What to do today (01/23)?

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)

2.1.2B 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: Hypothesis testing on independence

- With a 2×2 table, it can be formulated into testing on $H_0 : OR = 1$
- How about a general approach, which works with any two-way table?

$\Longrightarrow$ **Consider two general testing procedures:**

- **Pearson's Chi-Squared (χ^2-) Test** (K. Pearson, 1900)
- **Likelihood Ratio Test (LRT)**

2.1.2B Hypothesis testing on independence: χ^2-test

In general, testing on H_0 vs H_1 with a two-way contingency table (2×2 table as a special case):

Cell Counts

	Variable Y	
Variable X	1	2
1	N_{11}	N_{12}
2	N_{21}	N_{22}

Suppose, when H_0 is true, the expected frequencies $E_{H_0}(N_{ij}) = \mu_{ij}$

Test on H_0 by comparing N_{ij} with μ_{ij}?

Pearson's Chi-Squared Test (K. Pearson, 1900) on H_0 vs H_1 with an $I \times J$ contingency table

Consider the **Pearson χ^2-statistic**:

$$\mathcal{X}^2 = \sum_{i,j} \frac{(N_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

Approximately, $\mathcal{X}^2 \sim \chi^2(df)$ under H_0 for large n.

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

Remarks:

- The χ^2-approximation is good usually when $\mu_{ij} \geq 5$
- The degrees of freedom:
  $df = \#$(parameters under H_1) - $\#$(parameters under H_0)
- What are μ_{ij}? How to implement the procedure?

2.1.2B Hypothesis testing on independence: χ^2-test

Consider specifically ...

To test on $H_0 : X \perp\!\!\!\perp Y$ vs $H_1 : X \not\perp\!\!\!\perp Y$ with an $I \times J$ contingency table by the multinomial sampling with $N_{++} = n$

- **Reformulate the hypotheses according to the sampling ...**
  $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j vs H_1 : otherwise
- **Getting μ_{ij} or their best estimates ... ...**
  $\mu_{ij} = E_{H_0}(N_{ij}) = n\pi_{i+}\pi_{+j}$
  the MLE under H_0 is $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$
- **Applying Pearson's χ^2-test ...**
  - determine $df = (IJ - 1) - ([I - 1] + [J - 1]) = (I - 1)(J - 1)$ by Fisher (1922)
  - calculate $\mathcal{X}^2_{obs} = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$
  - calculate $p - value = P_{H_0}(\mathcal{X}^2 \geq \mathcal{X}^2_{obs})$ based on $\mathcal{X}^2 \sim \chi^2(df)$ approximately when $n >> 1$
  - draw conclusion

2.1.2B Hypothesis testing on independence: LRT

Likelihood Ratio Test (LRT) on H_0 vs H_1 with a two-way contingency table

Consider the **likelihood ratio test statistic**:

$$-2\log\left(\frac{\max L_{H_0}(parameter|data)}{\max L(parameter|data)}\right) \propto G^2 = 2\sum_{i,j} N_{ij}\log\left(\frac{N_{ij}}{\mu_{ij}}\right)$$

Approximately, $G^2 \sim \chi^2(df)$ under H_0 for large n

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

- the χ^2-approximation is good usually when $\mu_{ij} \geq 5$
- the degrees of freedom
  df = #(parameters under H_1) - #(parameters under H_0)
- What are μ_{ij}? How to implement the procedure?

2.1.2B Hypothesis testing on independence: LRT

Consider specifically ...

To test on $H_0 : X \perp\!\!\!\perp Y$ vs $H_1 : X \not\perp\!\!\!\perp Y$ with an $I \times J$ contingency table by the multinomial sampling with $N_{++} = n$

- **Reformulate the hypotheses according to the sampling ...**
  $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j vs H_1 : otherwise

- **Getting μ_{ij} or their best estimates ... ...**
  $\mu_{ij} = E_{H_0}(N_{ij}) = n\pi_{i+}\pi_{+j}$: the MLE under H_0 is
  $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$

