UPPSALA UNIVERSITET

Lecture Notes

Analysis of Categorical Data

Rami Abou Zahra

Inlämningsdatum
October 23, 2025

Contents

- Nominal: no ordering behind the categories

Note: The features can be continuous, but in this course the categorical variable is discrete.

Slide 33
One can always construct a table whose partial tables has odds ratio 1. For the Berkley data, looking at the universiuty as a whole we had independence but dependence when looking departmentwise. Just because the odds ratio is 1, does not mean that the marginal odds will also be 1.

		Y	
Z	X	0	1
Z_1	0	100	10
Z_2	1	200	20
Z_3	0	100	50
Z_4	1	60	30

Here, the odds ratio is $\frac{100 \cdot 20}{10 \cdot 200} = 1 = \frac{100 \cdot 30}{50 \cdot 60}$, but the marginal table looks like this:

	Y	
X	0	1
0	100+100	10+50
1	200+60	20+30

We can see that $\theta_{xy} = \frac{200 \cdot 50}{60 \cdot 260} \neq 1$

Slide 38
Odds ratio can be computed by pairwise computation.
Local odds ratio: *only* adjacent, eg $X = 0$ and $Y = 2$ columns will not be included. Only this is needed.

Slide 41
For the following table:

	Y	
X	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

Assuming multinomial sampling with the total sum being fixed to nm we wish to find the distribution of all n_{ij}. Since n is known, we normalize:

	Y	
X	1	2
1	n_{11}/n	n_{12}/n
2	n_{21}/n	n_{22}/n

Note that this is indeed a valid estimation, since they all sum to 1. Also, since they sum to 1, we only need to know three of them. When we want to estimate the distribution of $\frac{1}{n}\begin{bmatrix} n_{11} \\ n_{12} \\ n_{21} \end{bmatrix}$, we use the CLT:

$$\frac{1}{\sqrt{n}}\left(\frac{1}{n}\begin{bmatrix} n_{11} \\ n_{12} \\ n_{21} \end{bmatrix} - \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \end{bmatrix}\right) \approx N\left(0, \begin{bmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & \pi_{11}\pi_{21} \\ & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21} \\ & & \pi_{21}(1-\pi_{21}) \end{bmatrix}\right)$$

Example: Consider $g(x_1, x_2, x_3) = \ln(x_1) - \ln(x_2) - \ln(x_3) + \ln(1 - x_1 - x_2 - x_3)$

If:

$$\left.\begin{aligned}\frac{n_{11}}{n} &= x_1\\ \frac{n_{12}}{n} &= x_2\\ \frac{n_{21}}{n} &= x_3\end{aligned}\right\} \quad \ln\left(\frac{n_{11}}{n}\right) - \ln\left(\frac{n_{12}}{n}\right) - \ln\left(\frac{n_{21}}{n}\right) + \underbrace{\ln\left(1 - \frac{n_{11}}{n} - \frac{n_{12}}{n} - \frac{n_{21}}{n}\right)}_{\ln\left(n_{22}/n\right)} = \ln\left(\widehat{\theta}\right)$$

To find the distribution of $\ln\left(\widehat{\theta}\right)$, apply the delta method to g:

$$\frac{\partial g}{\begin{bmatrix}\partial x_1\\ \partial x_2\\ \partial x_3\end{bmatrix}} = \begin{bmatrix}\frac{1}{x_1} - \frac{1}{1 - x_1 - x_2 - x_3}\\ \vdots\end{bmatrix}$$

$$\Rightarrow \ln\left(\widehat{\theta}\right) - \ln\left(\theta_0\right) \approx N(0, ?)$$

$$? = \left[\frac{1}{\pi_{11}} - \frac{1}{\pi_{22}}, -\frac{1}{\pi_{12}} - \frac{1}{\pi_{22}}, -\frac{1}{\pi_{21}} - \frac{1}{\pi_{22}}\right] \begin{bmatrix}\pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & \pi_{11}\pi_{21}\\ & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21}\\ & & \pi_{21}(1-\pi_{21})\end{bmatrix}$$

$$\begin{bmatrix}\frac{1}{\pi_{11}} - \frac{1}{\pi_{22}}\\ -\frac{1}{\pi_{12}} - \frac{1}{\pi_{22}}\\ -\frac{1}{\pi_{21}} - \frac{1}{\pi_{22}}\end{bmatrix} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

The last equality holds regardless of sampling method.