Introduction to categorical data and conditional independence

MSc Further Statistical Methods, Lecture 1 Hilary Term 2006

Steffen Lauritzen, University of Oxford; January 18, 2006

Categorical Data

Examples of categorical variables

- *Sex*: Male, Female;
- *Colour of Hair*: Blond, Red, Neutral, Dark;
- *Degree of Satisfaction with work*: Low, Medium, High
- *Yearly income*: Below 10,000, 10,001-20,000, 20,001-40,000, above 40,000;

Some are *nominal*, others *ordinal*. They have different number of *states.*

Contingency Table

Data often presented in the form of a *contingency table* or *cross-classification*:

	Sex	
Admitted	Male	Female
Yes	1198	557
No	1493	1278

This is a *two-way table* (or two-way classification) with categorical variables A: Admitted? and S: Sex. In this case it is a 2×2*-table*.

The numerical entries are *cell counts* n_{ij}, the number of cases in the category $A = i$ and $S = j$. The *total number of cases* is $n = \sum_{ij} n_{ij}$.

Data in list form

Data can also appear in the form of a *list of cases:*

case	Admitted	Sex
1	Yes	Male
2	Yes	Female
3	No	Male
4	Yes	Male
⋮	⋮	⋮

The contingency table is then formed from the list of cases by counting the number of cases in each cell of the table.

Multinomial sampling model

The standard sampling model for data of this form specifies that cases are independent and $p_{ij} = P(A = i, S = j)$ is the probability that a given case belongs to cell ij.

The cell counts then follow a *multinomial distribution*

$$P(N_{ij} = n_{ij}, i = 1, \ldots I, j = 1, \ldots J) = \frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij} p_{ij}^{n_{ij}}.$$

The *expected cell counts* are

$$m_{ij} = \mathbf{E}(N_{ij}) = np_{ij}.$$

Other sampling schemes *fixes certain marginal totals* or have a *Poisson total* N, leading to cell counts being independent Poisson.

Hypothesis of independence

A typical hypothesis of interest is that of *independence* between the two variables, i.e. that

$$p_{ij} = P(A = i, S = j) = P(A = i)P(S = j) = p_{i+}p_{+_j},$$

where

$$p_{i+} = P(A = i) = \sum_j p_{ij}, \quad p_{+j} = P(S = j) = \sum_i p_{ij}$$

are the *marginal probabilities*.

Likelihood ratio test

Without assuming independence, the MLE of the cell probabilities and expected cell counts are

$$\hat{p}_{ij} = n_{ij}/n, \quad \hat{m}_{ij} = n\hat{p}_{ij} = n_{ij}.$$

Similarly, assuming independence, the MLE becomes

$$\hat{\hat{p}}_{ij} = n_{i+}n_{+j}/n^2, \quad \hat{\hat{m}}_{ij} = n\hat{\hat{p}}_{ij} = n_{i+}n_{+j}/n,$$

where

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_i n_{ij}$$

are the *marginal counts*. Hence we get

$$\begin{aligned} G^2 &= -2\log\Lambda = -2\log\frac{L(\hat{\hat{p}})}{L(\hat{p})} \\ &= 2\sum_{ij} n_{ij}\log\frac{\hat{p}_{ij}}{\hat{\hat{p}}_{ij}} = 2\sum_{ij} n_{ij}\log\frac{\hat{m}_{ij}}{\hat{\hat{m}}_{ij}} \\ &= 2\sum_{ij} n_{ij}\log\frac{n_{ij}}{\hat{\hat{m}}_{ij}} = 2\sum \mathsf{OBS}\log\frac{\mathsf{OBS}}{\mathsf{EXP}}, \end{aligned}$$

Here OBS refers to the *observed cell counts* and EXP to the *expected cell counts* under the hypothesis.

It can be shown that for large cell counts, *G^2 is approximately χ^2-distributed with degrees of freedom equal to $(I-1)(J-1)$* which is equal to 1 in this case.

Pearson's χ^2 statistic

An alternative to the LRT statistic or *deviance* G^2, one can use the statistic

$$\chi^2 = \sum \frac{(\text{OBS} - \text{EXP})^2}{\text{EXP}},$$

which is an approximation to the deviance and also has approximately the same distribution, under the null hypothesis, for large cell counts.

For the approximations to be valid, it is a *common rule of thumb for both G^2 and χ^2 that the expected cell counts $\hat{\hat{m}}_{ij}$ must be larger than 5.*

This condition is often *not* satisfied, in particular in multi-way tables with many variables.

Sparse tables

Data on oral lesions by region in India:

	Kerala	Gujarat	Andhra
Labial Mucosa	0	1	0
Buccal Mucosa	8	1	8
Commisure	0	1	0
Gingiva	0	0	1
Hard Palate	0	1	0
Soft palate	0	1	0
Tongue	0	1	1
Floor of Mouth	1	0	1
Alveolar Ridge	1	0	1

Exact testing methods

In sparse tables such as the data on oral lesions, asymptotic results can be very misleading.

Instead one can exploit that, under the hypothesis of independence, *the marginals are sufficient* and the conditional distribution of the counts $\{N_{ij}\}$ is:

$$P\left\{(n_{ij}) \,|\, (n_{i+}), (n_{+j})\right\} = \frac{\prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} n_{+j}!}{n! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!}. \qquad (1)$$

Fisher's exact test rejects for small values of the *observed value* of $P\left\{(n_{ij}) \,|\, (n_{i+}), (n_{+j})\right\}$ and evaluates the p-value in this distribution as well.

Monte-Carlo testing

In principle, exact testing requires enumeration of all possible tables with a given margin.

However, there is an *efficient algorithm* due to Patefield (1981) which generates samples $\{\tilde{n}_{ij}\}_k, k = 1, \ldots K$ from the distribution (1).

