Categorical Data Analysis

Chapter 2

Deyuan Li
School of Management
Fudan University

Fall 2015

Outline

1. 2.1 Probability Structure for Contingency Tables
2. 2.2 Comparing Two Proportions
3. 2.3 Partial Association in Stratified 2×2 Tables
4. 2.4 Extensions for $I \times J$ Tables

2.1 Probability Structure for Contingency Tables

2.1.1 Contingency tables（列联表）and their distributions

X: categorical response variable with I categories;
Y: categorical response variable with J categories.

For contingency table, the cells contain frequency counts of the outcomes for a sample. Also called cross-classification table, $I \times J$ table.

Table 2.1: Cross-Classification of Aspirin Use and Myocardial Infarction.

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

2.1.1 Contingency tables and their distributions

- Let $\pi_{ij} = P(X = i, Y = j)$ for $i = 1, ..., I$ and $j = 1, ..., J$.
- Marginal and joint probabilities:
  $\{\pi_{ij}\}$ is the joint distribution of (X, Y),
  $\{\pi_{i+}\}$ is the marginal distribution of X,
  $\{\pi_{+j}\}$ is the marginal distribution of Y,

  where
  $$\pi_{i+} = \sum_j \pi_{ij} \quad \text{and} \quad \pi_{+j} = \sum_i \pi_{ij}.$$
- $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1$.

2.1.2 Sensitivity and specificity

TABLE 2.2 Estimated Conditional Distributions for Breast Cancer Diagnoses

Breast Cancer	Diagnosis of Test		
	Positive	Negative	Total
Yes	0.82	0.18	1.0
No	0.01	0.99	1.0

X: true disease status (yes or no);
Y: diagnosis (positive or negative).

sensitivity（敏感度）: $P(Y = \text{positive}|X = \text{yes})$;
specificity（特异度）: $P(Y = \text{negative}|X = \text{no})$.

2.1.3 Independence of categorical variables

- Define

$$\pi_{j|i} = P(Y \text{ is in category } j \mid X \text{ is in category } i) = \pi_{ij}/\pi_{i+}.$$

$\{\pi_{1|i}, ..., \pi_{J|i}\}$: the cond. distribution of Y at category I of X.

- X and Y are independent if

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i = 1, 2, ..., I \text{ and } j = 1, 2, ..., J.$$

- Consequently, when X and Y are independent,

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \quad \text{for all } i, j.$$

2.1.4 Poisson and multinomial sampling

- Poisson sampling: treats cell counts, denoted by $\{Y_{ij}\}$, as independent Poisson variables with parameters $\{\mu_{ij}\}$.

  So,

$$P(Y_{11} = n_{11}, ..., Y_{ij} = n_{ij}, ..., Y_{IJ} = n_{IJ})$$
$$= \prod_i \prod_j \exp(-\mu_{ij})\mu_{ij}^{n_{ij}}/n_{ij}!.$$

2.1.4 Poisson and multinomial sampling

- Multinomial sampling:

  (1) total sample size n is fixed but row and column totals are not.

$$P(Y_{ij} = n_{ij}, i = 1, ..., I, j = 1, ..., J) = \frac{n!}{n_{11}! \cdots n_{IJ}!} \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

  Note $\sum_i \sum_j \pi_{ij} = 1$ and $\sum_i \sum_j n_{ij} = n$.

  (2) rows total sample sizes ($\{n_i = n_{i+}\}$) are fixed.

$$P(Y_{ij} = n_{ij}, j = 1, ..., J) = [n_i!/(n_{i1}! \cdots n_{iJ}!)] \prod_j \pi_{j|i}^{n_{ij}}.$$

  Note $\sum_j \pi_{j|i} = 1$ and $\sum_j n_{ij} = n_i$. Here I is fixed.

