STA4030: Categorical Data Analysis
Three-way Tables

Instructor: Bojun Lu

School of Data Science
CUHK(SZ)

October 20, 2020

Agenda

1. 6.1 Introduction
2. 6.2 Partial and Marginal Associations
3. 6.3 Example: Race and the Death Penalty
4. 6.4 Types of Independence
5. 6.5 Cochran-Mantel-Haenszel Methods

6.1 Introduction

6.1.1 Definition

Let X, Y, and Z denote three categorical response variables, X with I categories, Y with J categories, and Z with K categories. Classifications of subjects on three variables have $I \times J \times K$ possible combinations.

A contingency table with I rows, J columns, and K layers is called a $I \times J \times K$ table.

In most studies, it is important to investigate how several variables are interrelated. Inadequate and even incorrect conclusions can result from studying the variables only two at a time.

6.1 Introduction

6.1.2 Notation

Three-way tables result from a triplet of categories being collected for each unit in a sample of size n. Count n_{ijk} is the number of units belonging to category i of variable X and category j of variable Y and category k of variable Z.

Thus the data is contained in $(n_{111}, n_{112}, \ldots, n_{IJK})$ which in turn can be represented in an $I \times J \times K$ table. We may wish to sum over all categories of a variable. For example, if we are interested in the number of units belonging to category i of X and k of Z, we calculate $n_{i+k} = \sum_{j=1}^{J} n_{ijk}$. In general, we use a "$+$" to indicate summing over the categories of a variable.

6.1 Introduction

The counts $(n_{1+1}, n_{1+2}, \ldots, n_{I+K})$ can be represented in a two-way table with $I \times K$ dimensions. This is called the *marginal table* of X and Z - we have removed Y from consideration by *marginalizing* over it.

The distribution of the counts $n_{111}, n_{112}, \ldots, n_{IJK}$ is Multinomial with parameters n and $\boldsymbol{\pi}$, where $\boldsymbol{\pi} = (\pi_{111}, \pi_{112}, \ldots, \pi_{IJK})$. That is, π_{ijk} is the probability that a randomly selected member of the population belongs to category i of X, category j of Y and category k of Z. $\{\pi_{ijk}\}$ is the *joint distribution* of X, Y and Z. Of course, these probabilities must sum to one: $\sum_i \sum_j \sum_k \pi_{ijk} = 1$.

The MLEs for the π_{ijk}s are the sample proportions, $\hat{\pi}_{ijk} = n_{ijk}/n$.

6.1 Introduction

6.1.3 Sampling Schemes

The sampling schemes for three-way tables are the same as for two-way tables. We have

- Poisson sampling: n random, each cell count a Poisson random variable.
- Multinomial sampling: n fixed, sample allocated to cells according to $\{\pi_{ijk}\}$.
- Product-Multinomial Sampling: can happen in several different ways
  - e.g. n_{++k} fixed, fixing the sample size of two-way tables. So-called stratified sampling.
  - e.g. n_{i+k} fixed, fixing the row sizes of two-way tables

6.1 Introduction

6.1.4 Example: Treatments and Clinics

Table 6.1

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

where $Y =$ response (success, failure),

$X =$ drug treatment (A, B),

$Z =$ clinic (1, 2).

6.1 Introduction

6.1.5 Example: Husbands and Wives

Table 6.2

		Wife	
Age	Husband	Yes	No
30-45	Yes	1	9
	No	9	81
45-60	Yes	16	24
	No	24	36
60-75	Yes	49	21
	No	21	9

This study investigates if there is any association between a husband and wife having a disease.

6.2 Partial and Marginal Associations

6.2.1 Marginal Tables and Partial Tables

Given a three-way table, we have numerous ways of presenting the data.

Partial tables are cross-sections of the three-way table. They display the relationship between two variables, X and Y say, while holding the level of the third variable (Z in this case) constant.

Marginal tables are obtained by summing counts in the partial tables. They display the relationship between two variables, Y and Z say, after summing over all levels of the third variable (X in this case).

