STAT3015/4030/7030:

Generalised Linear Modelling

Contingency Tables

Semester 2 2016

Originally prepared by Bronwyn Loong

Reference

Faraway, Ch 4 (Ch 6 in the new edition with an additional section on Correspondence Analysis)

Ramsey and Schafer, Ch 19

Contingency Tables

A contingency table is used to show cross-classified categorical data on two or more variables.

The variables can be *nominal* or *ordinal*.

Example of nominal variable? Example of ordinal variable?

(For this class, we will be considering the analysis of contingency tables with nominal variables only)

Case Study -Hair and eye colour

Data was collected on the hair and eye colour of 592 students. The data is cross-classified in a contingency table as below:

		eye		
Hair	Green	Hazel	Blue	Brown
BLACK	5	15	20	68
BROWN	29	54	84	119
RED	14	14	17	26
BLONDE	16	10	94	7
Total	64	93	215	220

Q: What do you think is the question of interest the researchers hope to answer by collecting this data? How are you going to answer this question? What statistical model will you build and use?

Population Models for $R \times C$ Tables of Counts

Sampling Schemes for $R \times C$ tables

Is the data produced by

- a single population with $R \times C$ possible response categories?
- R populations, each with C categories of responses?

Population Models for $R \times C$ Tables of Counts

Sampling Schemes for $R \times C$ Tables

- **Poisson distribution** - a fixed amount of time (space, volume, money etc.) is devoted to collecting a random sample from a single population, and each member of the population falls into one of the $R \times C$ cells. None of the marginal totals are known in advance.
- **Multinomial distribution** - Similar to Poisson sampling but the total sample size is known in advance.
- **Product multinomial distribution** - row totals are fixed. Within each row, the response can fall into one of column j, (j=1,..,C). That is, each row defines a multinomial population. (Vice versa, can have column totals fixed and each column defines a multinomial population).

Population Models for $R \times C$ Tables of Counts

Hypotheses of Homogeneity and of Independence

Let p_{ij} be the probability the sampled subject contributes to the count in the (i,j)th cell (i=1,..,R; j=1,..,C)

- Test of **homogeneity** (for fixed column totals) (that is, homogeneity of the columns)

  $$H_0 : p_{i1} = p_{i2} = ... = p_{iC} = p_{i\bullet} \ \forall_i$$

  Example: The proportion of black hair people is the same within each eye colour population

- For the test of **independence**, the null hypothesis is:

  H_0: row categorization is independent of the column parametization. That is, is there an association between the column factor and row factor?

Which test?? Depends on the sampling scheme which produced the data.

Population Models for $R \times C$ Tables of Counts

Sampling scheme	**Marginal totals fixed in advance**	**Independence**	**Homogeneity**
Poisson	None	✓	✓
Multinomial	Grand Total	✓	✓
Product Multinomial	Row totals or column totals		✓

Poisson sampling model

Suppose students were sampled as they passed through the library entrance during one lunch period. Each student was categorized (by their hair and eye colour) into one of the cells in the table. (That is, we count the number of occurrences of the possible outcomes).

Let Y_{ij} be the count observed in cell $\{i, j\}$ of the table. We can think about each outcome occurring at different rates $\lambda_{ij} = \mu_{ij}$ and fit a Poisson GLM to model $Y_{ij} \sim Pois(\lambda_{ij})$

Exercise: What are the predictors in the model? Write down the equation which relates the mean response, μ_{ij}, to these predictors. (see R code)

Multinomial sampling model

Suppose we assume that the total sample size was fixed at 592 and the frequency of the 16 possible outcomes was recorded.

Define

- y_{ij}: the observed response in cell (i, j)
- p_{ij}: the probability than an observation falls in that cell.
- n be the sample size.

The probability of the observed data is then::

$$\frac{n!}{\prod_i \prod_j y_{ij}!} \prod_i \prod_j p_{ij}^{y_{ij}}$$

What are the parameters of the multinomial model?

