What to do today (Mar 22)?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analysis with Count Variables (Chp 4)*
5. *Model Selection and Evaluation (Chp 5)*

6. **Additional Topics (Chp 6)**
   - **6.1 Exact inference (Chp 6.2)**
   - **6.2 Revisit to Loglinear and Logistic Models for Contingency Tables: the Loglinear-Logit Connection** (Supplementary)
   - *6.3 Revisit II to GLM and Some Advanced Topics (Chp 5.3, Chp 6.5)*

6.1A Exact Inference: Introduction

Recall the discussion about estm the prob of success with small sample in **Chp 1**

- the "plus-4" approach;
- the exact method for constructing CI

Exact Confidence interval (CI) for π (Clopper-Pearson CI) with confidence level $1-\alpha$:

- By the *exact* distribution of $W \sim B(n, \pi)$, with observation w,

$$\{\pi : P(W \leq w) > \alpha/2 \text{ and } P(W \geq w) > \alpha/2\}$$

- By the relationship between the cumulative binomial distribution and the beta distribution, the CI is

$$B(\alpha/2; w, n-w+1) < \pi < B(1-\alpha/2; w+1, n-w)$$

conservative but applicable ... What other exact inference procedures?

6.1B Exact Inference: Fisher's Exact Test

Fisher's Tea Tasting Experiment ("Lady Tasting Tea") (RA Fisher, 1935) Fisher designed an experiment to test if his colleague could really tell whether milk or tea was added to the cup first. She was told there were four cups of each type before starting her try, and the tea cups were presented to her in a random order.

	Guess Poured First		
Poured First	milk	tea	Total
milk	3	1	4
tea	1	3	4
Total	4	4	8

Did she really tell the difference?

$\Longrightarrow$ is there a strong evidence against that her guess is indpt of the actual order?

6.1B Fisher's Exact Test

- **formulation.**
  X=actual order with 2 levels; Y=her guess with 2 levels
  $\Longrightarrow$ to test $H_0 : X \perp\!\!\!\perp Y$ vs H_1: otherwise
  (Better with $H_0 : \theta = 1$ vs $\theta > 1$!)

- **data.** tabulated as the 2×2 contingency table
  row totals $n_{1.}, n_{2.} = 4$; column totals $n_{.1}, n_{.2} = 4$
  $\Longrightarrow$ only one not-predetermined cell count, say, N_{11} [# in (milk,milk) category]: $N_{11,obs} = n_{11} = 3$

6.1B Fisher's Exact Test

Recall the **hypergeometric distribution** ... ...

Suppose an urn has $n = a + b$ balls, a red and b blue balls. Draw randomly k balls from the urn, $M =$ the number of red balls drawn out:

$$P(M = m) = \frac{\binom{a}{m}\binom{b}{a-m}}{\binom{n}{k}}$$

	Urn		
Color	drawn out	remaining	Total
red	m	a-m	n_{1+}=a
blue	k-m	b-k+m	n_{2+}=b
Total	n_{+1}=k	n_{+2}=n-k	n

6.1B Fisher's Exact Test

- **test statistic.** Under H_0, $N_{11} \sim$ hypergeometric disn:

$$P_{H_0}(N_{11} = m) = \frac{\binom{n_{1+}}{m}\binom{n_{2+}}{n_{+1}-m}}{\binom{n}{n_{+1}}}$$

- **p-value.** $P_{H_0}(N_{11} \geq N_{11,obs}) = P(N_{11} = 3) + P(N_{11} = 4) = 0.229 + 0.014$; or, the mid-p-value$= 0.229/2 + 0.014$
- **conclusion.** Association between the actual order and the guess could not be established.

6.1B Fisher's Exact Test

- e.g. R: *"fisher.test(data, ..., conf.int=TRUE, conf.level=0.95,...)"*;
  e.g. other packages such as StatXact; *PROC FREQ* in SAS

- What if to test for independence with $I \times J$ tables?
  Under the independence assumption, the prob of having a specific set of cell counts $n_{ij} : i = 1, \ldots, I;\ j = 1, \ldots, J$ with fixed row and column totals is

$$\frac{\prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} n_{+j}!}{n! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!}$$

  $\Longrightarrow (N_{ij} : i = 1, \ldots, I;\ j = 1, \ldots, J)$ with fixed row and column follows the *multiple hypergeometric distribution*