By choosing K large, the correct p-value *for any test statistic* T can be calculated to any degree of accuracy as

$$\tilde{p} = \frac{|\{k : \tilde{t}_k \geq t_{\text{obs}}\}|}{K},$$

where $\tilde{t}_k$ is calculated from the table $\{\tilde{n}_{ij}\}_k$.

This may well be preferable to using asymptotic results.

Three-way tables

Admissions to Berkeley by department

Department	Sex	Whether admitted	
		Yes	No
I	Male	512	313
	Female	89	19
II	Male	353	207
	Female	17	8
III	Male	120	205
	Female	202	391
IV	Male	138	279
	Female	131	244
V	Male	53	138
	Female	94	299
VI	Male	22	351
	Female	24	317

Here are three variables A: Admitted?, S: Sex, and D: Department.

Conditional independence

For three variables it is of interest to see whether independence holds for fixed value of one of them, e.g. *is the admission independent of sex for every department separately?* We denote this as $A \perp\!\!\!\perp S \mid D$ and graphically as

Algebraically, this corresponds to the relations

$$p_{ijk} = p_{i+\,|\,k} p_{+j\,|\,k} p_{++k} = \frac{p_{i+k} p_{+jk}}{p_{++k}}.$$

Marginal and conditional independence

Note that there the two conditions

$$A \perp\!\!\!\perp S, \quad A \perp\!\!\!\perp S \mid D$$

are very different and will typically not both hold unless we either have $A \perp\!\!\!\perp (D, S)$ or $(A, D) \perp\!\!\!\perp S$, i.e. if one of the variables are completely independent of both of the others.

This fact is a simple form of what is known as *Yule–Simpson paradox.*

It can be much worse than this:

A *positive conditional association can turn into a negative marginal association* and vice-versa.

Admissions revisited

Admissions to Berkeley

Sex	Whether admitted	
	Yes	No
Male	1198	1493
Female	557	1278

Note this marginal table shows much lower admission rates for females.

Considering the departments separately, there is only a difference for department I, and it is the other way around...

Florida murderers

Sentences in 4863 murder cases in Florida over the six years 1973-78

	Sentence	
Murderer	Death	Other
Black	59	2547
White	72	2185

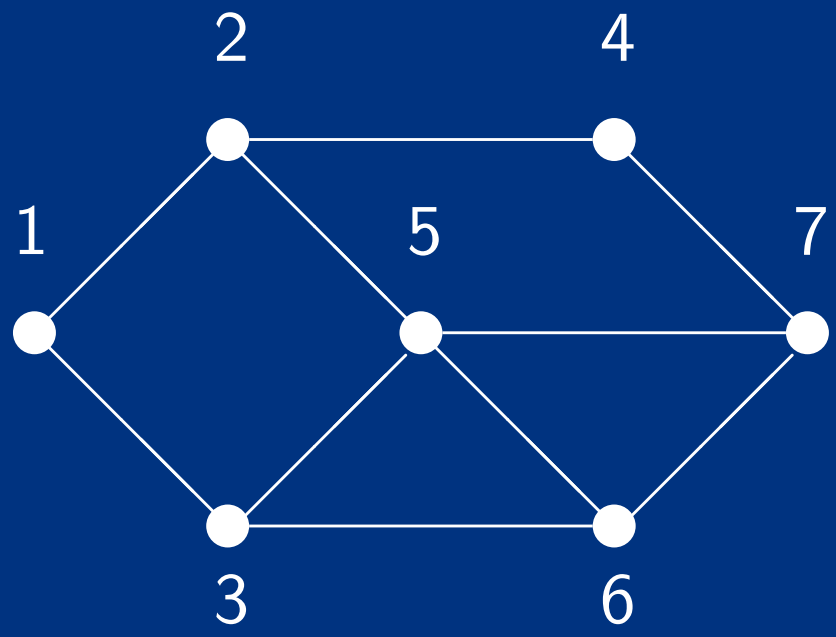
The table shows a greater proportion of white murderers receiving death sentence than black (3.2% vs. 2.3%), although the difference is not big, the picture seems clear.

Controlling for colour of victim

Victim	Murderer	Sentence	
		Death	Other
Black	Black	11	2309
	White	0	111
White	Black	48	238
	White	72	2074

Now the table for given colour of victim shows a very different picture. In particular, note that 111 white murderers killed black victims and none were sentenced to death.

Graphical models



For several variables, complex systems of conditional independence can be described by undirected graphs.

Then a set of variables A is conditionally independent of set B, given the values of a set of variables C if C *separates* A *from* B.

Conditional independence

Random variables X and Y are *conditionally independent* given the random variable Z if

$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z).$$

We then write $X \perp\!\!\!\perp Y \mid Z$

Intuitively:

Knowing Z renders Y *irrelevant* for predicting X.

Factorisation of probabilities:

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \quad &\iff \quad p_{xyz} p_{++z} = p_{x+z} p_{+y,z} \\ &\iff \quad \exists a, b : p_{xyz} = a_{yz} b_{yz}. \end{aligned}$$

Fundamental properties

For any random variables X, Y, Z, and W it holds

(C1) if $X \perp\!\!\!\perp Y \,|\, Z$ then $Y \perp\!\!\!\perp X \,|\, Z$;

(C2) if $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \,|\, Z$;

(C3) if $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \,|\, (Z, U)$;

(C4) if $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \,|\, Z$;

If all joint probabilities p_{xyzw} are strictly positive also

(C5) if $X \perp\!\!\!\perp Y \,|\, (Z, W)$ and $X \perp\!\!\!\perp Z \,|\, (Y, W)$ then $X \perp\!\!\!\perp (Y, Z) \,|\, W$.