Discrete Response Model
Lecture 4

Subtitle

datascience@berkeley

Part 2:
I×J Contingency Tables and Inference Procedures

(6 minutes)

Introduction

- In Lecture 1, we introduced a 2×2 contingency table.
- Now we extend this concept to an I×J contingency table.
- We will begin by focusing on two separate ways that one can think of how the counts arise in a contingency table structure through using a multinomial distribution.
- In the next lecture, we will consider another way through using a Poisson distribution.
- The text also considers a third way that uses a hypergeometric distribution, but we will not cover it in this course.

One Multinomial Distribution

Set-up:

- X denotes the row variable with levels i = 1, …, I
- Y denotes the column variable with levels j = 1, …, J
- $P(X = i, Y = j) = \pi_{ij}$
- $\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1$
- n_{ij} denotes the cell count for row i and column j
- $\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij} = n$

Contingency tables summarizing this information are shown below:

		Y				
		1	2	…	J	
X	1	π_{11}	π_{12}	…	π_{1J}	π_{1+}
	2	π_{21}	π_{22}	…	π_{2J}	π_{2+}
	⋮	⋮	⋮	⋱	⋮	⋮
	I	π_{I1}	π_{I2}	…	π_{IJ}	π_{I+}
		π_{+1}	π_{+2}	…	π_{+J}	1

		Y				
		1	2	…	J	
X	1	n_{11}	n_{12}	…	n_{1J}	n_{1+}
	2	n_{21}	n_{22}	…	n_{2J}	n_{2+}
	⋮	⋮	⋮	⋱	⋮	⋮
	I	n_{I1}	n_{I2}	…	n_{IJ}	n_{I+}
		n_{+1}	n_{+2}	…	n_{+J}	n

Multinomial Probability Model

The setup given for these contingency tables fits right into the multinomial setting of Section 3.1. We now just categorize the responses with respect to X and Y. The probability mass function for observing particular values of $n_{11}, \ldots, n_{IJ}$ is

$$\frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!}\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{n_{ij}}$$

The MLE of π_{ij} is the estimated proportion $\hat{\pi}_{ij} = n_{ij}/n$.

We can also discuss marginal distributions for X and for Y as well:

- X is multinomial with counts n_{i+} for $i = 1, \ldots, I$ and corresponding probabilities π_{i+}.
  The maximum likelihood estimate of π_{i+} is $\hat{\pi}_{i+} = n_{i+}/n$.
- Y is multinomial with counts n_{+j} for $j = 1, \ldots, J$ and corresponding probabilities π_{+j}.
  The MLE of π_{+j} is $\hat{\pi}_{+j} = n_{+j}/n$

Example

As a quick way to see what a sample looks like in a 2×3 contingency table setting, consider the situation with n = 1,000 observations, $\pi_{11} = 0.2$, $\pi_{21} = 0.3$, $\pi_{12} = 0.2$, $\pi_{22} = 0.1$, $\pi_{13} = 0.1$, and $\pi_{23} = 0.1$. Below is how we can simulate a sample:

```
> pi.ij<-c(0.2, 0.3, 0.2, 0.1, 0.1, 0.1)
> pi.table<-array(data = pi.ij, dim = c(2,3), dimnames =
+     list(X = 1:2, Y = 1:3))
> pi.table
   Y
X     1   2   3
  1 0.2 0.2 0.1
  2 0.3 0.1 0.1
```

```
> set.seed(9812)
> save<-rmultinom(n = 1, size = 1000, prob = pi.ij)
> c.table1<-array(data = save, dim = c(2,3), dimnames =
+ list(X = 1:2, Y = 1:3))
> c.table1
   Y
X     1   2   3
  1 191 206  94
  2 311  95 103
> c.table1/sum(c.table1)
   Y
X       1     2     3
  1 0.191 0.206 0.094
  2 0.311 0.095 0.103
```

Berkeley SCHOOL OF INFORMATION