Reminder

- Assignment 1 due 4pm tomorrow
Summary

Last lecture:

- The $I \times J$ table
- Started with 3-way tables

Today:

- Continue: 3-way tables; testing for conditional independence
- Partial and marginal independence
- Simpson’s paradox
Notation for Cell Frequencies and Probabilities

\[ X_{ijk} = \text{observations in cell } (i, j, k): \text{ row } i, \text{ column } j, \text{ layer } k \]

\[
\sum_{i,j,k} X_{ijk} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk} = n
\]

\[ X_{ij.} = \sum_{k} X_{ijk} = \text{number of observations in } i\text{th row and } j\text{th column}, \text{ and similarly for } X_{i.k} \text{ and } X_{.jk} \]

\[ X_{i..} = \sum_{j,k} X_{ijk} = \text{number of observations in } i\text{th row, and similarly for } X_{.j.} \text{ and } X_{..k} \]
\[ p_{ijk} = P(X_1 = i, X_2 = j, X_3 = k), \quad p_{ij.} = \sum_k P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i, X_2 = j), \text{ and similarly for } p_{i..} \text{ and } p_{..k} \]

\[ p_{i..} = \sum_{j,k} P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i), \text{ and similarly for } p_{.j.} \text{ and } p_{..k} \]

Constraints:

\[ \sum_{i,j,k} p_{ijk} = 1, \quad \sum_{i,j} p_{ij.} = \sum_{i,k} p_{i..} = \sum_{j,k} p_{..k} = 1 \]

\[ \sum_i p_{i..} = \sum_j p_{.j.} = \sum_k p_{..k} = 1 \]
Hypothesis Testing

To test a particular null hypothesis $H_0$, estimate the expected cell frequencies $E_{ijk}$ under $H_0$ and use Pearson’s $\chi^2$ statistic

$$Z^2 = \sum_{i,j,k} \frac{(X_{ijk} - E_{ijk})^2}{E_{ijk}}$$

where $X_{ijk}$ are observed cell frequencies

Distribution of $Z^2$ under $H_0$ is $\chi^2$ with degrees of freedom

df = number of cells $- 1 -$ number of parameters estimated under $H_0$
Hypothesis Testing

For a 3-way table, different types of independence are possible, placing different levels of restriction on the expected cell frequencies “Hierarchy” of independence (strongest to weakest):

1. **Mutual independence** of $X_1$, $X_2$ and $X_3$

2. **Joint independence** of two variables from the third variable

3. **Conditional independence** of two variables given the third variable

**Mutual independence** of $X_1$, $X_2$ and $X_3$ $\Rightarrow$ **Joint independence** of $X_1$ and $X_2$ from $X_3$ $\Rightarrow$ **Conditional independence** of $X_1$ and $X_2$ given $X_3$
Model 1: Mutual Independence of $X_1$, $X_2$, $X_3$

\[
P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i)P(X_2 = j)P(X_3 = k)
\]
\[
p_{ijk} = p_{i..}p_{..j}p_{..k} \text{ for all } i, j, k
\]

Expected $(i, j, k)$ cell frequency = $np_{ijk}$

Under $H_0$: $np_{i..}p_{..j}p_{..k}$

MLE’s:
\[
\hat{p}_{i..} = \frac{X_{i..}}{n}, \hat{p}_{..j} = \frac{X_{..j}}{n}, \hat{p}_{..k} = \frac{X_{..k}}{n}
\]
Fitted \((i, j, k)\) cell frequency:

\[ E_{ijk} = np_{i..}p_{..j}p_{...k} = \frac{X_{i..}X_{..j}X_{...k}}{n^2} \]

Degrees of freedom for \(Z^2\): number of cells \(-1\) – number of parameters estimated under \(H_0\)

\[ = IJK - 1 - (I - 1) - (J - 1) - (K - 1) = IJK - I - J - K + 2: \]

\(I p_{i..}'s, J p_{..j}'s, K p_{...k}'s\) less 1 each for the constraints

\(\sum_i p_{i..} = 1, \sum_j p_{..j} = 1, \sum_k p_{...k} = 1\)
Model 2: Joint Independence of $X_1$, $X_2$ from $X_3$

$$P(X_1 = i, X_2 = j, X_3 = k) = P(X_1 = i, X_2 = j)P(X_3 = k)$$

$$p_{ijk} = p_{ij} \cdot p_{\cdot \cdot k} \text{ for all } i, j, k$$

Expected $(i, j, k)$ cell frequency = $n p_{ijk}$

Under $H_0$: $n p_{ij} \cdot p_{\cdot \cdot k}$

MLE’s:

$$\hat{p}_{ij} = \frac{X_{ij}}{n}, \hat{p}_{\cdot \cdot k} = \frac{X_{\cdot \cdot k}}{n}$$
Fitted \((i, j, k)\) cell frequency:

\[
E_{ijk} = n \hat{p}_{ij} \hat{p}_{..k} = \frac{X_{ij} X_{..k}}{n}
\]

