Discrete Response Model Lecture 4

Subtitle

datascience@berkeley

Part 2: IxJ 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 IxJ 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
- $P(X = \underline{i}, Y = \underline{j}) = \pi_{\underline{i}\underline{j}}$
- $\bullet \quad \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} \, = 1$
- 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:

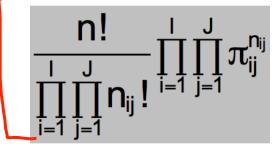
4							•
			1	2		J	
		1	π_{11}	π_{12}		π 1J	π_{1+}
	v	2	π_{21}	π_{22}		$\pi_{2\mathtt{J}}$	π_{2+}
	Х	:	:	:	٠.	:	:
		I	$\pi_{{\scriptscriptstyle \mathtt{I}\mathtt{1}}}$	π_{12}		$oldsymbol{\pi}_{ exttt{IJ}}$	$\pi_{{\scriptscriptstyle \mathtt{I}}^+}$
			π_{+1}	π_{+2}	•••	$oldsymbol{\pi}_{+\mathtt{J}}$	1

1,							Ī
	Y						
_			1	2		J	
		1	n ₁₁	n ₁₂	:	n_{1J}	n ₁₊
	Х	2	n ₂₁	n ₂₂		n _{2J}	n ₂₊
	Λ .	:	:	:	٠.	:	:
		I	n _{I1}	n _{I2}		n_{IJ}	$n_{ m J^+}$
_			n ₊₁	n ₊₂		$n_{\pm J}$	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} , ..., n_{IJ} is



The MLE of $\pi_{i,i}$ is the estimated proportion $\hat{\pi}_{ij} = \underline{n}_{i,j}/n$.

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

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

Example

2 0.311 0.095 0.103

```
As a quick way to see what a sample looks like in a 2\times3
contingency table setting, consider the situation with n =
1,000 observations, \pi_{1} = 0.2, \pi_{21} = 0.3, \pi_{12} = 0.2, \pi_{22} = 0.2
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 =</pre>
 + list(X = 1:2, Y = 1:3))
> pi.table
  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)</pre>
 > c.table1<-array(data = save, dim = c(2,3), dimnames =</pre>
 + list(X = 1:2, Y = 1:3))
 > c.table1
      1 2
   1 191 206 94
   2 311 95 103
 > c.table1/sum(c.table1)
   1 0.191 0.206 0.094
```

Berkeley school of information