Lecture 3: Contingency Tables

Author: Nick Reich

Course: Categorical Data Analysis (BIOSTATS 743)

Contingency tables of counts

Let X and Y be categorical variables with I and J categories, respectively.

IxJ contingency tables of counts can be used to represent the cross-tabulation or joint distribution of two such categorical variables.

n_{ij} = the number of observations with $X = i$ and $Y = j$

$n = \sum_{i,j} n_{ij}$ = the total number of observations

	$Y=1$	$Y=2$	$\cdots$	$Y=i$	$\cdots$	$Y=J$	
$X=1$	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
$X=2$	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\ldots$	$\ldots$	$\ddots$	$\ldots$	$\ldots$	$\ldots$	$\vdots$
$X=i$	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ddots$	$\ldots$	$\vdots$
$X=I$	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
	n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

Contingency table probabilities

A contingency table can be represented by probabilities as well.

We define π_{ij} to be population parameter representing the true probability of being in the ij^{th} cell, i.e. the probability that both $X = i$ and $Y = j$). Formally, $\pi_{ij} = Pr(X = i, Y = j)$ and is called the joint probability of X and Y for all $i = 1, ..., I$ and $j = 1, ..., J$.

	$Y = 1$	$Y = 2$	$\cdots$	$Y = i$	$\cdots$	$Y = J$	
$X = 1$	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1J}	π_{1+}
$X = 2$	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\ldots$	$\ldots$	$\ddots$	$\ldots$	$\ldots$	$\ldots$	$\vdots$
$X = i$	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}	π_{i+}
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ddots$	$\ldots$	$\vdots$
$X = I$	π_{I1}	π_{I2}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}	π_{I+}
	π_{+1}	π_{+2}	$\cdots$	π_{+j}	$\cdots$	π_{+J}	π

Example contingency table

TABLE 2—Cell Type and Stage of Screen-Detected Lung Cancers Among Nuclear Weapons Workers: United States, 2000–2013

	Screen-Detected Lung Cancers	
	Baseline, No. (%)	Annual, No. (%)
Cell type		
Adenocarcinoma	30 (50.0)	9 (45.0)
Squamous cell carcinoma	12 (20.0)	6 (30.0)
Adenosquamous carcinoma	2 (3.3)	1 (5.0)
Large cell carcinoma	2 (3.3)	1 (5.0)
Other non–small cell—unspecified	3 (5.0)	0 (0.0)
Sarcomatoid carcinoma	1 (1.7)	0 (0.0)
Small cell	6 (10.0)	2 (10.0)
Large cell endocrine	3 (5.0)	0 (0.0)
Missing	1 (1.7)	1 (5.0)
Total	60 (100)	20 (100.0)

Markowitz et al. (2018), Yield of Low-Dose Computerized Tomography Screening for Lung Cancer in High-Risk Workers: The Case of 7189 US Nuclear Weapons Workers, *AJPH*

Notation for contingency table probabilities

One important probabilistic quantity from the contingency table is $\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = Pr(Y = j|X = i)$ or the conditional probability of j given i.

Note that there are similarities here to a regression-like problem, as we are trying to describe an `outcome'' variable as a function of a`predictor" variable. This is similar to the conditional formulation of $E[Y|X]$ in regression where we are modeling an outcome of Y conditional on observed X.

Sampling methodologies

Data arise from different sampling strategies. Different methods are appropriate for each strategy, so it's important to be able to identify the key features of different strategies.

1. Poisson

- The overall **n** is not fixed
- There is generally a time interval implied
- Example: A prospective longitudinal cohort study about developing a disease

	Disease
X1	n_1
X2	n_2
X3	n_3

$n_1 =$ total # of people in catergory X1 with the disease

- Example: # of accidents at an intersection over a year

Multinomial

a. with fixed **n**
   - Example: A cohort study with 3 categories of socioeconomic status and a binary outcome of illness (a fixed # of people are enrolled in the study)

	Sick	Not Sick	Total
SE_1	n_11	n_12	n_1+
SE_2	n_21	. . .	. . .
SE_3	. . .	. . .	. . .
Total	n_+1	. . .	2000

b. **row or column totals** are fixed
   - Example: A case-control study

	Case	Control
SE_1	. . .	. . .
SE_2	. . .	. . .
SE_3	. . .	. . .
Total	1000	1000