STAT 226 Lecture 8

Section 2.7 Association In Three-Way Tables

Yibi Huang

Textbook Coverage

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

The the two topics below are in Section 4.3.4 of the 2nd edition of the textbook but not in the 3rd edition

- CMH Test for Conditional Independence
- Mantel-Haenszel Estimate for the Common Odds Ratio

Example — Kidney Stone Treatments

A study[1] in 1986 compared 2 treatments for reducing or eliminating kidney stones.

	Outcome (Y)	
Treatment (X)	Success	Failure
Open Surgery	273	77
PCNL	289	61

- PCNL = percutaneous nephrolithotomy
  - cheaper, less invasive, but is it better?
- "Success" means no stone of size $>$ 2 mm three month later

[1] Charig, R., Webb, D.R., Payne, S.(1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". *British Medical Journal* (Clinical Residents Edition), 292(6524): 879–882.

Example — Kidney Stone Treatments

A study[1] in 1986 compared 2 treatments for reducing or eliminating kidney stones.

	Outcome (Y)	
Treatment (X)	Success	Failure
Open Surgery	273	77
PCNL	289	61

- PCNL = percutaneous nephrolithotomy
  - cheaper, less invasive, but is it better?
- "Success" means no stone of size $>$ 2 mm three month later
- It's an observational study, cannot conclude PCNL is better
  - need to control for *confounders*.
- 3-way contingency tables can control for a *single* confounder. Can control for more confounders by models in later chapters

[1] Charig, R., Webb, D.R., Payne, S.(1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". *British Medical Journal* (Clinical Residents Edition), 292(6524): 879–882.

Example — Kidney Stone Treatments (Cont'd)

Breaking the XY table down by a control variable Z = initial size of kidney stones, we get the following three-way table.

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

Y = Outcome (response)

X = Treatment (explanatory variable)

Z = Initial size of kidney stones (control variable)

Initial		*Outcome* (Y)	
Stone Size (Z)	*Treatment* (X)	Success	Failure
Small	Open Surgery	81	6
	PCNL	234	36
Large	Open Surgery	192	71
	PCNL	55	25

Partial Tables & Marginal Tables

In each XY-partial table, the effect of Z is fixed/controlled.

Stone Size (Z)	*Treatment* (X)	*Outcome* (Y) Success	Failure	
Small	Open Surgery	81	6	$\rightarrow XY$-*partial table* given Z = Small
	PCNL	234	36	
Large	Open Surgery	192	71	$\rightarrow XY$-*partial table* given Z = Large
	PCNL	55	25	

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

Treatment (X)	*Outcome* (Y) Success	Failure
Open Surgery	273	77
PCNL	289	61

Simpson's Paradox

Association in the marginal table might be reversed or disappear in each partial table after controlling for a third variable. This is called *Simpson's paradox*.

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

- Cause?

Simpson's Paradox

Association in the marginal table might be reversed or disappear in each partial table after controlling for a third variable. This is called *Simpson's paradox*.

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

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

Conditional Odds Ratio

The (estimated) *conditional odds ratio* of XY given $Z = k$ is the odds ratio of the XY partial table given $Z = k$.

Stone		*Outcome* (Y)	
Size (Z)	*Treatment* (X)	Success	Failure
Small	Open Surgery	81	6
	PCNL	234	36
Large	Open Surgery	192	71
	PCNL	55	25

$$\rightarrow \widehat{\theta}_{XY(1)} = \frac{81 \times 36}{6 \times 234} \approx 2.08$$

$$\rightarrow \widehat{\theta}_{XY(2)} = \frac{192 \times 25}{71 \times 55} \approx 1.23$$

- For patients with small kidney stones, the odds of success for open surgery are 2.08 times as large as the odds for PCNL
- For patients with large kidney stones, the odds of success for open surgery are 1.23 times as large as the odds for PCNL
- Controlling for the initial size of kidney stone, open surgery has higher odds of success than PCNL.

Marginal Odds Ratio

The *XY-marginal odds ratio* is the odds ratio of the XY-marginal table.

	Outcome (Y)	
Trt (X)	Success	Failure
Open	273	77
PCNL	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, open surgery has lower odds of success than PCNL.

Conditional Independence

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 (Z)	Treatment (X)	*Outcome* (Y) 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

If X and Y have an identical associations at each level of Z, we say X and Y have *homogeneous association* given 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. To compare the effectiveness ($Y = S$ or F) of two treatments ($X = A$ or B), we use 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 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

Proof. 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” example has 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

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

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

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

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

Under H_0, the CMH statistic is approximately $N(0, 1)$.
(Or equivalently $(\text{CMH})^2$ is approx. chi-squared w/ 1 degree of freedom.)

Example: Lung Cancer and Passive Smoking

To study the effect of passive smoking and lung cancer, a case-control study was done in each of the 3 countries: Japan, UK, and US, using nonsmoking women married to smokers[2].

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	

[2]Source: Exercise 3.8 on p.68 of *An Introduction to Categorical Data Analysis*, 1ed, 1996, by A. Agresti

Example: Lung Cancer and Passive Smoking

To study the effect of passive smoking and lung cancer, a case-control study was done in each of the 3 countries: Japan, UK, and US, using nonsmoking women married to smokers[2].

