

By default, R will conduct two-sided tests.
R can also perform one-sided CHM test.

```
mantelhaen.test(PSM, correct = F, alternative = "greater")

mantelhaen.test(PSM, correct = F, alternative = "less")
```

CMH Test and Sparse Data

- The normal approximation for CMH statistic requires only overall sample size (sum over all tables) to be big enough.
- CMH test can be used when there are big numbers of partial tables with only a few observations each, provided the total number of observations is big enough.
- The number of observations in a partial table can be as small as 2, but the marginal counts (R_1, R_2, C_1, C_2) must be non-zero. Otherwise the marginal counts will completely determines the cell counts, making $n_{11} - \mathbb{E}(n_{11}) = \mathrm{Var}(n_{11}) = 0$, and the partial table will have no contribution to the CMH statistic.

Remarks About CMH Test

- The formula for the CMH statistic is given using the n_{11} cell in the partial tables. In fact, CMH statistic can be calculated using any of the other three cells: n_{21}, n_{21}, or n_{22}. The value of CMH statistic does not depend on the choice of which cell to use, which makes sense any of them will determine the value of the other three.
- CMH test can be applied to both prospective and retrospective study.
- The textbook introduces CMH test in Section 4.3.4 along with two other tests of conditional independence from logistic models.

After Rejecting the H_0 of Conditional Independence...

When the H_0 of XY conditional independence is rejected, we may examine the estimated odds ratios in the partial tables.

- If estimated odds ratios varies a lot (several times larger) from table to table, i.e, no homogeneous XY association, this means how X is associated Y depends on Z. We'll have to describe XY association separately for each levels of Z.
- If estimated odds ratios do not change much from table to table, we might suspect if XY is homogeneously associated and want to estimate the common odds ratio.
- In fact, we can test homogeneous association (in Chapter 4).

Estimate of the Common Odds Ratio

Suppose the kth XY partial table is

	$Z = k$		
	$Y = 1$	$Y = 2$	row total
$X = 1$	n_{11k}	n_{12k}	R_{1k}
$X = 2$	n_{21k}	n_{22k}	R_{2k}
column total	C_{1k}	C_{2k}	T_k

Mantel-Haenszel's estimate of the common odds ratio from several tables

$$\widehat{\theta}_{MH} = \frac{\text{Sum of } n_{11k}n_{22k}/T_k \text{ over all partial tables}}{\text{Sum of } n_{12k}n_{21k}/T_k \text{ over all partial tables}}$$

Example: Lung Cancer and Passive Smoking (CMH-test)

Spouse	Japan			UK			US		
Smoked	Case	Control	total	Case	Control	total	Case	Control	total
Yes	73	188	261	19	38	57	137	363	500
No	21	82	103	5	16	21	71	249	320
total	94	270	364	24	54	78	208	612	820

Mantel-Haenszel's estimate of the common odds ratio is

$$\widehat{\theta}_{MH} = \frac{(73 \cdot 82)/364 + (19 \cdot 16)/78 + (137 \cdot 249)/820}{(188 \cdot 21)/364 + (38 \cdot 5)/78 + (363 \cdot 71)/820} \approx 1.4$$

The odds of getting lung cancer for nonsmoking wives were about 1.4 times higher if their husbands smoked.

Confidence Interval for the Common Odds Ratio (in R)

In fact, the R function `mantelhaen.test()` that performs the CHM test also reports the MH estimate for the common odds ratio (1.385 as follows, which agrees with our calculation) and provides a confidence interval for it (1.05 to 1.82). The formula for the CI is complex and will not be described here.

```
> mantelhaen.test(PSM, correct = F)

        Mantel-Haenszel chi-squared test without continuity correction

data:  PSM
Mantel-Haenszel X-squared = 5.4497, df = 1, p-value = 0.01957
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.053554 1.821709
sample estimates:
common odds ratio
         1.385377
```

With 95% confidence, the odds of getting lung cancer for nonsmoking wives were about 1.05 to 1.82 times higher if their husbands smoked.