Aims of today’s lecture

In today’s lecture we begin our study of contingency tables. We will apply Poisson regression to the analysis of one dimensional contingency tables. Topics covered are

- R trick of the day
- Contingency tables
- Sampling models
- Equivalence of Poisson and multinomial models
R trick of the day: random number generation

> rnorm(5)
[1]  0.7346887  -1.4041144  -1.9004288
[4]  0.6630558  -1.2165798
> means = 1:8
> rpois(8,means)
[1] 0 0 3 2 9 11 4 10
> probs = rep(1/12, 12)
> y = rmultinom(1,16978, probs)
> sum(y)
[1] 16978
> t(y)
[1,] 1413 1453 1430 1414 1511 1323 1348 1347 1411 1424 1488 1416
Contingency tables

Contingency tables arise when we classify a number of individuals into categories using one or more criteria.

Result is a cross-tabulation or contingency table:

- Classifying by one criterion gives a 1-dimensional table
- Classifying using two criteria give a 2-dimensional table
- and so on ...
Example: one dimension

Income distribution of New Zealanders 15+ 2006 census:

<table>
<thead>
<tr>
<th></th>
<th>$5,000 or Less</th>
<th>$5,001-$10,000</th>
<th>$10,001-$20,000</th>
<th>$20,001-$30,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>383,574</td>
<td>226,800</td>
<td>615,984</td>
<td>434,958</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$30,001-$50,000</th>
<th>$50,001 or more</th>
<th>Not Stated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>666,372</td>
<td>511,803</td>
<td>320,892</td>
<td>3,160,371</td>
</tr>
</tbody>
</table>
### New Zealanders 15+ by Gender and Income, 2006 census:

<table>
<thead>
<tr>
<th></th>
<th>$10,000 or Less</th>
<th>$10,001-$20,000</th>
<th>$20,001-$30,000</th>
<th>$40,001-$50,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>232,620</td>
<td>232,884</td>
<td>407,466</td>
<td>331,320</td>
</tr>
<tr>
<td>Female</td>
<td>377,754</td>
<td>383,097</td>
<td>431,565</td>
<td>212,139</td>
</tr>
<tr>
<td>Total</td>
<td>610,374</td>
<td>615,981</td>
<td>839,031</td>
<td>543,459</td>
</tr>
</tbody>
</table>

$70,001-$50,000 or more:

<table>
<thead>
<tr>
<th></th>
<th>$70,001-$50,000</th>
<th>$100,001 or more</th>
<th>Not Stated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>90,177</td>
<td>82,701</td>
<td>144,420</td>
<td>1,521,591</td>
</tr>
<tr>
<td>Female</td>
<td>34,938</td>
<td>22,821</td>
<td>176,472</td>
<td>1,638,783</td>
</tr>
<tr>
<td>Total</td>
<td>125,115</td>
<td>105,522</td>
<td>320,892</td>
<td>3,160,374</td>
</tr>
</tbody>
</table>
Censuses and samples

- Sometimes tables come from a census
- Sometimes the table comes from a random sample from a population
- In this case we often want to draw inferences about the population on the basis of the sample e.g. is income independent of sex?
Example: death by falling

The following is a random sample of 16,976 persons who met their deaths in fatal falls, selected from a larger population of deaths from this cause. The falls classified by month are

<table>
<thead>
<tr>
<th>Month</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>1688</td>
</tr>
<tr>
<td>Feb</td>
<td>1407</td>
</tr>
<tr>
<td>Mar</td>
<td>1370</td>
</tr>
<tr>
<td>Apr</td>
<td>1309</td>
</tr>
<tr>
<td>May</td>
<td>1341</td>
</tr>
<tr>
<td>Jun</td>
<td>1388</td>
</tr>
<tr>
<td>July</td>
<td>1406</td>
</tr>
<tr>
<td>Aug</td>
<td>1446</td>
</tr>
<tr>
<td>Sep</td>
<td>1322</td>
</tr>
<tr>
<td>Oct</td>
<td>1363</td>
</tr>
<tr>
<td>Nov</td>
<td>1410</td>
</tr>
<tr>
<td>Dec</td>
<td>1526</td>
</tr>
</tbody>
</table>
The question

- Question: if we regard this as a random sample from several years, are the months in which death occurs equally likely?

- Or is one month more likely than the others?

- To answer this, we use the idea of maximum likelihood.

- First, we must detour into some sampling theory in order to work out the likelihood.
Sampling models

There are two common models used for contingency tables

- The **multinomial sampling model** assumes that a fixed number $n$ of individuals are classified into $M$ categories at random with fixed probabilities of being assigned to the $M$ different categories.

