Aims of today’s lecture

In today’s lecture we continue our discussion of contingency tables. Topics:

- Today’s R trick
- Simpson’s Paradox
- Collapsing tables
- Adequacy of the chi-square approximation

See also Tutorial 10
# overlaying a density on a histogram

```r
y = rnorm(10000)
hist(y, nclass=50, freq=FALSE)
x = seq(-4,4, length=100)
lines(x, dnorm(x), col="blue", lwd=2)
```
R trick of the day: picture

Histogram of y

Density

-4 -2 0 2 4

0.0 0.1 0.2 0.3 0.4
# estimating Poisson means under a model

# assume model 7:
> model7 = glm(counts~Defendant*Victim+ DP*Victim,  
         family=poisson, data=murder.df)
> means = predict(model7, type="response")
> means

       1       2       3       4
 5.517857 0.482143 97.482143 8.517857
       5       6       7       8
 8.831776 21.16822 54.16822 129.83178
Example: the Florida murder data

If we “collapse” the table over victim, we get the 2x2 table of defendant by death penalty:
Odds of suffering DP for the black defendants are 17/149 Odds of suffering DP for white defendants are 19/149

\[
\text{OR} = \frac{17/149}{19/149} = \frac{141 \times 17}{149 \times 19}/
\]
\[
= 0.847
\]

The odds of suffering the DP are about 15% lower for black defendants than white defendants.
Florida murders: conditional tables

The tables conditional on victim are

<table>
<thead>
<tr>
<th></th>
<th>Death Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defendant</td>
<td>DP=Yes</td>
</tr>
<tr>
<td>Black</td>
<td>6</td>
</tr>
<tr>
<td>White</td>
<td>0</td>
</tr>
</tbody>
</table>

Black Victim: \( \text{OR} = 1.113 \)

<table>
<thead>
<tr>
<th></th>
<th>Death Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defendant</td>
<td>DP=Yes</td>
</tr>
<tr>
<td>Black</td>
<td>11</td>
</tr>
<tr>
<td>White</td>
<td>19</td>
</tr>
</tbody>
</table>

White Victim \( \text{OR} = 1.470 \)

Now odds of being black are greater in the DP group!

NB: we have added 0.5 to the zero cell.
Association Reversal

In our regression lectures, we saw that regression coefficients could change sign when new variables were added to the model. i.e. the coefficient of $x_1$ in the model

$$y \sim x_1$$

can have the opposite sign to the coefficient of $x_1$ in the model

$$y \sim x_1 + x_2$$

See e.g. Lecture 5, Slides 12 and 13.
The reason

- The coefficient of $x_1$ in the model $y \times x_1$ is the slope of the fitted line in the scatter plot, summarising the marginal relationship between $y$ and $x_1$.

- The coefficient of $x_1$ in the model $y \times x_1 + x_2$ is the slope of the line in the coplot, which summarises the relationship between $y$ and $x_1$, conditional on $x_2$.
Marginal and conditional association

- In the first case, we are looking at association between $y$ and $x_1$

- In the second case, we are looking at the association between $y$ and $x_1$, with $x_2$ held fixed (conditional on $x_2$)

- When these are different (of opposite signs) we say *association reversal* has occurred.

- We can do the same sort of thing in contingency tables
Association in contingency tables

- In 2 x 2 tables, we measure association using the odds ratio.

- If we have 3 factors A, B and C, each at 2 levels, we can look at the marginal odds ratio of A and B, using the AxB table, ignoring C.

- Or, the 2 conditional odds ratios, one the AxB table corresponding to C=1, and the other corresponding to C=2.
When will these be the same?

- In ordinary regression, the coefficient of $x_1$ in the model $y \times x_1$ will be the same as the coefficient of $x_1$ in the model $y \times x_1 + x_2$ if and only if $x_1$ and $x_2$ are uncorrelated.

- In contingency tables, the marginal and conditional population OR’s for the AxB tables will be the same if A and C are independent, given B, or if B and C are independent, given A. See Math Sheet 4 on the web page for the mathematical justification of these statements.

- Thus, we can “collapse the table over C” (get the same OR in the marginal and conditional tables) if the ABC interaction is zero, and either the AC interaction is zero, or the BC interaction is zero.

- If this is not the case, association reversal (Simpson’s paradox) may occur. The more associated A and/or B is with C, the more likely it is to happen.
Florida data

- Very strong association between DP and race of victim: if victim is black, small chance of DP

- Very strong association between race of victim and race of defendant

- Since blacks murder blacks, (and whites murder whites), not so many DPs for black defendants overall
When can we collapse

Thus, we can collapse the table over C if

- The ABC interaction is zero, and
- Either the AC interaction is zero, or The BC interaction is zero.