6.2 Partial and Marginal Associations

For example, recall Table 6.1, the Treatments and Clinics data. These are the two Partial Tables between Treatment (X) and Response (Y) given a level of Clinic (Z):

Clinic 1 ($Z = 1$):

	Y	Response		
X		Success	Failure	
Treatment	A	18	12	30
	B	12	8	20
		30	20	50

Clinic 2 ($Z = 2$):

	Y	Response		
X		Success	Failure	
Treatment	A	2	8	10
	B	8	32	40
		10	40	50

6.2 Partial and Marginal Associations

Here is the Marginal Table obtained by summing over the levels of the Clinic variable Z:

	Y	Response		
X		Success	Failure	
Treatment	A	20	20	40
	B	20	40	60
		40	60	100

6.2 Partial and Marginal Associations

6.2.2 Conditional and Marginal Odds Ratios

Having arranged data into partial or marginal two-way tables, we have all the tests and measures of association available to us to investigate any relationship between the variables.

Partial association is the association obtained from a partial table. It is also called *conditional association* because it refers to an association between two variables conditional on the third being fixed at some level.

Marginal association is the association obtained from a marginal table. The marginal association refers to an association between two variables while ignoring the third.

6.2 Partial and Marginal Associations

Partial and marginal associations can be measured by the appropriate odds ratios.

Consider a $2 \times 2 \times K$ table, where K denotes the number of categories of the control variable, Z. Since Z is a control variable, it makes sense to investigate the association between X and Y conditional on levels of Z and marginalizing over Z.

Let $\{\mu_{ijk}\}$ denote cell expected frequencies for some sampling model, such as binomial, multinomial, product-multinomial or Poisson sampling model.

6.2 Partial and Marginal Associations

Then for each level of Z we have the XY conditional odds ratios:

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}},\ k = 1, \ldots, K$$

and marginalizing over Z we have the XY marginal odds ratio:

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}},\ \mu_{ij+} = \sum_k \mu_{ijk}$$

The sample analogues are

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}, k = 1, \ldots, K\ \ \hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$$

6.2 Partial and Marginal Associations

Recall the Treatments and Clinics data. From Table 6.1,

$$\hat{\theta}_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0,$$

$$\hat{\theta}_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1.0.$$

Given the clinic, response and treatment are conditionally independent.

$$\hat{\theta}_{XY} = \frac{20 \times 40}{20 \times 20} = 2.0.$$

Ignoring the clinic, the odds of a success for treatment A are twice those for treatment B.

6.2 Partial and Marginal Associations

Why? The conditional XZ and YZ odds ratios give a clue.

The partial tables for XZ and YZ are:

$Y = 1$ (Success)

$Z \backslash X$	A	B
1	18	12
2	2	8

$Y = 2$ (Failure)

$Z \backslash X$	A	B
1	12	8
2	8	32

$$\hat{\theta}_{XZ(1)} = \frac{18 \times 8}{12 \times 2} = 6.0, \quad \hat{\theta}_{XZ(2)} = \frac{12 \times 32}{8 \times 8} = 6.0.$$

6.2 Partial and Marginal Associations

$X = 1$ (A)
S – Success
F – Failure

$Z\backslash Y$	S	F
1	18	12
2	2	8

$X = 2$ (B)

$Z\backslash Y$	S	F
1	12	8
2	8	32

$$\hat{\theta}_{YZ(1)} = \frac{18 \times 8}{12 \times 2} = 6.0, \qquad \hat{\theta}_{YZ(2)} = \frac{12 \times 32}{8 \times 8} = 6.0.$$

Clinic 1 tends to use treatment A more often, and clinic 1 also tends to have more successes regardless of the treatment received. So, treatment A having a better successful rate may be due to other reasons, e.g. patients at clinic 1 tended to be younger or in better health than those at clinic 2.

6.2 Partial and Marginal Associations

Recall Table 6.2. Husbands, wives, diseases - any association? Let's look at some partial tables.

Age 30-45 ($x_3 = 1$):

x_1 \ x_2		Wife Yes	Wife No	
Husband	Yes	1	9	10
	No	9	81	90
		10	90	100

Observed odds ratio: $\hat{\theta} = 1$, x_1 and x_2 are not associated.