2.1.4 Poisson and multinomial sampling

Relation between Poisson and multinomial samplings:

conditional on $\{n_i\}$, the cell counts $\{n_{ij}, j = 1, 2, ..., J\}$ of Poisson sampling have the multinomial sampling (2) with $\{\pi_{j|i} = \mu_{ij}/\mu_{i+}, j = 1, 2, ..., J\}$.

2.1.4 Poisson and multinomial sampling

To see this, let $\mu_{i+} = \sum_j \mu_{ij}$ and $n_{i+} = \sum_j n_{ij}$. Independence of Y_{ij}'s implies that $\sum_j Y_{ij} \sim$ Poisson(μ_{i+}). Thus

$$P(\sum_j Y_{ij} = n_{i+}) = \exp(-\mu_{i+})\mu_{i+}^{n_{i+}}/n_{i+}!$$

and

$$\begin{aligned} & P(Y_{ij} = n_{ij},\ j = 1, 2, ..., J | \sum_j Y_{ij} = n_{i+}) \\ = \quad & P(Y_{ij} = n_{ij},\ j = 1, 2, ..., J)/P(\sum_j Y_{ij} = n_{i+}) \\ = \quad & \frac{\prod_j \exp(-\mu_{ij})\mu_{ij}^{n_{ij}}/n_{ij}!}{\exp(-\mu_{i+})\mu_{i+}^{n_{i+}}/n_{i+}!} = \frac{n_{i+}!}{\prod_j n_{ij}!}\prod_j \left(\frac{\mu_{ij}}{\mu_{i+}}\right)^{n_{ij}}. \end{aligned}$$

2.1.5 Seat belt example

Study the relation between seat-belt use (yes/no) and outcome of an automobile crash (fatality/nonfatality) for drives involved in accidents on the Massachusetts Turnpike.

The results will be summarized by the following table

Table 2.4 Seat-Belt Use and Results of Automobile Crashes

	Result of Crash	
Seat-Belt Use	Fatality	Nonfatality
Yes		
No		

2.1.5 Seat belt example

- If studying the results of next year, the total sample size is unknown. So apply Poisson sampling with unknown $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$.
- If 200 observations are random sampled from the records of crashes on the turnpike in the past year, the apply multinomial sampling with sample size $n = 200$ and unknown joint probability $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$.
- If the records have been classified according to fatality/nonfatality, and we random sample 100 observations from fatality records and 100 observations from nonfatality records. So apply conditional multinomial sampling method.

Summary: Model depends on the way of collecting data.

Outline

1 2.1 Probability Structure for Contingency Tables

2 2.2 Comparing Two Proportions

3 2.3 Partial Association in Stratified 2×2 Tables

4 2.4 Extensions for $I \times J$ Tables

2.2 Comparing two proportions

We focus on 2×2 contingency table.

The rows are groups (X) and the columns are the categories of Y.

Let $\pi_{ij} = P(X = i, Y = j)$ for $i, j = 1, 2$.

	Y	
X	Success	Failure
group 1	π_{11}	π_{12}
group 2	π_{21}	π_{22}

2.2.1 Difference of proportion（比例差）

Denote π_i for $\pi_{1|i} = \pi_{i1}/(\pi_{i1} + \pi_{i2})$.

The difference of proportion of successes is $\pi_1 - \pi_2$.

The difference of proportion of failure is $\pi_2 - \pi_1$.

Of course, $|\pi_1 - \pi_2| \in [0, 1]$.

X and Y are independent when $\pi_1 = \pi_2$.

2.2.2 Relative risk (相对风险, RR)

Relative risk (*for success*) is defined by π_1/π_2.

Relative risk (*for failure*) is defined by $(1-\pi_1)/(1-\pi_2)$.

X and Y are independent when relative risk equals to 1.

