What to do today (Feb 20)?

1. *Introdution and Preparation*

2. *Analysis with Binary Variables (Chp1-2)*

3. **Analysis with Multicategory Variables (Chp3)**
   - **3.1 Analysis of larger contingency tables**
     - *3.1.1 Review of two-way contingency tables*
     - *3.1.2 Analysis of $I \times J$ contingency table*
     - **3.1.3 Multi-way contingency tables (supplementary)**
   - *3.2 Analysis with Multicategory Response*

3.1.3D Statistical inference: Analogy to Procedures with Two-Way Tables

Estimation.

With the multinomial sampling (with fixed overall total n)

- the MLE $\hat{\pi}_{ijk} = n_{ijk}/n$; $\hat{\mu}_{ijk} = n\hat{\pi}_{ijk} = n_{ijk}$
- the MLE $\hat{\pi}_{ij+} = n_{ij+}/n$, $\hat{\pi}_{i++} = n_{i++}/n$, etc; $\hat{\mu}_{ij+} = n\hat{\pi}_{ij+} = n_{ij+}$, etc

- with a $2 \times 2 \times K$ table
  - the MLE of the conditional OR: $\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$
  - the MLE of the marginal OR: $\hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$
  - confidence intervals? (the strategy of estimating $\log(\theta)$ first)

3.1.3D Statistical inference: Analogy to Procedures with Two-Way Tables

Hypothesis Testing.

- Regarding a parameter:
  - e.g. $H_0 : \pi_{111} = 1/2$ vs $H_1 : \pi_{111} > 1/2$

  the Wald type, score, and LRT tests

- Regarding independence: H_0 : X,Y,Z are independent ($H_0 : \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$) vs H_1 : otherwise
  with $n_{ijk} \geq 5$, $\hat{\mu}_{ijk} = n\left(\frac{N_{i++}}{n}\right)\left(\frac{N_{+j+}}{n}\right)\left(\frac{N_{++k}}{n}\right)$ under H_0, and $df = (I-1)(J-1)(K-1)$
  - the Pearson's χ^2-test:
    $\mathcal{X}^2 = \sum_{i,j,k} \frac{(N_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}} \sim \chi^2(df)$ *approximately*
  - the LRT-test:
    $G^2 = 2\sum_{i,j,k} N_{ijk} \log\left(\frac{N_{ijk}}{\hat{\mu}_{ijk}}\right) \sim \chi^2(df)$ *approximately*

3.1.3D Statistical inference: Procedures New to the Ones with Two-Way Tables

Cohran-Mantel-Haenszel Test. with a $2 \times 2 \times K$ table, to test $X \perp Y | Z$ (R: mantelhaen.test(...))
H_0 : "$\theta_{XY(k)} = 1$ for all $k = 1, \ldots, K$" vs H_1 : otherwise

$$CMH = \frac{\left[\sum_k \left\{ N_{11k} - E_{H_0}(N_{11k}) \right\} \right]^2}{\sum_k Var(N_{11k})} \sim \chi^2(1)$$

approximately under H_0 when $n >> 1$.

3.1.3D Statistical inference: Procedures New to the Ones with Two-Way Tables

Mantel-Haenszel Estimator. with a $2 \times 2 \times K$ table, when $\theta_{XY(1)} = \ldots = \theta_{XY(K)}$, to estimate the common conditional odds ratio

The Mantel-Haenszel estimator is

$$\hat{\theta}_{XY,MH} = \frac{\sum_k N_{11k} N_{22k} / N_{++k}}{\sum_k N_{12k} N_{21k} / N_{++k}}$$

3.1.3D Statistical inference: Procedures New to the Ones with Two-Way Tables

Breslow-Day Test. with a $2 \times 2 \times K$ table, to test for homogeneity of conditional odds ratios (R: breslowday.test(...))
$H_0 : \theta_{XY(1)} = \ldots = \theta_{XY(K)}$ vs H_1 : otherwise
With $\hat{\mu}_{ijk}$, the MLE of $\mu_{ijk} = E_{H_0}(N_{ijk})$,

$$BD = \sum_{i,j,k} \frac{(N_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}} \sim \chi^2(K-1) \ \ \textit{approximately}$$

Example. Chinese Smoking vs Lung Cancer Study (meta analysis): a $2 \times 2 \times 8$ contingency table (Agresti 2006)

