Analysis of Categorical Data

Two-Way Contingency Table
### Contingency Table

Table 1: $I \times J$ Table

\[
\begin{array}{cccc}
 & 1 & 2 & \ldots & J \\
1 & \pi_{11} & \pi_{12} & \ldots & \pi_{1J} & \pi_{1+} \\
2 & \pi_{21} & \pi_{22} & \ldots & \pi_{2J} & \pi_{2+} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
I & \pi_{I1} & \pi_{I2} & \ldots & \pi_{IJ} & \pi_{I+} \\
\end{array}
\]

- $\pi_{i+} = \sum_{j=1}^{J} \pi_{ij} =$ row $i$ marginal prob.
- $\sum_{i=1}^{I} \pi_{i+} = 1$

- $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij} =$ colimn $j$ marginal prob.
- $\sum_{j=1}^{J} \pi_{+j} = 1$
Contingency Table

Table 2: Observed Counts

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$I$</th>
<th>$1+$</th>
<th>$2+$</th>
<th>...</th>
<th>$I+$</th>
<th>$++$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>...</td>
<td>$n_{1J}$</td>
<td>$n_{1+}$</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>...</td>
<td>$n_{2J}$</td>
</tr>
<tr>
<td>Explanatory Variable</td>
<td>$n_{J1}$</td>
<td>$n_{J2}$</td>
<td>...</td>
<td>$n_{JJ}$</td>
<td>$n_{J+}$</td>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
<td>...</td>
<td>$n_{+J}$</td>
</tr>
</tbody>
</table>

$n_{ij} = \#$ observed in $i,j$ cell

$n = \#$ total sample size
Basic Sampling Distributions

- Binomial: each row defines different groups and the sample size \((n_{1+}, n_{2+})\) are fixed by design. Need conditional distribution.

- Multinomial: When the total sample size is fixed and the response has \(k\) categories.

- Poison: Used for counts of events that occur randomly over time or space, when outcomes in disjoint periods are independent.
Analysis of the Table

■ Sample Proportions-

■ Conditional Probabilities

■ Under Independent Assumptions
Example 1: Cancer vs. Dose
Popular Measures of Association

- Difference in Proportions
- Relative Risk
- Odds Ratio
## Notation for $2 \times 2$ Tables

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Risk Group 1</th>
<th>Risk Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
</tr>
<tr>
<td></td>
<td>$1 - \pi_1$</td>
<td>$1 - \pi_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
<th>Response Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Success</td>
</tr>
<tr>
<td>Explanatory Variable</td>
<td>$n_{11}$</td>
</tr>
<tr>
<td></td>
<td>$n_{21}$</td>
</tr>
<tr>
<td></td>
<td>$n_{+1}$</td>
</tr>
</tbody>
</table>
Difference in Proportions

- Want to make inference about $\pi_1 - \pi_2$
- Assumptions

- Estimation:
  - Properties of estimators
    - Mean
    - Variance

- Confidence interval
Example 1 (Cont’d)
Relative Risk

- Define Relative Risk:
- Possible Values:
- Estimation:
- Variance:

- Confidence Interval:

- Side Comment:
Example 1 (Cont’d)
Odds Ratio

- Odds and Odds Ratio $\theta$:
- Properties of $\theta$

- Estimation
- Variance

- Confidence Interval
Example 1 (Cont’d)
Relationship Between R and $\theta$

odds ratio =

$$\theta = \frac{\pi_1(1 - \pi_2)}{\pi_2(1 - \pi_1)} = \left(\frac{\pi_1}{\pi_2}\right)\left(\frac{1 - \pi_2}{1 - \pi_1}\right) \approx \frac{\pi_1}{\pi_2}$$

= Relative Risk

The approximation is good if both $\pi_1$ and $\pi_2$ are small.
Chi-square Test for Independence

- Expected cell counts assuming no association
- Pearson’s Chi-square statistics
- Yates’ corrected chi-square
Example 2: Spouses’ Heights
Fisher’s Exact Test

- Useful for small samples
- Condition on both sets of marginal values
- Use Hypergeometric Distribution
  - Under independence, probability for the observed data:
  
  - Margin probability of the columns:
  
  - Conditional distribution of observed data given the margin:
Example 3: Non-Smoking Males