Degrees of freedom for \(Z^2\): number of cells \(-1\) − number of parameters estimated under \(H_0\)

\[
= IJK - 1 - (IJ - 1) - (K - 1) = IJK - IJ - K + 1: IJ \ p_{ij}.'s, \ K \ p_{..k}.'s
\]

less 1 each for the constraints

\[
\sum_{i,j} p_{ij} = 1, \sum_k p_{..k} = 1
\]
Model 3: Conditional Independence of $X_1, X_2$ given $X_3$

\[
P(X_1 = i, X_2 = j | X_3 = k) = P(X_1 = i | X_3 = k) P(X_2 = j | X_3 = k)
\]

\[
\frac{p_{ijk}}{p_{..k}} = \frac{p_{i.k}}{p_{..k}} \times \frac{p_{.jk}}{p_{..k}}
\]

\[
p_{ijk} = \frac{p_{i.k}p_{.jk}}{p_{..k}} \quad \text{for all } i, j, k
\]

Expected $(i, j, k)$ cell frequency = $np_{ijk}$; under $H_0$: $np_{i.k}p_{.jk}/p_{..k}$ MLE’s:

\[
\hat{p}_{i.k} = \frac{X_{i.k}}{n}, \hat{p}_{.jk} = \frac{X_{.jk}}{n}, \hat{p}_{..k} = \frac{X_{..k}}{n}
\]
Fitted frequency:

\[ E_{ijk} = n \times \frac{X_{i,k}}{n} \times \frac{X_{.jk}}{n} = \frac{X_{i,k}X_{.jk}}{X_{..k}} \]

DoF for \( Z^2 \) = number of cells \( -1 \)– number of parameters estimated under \( H_0 = IJK - 1 - (IK - K) - (JK - K) - (K - 1) = IJK - IK - JK + K \):

• \( K \ p_{..k} \)'s less 1 for the constraint \( \sum_k p_{..k} = 1 \)

• \( IK \ p_{i,k} \)'s less \( K \) for the constraints \( \sum_i p_{i,k} = p_{..k} \) for \( k = 1, \ldots, K \)

• \( JK \ p_{.jk} \)'s less \( K \) for the constraints \( \sum_j p_{.jk} = p_{..k} \) for \( k = 1, \ldots, K \)
Testing for Conditional Independence of $X_1, X_2$ given $X_3$

Idea: Separate tests for each of the $K I \times J$ tables:

- Pearson’s $\chi^2$ separately for each table
- Calculate $\hat{OR}$ for each table and test $H_0 : OR = 1$ directly using the test statistic $\frac{\log \hat{OR}}{SE(\log \hat{OR})}$ (null distribution $N(0, 1)$)
- Calculate 95% confidence interval for $\log OR \rightarrow 95\%$ confidence interval for $OR$, and reject $H_0 : OR = 1$ if 1 is not in the interval
Example: Asbestos and lung cancer, controlling for smoking. Here we should test both hypotheses:

$H_0$: Asbestos exposure and Cancer are independent for Smokers.

$H'_0$: Asbestos exposure and Cancer are independent for Nonsmokers.

Only if both $H_0$ and $H'_0$ are accepted can we claim that Asbestos exposure and Lung cancer are conditionally independent given Smoking (that is Model 3 holds). This turns out not to be the case (CI based on the estimate of 1.61 excludes the value 1).
Partial and Marginal Tables

Partial tables: the $K$ two-way $I \times J$ tables for $X_1, X_2$ at each level of $X_3$

Partial tables control the effect of $X_3$ by holding its value constant

Marginal table: single two-way table obtained by summing the partial tables

The marginal table, rather than controlling for $X_3$, ignores it
Conditional versus Marginal Independence

Conditional associations are derived from the partial tables: association between $X_1$ and $X_2$ conditional on fixed value for $X_3$. Marginal association is derived from the marginal table: association between $X_1$ and $X_2$ ignoring $X_3$. Conditional and marginal associations can be quite different.
Simpson’s Paradox

An extreme example of confounding, where the direction of association is reversed in the marginal table compared to the partial tables.

Example: Death penalty verdict by defendant’s race and victim’s race.
Percent receiving death penalty by defendant's race

- White victims
  - 11.3%
- Black victims
  - 22.9%
- All victims
  - 0%
  - 2.8%
  - 11%
  - 7.9%
Another Example of Simpson’s Paradox

Survival following surgery at 2 hospitals

Confounder is patient condition
Mortality rate % by hospital

<table>
<thead>
<tr>
<th>Condition</th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good condition</td>
<td>1</td>
<td>1.33</td>
</tr>
<tr>
<td>Poor condition</td>
<td>4</td>
<td>3.8</td>
</tr>
<tr>
<td>All patients</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Mortality rate % by patient condition

Percent in good condition by hospital

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Percent in Good Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.6</td>
</tr>
<tr>
<td>B</td>
<td>75</td>
</tr>
</tbody>
</table>