If these conditions are not met, then the marginal and conditional tables may have different (even reversed) degrees of association.
Florida data

Recall that there are significant interactions between

▶ Race of victim and race of defendant
▶ Race of victim and death penalty

Thus, marginal and conditional OR’s will be different

> anova(murder.glm, test="Chisq")

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>7</td>
<td>395.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defendant</td>
<td>1</td>
<td>0.110</td>
<td>6</td>
<td>395.80</td>
</tr>
<tr>
<td>DP</td>
<td>1</td>
<td>225.419</td>
<td>5</td>
<td>170.39</td>
</tr>
<tr>
<td>Victim</td>
<td>1</td>
<td>32.456</td>
<td>4</td>
<td>137.93</td>
</tr>
<tr>
<td>Defendant:DP</td>
<td>1</td>
<td>0.221</td>
<td>3</td>
<td>137.71</td>
</tr>
<tr>
<td>Defendant:Victim</td>
<td>1</td>
<td>129.798</td>
<td>2</td>
<td>7.91</td>
</tr>
<tr>
<td>DP:Victim</td>
<td>1</td>
<td>7.209</td>
<td>1</td>
<td>0.70</td>
</tr>
<tr>
<td>Defendant:DP:Victim</td>
<td>1</td>
<td>0.701</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Marginal or conditional: Which should we use?

- If we are trying to investigate a causal relationship between factors A and B (say if A causes B), and C is a confounding factor, we should always condition on C (i.e. hold C fixed when exploring the connection between A and B).

- Otherwise, the effect of C on A and/or B is not recognised, and the effect of C on B will be ascribed to A.
In graphical form

- Possible association reversal if we collapse over A
- Collapsing over B and C OK.
Admission data from UC Berkeley grad school:

<table>
<thead>
<tr>
<th>Dept</th>
<th>Female Admitted</th>
<th>Male Admitted</th>
<th>Female Gender</th>
<th>Male Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>89</td>
<td>512</td>
<td>19</td>
<td>313</td>
</tr>
<tr>
<td>B</td>
<td>17</td>
<td>353</td>
<td>8</td>
<td>207</td>
</tr>
<tr>
<td>C</td>
<td>202</td>
<td>120</td>
<td>391</td>
<td>205</td>
</tr>
<tr>
<td>D</td>
<td>131</td>
<td>138</td>
<td>244</td>
<td>279</td>
</tr>
<tr>
<td>E</td>
<td>94</td>
<td>53</td>
<td>299</td>
<td>138</td>
</tr>
<tr>
<td>F</td>
<td>24</td>
<td>24</td>
<td>317</td>
<td>317</td>
</tr>
</tbody>
</table>

OR's are (Odds admission female/odds admission male):

<table>
<thead>
<tr>
<th>Dept</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.86</td>
</tr>
<tr>
<td>B</td>
<td>1.25</td>
</tr>
<tr>
<td>C</td>
<td>0.88</td>
</tr>
<tr>
<td>D</td>
<td>1.09</td>
</tr>
<tr>
<td>E</td>
<td>0.82</td>
</tr>
<tr>
<td>F</td>
<td>1.21</td>
</tr>
</tbody>
</table>
Collapse over departments

- \( OR = 0.54 \) (odds admission female / odds admission male)
- Seems to be strong evidence of bias against women
- Reason: females apply for programs that have low admission rates, males for programs that have high admission rates, so strong dependence between dept and the other 2 factors

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Yes 557</td>
</tr>
<tr>
<td>Male</td>
<td>Yes 1198</td>
</tr>
</tbody>
</table>
> berkeley.glm = glm(counts~Department*Gender*Admitted, + family=poisson,data=berkeley.df)
> anova(berkeley.glm, test="Chisq")

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>23</td>
<td>2650.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Department</td>
<td>5</td>
<td>159.52</td>
<td>18</td>
<td>2490.57</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>162.87</td>
<td>17</td>
<td>2327.70</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Admitted</td>
<td>1</td>
<td>230.03</td>
<td>16</td>
<td>2097.67</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Department:Gender</td>
<td>5</td>
<td>1220.61</td>
<td>11</td>
<td>877.06</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Department:Admitted</td>
<td>5</td>
<td>855.32</td>
<td>6</td>
<td>21.74</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Gender:Admitted</td>
<td>1</td>
<td>1.53</td>
<td>5</td>
<td>20.20</td>
<td>0.215928</td>
</tr>
<tr>
<td>Department:Gender:Admitted</td>
<td>5</td>
<td>20.20</td>
<td>0</td>
<td>0.00</td>
<td>0.001144 **</td>
</tr>
</tbody>
</table>

Significant 3 factor interaction, so no conditional independence, collapsibility conditions not satisfied.
Moral: Collapsing tables

- Dangerous to collapse tables when factor being collapsed over (e.g., victim’s race, Berkeley department) is strongly associated with the other factors.

