

# Introduction to categorical data

**MSc Further Statistical Methods, Lecture 1**  
**Hilary Term 2005**

Steffen Lauritzen, University of Oxford; January 21, 2005

# 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*:

Admitted	Sex	
	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  $2 \times 2$ -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  $ij$ .

This implies that the cell counts follow a *multinomial distribution*

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

The *expected cell counts* are

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

## 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

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

where

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

are the *marginal counts*. Hence we get

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

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  $\chi^2$ -distributed with degrees of freedom equal to  $(I - 1)(J - 1)$*  which is equal to 1 in this case.



## Pearson's $\chi^2$ 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  $\chi^2$  that the expected cell counts  $\hat{m}_{ij}$  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

## Monte-Carlo testing

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  $\{N_{ij}\}$  has a known form:

$$P\{(n_{ij}) \mid (n_{i+}), (n_{+j})\} = \frac{\prod_{i=1}^I n_{i+}! \prod_{j=1}^J n_{+j}!}{n! \prod_{i=1}^I \prod_{j=1}^J n_{ij}!}. \quad (1)$$

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

By choosing  $K$  large enough, the correct  $p$ -value can then be calculated to any desired degree of accuracy as

$$\tilde{p} = \frac{|\{k : \tilde{G}_k^2 \geq G_{\text{obs}}^2\}|}{K},$$

where  $\tilde{G}_k^2$  is calculated from the table  $\{\tilde{n}_{ij}\}_k$ , and similarly for  $\chi^2$ .

This method may well be preferable to the asymptotic result unless the cell counts are very large.