Spouse Smoked	Japan Case	Japan Control	UK Case	UK Control	US Case	US 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 3 partial tables all have conditional odds ratios > 1, none is significant by Pearson's X^2 test or Fisher's exact test.

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

[2]Source: Exercise 3.8 on p.68 of *An Introduction to Categorical Data Analysis*, 1ed, 1996, by A. Agresti

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 applying Pearson's X^2 or Fisher's exact test on the combined table (marginal table) would ignore the country effect, and might result in Simpson's paradox if Country is associated with both passive smoking & lung cancer, not revealing the true association between lung cancer and passive smoking
- CMH test can 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
$E(n_{11})$	$\frac{261\cdot 94}{364} \approx 67.4$			$\frac{57\cdot 24}{78} \approx 17.5$			$\frac{500\cdot 208}{820} \approx 126.8$		
$\mathrm{Var}(n_{11})$	$\frac{261\cdot 103\cdot 94\cdot 270}{364^2(364-1)} \approx 14.2$			$\frac{57\cdot 21\cdot 24\cdot 54}{78^2(78-1)} \approx 3.3$			$\frac{500\cdot 320\cdot 208\cdot 612}{820^2(820-1)} \approx 37.0$		

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 $2P(Z > 2.34) \approx 2\%$, showing significant association between passive smoking and lung cancer.

Three Way Tables in R

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

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

Within each XY table, the counts are entered **by column**.

For the “lung cancer and passive smoking” study,

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

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")))
```

```
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 = 5.4, df = 1, p-value = 0.02
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.054 1.822
sample estimates:
common odds ratio
            1.385
```

- By default, R performs CMH test with a continuity correction. To go without the correction, need to add `correct=F`.
- R use $(\text{CHM})^2 = (2.34)^2 = 5.4756$ as the test statistic, which has a chi-squared distribution with df $= 1$.

CMH Test in R

By default, R performs two-sided tests.

R can also perform one-sided CHM test.

```r
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 $n_{11} - \mathrm{E}(n_{11}) = \mathrm{Var}(n_{11}) = 0$, 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 (2nd edition) 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 estimated to be 1.4 times as high if their husbands smoked, compared to those nonsmoking wives in the same country with nonsmoking husbands.

Confidence Interval for the Common Odds Ratio (in R)

The R function `mantelhaen.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 hence is not described here.

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

    Mantel-Haenszel chi-squared test without continuity correction

data:  PSM
Mantel-Haenszel X-squared = 5.4, df = 1, p-value = 0.02
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.054 1.822
sample estimates:
common odds ratio
            1.385
```

Interpretation of the 95% CI (1.05 to 1.82) for the common odds ratio:

With 95% confidence, the odds of getting lung cancer for nonsmoking wives with smoking husbands were about 1.05 to 1.82 times as high, compare to nonsmoking wives in the same country with nonsmoking husbands.

Back to Kidney Stone Treatments

Stone		*Outcome* (Y)	
Size (Z)	*Treatment* (X)	Success	Failure
Small	Open Surgery	81	6
	PCNL	234	36
Large	Open Surgery	192	71
	PCNL	55	25

$$\rightarrow \widehat{\theta}_{XY(1)} = \frac{81 \times 36}{6 \times 234} \approx 2.08$$

$$\rightarrow \widehat{\theta}_{XY(2)} = \frac{192 \times 25}{71 \times 55} \approx 1.23$$

Does Open Surgery have higher odds of success than PCNL, controlling for initial stone size?

```
KS = array(c(81, 234, 6, 36,
             192, 55, 71, 25),
           dim=c(2,2,2),
           dimnames = list(
             Treatment = c("OpenSurgery", "PCNL"),
             Outcome = c("S", "F"),
             StoneSize = c("Small", "Large")
             )
           )
```

```
KS
, , StoneSize = Small

             Outcome
Treatment       S  F
  OpenSurgery  81  6
  PCNL        234 36

, , StoneSize = Large

             Outcome
Treatment       S  F
  OpenSurgery 192 71
  PCNL         55 25
```

```
options(digits=6)
mantelhaen.test(KS, correct = F)


	Mantel-Haenszel chi-squared test without continuity correction

data:  KS
Mantel-Haenszel X-squared = 2.434, df = 1, p-value = 0.119
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.915793 2.285849
sample estimates:
common odds ratio
          1.44685
```

No significant difference in the odds of success (two-sided P-value 0.119)

At 95% confidence, the odds of success for Open surgery were 0.916 to 2.286 times the odds for PCNL.

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


    Mantel-Haenszel chi-squared test without continuity correction

data:  KS
Mantel-Haenszel X-squared = 2.434, df = 1, p-value = 0.0594
alternative hypothesis: true common odds ratio is greater than 1
95 percent confidence interval:
 0.985669      Inf
sample estimates:
common odds ratio
          1.44685
```

The one-sided P-value is 0.059, still not small enough to claim that Open Surgery had higher odds of success than PCNL, controlling for initial size of stone.