Contingency Tables Introduction

Demetris Athienitis



Section 1

Introduction

2 Notation

Independence

Key Points

Interested in understanding various relationships and probabilities between two categorical random variables.

- Let X and Y be categorical random variables
- X has I categories and Y has J categories
- Display the IJ possible combinations of outcomes in a rectangular table having I rows for the categories of X and J columns for the categories of Y

Definition (Contingency table)

A table that displays the possible combinations of outcomes in a rectangular (array) table in which the cells contain frequency counts of outcomes.

Key Points

Example (Physicians' Health Study)

A study on Myocardial Infraction (MI) and treatment. We consider

- Y = heart attack: yes/no, response variable
- X = group: placebo/aspirin, explanatory variable

Group	MI	
	Yes	No
Placebo	189	10845
Aspirin	104	10933

Is aspirin use correlated with a reduction in heart attacks?

Section 2

Introduction

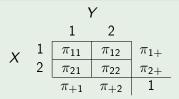
2 Notation

Independence

Notation

- $\pi_{ij} = P(X=i,Y=j)
 ightarrow \{\pi_{ij}\}$ form the joint distribution of X and Y
- $\pi_{i+} = \sum_{i=1}^J \pi_{ij} = P(X=i) \to \{\pi_{i+}\}$ marginal distribution of X
- $\pi_{+j} = \sum_{i=1}^{J} \pi_{ij} = P(Y = j) \rightarrow \{\pi_{+j}\}$ marginal distribution of Y

Example



Notation

- Similarly, let $\{n_{ij}\}, \{n_{i+}\}, \{n_{+j}\}$ denote the cell counts, row and column totals respectively.
- Let

$$p_{ij} = \frac{n_{ij}}{n}, \quad p_{i+} = \frac{n_{i+}}{n}, \quad p_{+j} = \frac{n_{+j}}{n}$$

Probability distribution consisting of conditional probabilities for Y
given the level of X is called a conditional distribution

$$\pi_{j|i} = rac{\pi_{ij}}{\pi_{i+}}$$
 estimated by $p_{j|i} = rac{n_{ij}}{n_{i+}}$

Sensitivity and Specificity

For many diseases there are tests to detect the disease but such tests are not foolproof. A 2×2 contingency table helps explore the effectiveness of the test.

- $Y = \text{outcome of the test with } \begin{cases} 1 & \text{positive} \\ 2 & \text{negative} \end{cases}$
- $X = \text{actual condition with } \begin{cases} 1 & \text{diseased} \\ 2 & \text{not diseased} \end{cases}$

The following two terms are important

- Sensitivity: P(Y = 1|X = 1) (True positive)
- Specificity: P(Y = 2|X = 2) (True negative)

Section 3

Introduction

2 Notation

Independence

Independence

Definition (Independence)

Variables X and Y are statistically independent if the true conditional distribution of Y is the same at each level of X.

That is

$$\pi_{j|i} = \pi_{j|i'} \quad \forall i, i'$$

Lemma

X and Y are independent if and only if

$$\pi_{ij} = \pi_{i+}\pi_{+i} \quad \forall i,j$$

Example

for illustration Here is a 2 x 2 table where independence holds

$$\begin{array}{c|ccccc}
 & Y & & & & \\
 & 1 & 2 & & & \\
 & .42 & .28 & .7 & & \\
 & .18 & .12 & .3 & & \\
 & .6 & .4 & 1 & & \\
\end{array}$$

All joint probabilities are products of their respective marginal probabilities (0.28 $=0.7\times0.4,\mbox{ etc.})$

We learned

- What are contingency tables
- Independence of two categorical variables