6.1C Exact Inference: the Permutation Test for Independence

When the Pearson's chi-square test is conducted with a two-way contingency table for independence, the test statistic

$$\mathcal{X}^2 = \sum \frac{(\text{observed-fitted})^2}{\text{fitted}} \sim \chi^2((I-1)(J-1))$$

approximately when $n >> 1$ and $n_{ij} > 5$ in general. When n is small, the chi-square distn approximation is not good: what is the exact distn of $\mathcal{X}^2$? Let's see the following table.

the Fisher's experiment

M	P(M=m)	$\hat{OR}$	$\mathcal{X}^2_{obs}$
0	.0143	0	8
1	.2286	1/9	2
2	.5143	1	0
3	.2286	9	2
4	.0143	∞	8

in the Fisher's experiment

M	P(M=m)	$\hat{OR}$	$\mathcal{X}^2_{obs}$	prob
0	.0143	0	8	1 out of 70
1	.2286	1/9	2	16 out of 70
2	.5143	1	0	36 out of 70
3	.2286	9	2	16 out of 70
4	.0143	∞	8	1 out of 70

p-value=$P_{H_0}(\mathcal{X}^2 \geq \mathcal{X}^2_{obs})$

$\Longrightarrow$ a general test procedure: **permutation test** for independence – to calculate the p-value based on the *permutation distribution* of the test statistic $\mathcal{X}^2$

6.1C the Permutation Test for Independence

- It is possible to obtain the permutation distn of the test statistic directly.

- The permutation distn can be estimated by simulation.
  e.g. In Fisher's experiment,
  - (i) randomly permute the "guess" of the lady (e.g. using $sample(..., replace = FALSE)$ in R) and obtain the evaluation of $\mathcal{X}^2$;
  - (ii) repeat (i) B ($>> 1$) times and have $\mathcal{X}_b^2 : b = 1, \ldots, B$;
  - (iii) calculate $[\#\{\mathcal{X}_b^2 \geq \mathcal{X}_{obs}^2\}]/B$ and use it as an approximated p-value

- e.g. Use the R function
  $chisq.test(x, simulat.p.value = TRUE, B)$ to implement

6.2A Correspondence between Logit and Loglinear Models

In general, there are following correspondence with 3 categorical variables X, Y, Z:

- Saturated:
  Loglinear(XYZ) $\Leftrightarrow$ Y$\sim$ Logit(XZ) or Y$\sim$ Multi-Logit(XZ)
- Homogeneous Association I:
  Loglinear(XY,YZ,XZ) $\Leftrightarrow$
  Y$\sim$ Logit(X,Z) or Y$\sim$ Multi-Logit(X,Z)
  - XZ association term in loglinear model is cancelled out in the logit models
  - Logit models don't have description about relationship between predictors but only about how X,Z, and XZ affect Y.
  - Caution with *collinearity*
- Homogeneous Association II (conditional indpt of $X \perp Z|Y$):
  Loglinear(XY,YZ) $\Leftrightarrow$ Y$\sim$ Logit(X,Z) or Y$\sim$ Multi-logit(X,Z)

6.2B Example for Logit-Loglinear Connection

Example. The table below summarizes admissions to the graduate school at UC-Berkeley in 1973: it cross-classifies the admission decisions by gender of applicant and type of department. Answer the following questions based on the table.

Table. Graduate School Admission

department type (Z)	gender (X)	whether admitted (Y)				
		Male		Female		
		Yes	Not	Yes	Not	Total
A		865	520	106	27	1518
B		258	484	333	635	1710
C		75	489	118	616	1298
	Total	1198	1493	557	1278	4526

- the sample marginal odds ratio (OR) of admission between male and female applicants: 1.84
- the sample conditional odds ratio (OR) of admission between male and female applicants for departments of types A,B and C: 0.43, 1.02, and 0.80

Data with Fitted Counts

observed	gender	admt	dept	LL(X,Y,Z)	LL(XY,XZ,YZ)	LL(XYZ)
27	0	0	1	376.8	43.2	27
635	0	0	2	424.5	617.0	635
616	0	0	3	322.2	617.9	616
520	1	0	1	552.6	503.8	520
484	1	0	2	622.5	502.0	484
489	1	0	3	472.5	487.1	489
106	0	1	1	238.6	89.8	106
333	0	1	2	268.8	351.0	333
118	0	1	3	204.1	116.1	118
865	1	1	1	350.0	881.2	865
258	1	1	2	394.2	240.0	258
75	1	1	3	299.3	76.9	75

How to obtain the fitted counts?