6.2 Partial and Marginal Associations

Age 45–60 ($x_3 = 2$):

	x_2	Wife		
x_1		Yes	No	
Husband	Yes	16	24	40
	No	24	36	60
		40	60	100

Observed odds ratio: $\hat{\theta} = 1$, x_1 and x_2 are not associated.

Age 60-75 ($x_3 = 3$):

	x_2	Wife		
x_1		Yes	No	
Husband	Yes	49	21	70
	No	21	9	30
		70	30	100

Observed odds ratio: $\hat{\theta} = 1$, x_1 and x_2 are not associated.

6.2 Partial and Marginal Associations

If we collapse over x_3, i.e. marginalize over the age variable, we have:

	x_2	Wife		
x_1		Yes	No	
Husband	Yes	66	54	120
	No	54	126	180
		120	180	300

Observed odds ratio: $\hat{\theta} = 2.85$

$\therefore X_1$ and X_2 appear associated. In fact, upon calculating the Wald confidence interval for $\theta_{X_1X_2}$ we do reject the null hypothesis of $\theta_{X_1X_2} = 1$.

6.2 Partial and Marginal Associations

6.2.3 Simpson's Paradox

In the Treatments and Clinics data, conditional on the clinic, it appeared that Treatment A and Treatment B were equally effective. However, when we combined the data from the two clinics, it seemed that Treatment A was much more effective.

The Husbands and Wives data suggested that, ignoring age, there was an association between a husband having the disease and the wife having the disease. However, this association disappeared upon examining the data conditional on the age of the husband and wife.

In both examples, the marginal association we see for X and Y is different to the association we see after controlling for the different levels of a third variable Z. This apparent contradiction is called *Simpson's Paradox*.

6.2 Partial and Marginal Associations

As shown in the Husbands and Wives example, in studying the effect of X on Y, one should control any covariate (e.g. age) that can influence that relationship (e.g. husband and wife).

This involves using some mechanism to hold the covariate constant. Otherwise, an observed effect of X and Y may actually reflect effects of that covariate on both X and Y. The relationship between X and Y then shows *confounding*.

The association comes from age being associated with the husband having the disease and with the wife having the disease.

6.2 Partial and Marginal Associations

For example, consider investigating whether passive smoking is associated with lung cancer.

A cross-sectional (partial) table might be designed as:

	x_2	Lung cancer	
x_1		Yes	No
spouse	smoke	n_{11}	n_{12}
	do not smoke	n_{21}	n_{22}

The study should attempt to control for age, socioeconomic status, or other factors that might relate both to spouse smoking and to developing lung cancer.

6.3 Example: Race and the Death Penalty

6.3.1 Introduction

The 674 subjects classified in Table 6.3 were defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987. The variables in the table are

Y = death penalty verdict (yes, no)
X = race of defendants (white, black)
Z = race of victims (white, black)

We study the effect of defendants' race on the death penalty verdict, treating victims' race as a control variable.

Table 6.3 has a 2×2 partial table relating defendants' race and the death penalty verdict at each category of victims' race.

6.3 Example: Race and the Death Penalty

Table 6.3

Victims' Race (Z)	Defendants' Race (X)	Death Penalty (Y) Yes	Death Penalty (Y) No	Percent 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

6.3 Example: Race and the Death Penalty

6.3.2 Sample associations

The conditional association:

When Z(victim) $=$ White, the death penalty was imposed $22.9\% - 11.3\% = 11.6\%$ more often for black defendants than for white defendants.

When Z(victim) $=$ Black, the death penalty was imposed $2.8\% - 0\% = 2.8\%$ more often for black defendants than for white defendants.

Controlling for victims' race by keeping it fixed, the death penalty was imposed more often on black defendants than on white defendants.

6.3 Example: Race and the Death Penalty

The marginal association:
Ignoring victims' race, the death penalty was imposed less often on black defendants than on white defendants ($7.9\% - 11.0\% = -3.1\%$).

The marginal association reverses direction compared to the partial tables! We can show this more formally by looking at the conditional odds ratios and marginal odds ratio:
$\hat{\theta}_{XY(1)} = 0.43 < 1,\ \hat{\theta}_{XY(2)} = 0.94^{*} < 1,\ \hat{\theta}_{XY} = 1.45 > 1.$

Simpson's paradox appears again!

