Analysis of Categorical Data

Two-Way Contingency Table

Contingency Table

Table 1: $I \times J$ Table

		1	2	...	J	
	1	π_{11}	π_{12}	$\ldots$	π_{1J}	π_{1+}
rows	2	π_{21}	π_{22}	$\ldots$	π_{2J}	π_{2+}
	...	$\ldots$			$\ldots$	$\ldots$
	I	π_{I1}	π_{I2}	$\ldots$	π_{IJ}	π_{I+}
		π_{+1}	π_{+2}	$\ldots$	π_{+J}	

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

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

Contingency Table

Table 2: Observed Counts

		Response Variable				
	1	n_{11}	n_{12}	$\ldots$	n_{1J}	n_{1+}
Explanatory	2	n_{21}	n_{22}	$\ldots$	n_{2J}	n_{2+}
Variable	$\ldots$	$\ldots$			$\ldots$	$\ldots$
	I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	n_{I+}
		n_{+1}	n_{+2}	$\ldots$	n_{+J}	$n_{++} = n$

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

$$n = \text{total sample size}$$

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 occure 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 2×2 Tables

Proportion

		Response Variable	
		Success	Failure
Explanatory	Risk Group 1	π_1	$1 - \pi_1$
Variable	Risk Group 2	π_2	$1 - \pi_2$

Data

		Response Variable		
		Success	Failure	
Explanatory	Risk Group 1	n_{11}	n_{12}	n_{1+}
Variable	Risk Group 2	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

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 θ:
- Properties of θ
- Estimation
- Variance
- Confidence Interval

Example 1 (Cont'd)

Relationship Between R and θ

odds ratio =

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

= Relative Risk

The approximation is good if both π_1 and π_2 are small.

Chi-square Test for Independence

- Expected cell counts assuming no association

- Pearson's Chi-sqaure statsitcs

- 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