2.2.3 Odds ratio (优势比, OR)

- For a probability π, odd is defined to be $\Omega = \pi/(1-\pi)$. Then (1) $\Omega \geq 0$; (2) $\pi = \Omega/(1+\Omega)$; (3) $\Omega > 1$ means a success is more likely than a failure.
- The odds ratio is defined to be

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}, \qquad (2.4)$$

  where Ω_i is the odd in row *I*.
- In 2×2 contingency table, the odds in row *i* is equivalently defined as $\Omega_i = \pi_{i1}/\pi_{i2}$ for $i = 1, 2$ and thus the odds ratio is

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

  θ is also called *cross-product ratio*.

2.2.4 Properties of the odds ratio

- $\theta \geq 0$; $\theta = 1 \Leftrightarrow$ independence;
- $\theta > 1$ means that subjects in row 1 is more likely to have a success than those in row 2;
- θ does not change if rows and columns are transformed.

The sample odds ratio is $\hat{\theta} = n_{11}n_{22}/(n_{12}n_{21})$.

Relation between odds ratio and relative risk, by $\Omega = \pi/(1-\pi)$ and (2.4), is

$$\text{odds ratio} = \text{relative risk} \left(\frac{1-\pi_2}{1-\pi_1}\right).$$

Outline

1 2.1 Probability Structure for Contingency Tables

2 2.2 Comparing Two Proportions

3 2.3 Partial Association in Stratified 2 × 2 Tables

4 2.4 Extensions for $I \times J$ Tables

2.3 Partial Association in Stratified 2 × 2 Tables

Discuss the association between categorical variables X and Y while controlling for a variable Z.

2.3.1 Partial table

Three-way contingency table (XYZ-table) includes K (the number of categories for Z) partial tables.

The associations in partial tables are called partial associations.

XY-marginal table: combining the partial tables such that no information on Z in the new table.

2.3.2 Death penalty example

Study the effects of racial characteristics on whether persons convicted of homicide received the death penalty.

Table 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race(Z)	Defendant's Race(X)	Death Penalty(Y) Yes	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

Two partial tables, one *XY* marginal table.

2.3.3 Conditional and marginal odds ratios

We illustrate for $2 \times 2 \times K$ tables, where K denotes the number of categories of a control variable Z.

- Let $\{\mu_{ijk}\}$ denote cell expected frequencies for some sampling models, e.g. binomial, multinomial or Poisson sampling. Note that for multinomial sampling, $\mu_{ijk} = n\pi_{ijk}$.
- For fixed k, the odds ratio

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

describes conditional XY association in partial table k.

2.3.3 Conditional and marginal odds ratios

- The XY marginal odds ratio is

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

where $\mu_{ij+} = \sum_k \mu_{ijk}$.

In Table 2.6,

$$\theta_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43,$$

$$\theta_{XY(2)} = \frac{0 \times 139}{16 \times 4} = 0,$$

and

$$\theta_{XY} = (53 \times 176)/(430 \times 15) = 1.45.$$

2.3.4 Marginal versus conditional independence

- If X and Y are independent in partial table k, then X and Y are called conditional independent at level k of Z, i.e.

$$P(X = i, Y = j|Z = k) = P(X = i|Z = k)P(Y = j|Z = k). \quad (1)$$

- If (1) holds for all k, then X and Y are called conditional independent given Z.
- Let $\pi_{ijk} = P(X = i, Y = j, Z = k)$. Then conditional independence is equivalent to

$$\begin{aligned} \pi_{ijk} &= P(X = i, Y = j|Z = k)P(Z = k) \\ &= P(X = i|Z = k)P(Y = j|Z = k)P(Z = k) \\ &= P(X = i, Z = k)P(Y = j, Z = k)/P(Z = k) \\ &= \pi_{i+k}\pi_{+jk}/\pi_{++k} \end{aligned}$$

for all i, j and k.