- The **Poisson sampling model** assumes that the table counts have independent Poisson distributions.
Multinomial sampling

Situation (b): Single sample, several response categories

Single sample

Cat. 1  Cat. 2  Cat. 3  Cat. 4  Cat. 5  Cat. 6

Compare proportions
Multinomial model

- Suppose a table has $M$ “cells” (categories)

- $n$ individuals are classified independently (a.k.a. sampling with replacement)

- Each individual has probability $\pi_i$ of being in cell $i$

- Every individual is classified into exactly 1 cell, so

$$\pi_1 + \pi_2 + \ldots + \pi_M = 1$$

- Let $Y_i$ be the number in the $i$th cell, so

$$Y_1 + Y_2 + \cdots + Y_M = n$$
Multinomial model (cont)

Under multinomial sampling, $Y_1, Y_2, \cdots, Y_M$ have a multinomial distribution

$$Pr(Y_1 = y_1, \ldots, Y_M = y_M) = \frac{n!}{y_1! y_2! \cdots y_M!} \pi_1^{y_1} \cdots \pi_M^{y_M}$$

This is the maximal model, as in logistic regression, making no assumptions about the probabilities.
The log-likelihood is

\[
\log \Pr(Y_1 = y_1, \ldots, Y_M = y_M) = \log \frac{n!}{y_1!y_2! \cdots y_M!} \pi_1^{y_1} \cdots \pi_M^{y_M}
\]

\[
= y_1 \log \pi_1 + \cdots + y_M \log \pi_M + \text{const}
\]
MLE’s

- If there are no restrictions on the $\pi$’s (except that they add to one) the log-likelihood is maximised when $\pi_i = y_i / n$

- These are the MLE’s for the maximal model

- Substitute these into the log-likelihood to get the maximum value of the maximal log-likelihood. Call $\log L_{max}$
Deviance

- Suppose we have some model for the probabilities: this will specify the form of the probabilities, perhaps as functions of other parameters. In our death by falling example, the model says that each probability is $1/12$.

- Let $\log L_{\text{mod}}$ be the maximum value of the log-likelihood, when the probabilities are given by the model.

- As for logistic regression, we define the deviance as

\[ D = 2(\log L_{\text{max}} - \log L_{\text{mod}}) \]
Deviance: Testing adequacy of the model

- If $n$ is large (all cells more than 5) and $M$ is small, and if the model is true, the deviance will have (approximately) a chi-square distribution with $M - k - 1$ df, where $k$ is the number of unknown parameters in the model.

- Thus, we accept the model if the deviance p-value is more than 0.05. This is almost the same as the situation in logistic regression with strongly grouped data.

- These tests are an alternative form of the chi-square tests used in stage 2.
Example: death by falling

- Are deaths equally likely to fall in every month? In statistical terms, is
  \[ \pi_i = \frac{1}{12}, \quad i = 1, 2, \ldots, 12? \]
- \[ \log L_{max} \] is, up to a constant,
  \[ y_1 \log \left( \frac{y_1}{n} \right) + \cdots + y_{12} \log \left( \frac{y_{12}}{n} \right) \]

- Calculated in R by
  ```r
  > y<-c(1688,1407,1370,1309,1341,1388,1406,1446,1322,1363,1410,1526)
  > sum(y*log(y/sum(y)))
  [1] -42143.23
  ```
Example: death by falling, continued

Our model is $\pi_i = 1/12, \ i = 1, 2, \ldots, 12$ (This completely specifies the probabilities, so $k=0, M = 12, M - k - 1 = 11$. $\log L_{\text{max}}$ is, up to a constant,

$$y_1 \log(1/12) + \cdots + y_{12} \log(1/12)$$

Calculated in R by

```r
> sum(y*log(1/12))
[1] -42183.78 # now calculate deviance
> D<-2*sum(y*log(y/sum(y)))-2*sum(y*log(1/12))
> D
[1] 81.09515
> 1-pchisq(D,11)
```

The model is not plausible!
Barplot

```r
> Months<-c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
> barplot(y/sum(y),names.arg=Months, col="red", xlab = "Month", ylab = "Proportion")
```
Poisson model

- Assume that each cell count has a Poisson distribution with a different mean: this is just a Poisson regression model, with a single explanatory variable “month”.

- This is the maximal model.