*we added 0.5 to each entry of the partial table in order to estimate $\theta_{XY(2)}$.

6.3 Example: Race and the Death Penalty

6.3.3 Digging deeper

Why? The causes are:

1.

Z \ X	W	B
W	53+414	11+37
B	0+16	4+139

$$\hat{\theta}_{XZ} = \frac{467\times 143}{16\times 48} = 87.0 \gg 1$$

The association between victims' and defendants' races is extremely strong. i.e. Whites tended to kill whites.

6.3 Example: Race and the Death Penalty

2.

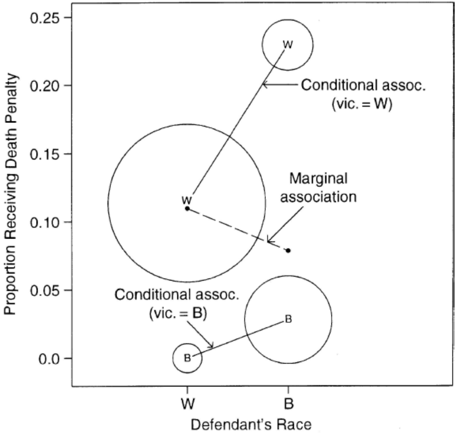
Z \ Y	Yes	No
W	53+11	414+37
B	0+4	16+139

$$\hat{\theta}_{YZ} = \frac{64\times 155}{4\times 451} = 5.5 > 1$$

Regardless of defendants' race, the death penalty was much more likely when the victims were white than when the victims were black. i.e. Killing whites is more likely to result in the death penalty.

6.3 Example: Race and the Death Penalty

So, the marginal association show a greater tendency than the conditional associations for white defendants to receive the death penalty. The following figure further illustrates the reason:



6.3 Example: Race and the Death Penalty

Proportion represented by each circle:

Z \ X	W	B	W	B
W	467	48	69.3%	7.1%
B	16	143	2.4%	21.2%

Proportion of “Yes" in each circle (conditional proportion):

	Percent of Yes		Population size	
Z \ X	W	B	W	B
W	11.3%	22.9%	69.3%	7.1%
B	0%	2.8%	2.4%	21.2%

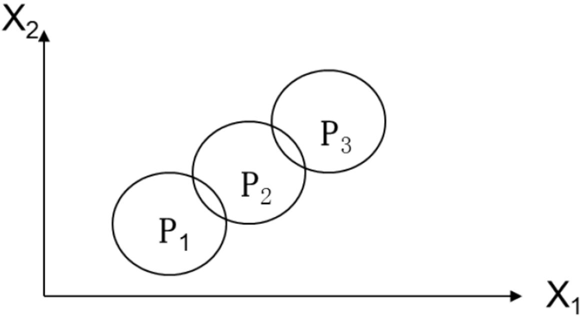
Marginal proportions (collapse victims’ race Z):

$X = W$: $(11.3\% \times 69.3\% + 0\% \times 2.4\%)/(69.3\% + 2.4\%) = 11\%$,
$X = B$: $(22.9\% \times 7.1\% + 2.8\% \times 21.2\%)/(7.1\% + 21.2\%) = 7.9\%$.

6.3 Example: Race and the Death Penalty

This situation also appears when continuous variables are encountered.

Within each population P_i, X_1 and X_2 are uncorrelated. However, as a whole, the correlation is positive.



6.4 Types of Independence

6.4.1 Conditional Independence

With two-way tables, our intuitive definition of independence was

$$P(Y=j|X=i)=P(Y=j), \text{ for all levels } i,j.$$

With three-way tables, we have more choices to condition on and therefore more types of independence. For example, we say Y and X are conditionally independent at level k of Z if

$$P(Y=j|X=i,Z=k)=P(Y=j|Z=k) \text{ for all levels } i,j.$$

6.4 Types of Independence

We say Y and X are *conditionally independent given* Z if Y and X are conditionally independent at every level of Z.

In terms of the population proportions, if $\pi_{ijk} = \Pr(X = i, Y = j, Z = k)$, then X and Y conditionally independent given $Z \Leftrightarrow \pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$, for all i, j, k.