2.3.4 Marginal versus conditional independence

Conditional independence does not imply marginal independence, i.e.

$$P(X \in A, Y \in B|Z \in C) = P(X \in A|Z \in C)P(Y \in B|Z \in C)$$

does not imply

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

2.3.5 Homogeneous association（齐次关联）

A $2 \times 2 \times K$ table has homogeneous XY association if

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

Recall that

(1) $\theta_{XY(k)} = \mu_{11k}\mu_{22k}/(\mu_{12k}\mu_{21k})$;

(2) conditional independence of X and Y means $\theta_{XY(k)} = 1.0$ for all k.

Outline

1 2.1 Probability Structure for Contingency Tables

2 2.2 Comparing Two Proportions

3 2.3 Partial Association in Stratified 2×2 Tables

4 2.4 Extensions for $I \times J$ Tables

2.4 Extensions for $I \times J$ Tables

For 2×2 table, a single number (e.g. odds ratio) summary the association;

for $I \times J$ table, a set of numbers (e.g. a set of ratios) or a single index can describe the association.

2.4.1 Odds ratios in $I \times J$ Tables

- There exists $C_I^2 C_J^2 = I(I-1)J(J-1)/4$ odds ratios. Too much!
- Consider the subset of $(I-1)(J-1)$ local odds ratios

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}, \quad i = 1, 2, ..., I-1, \ j = 1, 2, ..., J-1.$$

- Other odds ratios can be expressed by these $(I-1)(J-1)$ basic odds ratios. Independence is equivalent to all $(I-1)(J-1)$ odds ratios equaling 1.0.
- Another basic odds ratios are

$$\alpha_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}}, \quad i = 1, 2, ..., I-1, \ j = 1, 2, ..., J-1.$$