- The null model is the model with all means equal

- Null model deviance will give us a test that all months have the same mean
R-code for Poisson regression

> Months<-factor(Months,levels=Months)
> falls(glm<-glm(y~Months,family=poisson)
> summary(falls glm)

Deviances are
Null deviance: 8.1095e+01 on 11 degrees of freedom
Residual deviance: -7.5051e-14 on 0 degrees of freedom
> D # deviance calculated on previous slide
[1] 81.09515

Same as before? Why? Why is residual deviance effectively zero?
Estimated regression coefficients

Coefficients:

|                         | Estimate | Std. Error | z value | Pr(>|z|)  |
|-------------------------|----------|------------|---------|----------|
| (Intercept)             | 7.43130  | 0.02434    | 305.317 | < 2e-16  |
| months.facFeb           | -0.18208 | 0.03610    | -5.044  | 4.56e-07 |
| months.facMar           | -0.20873 | 0.03636    | -5.740  | 9.46e-09 |
| months.facApr           | -0.25428 | 0.03683    | -6.904  | 5.04e-12 |
| months.facMay           | -0.23013 | 0.03658    | -6.291  | 3.15e-10 |
| months.facJun           | -0.19568 | 0.03623    | -5.401  | 6.64e-08 |
| months.facJul           | -0.18280 | 0.03611    | -5.063  | 4.13e-07 |
| months.facAug           | -0.15474 | 0.03583    | -4.318  | 1.57e-05 |
| months.facSep           | -0.24440 | 0.03673    | -6.655  | 2.84e-11 |
| months.facOct           | -0.21386 | 0.03642    | -5.873  | 4.29e-09 |
| months.facNov           | -0.17995 | 0.03608    | -4.988  | 6.10e-07 |
| months.facDec           | -0.10089 | 0.03532    | -2.856  | 0.00429  |
General Principle

▶ To every Poisson model for the cell counts, there corresponds a multinomial model, obtained by conditioning on the table total.

▶ Suppose the Poisson means are $\mu_1, \ldots, \mu_M$. The cell probabilities in the multinomial sampling model are related to the means in the Poisson sampling model by the relationship

$$\pi_i = \frac{\mu_i}{(\mu_1 + \cdots + \mu_M)}.$$ 

▶ We can estimate the parameters in the multinomial model by fitting the Poisson model and using the relationship above.

▶ We can test hypotheses about the multinomial model by testing the equivalent hypothesis in the Poisson regression model.
Death by falling

The Poisson means are

\[
\begin{align*}
\mu_{Jan} &= \exp((\text{Int})) \\
\mu_{Feb} &= \exp((\text{Int}) + \text{monthsFeb}) \\
\mu_{Mar} &= \exp((\text{Int}) + \text{monthsMar}) \\
\ldots\\
\mu_{Dec} &= \exp((\text{Int}) + \text{monthsDec})
\end{align*}
\]
Relationship between Poisson means and binomial probabilities

Recall that

$$\pi_i = \frac{\mu_i}{(\mu_1 + \cdots + \mu_M)}.$$ 

Thus, to calculate the means and probabilities, we can type

```r
> params = coef(falls.glm)
> means = exp(params[1] + c(0, params[2:12]))
> means

    MonthsFeb MonthsMar MonthsApr MonthsMay MonthsJun MonthsJul
  1688   1407   1370   1309   1341   1388   1406

    MonthsAug MonthsSep MonthsOct MonthsNov MonthsDec
  1446   1322   1363   1410   1526

> probs = means/sum(means)
> probs

    MonthsFeb MonthsMar MonthsApr MonthsMay MonthsJun
  0.09943450 0.08288172 0.08070217 0.07710886 0.07899387 0.08176249

    MonthsJul MonthsAug MonthsSep MonthsOct MonthsNov MonthsDec
  0.08282281 0.08517908 0.07787465 0.08028982 0.08305844 0.08989161
```

Note that these are the same as the multinomial MLE’s under the maximal model.
Testing all months equal, Poisson method

- Clearly, all months have equal probabilities if and only if all the means are equal.

- Thus, testing for equal months in the multinomial model is the same as testing for equal means in the Poisson model.

- This is done using the null model deviance, or, equivalently, the anova table.

- Recall: The null deviance was 8.1095e+01 on 11 degrees of freedom, with a p-value of 9.05831e-13, so months are not equally likely.
> anova(falls(glm, test="Chisq")
Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance Resid.</th>
<th>Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>11</td>
<td>81.095</td>
<td>11</td>
<td>81.095</td>
</tr>
</tbody>
</table>

> null(glm = glm(y~1,family=poisson)
> anova(null(glm, falls glm, test="Chisq")
Analysis of Deviance Table
Model 1: y ~ 1
Model 2: y ~ Months
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 11 81.095
2 0 0.000 11 81.095 9.059e-13 ***