For $2 \times 2 \times K$ tables, X and Y are conditionally independent $\Leftrightarrow$ $\theta_{XY(k)} = 1, \; k = 1, \ldots, K$.

6.4 Types of Independence

6.4.2 Marginal Independence

Marginal independence refers to the independence between two variables once the the third has been ignored by marginalization.

In terms of the population proportions, then X and Y are marginally independent if $\pi_{ij+} = \pi_{i++}\pi_{+j+}$.

X and Y are marginally independent $\Leftrightarrow \theta_{XY} = 1$.

Conditional independence does not imply marginal independence.

6.4 Types of Independence

6.4.3 Homogeneous Association

A $2 \times 2 \times K$ table has *homogeneous XY association* when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

Then, the effect of X on Y is the same at each category of Z.

Conditional independence of X and Y is the special case:

$$\theta_{XY(k)} = 1.0, \quad k = 1, \ldots, K.$$

When homogeneous XY association occurs, there is no interaction between X and Y in their effects on Z. The conditional odds ratios don't depend on the level of the third variable, Z.

6.4 Types of Independence

When interaction exists, the conditional odds ratio for any pair of variables changes across categories of the third.

e.g. For $X =$ smoking (yes, no)
$Y =$ lung cancer (yes, no)
$Z =$ age $(< 45,\ 45 - 65,\ > 65)$

Suppose that $\theta_{XY(1)} = 1.2$, $\theta_{XY(2)} = 3.9$, and $\theta_{XY(3)} = 8.8$.

Then, smoking has a week effect on lung cancer for young people, but the effect strengthens considerably with age.

In this example, age is called an *effect modifier*: the effect of smoking is modified depending on the age.

6.5 Cochran-Mantel-Haenszel Methods

6.5.1 Introduction

This section introduces inferential analyses for three-way tables, which include:

1. Tests of conditional independence and homogeneous association with the K conditional odds ratios in $2 \times 2 \times K$ tables.
2. Combine the sample odds ratios from the K partial tables into a single summary measure of partial association.

Analyses of conditional association are very relevant in many applications involving multivariate data. To illustrate, we analyze the data in Table 6.4, which summarizes eight studies in China about smoking and lung cancer.

6.5 Cochran-Mantel-Haenszel Methods

Table 6.4 Lung cancer data.

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

6.5 Cochran-Mantel-Haenszel Methods

Table 6.4 is a $2 \times 2 \times K$ table, which includes original data (Columns 1-4), the sample odds ratios (Column 5), and the expected values and variances of n_{11k} (the number of lung cancer cases who were smokers) (Columns 6-7).
Let

$$\begin{aligned} X &= \text{status of smoking (smoker, nonsmoker)} \\ Y &= \text{outcomes for lung cancer (yes, no)} \\ Z &= \text{different cities (levels } k = 1, \ldots, 8) \end{aligned}$$

Subjects may vary among cities on relevant characteristics such as socioeconomic status, which may cause heterogeneity among the cities in smoking rate and in the lung cancer rate. Therefore, we investigate the association between X and Y while controlling for Z.

6.5 Cochran-Mantel-Haenszel Methods

6.5.2 The Cochran-Mantel-Haenszel (CMH) test

For $2 \times 2 \times K$ tables, the null hypothesis that X and Y are conditionally independent, given Z, means that

$$H_0 : \theta_{XY(k)} = 1,\ k = 1, \ldots, K.$$

In partial table k, the row totals are $\{n_{1+k}, n_{2+k}\}$, the column totals are $\{n_{+1k}, n_{+2k}\}$.

When Z is controlled, both these totals are given. Hence, the cell count n_{11k} determines all other counts in the partial table. As a result, n_{11k} follows a hypergeometric distribution. The test statistic utilizes this cell in each partial table.

6.5 Cochran-Mantel-Haenszel Methods

Under H_0, it can be shown that the mean and variance of n_{11k} are

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

$$\mathrm{Var}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}.$$

When the true odds ratio $\theta_{XY(k)}$ exceeds 1.0 in partial table k, we expect $(n_{11k} - \mu_{11k}) > 0$. The test statistic combines these differences across all K partial tables.

