What to do today (01/18)?

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

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

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

- **Relative Risk**

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

- **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

$$\theta = odds_1 / odds_2$$

How to make inference on RR/OR with the 2 by 2 contingency data?

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

Example of "belief in afterlife" cont'd

	Belief in Afterlife		
Gender	yes	no/undecided	n_{i+}
female	509	116	625
male	398	104	502
n_{+j}	907	220	$n_{++} = 1127$

2.1.2B Inference with two binary variables: Estimating Probabilities

- **To estm** $p_1 = \pi_{11}/\pi_{1+}, p_2 = \pi_{21}/\pi_{2+}$ with data from case-control studies by binomial sampling: (e.g. Example of "lung cancer")

Given n_{1+}, n_{2+}, $N_{11} \sim B(n_{1+}, p_1)$ and $N_{21} \sim B(n_{2+}, p_2)$

	Smoked		
Lung Cancer	Y	N	Total
Y	n_{11}	n_{12}	n_{1+}
N	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- the likelihood functions:
  $L(p_1|data\ in\ line\ 1) \propto p_1^{n_{11}}(1-p_1)^{n_{1+}-n_{11}}$,
  $L(p_2|data\ in\ line\ 2) \propto p_2^{n_{21}}(1-p_2)^{n_{2+}-n_{21}}$

- **To estm** $p_1 = \pi_{11}/\pi_{1+}, p_2 = \pi_{21}/\pi_{2+}$ with data from case-control studies by binomial sampling: (e.g. Example of "lung cancer")

Given row totals n_{1+}, n_{2+}, $N_{11} \sim B(n_{1+}, p_1)$ and $N_{21} \sim B(n_{2+}, p_2)$

	Smoked		
Lung Cancer	Y	N	Total
Y	n_{11}	n_{12}	n_{1+}
N	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

$\Longrightarrow$ the MLE $\hat{p}_1 = n_{11}/n_{1+}$, and $\hat{p}_2 = n_{21}/n_{2+}$
$\Longrightarrow$ confidence intervals: Wald-type, score-based, LRT-based with large sample
e.g. Wald type CI: $\hat{p}_1 \pm 1.96\sqrt{\hat{p}_1[1-\hat{p}_1]/n_{1+}}$

Example of "lung cancer" cont'd

	Have Smoked		
Lung Cancer	yes	not	total
case	688	21	709
control	650	59	709
total	1338	80	1418

2.1.2B Inference with two binary variables: Estimating RR and OR

With data from cross-sectional studies by multinomial sampling:
(e.g. Example of "belief in afterlife")

Given $N_{++} = n$, $(N_{11}, N_{12}, N_{21}, N_{22}) \sim multinomial(n, \pi'_{ij}s)$

Recall the MLE $\hat{\pi}_{ij} = n_{ij}/n$, $i = 1, 2$ and $j = 1, 2$

$\Longrightarrow$ the MLE $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$

$\Longrightarrow$ the MLE $\hat{RR} = \frac{\hat{\pi}_{11}/\hat{\pi}_{1+}}{\hat{\pi}_{21}/\hat{\pi}_{2+}} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$

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

2.1.2B Inference with two binary variables: Estimating RR and OR

With data from case-control studies by binomial sampling: (e.g. Example of "lung cancer")

Given $N_{1+} = n_{1+}$, $N_{2+} = n_{2+}$, $N_{11} \sim B(n_{1+}, p_1)$ and $N_{21} \sim B(n_{2+}, p_2)$

Recall the MLE $\hat{p}_1 = n_{11}/n_{1+}$, and $\hat{p}_2 = n_{21}/n_{2+}$

$\Longrightarrow$ the MLE $\widehat{RR} = \frac{\widehat{\pi_{11}/\pi_{1+}}}{\widehat{\pi_{21}/\pi_{2+}}} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$,

$\Longrightarrow$ the MLE $\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$

The MLEs of RR and OR are the same as the corresponding ones with the multinomial sampling!

Recall the MLE of OR: $\hat{\theta} = \frac{n_{11} n_{22}}{n_{21} n_{12}}$, the "cross-product"

Note the following facts about OR:

- $0 \le \theta < \infty$
- $\log \hat{\theta} \sim N(\log \theta, \sigma^2)$ approximately, with $\hat{\sigma}^2 = \sum_{i,j} \frac{1}{n_{ij}}$

$\Longrightarrow$ an approximate $(1-\alpha)$ CI of $\log \theta$: $\log \hat{\theta} \pm z_{\alpha/2} \hat{\sigma}$

$\Longrightarrow$ an approximate $(1-\alpha)$ CI of θ

$$\exp\{\log \hat{\theta} \pm z_{\alpha/2} \hat{\sigma}\} = \left(\hat{\theta} e^{-z_{\alpha/2}\hat{\sigma}}, \hat{\theta} e^{z_{\alpha/2}\hat{\sigma}}\right)$$

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

	Myocardial Infarction		
Group	yes	no	Total
placebo	189	10,845	11,034
aspirin	104	10,933	11,034

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)*