- **Applying LRT-test ...**
  - determine $df = (IJ - 1) - ([I - 1] + [J - 1]) = (I - 1)(J - 1)$
  - calculate $G^2_{obs} = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$
  - calculate $p - value = P_{H_0}(G^2 \geq G^2_{obs})$
  - draw conclusion

2.1.2B Hypothesis testing on independence

Example. Gender Gap in Political Affiliation

	Party Identification			
Gender	democrat	independent	republican	Total
female	762	327	468	1557
male	484	239	477	1200
Total	1246	566	945	2757

Data from 2000 General Social Survey

X = *gender* with $I = 2$ levels, female vs males; Y = *party* with $J = 3$ levels, democrat vs indept vs republican

To test on $H_0 : X \perp\!\!\!\perp Y$ vs $H_1 : X \not\perp\!\!\!\perp Y$ with the 2×3 contingency table by the multinomial sampling with $N_{++} = n = 2575$

- **Reformulate the hypotheses according to the sampling ...**

  $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ vs H_1 : otherwise
- **Getting μ_{ij} or their best estimates ... ...**

  $\mu_{ij} = E_{H_0}(N_{ij}) = n\pi_{i+}\pi_{+j}$, with the MLE under H_0
  $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$

Applying Pearson's χ^2-Test ...

- determine $df = (I-1)(J-1) = (1)(2)$
- calculate $\mathcal{X}^2_{obs} = \sum \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = 30.1$ with $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$
- calculate $p-value = P_{H_0}(\mathcal{X}^2 \geq 30.1) < .0001$ based on $\mathcal{X}^2 \sim \chi^2(2)$ approximately when $n >> 1$
- concluding: strong evidence against H_0 – there is a signficant association between gender and political affiliation

Applying LRT ...

- determine $df = (I-1)(J-1) = 2$
- calculate $G^2_{obs} = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) = 30.0$ with $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$
- calculate $p-value = P_{H_0}(G^2 \geq 30.0) < .0001$ based on $G^2 \sim \chi^2(2)$ approximately when $n >> 1$
- concluding: strong evidence against H_0 – there is a signficant association between gender and political affiliation

2.1.2B Hypothesis testing on independence

To test on $H_0 : X \perp\!\!\!\perp Y$ vs $H_1 : X \not\perp\!\!\!\perp Y$ with an $I \times J$ contingency table by the purposive sampling with $N_{i+} = n_{i+}$

- **Reformulate the hypotheses according to the sampling ...**
  $H_0 : p_j = \pi_{+j}$ for all i, j vs H_1 : otherwise
  $(p_j = \frac{\pi_{ij}}{\pi_{i+}} = P(Y = j|X = i))$
- **Getting μ_{ij} or their best estimates ... ...**
  $\mu_{ij} = E_{H_0}(N_{ij}) = n_{i+}p_j$
  the MLE under H_0 $\hat{\mu}_{ij} = n_{i+}\hat{p}_j = n_{i+}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$ (the same as it with multinomial sampling)

2.1.2B Hypothesis testing on independence

Applying Pearson's χ^2-Test ...

- determine $df = (IJ - I) - (J - 1) = (I - 1)(J - 1)$
- calculate $\mathcal{X}^2_{obs} = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$
- calculate $p - value = P_{H_0}(\mathcal{X}^2 \geq \mathcal{X}^2_{obs})$ based on $\mathcal{X}^2 \sim \chi^2(df)$ approximately when $n >> 1$
- conclude ...

Applying LRT ...

- determine $df = (IJ - I) - (J - 1) = (I - 1)(J - 1)$
- calculate $G^2_{obs} = 2 \sum_{i,j} n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right)$
- calculate $p - value = P_{H_0}(G^2 \geq G^2_{obs})$
- conclude ...

The same test statistics $\mathcal{X}^2$, G^2 as used with tables by the multinominal sampling.

What will we study in the next class?

1. *Introduction and Preparation*

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

   - *2.1 Analysis with binary variables (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 (Chp 2)**
     - **2.2.1 Regression models (Chp2.1, Chp2.2.1)**
     - *2.2.2 Inference with logistic regression models (Chp2.2.1-7)*
     - 2.2.3 Further topics (Chp2.2.8, Chp2.3)