		Lung Cancer		
City	Smoking	Yes	No	Odds Ratio
Beijing	Smokers	126	100	2.20
	Nonsmokers	35	61	
Shanghai	Smokers	908	688	2.14
	Nonsmokers	497	807	
Shenyang	Smokers	913	747	2.18
	Nonsmokers	336	598	
Nanjing	Smokers	235	172	2.85
	Nonsmokers	58	121	
Harbin	Smokers	302	308	2.32
	Nonsmokers	121	215	
Zhengzhou	Smokers	182	156	1.59
	Nonsmokers	72	98	
Taiyuan	Smokers	60	99	2.37
	Nonsmokers	11	43	
Nanchang	Smokers	104	89	2.00
	Nonsmokers	21	36	

An analysis that combines information from several studies is called a *meta analysis*: it usually provides stronger evidence of an association than any single partial table.

```
DATAex3.3< −as.table(array(
c(126 , 100, 35 , 61, 908 , 688, 497 , 807, 913 , 747,
336 , 598, 235 , 172, 58 , 121, 302 , 308, 121 , 215,
182 , 156, 72 , 98, 60 , 99, 11 , 43, 104 , 89, 21 , 36 ),
dim = c(2, 2, 8),
dimnames = list(”Lung Cancer” = c(”yes”, ”no”),
Smoker = c(”yes”, ”no”),
City = c(”Beijing”, ”Shanghai”, ”Shenyang”,”Nanjing”,
”Harbin”,”Zhengzhou”,”Taiyuan”,”Nanchang”))))

mantelhaen.test(DATAex3.3)

breslowday.test(DATAex3.3)
```

```
> mantelhaen.test(DATAex3.3)
Mantel-Haenszel chi-squared test with continuity correction
data: DATAex3.3
Mantel-Haenszel X-squared = 254.8175, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
1.919668 2.306974
sample estimates:
common odds ratio

2.10443
> breslowday.test(DATAex3.3)
```

	Beijing	Shanghai	Shenyang	Nanjing	Harbin
log OR	0.78663752	0.762189183	0.777150289	1.04743857	0.5551747
Weight	0.06191318	0.005791079	0.007044785	0.03366529	0.0199274
	Zhengzhou	Taiyuan	Nanchang		
log OR	0.46245204	0.8625296	0.69475103		
Weight	0.03682993	0.1310932	0.09542161		
	Common OR	Stat	df	p-value	
	2.1044297	6.9674152	7.0000000	0.4322805	

- 1. CMH test:
  $CMH_{obs} = 254.82$ with df $=1 \Longrightarrow p < .001$
  extremely strong evidence against conditional independence.
  study with large sample size n=8419

- 2. Estimate of the common odds ratio:

$$\hat{\theta}_{XY,MH} = 2.10.$$

- 3. Breslow-Day test:
  $BD_{obs} = 6.97$ with df$=7 \Longrightarrow p = .43$
  not contradict to the hypothesis of equal odds ratios

3.1.3D Statistical inference: Regression Analysis

What if Y of the three categorical variables X, Y, Z with the three-way contingency table is the response?

Example. AZT Use and AIDS (NY Times, 1991): a clinical trial with n=338 HIV infected subjects

Development of AIDS by AZT Use and Race

		AIDS Symptoms	
Race	AZT Use	yes	no
white	yes	14	93
	no	32	81
black	yes	11	52
	no	12	43

- binary response Y: AIDS developed or not
- two factors X= AZT Use: received immediately or not, and Z=Race: white or black

3.1.3D Statistical inference: Regression Analysis

Example. AZT Use and AIDS (NY Times, 1991): a clinical trial with n=338 HIV infected subjects

Development of AIDS by AZT Use and Race

		AIDS Symptoms	
Race	AZT Use	yes	no
white	yes	14	93
	no	32	81
black	yes	11	52
	no	12	43

- binary response Y: AIDS developed or not
- two factors $X=$ AZT Use: received immediately or not, and Z=Race: white or black
- multiple logistic model: (i) $logit\left[\pi(x,z)\right] = \alpha + \beta_1 X + \beta_2 Z$; (ii) $logit\left[\pi(x,z)\right] = \alpha + \beta_1 X + \beta_2 Z + \beta_{12} XZ$

3.1.3E General multi-way tables

Recall what we've considered: one-way to three-way contingency tables … …

What do we do with one-way contingency tables? e.g. Chp 1

- One categorical variable X's observed frequencies $\{n_i : i = 1, \ldots, I\}$:
  the distn of X $P(X = i) = \pi_i$; the expected frequencies $\mu_i = E(N_i)$
- Inference with a one-way contingency table:
  - estm/test on $\pi_i = P(X = i)$ and then $\mu_i = E(N_i)$

    the three likelihood based procedures

3.1.3E General multi-way tables

What do we do with two-way contingency tables? e.g. Chp 2

- Two categorical variables X, Y's observed frequencies $\{n_{ij} : i = 1, \ldots, I; j = 1, \ldots, J\}$:
  the joint distn of (X, Y) $\pi_{ij} = P(X = i, Y = j)$; the expected frequencies $\mu_{ij} = E(N_{ij})$

