What to do today (01/16)?

1. Introdution and Preparation

2. Analysis with Binary Variables (Chp 1-2)

2.1 Analysis with binary variables I (Chp 1)

2.1.1 On one binary variable (Chp1.1)

2.1.2 On two binary variables (Chp1.2)

2.1.2A Introduction

2.1.2B Inference with two binary variables

2.1.2C Further topics

2.2 Analysis with binary response II (Chp 2)

2.1.2A On two binary variables (Chp1.2): Introudction

Often it is of interest to jointly study two binary variables, say X and Y.

- Two binary variables: X and Y
  - joint prob:
    $P(X = i, Y = j) = \pi_{ij}$ for $i = 1, 2$ and $j = 1, 2$
  - marginal prob:
    $P(X = i) = \pi_{i1} + \pi_{i2} = \pi_{i+}$ for $i = 1, 2$;
    $P(Y = j) = \pi_{1j} + \pi_{2j} = \pi_{+j}$ for $j = 1, 2$
  - conditional prob:
    $P(X = i|Y = j) = \pi_{ij}/\pi_{+j}$ for $i = 1, 2, j = 1, 2$;
    $P(Y = j|X = i) = \pi_{ij}/\pi_{i+}$ for $i = 1, 2, j = 1, 2$
  - X and Y are independent ($X \perp\!\!\!\perp Y$) iff
    - $\pi_{ij} = \pi_{i+}\pi_{+j}$, or
    - $P(X = i|Y = j) = \pi_{i+}$ for all i, j, or
    - $P(Y = j|X = i) = \pi_{+j}$ for all i, j

2.1.2A On two binary variables (Chp1.2): Introudction

Often it is of interest to jointly study two binary variables, say X and Y.

- Data: iid $(X_k, Y_k) : k = 1, \ldots, n$
  - tabulate the data by 2×2 contingency table:

	Y		
X	y=1	y=2	total
x=1	n_{11}	n_{12}	n_{1+}
x=2	n_{21}	n_{22}	n_{2+}
total	n_{+1}	n_{+2}	n

How to analyze the contingency table?

2.1.2A On two binary variables (Chp1.2): Introudction

Example. Cross-classification of belief in afterlife by gender based on data from 1998 general social survey

	Belief in Afterlife	
Gender	yes	no/undecided
female	509	116
male	398	104

Setting. a random sample of $n = 1127$ subjects were classified according to presence/absence of two characteristics, yes/no of belief in afterlife and female or male: the table presents the frequency counts in the 2×2 categories

2.1.2A On two binary variables (Chp1.2): Introudction

Example. Cross-classification of lung cancer or not by ever smoked or not based on data from an early cancer study

	Have Smoked	
Lung Cancer	yes	no
case	688	21
control	650	59

Setting. a random sample of 709 lung cancer patients, and a random sample of 709 non-lung cancer patients were respectively categorized according to ever smoked or not: the table presents the frequency counts in the 2×2 categories

2.1.2A On two binary variables (Chp1.2): Introudction

Example. Cross-classification of aspirin use and heart attack based on data from a Harvard physicians' health study

	Myocardial Infarction	
Group	yes	no
placebo	189	10,845
aspirin	104	10,933

Setting. enrolled subjects were randomized to placebo or aspirin group, and whether they had any heart attacks during the 5-year study were recorded: the table presents the frequency counts in the 2×2 categories

Types of practical studies:

- retrospective vs prospective: Examples of "belief in afterlife" and "lung cancer" vs Example of "aspirin"
- observational (e.g. cohort study, case-control study) vs experimental (e.g. clinical trial)
- cross-sectional vs longitudinal studies

Types of sampling in observational studies:

- Example of "belief in afterlife" : Subjects were a random sample from the population
- Example of "lung cancer": purposive sampling

The data in the examples are all presented using a 2×2 table.

Contingency Table

- a table with cells contain *frequency counts* of outcome according to categorical variables

2-Way Contingency Table

- a table with cells contain *frequency counts* of outcome according to 2 categorical variables

I× J Contingency Table

- a table with cells contain *frequency counts* of outcome according to 2 categorical variables, one with I levels and one with J levels

2.1.2A On two binary variables (Chp1.2): Introudction

Basic concepts related to 2×2 contingency table: Relative Risk and Odds Ratio

Given a 2×2 table,

probabilities

	Disease		
Group	Yes	Not	
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}

- **Relative Risk**

$$RR = \frac{Pr(disease\ in\ M|M)}{Pr(disease\ in\ F|F)} = \frac{\pi_{11}/\pi_{1+}}{\pi_{21}/\pi_{2+}}$$

2.1.2A On two binary variables (Chp1.2): Introudction

Given a 2×2 table,

probabilities

	Disease		
Group	Yes	Not	
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}

- **Odds Ratio** (OR)
  disease odds in Male(1st)-group/Female(2nd)-group:

$$odds_1 = \pi_{11}/\pi_{12}; \quad odds_2 = \pi_{21}/\pi_{22}$$

the odds ratio is

$$\theta = odds_1 \big/ odds_2$$

2.1.2A On two binary variables (Chp1.2): Introudction

Basic concepts related to 2×2 contingency table: Relative Risk and Odds Ratio

Remarks.

- $RR \approx \theta$ (OR) when $\pi_{11} << \pi_{12}$ and $\pi_{21} << \pi_{22}$ (for rare disease)
- $X \perp\!\!\!\perp Y \Longrightarrow \theta = 1$ and $RR = 1$

  How about "$\Longleftarrow$"?