Multinomial sampling model

Log-likelihood ($\mu_{ij} = E[Y_{ij}] = np_{ij}$):

$$\log L = \sum_i \sum_j y_{ij} \log p_{ij} + d(Y) = \sum_i \sum_j y_{ij} \log \frac{\mu_{ij}}{n} + d(Y)$$

(d(Y) - only a function of Y, not parameters p_{ij})
Main hypothesis of interest: are row and column factors independent? Define

- $p_{i\bullet}$ (for i=1,..,R) - the probabilities of row outcomes
- $p_{\bullet j}$ (for $j = 1, .., C$) - the probabilities of column outcomes

Under independence $p_{ij} = p_{i\bullet} p_{\bullet j}$.

$$\hat{p}_{i\bullet} = \sum_j y_{ij}/n = \frac{y_{i\bullet}}{n} \text{ and } \hat{p}_{\bullet j} = \sum_i y_{ij}/n = \frac{y_{\bullet j}}{n}$$

Multinomial sampling model

The fitted values for the independence model are $\mu^*_{ij} = n\hat{p}_{i\bullet}\hat{p}_{\bullet j} = \sum_i y_{ij} \sum_j y_{ij}/n = \frac{y_{i\bullet}y_{\bullet j}}{n}$.

Multinomial sampling model

The fitted values for the independence model are
$\mu^*_{ij} = n\hat{p}_{i\bullet}\hat{p}_{\bullet j} = \sum_i y_{ij} \sum_j y_{ij}/n = \frac{y_{i\bullet}y_{\bullet j}}{n}$.
The fitted values in the saturated model are y_{ij}

So the deviance is
$D = 2\sum_i \sum_j y_{ij} \log(y_{ij}/\mu^*_{ij}) = 2\sum_i \sum_j O_{ij} \log(O_{ij}/E_{ij})$

O_{ij}: observed count; E_{ij}: expected count

Run a drop in deviance test to decide if independence holds. How many parameters in saturated model? . ??
How many parameters in the independence model? ??

p-value: $Pr(\chi^2_{(R-1)(C-1)} \geq D)$

Multinomial sampling model - Pearsons χ^2

We can show that

$$2\sum_i \sum_j O_{ij} \log(O_{ij}/E_{ij}) \approx \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \chi^2$$

This is the Pearson χ^2 statistic. The Pearson χ^2 is more commonly used to test for independence in contingency tables.

The p-value is the proportion of values from a chi-squared distribution on (R-1)(C-1) degrees of freedom that are greater than the calculated Pearson χ^2 statistic.

Multinomial sampling model - Pearsons χ^2

We can use the Pearson Chi-squared statistic to test quite general hypothesised structures for the p_{ij}'s. All that is required is to determine the appropriate E_{ij}'s under the null and the appropriate degrees of freedom.

(NB: to get the appropriate degrees of freedom, think of the minimum number of p_{ij}'s which if known determine the value of the rest of the p_{ij}'s).

Caution: The distribution of Pearsons Chi-squared statistic is an approximation only, unreliable if a large majority of E_{ij}'s are less than 5.
How to get around this??
Other limitations:...

Poisson sampling model - Pearsons χ^2

Note: If we assume the data arise from a Poisson sampling structure, then the Pearson residuals are:

$$z_{ij} = \frac{Y_{ij} - \hat{Y}_{ij}}{\sqrt{V(\hat{Y}_{ij})}} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

Thus

$$\text{Pearsons } \chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j z_{ij}^2$$

Test of Homogeneity

Consider the multinomial sampling scheme. What are the expected cell counts under the homogeneity assumption?

Suppose column totals are fixed, we can show that

$$\log Lik = \sum_i \sum_j y_{ij} \log p_{ij} + d(Y) = \sum_i \sum_j y_{ij} \log \frac{\mu_{ij}}{y_{\bullet j}} + d(Y)$$

(similar to fixed total case); moreover, under homogeneity,

$$E_{ij} = y_{\bullet j} \hat{p}_{i\bullet} = \frac{y_{i\bullet} y_{\bullet j}}{n}$$

same E_{ij}'s from independence model → test for homogeneity is the same as the test for independence!

Pearsons χ^2 - Limitations

Limitations of chi-squared tests

- Expected cell counts larger than 5.
- Not very informative, we only obtain a p-value as output. Does not describe degree of dependence.
- The alternative hypothesis - that row and column are not independent - is very general.

Higher-Dimensional Tables of Counts

Example: three-way contingency table.
Let p_{ijk} be the probability that an observation falls into the (i, j, k) cell.

- **Mutual independence** If all three variables are independent then $p_{ijk} = p_i p_j p_k$
- **Joint independence** If the first and second variables are dependent, but jointly independent of the third then, $p_{ijk} = p_{ij} p_k$
- **Conditional independence** suppose the first and second variables are independent given the third. Then $p_{ij|k} = p_{i|k} p_{j|k}$ which leads to $p_{ijk} = p_{ik} p_{jk} / p_k$

How to test these? Fit Poisson linear regression models, run drop in deviance tests.

(see R code)