- Inference with a two-way contingency table:
  - estm/test on
    - joint prob $\pi_{ij} = P(X = i, Y = j)$
    - marginal prob $\pi_{i+} = P(X = i)$ and $\pi_{+j} = P(Y = j)$
    - conditional prob $\pi_{j|i} = P(Y = j|X = i)$ and $\pi_{i|j} = P(X = i|Y = j)$
    - the OR, RR, etc with 2×2 tables

    the three likelihood based procedures
  - test on independence $X \perp\!\!\!\perp Y$; with 2×2 tables, test on OR=1
    the Pearson's χ^2-test, the LRT-test
  - regression analysis: e.g. Y's the response and X's the predictor

3.1.3E General multi-way tables

What do we do with three-way contingency tables? Chp 3

- Three categorical variables X, Y, Z's observed frequencies $\{n_{ijk} : i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\}$:
  the joint distn of (X, Y, Z) $\pi_{ijk} = P(X = i, Y = j, Z = k)$; the expected frequencies $\mu_{ijk} = E(N_{ijk})$
- Inference with a three-way contingency table:
  - estm/test on
    - joint prob π_{ijk}; marginal prob π_{i++}, etc; $\pi_{ij+} = P(X = i, Y = j)$, etc; conditional prob $P(X = i, Y = j|Z = k)$,etc; $P(X = i|Y = j, Z = k)$, etc
    - with $2 \times 2 \times K$ tables, the conditional OR $\theta_{XY(k)}$, the marginal OR θ_{XY+}: the Simpson's paradox

  the three likelihood based procedures

3.1.3E General multi-way tables

- Inference with a three-way contingency table: (cont'd)
  - test on independence of X, Y, Z;
    the Pearson's χ^2-test, the LRT-test
  - with $2 \times 2 \times K$ tables,
    - test on conditional indep $\theta_{XY(k)} \equiv 1$ for all $k = 1, \ldots, K$ *the Cohran-Mantel-Haenszel test*
    - estm the common conditional OR *the Mantel-Haenszel estm* $\hat{\theta}_{XY,MH}$
    - test on homogeneous conditional associations $\theta_{XY(k)} \equiv$ *const* for all $k = 1, \ldots, K$ *the Breslow-Day test*
  - regression analysis: Y's the response and X, Z are the predictors

3.1.3E General multi-way tables

How about four-way contingency tables?

- What is a four-way contingency table?

  Four categorical variables X, Y, Z, W's observed frequencies $\{n_{ijkl} : i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K; l = 1, \ldots, L\}$:
  the joint distn of (X, Y, Z, W)
  $\pi_{ijkl} = P(X = i, Y = j, Z = k, W = l)$, and the expected frequencies $\mu_{ijkl} = E(N_{ijkl})$

- Inference with a four-way contingency table:

  - estm/test on
    - joint prob π_{ijkl}; marginal prob $\pi_{i++} = P(X = i)$, etc; the conditional prob $P(X = i | Y = j, Z = k, W = l)$, etc;
    - with $2 \times 2 \times K \times L$ tables, the conditional OR $\theta_{XY(k,l)}$; the marginal OR θ_{XY++}, $\theta_{XY(k)+}$

  the three likelihood based procedures

3.1.3E General multi-way tables

- Inference with a four-way contingency table: (cont'd)
  - test on independence of (X, Y, Z, W), and conditional independence $X \perp\!\!\!\perp Y \perp\!\!\!\perp Z|W$;
    the Pearson's χ^2-test, the LRT-test
  - with $2 \times 2 \times K \times L$ tables,
    - test on conditional independence $\theta_{XY(k,l)} \equiv 1$ for all k and l – *the Cohran-Mantel-Haenszel test*
    - estm the common conditional OR – *the Mantel-Haenszel estimator* $\hat{\theta}_{XY,MH}$
    - test on homogeneous conditional associations $\theta_{XY(k,l)} \equiv$ *constant* for all k and l – *the Breslow-Day test*
  - regression analysis: Y's the response and X, Z, W are the predictors

3.1.3E General multi-way tables

How about G-way contingency tables, $G =$**5, or 6, ...?**

What will we study next?

1. *Introduction and Preparation*

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

3. **Analysis with Multicategory Variables (Chp 3)**

   - *3.1 Revisit to Analysis with Contingency Table*

   - **3.2 Analysis with Multicategory Response**
     - **3.2.1 Multicategory logit models: nominal response**
     - **3.2.2 Multicategory logit models: ordinal response**
     - **3.2.3 Additional regression models**