Step 3. Logistic Regression:
$Y \sim 1$; $Y \sim X$, $Y \sim Z$; $Y \sim X + Z$; $Y \sim X * Z$

- Read in Data:
  - n=4526;
  - 1755, 2771 admitted, not;
  - 2691, 1835 male, female;
  - 1518, 1710,1298 deptA, deptB, deptC

```
>glm(admt ~ gender, family = binomial)
Call:
glm(formula = admt ~ gender, family = binomial)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83049    0.05077 -16.357   <2e-16 ***
gender1      0.61035    0.06389   9.553   <2e-16 ***
---
    Null deviance: 6044.3  on 4525  degrees of freedom
Residual deviance: 5950.9  on 4524  degrees of freedom
AIC: 5954.9

>glm(admt ~ dept, family = binomial)
Call:
glm(formula = admt ~ dept, family = binomial)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.57388    0.05346   10.73   <2e-16 ***
dept2       -1.21225    0.07378  -16.43   <2e-16 ***
dept3       -2.31879    0.09457  -24.52   <2e-16 ***
---
    Null deviance: 6044.3  on 4525  degrees of freedom
Residual deviance: 5280.6  on 4523  degrees of freedom
AIC: 5286.6
```

```
>glm(admt ~ gender + dept, family = binomial)
Call:
glm(formula = admt ~ gender + dept, family = binomial)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.73331    0.09000   8.148  3.7e-16 ***
gender1     -0.17435    0.07897  -2.208   0.0273 *
dept2       -1.29720    0.08357 -15.521  < 2e-16 ***
dept3       -2.40508    0.10289 -23.375  < 2e-16 ***
---
    Null deviance: 6044.3  on 4525  degrees of freedom
Residual deviance: 5275.7  on 4522  degrees of freedom
AIC: 5283.7
```

```
>glm(admt ~ gender * dept, family = binomial)
Call:
glm(formula = admt ~ gender * dept, family = binomial)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.3676     0.2156   6.344 2.24e-10 ***
gender1          -0.8587     0.2226  -3.858 0.000114 ***
dept2            -2.0131     0.2259  -8.910  < 2e-16 ***
dept3            -3.0202     0.2378 -12.698  < 2e-16 ***
gender1:dept2     0.8751     0.2451   3.570 0.000357 ***
gender1:dept3     0.6364     0.2739   2.323 0.020160 *
---
    Null deviance: 6044.3  on 4525  degrees of freedom
Residual deviance: 5261.7  on 4520  degrees of freedom
AIC: 5273.7
```

Step 4. Loglinear-Logit Connection: e.g. Loglinear(XY,XZ,YZ)

• Loglinear model of homogeneous association in 3-way tables:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

The corresponding logit model: Admit (Y) as the response and X,Z explanatory variables

$$\mathit{logit}(\pi_{ik}) = \alpha + \beta_i^X + \beta_k^Z$$

Reporting the analysis outcomes ...

- $\beta_i^X = \lambda_{i1}^{XY} - \lambda_{i0}^{XY} = 0, -0.17435$ for $i = 0, 1$;
  $\beta_l^Z = \lambda_{1l}^{YZ} - \lambda_{0l}^{YZ} = 0, -1.29720, -2.40508$ for $l = 1, 2, 3$;
  no term of β's associated with λ_{ik}^{XZ}
- log OR of admission for males and females in deptA, deptB, deptC:
  $\beta_1^X - \beta_0^X = -0.17435$ and $\lambda_{11}^{XY} + \lambda_{00}^{XY} - [\lambda_{01}^{XY} + \lambda_{10}^{XY}] = -0.17435$

What will we study next?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analsyis with Count Response (Chp 4)*
5. *Model Selection and Evaluation (Chp 5)*
6. **Additional Topics (Chp 6)**
   - *6.1 Exact Inference (Chp 6.2)*
   - *6.2 Revisit to Loglinear and Logistic Models for Contingency Tables: the Loglinear-Logit Connection*
   - **6.3 Revisit III to GLM and Advanced Topics (Chp 5.3, Chp 6.5)**
     - **6.3.1 Revisit III to GLM**
     - **6.3.2 Marginal Modeling: Quasi-Score, Generalized Estimating Equation (GEE)**
     - *6.3.3 Mixed Effect Models for Correlated Data*