2.4.1 Odds ratios in $I \times J$ Tables

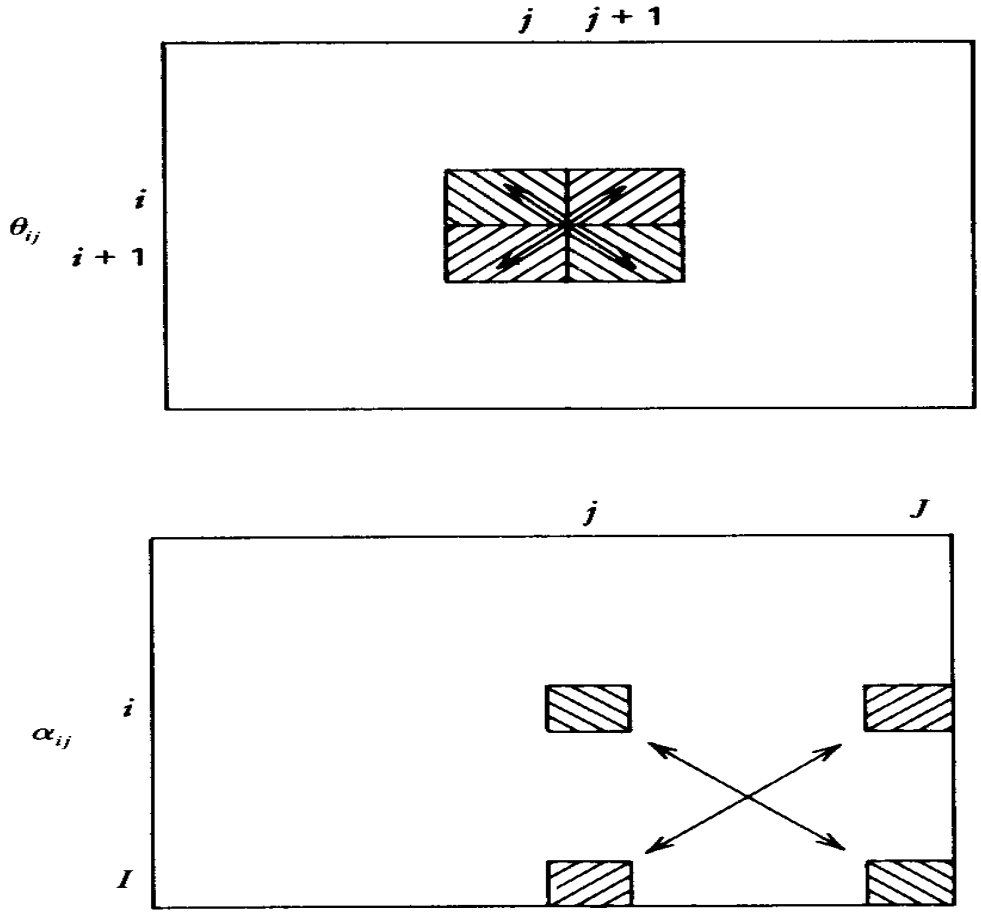

FIGURE 2.3 Odds ratios for $I \times J$ tables.

2.4.3 Ordinal trends: concordant and discordant pairs

When both X and Y are ordinal, a monotone trend association is common. As level of X increases, the level of response Y increases or decreases.

One single parameter can describe this trend.

A pair (of two cells) is

- *concordant*（同调） if the subject ranked higher on X also ranks higher on Y; or
- *discordant*（异调） if the subject ranking higher on X ranks lower on Y; or
- *tied*（紧） if the subjects have the same classification on X and/or Y.

2.4.3 Ordinal trends: concordant and discordant pairs

Example. Income and Job Satisfaction

Table 2.8: Cross-Classification of Job Satisfaction by Income.

Income (dollars)	Job Satisfaction: Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
<15,000	1	3	10	6
15,000-25,000	2	3	10	7
25,000-40,000	1	6	14	12
>40,000	0	1	9	11

The total number of concordant pairs, denoted by C, equals

$$\begin{aligned} C =& 1 \cdot (3 + 10 + 7 + 6 + 14 + 12 + 1 + 9 + 11) \\ &+ 3 \cdot (10 + 7 + 14 + 12 + 9 + 11) + 10 \cdot (7 + 12 + 11) \\ &+ 2 \cdot (6 + 14 + 12 + 1 + 9 + 11) + 3 \cdot (14 + 12 + 9 + 11) \\ &+ 10 \cdot (12 + 11) + 1 \cdot (1 + 9 + 11) + 6 \cdot (9 + 11) + 14 \cdot (11) \\ =& 1331. \end{aligned}$$

2.4.3 Ordinal trends: concordant and discordant pairs

The total number of discordant pairs, denoted by D, equals

$$\begin{aligned} D =& 3 \cdot (2 + 1 + 0) + 10 \cdot (2 + 3 + 1 + 6 + 0 + 1) \\ & + ... + 12 \cdot (0 + 1 + 9) = 849. \end{aligned}$$

In this example, $C > D$ suggests a tendency for low income to occur with low job satisfaction and high income with high job satisfaction.

For two independent observations from a joint probability distribution $\{\pi_{ij}\}$, the probability of concordance and discordance are

2.4.4 Ordinal measure of association: gamma

Given that a pair is untied, $\Pi_c/(\Pi_c + \Pi_d)$ is the probability of concordance and $\Pi_d/(\Pi_c + \Pi_d)$ is the probability of discordance. The difference

is called *gamma*. The estimator of γ is $\hat{\gamma} = (C - D)/(C + D)$.

About γ:

- symmetrical (X and Y could be exchanged, unnecessary to identify the response variable);
- $-1 \leq \gamma \leq 1$;
- reverses the category ordering of one variable, γ only changes the sign;
- $|\gamma| = 1$ implies X and Y are perfect linear;
- independence $\Rightarrow \gamma = 0$.

2.4.5 Gamma for job satisfaction example

For Table 2.8, $C = 1331$ and $D = 849$. Hence

$$\hat{\gamma} = (1331 - 849)/(1331 + 849) = 0.221.$$

Only a weak tendency exists for a job satisfaction to increase as income increase.

Of the untied pairs, the proportion of concordance pairs is 0.221 higher than the proportion of discordant pairs.