When the odds ratio exceeds 1.0 in every partial table, the sum of such differences tends to be a relatively large positive number; when the odds ratio is less than 1.0 in each table, the sum of such differences tends to be a relatively large negative number.

6.5 Cochran-Mantel-Haenszel Methods

The test statistic summarizes the information from the K partial tables using

$$\mathrm{CMH} = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k \mathrm{Var}(n_{11k})}.$$

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

The CMH statistic takes larger values when $(n_{11k} - \mu_{11k})$ is consistently positive or consistently negative for all partial tables, rather than positive for some and negative for others.

6.5 Cochran-Mantel-Haenszel Methods

When using CMH statistic, we should note that:

1. This test is inappropriate when the association varies dramatically among the partial tables. It works best when the XY association is similar in each partial table.
2. The CMH statistic combines information across partial tables. When the true association is similar in each table, this test is more powerful than separate tests within each table.
3. It is improper to combine results by adding the partial tables together to form a single 2×2 marginal table for the test. Simpson's paradox revealed the dangers of collapsing three-way tables.

6.5 Cochran-Mantel-Haenszel Methods

6.5.3 Example: Lung Cancer Meta Analysis

Table 6.4 summarizes eight case-control studies in China about smoking and lung cancer.

In each partial table k $(k = 1, \dots, 8)$, we test conditional independence between smoking and lung cancer (controlling $Z = k$), which is equivalent to test

$$H_0 : \theta_{XY(k)} = 1, \; k = 1, \dots, 8.$$

Under this hypothesis, the sample odds ratio for each partial table and the expected value and variances of the number of lung cancer cases who were smokers (the cell count n_{11k}) are calculated and reported in Table 6.4.

6.5 Cochran-Mantel-Haenszel Methods

In each partial table, the sample odds ratio shows a moderate positive association, so it makes sense to combine results through the *CMH* statistic:

$$\begin{aligned}\mathrm{CMH} &= \frac{\sum_k (n_{11k} - \mu_{11k})^2}{\sum_k \mathrm{Var}(n_{11k})} \\ &= \frac{(2930.0 - 2562.5)^2}{482.1} = 280.1\end{aligned}$$

with $df = 1$. There is extremely strong evidence against conditional independence ($P < 0.0001$).

A statistical analysis that combines information from several studies is called meta analysis. The meta analysis of Table 6.4 provides stronger evidence of an association between smoking and lung cancer than any single partial table gives.

6.5 Cochran-Mantel-Haenszel Methods

6.5.4 Estimation of common odds ratio

It is more informative to estimate the strength of association than simply to test a hypothesis about it. When the association seems stable across partial tables, we can estimate an assumed common value of the K true ratios.

In a $2 \times 2 \times K$ table, suppose that

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

The *Mantel-Haenszel estimator* of that common value equals

$$\hat{\theta}_{MH} = \frac{\sum_k (n_{11k} n_{22k} / n_{++k})}{\sum_k (n_{12k} n_{21k} / n_{++k})}.$$

6.5 Cochran-Mantel-Haenszel Methods

The standard error for $\log(\hat{\theta}_{MH})$ has a complex formula, so we shall not report it here. Its value can be given by computer software.

Table 6.4 continued

For the 'Chinese Smoking and Lung Cancer Study' summarized in Table 6.4, the Mantel-Haenszel odds ratio estimate equals

$$\hat{\theta}_{MH} = \frac{(126)(61)/(322) + \cdots + (104)(36)/(250)}{(35)(100)/(322) + \cdots + (21)(89)/(250)} = 2.17.$$

From computer software, the standard error of $\log(\hat{\theta}_{MH})$ is

$$\mathrm{ASE}(\log(\hat{\theta}_{MH})) = 0.046.$$

6.5 Cochran-Mantel-Haenszel Methods

The 95% confidence interval for the common log odds ratio is

$$\log(2.17) \pm 1.96 \times 0.046, \quad \text{or} \quad (0.686, 0.868).$$

The 95% confidence interval for the common odd ratio is

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

The odds of lung cancer for smokers equal about twice the odds for non-smokers.

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. Similarly, the CMH test is a powerful summary of evidence against the hypothesis of conditional independence, as long as the sample associations fall primarily in a single direction.