2.1.2A On two binary variables (Chp1.2): Introudction

Basic concepts related to 2×2 contingency table: Sensitivity and Specificity

For a diagnostic test:

	Diseased (Y)		
Test Outcome (X)	true	not	Total
positve	π_{11}	π_{12}	π_{1+}
negative	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

- **sensitivity** $Pr(X = positive|Y = true) = \frac{\pi_{11}}{\pi_{+1}}$
- **specificity** $Pr(X = negative|Y = not) = \frac{\pi_{22}}{\pi_{+2}}$

two conditional probabilities

2.1.2A On two binary variables (Chp1.2): Introudction

Probability Models for 2×2 Tables

- **multinomial sampling:** e.g. Example of "belief in afterlife" with fixed $N = n$,
  $(N_{11}, N_{12}, N_{21}, N_{22}) \sim multinomial\big(n; (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})\big)$

2.1.2A On two binary variables (Chp1.2): Introudction

Probability Models for 2×2 Tables

- **binomial sampling:** e.g. Example of "lung cancer"

  Given $N_{1+} = n_{1+}$, $(N_{11}, N_{12}) \sim B(n_{1+}, p_1)$ with $p_1 = \pi_{11}/\pi_{1+}$;

  Given $N_{2+} = n_{2+}$, $(N_{21}, N_{22}) \sim B(n_{2+}, p_2)$ with $p_2 = \pi_{21}/\pi_{2+}$

2.1.2A On two binary variables (Chp1.2): Introudction

Probability Models for 2×2 Tables

- **hyper-geometric distn:** e.g. select balls from a box with black and red balls

  Given the row and column totals n_{i+} and n_{+j},

$$Pr(N_{11} = x | n_{1+}, n_{2+}, n_{+1}, n_{+2}) = \frac{\begin{pmatrix} n_{+1} \\ x \end{pmatrix} \begin{pmatrix} n_{+2} \\ n_{1+} - x \end{pmatrix}}{\begin{pmatrix} n \\ n_{1+} \end{pmatrix}}$$

2.1.2B Inference with two binary variables

Likelihood-based and others approaches with 2×2 contingency tables:

- **Estimation**
  - estm probabilities of π_{ij}, π_{i+}, π_{+j}, $p_i = \pi_{i1}/\pi_{i+}$
  - estm RR and OR
- **Hypothesis Testing**
  - about a parameter: e.g. $p_1 - p_2$
  - about independence

2.1.2B Inference with two binary variables

Likelihood-based and others approaches with 2×2 **contingency tables:**

- **Estimation**
  - **estm probabilities** of π_{ij}, π_{i+}, π_{+j}, $p_i = \pi_{i1}/\pi_{i+}$
  - **estm RR and OR**
- **Hypothesis Testing**
  - *about a parameter: e.g.* $p_1 - p_2$
  - **about independence**

2.1.2B Inference with two binary variables: Estimating Probabilities

- **To estm** π_{ij} with data from cross-sectional studies by multinomial sampling: (e.g. Example of "belief in afterlife")

Given the grand total n, $(N_{11}, N_{12}, N_{21}, N_{22}) \sim multinomial(n, \pi_{ij}'s)$

	AfterLife		
Group	Y	N	total
F	n_{11}	n_{12}	n_{1+}
M	n_{21}	n_{22}	n_{2+}
total	n_{+1}	n_{+2}	n

- the likelihood function (with constraint $\sum \pi_{ij} = 1$):

$$L(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22} | data) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}} \propto \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}}$$

2.1.2B Inference with two binary variables: Estimating Probabilities

$\Longrightarrow$ the MLE $\hat{\pi}_{11} = n_{11}/n$, $\hat{\pi}_{12} = n_{12}/n$, $\hat{\pi}_{21} = n_{21}/n$, $\hat{\pi}_{22} = n_{22}/n$

Plus, $\hat{\pi}_{1+} = \hat{\pi}_{11} + \hat{\pi}_{12} = n_{1+}/n$, $\hat{\pi}_{2+} = \hat{\pi}_{21} + \hat{\pi}_{22} = n_{2+}/n$,
$\hat{\pi}_{+1} = \hat{\pi}_{11} + \hat{\pi}_{21} = n_{+1}/n$, $\hat{\pi}_{+2} = \hat{\pi}_{12} + \hat{\pi}_{22} = n_{+2}/n$.

and $\hat{p}_1 = \hat{\pi}_{11}/\hat{\pi}_{1+} = n_{11}/n_{1+}$, $\hat{p}_2 = \hat{\pi}_{21}/\hat{\pi}_{2+} = n_{21}/n_{2+}$,

the same as the corresponding sample proportions!

$\Longrightarrow$ confidence intervals: Wald-type, score-based, LRT-based with large sample

e.g. Wald type: $\hat{\pi}_{11} \pm (1.96)\sqrt{\frac{\hat{\pi}_{11}[1-\hat{\pi}_{11}]}{n}}$

What will we study next class?

1. *Introduction and Preparation*

2. **Analysis with Binary Variables (Chp 1-2)**

- **2.1 Analysis with binary variables I (Chp 1)**
  - *2.1.1 On one binary variable (Chp1.1)*
  - **2.1.2 On two binary variables (Chp1.2)**
    - *2.1.2A Introduction*
    - **2.1.2B Inference with two binary variables**
    - **2.1.2C Further topics**
- *2.2 Analysis with binary response II (Chp 2)*