- Marginal (collapsed) tables may give a very misleading picture, need to look at conditional tables.
How good is Chi-square?

- The Chi-square distribution used to compute p-values from the residual deviance is only an approximation to the true deviance distribution.

- Approximation is good when cell counts are large and the number of cells is small.

- Conventional wisdom is that cells counts should be at least 5. However, this is quite conservative.

- We will use R to investigate this using a simulation approach.
Example

Suppose we have a $2 \times 4$ table of probabilities shown below, corresponding to a cross-classification of two factors A and B. Thus, the probability that a randomly chosen individual is in category 1 of A and category 2 of B is 0.0299.

<table>
<thead>
<tr>
<th>Category A</th>
<th>Category B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B=1</td>
</tr>
<tr>
<td>A=1</td>
<td>0.0600</td>
</tr>
<tr>
<td>A=2</td>
<td>0.3039</td>
</tr>
<tr>
<td>A=3</td>
<td>0.1040</td>
</tr>
</tbody>
</table>
Suppose we select 100 individuals from this population and classify them in a $3 \times 4$ contingency table:

<table>
<thead>
<tr>
<th>Category A</th>
<th>Category B</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B=1</td>
<td>B=2</td>
<td>B=3</td>
<td>B=4</td>
</tr>
<tr>
<td>A=1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A=2</td>
<td>42</td>
<td>18</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>A=3</td>
<td>14</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
In actual fact, with the probabilities shown, A and B are independent.

Suppose we were to test the independence of A and B by fitting an additive model (i.e. no interactions) and examining the residual deviance. We would calculate a p-value using a chi-square distribution with \( 6 = (3 - 1) \times (4 - 1) \) degrees of freedom.

How accurate would the p-value be? How closely does the chi-squared distribution match the true distribution of the deviances? After all, some of the cell counts are small.

Since we know the probabilities, we can answer these questions by simulating a large number of deviances and comparing their distribution to a chi-square.
Simulation procedure

- Use the function `rmultinom` to simulate a contingency table from the given probabilities.
- Fit the independence model and record the residual deviance.
- Repeat say 10,000 times.
- Compare the distribution of the 10,000 deviances to the chi-square distribution with 6 degrees of freedom.
R code: simulating

Assume probabilities are in a 3 x 4 matrix pimat
Ntot = 100  # 100 individuals classified
Nsim=10000  # 10,000 simulations
devs=numeric(Nsim) make a place to store the results
# now create factors A and B in a data frame
factors = expand.grid(A=factor(1:3), B=factor(1:4))
# generate all the count data in one hit
X = rmultinom(Nsim, Ntot, pimat)
# now loop
for(i in 1:Nsim){
    counts = X[,i]
data.df= data.frame(counts,factors)
devs[i] = glm(counts~A+B, family=poisson,
              data=data.df)$deviance
}

The vector devs now contains the 10,000 deviances.
R code: comparing using a histogram

```r
# draw a histogram
hist(devs, nclass=50, freq=FALSE,
     main = "Histogram of deviances",
    xlab = "deviances", ylim=c(0,0.14))
# overlay the chi-squared 6 density function
x = seq(0,25, length=100)
lines(x, dchisq(x,6), lwd=2, col="blue")
```
The picture

Histogram of deviances

Density

Deviances
Thus, to see if a sample comes from a particular distribution, we

- Calculate the sample quantiles and the distribution quantiles for a large number of selected values of alpha.
- If the two sets of quantiles agree, it means that we have used the correct distribution in our calculation.
- If they agree, plotting the sample quantiles against the distribution quantiles should give a plot clustering around 45 degree line.
- Such a plot is called a Q-Q (Quantile-Quantile) plot.
R code: comparing using quantiles

alphas = ((1:Nsim)-0.5)/Nsim  # make 10,000 alpha values
quantiles = qchisq(alphas,6)  # calculate chi-square quantiles
# now plot. Sort(dev) gives the sample
# quantiles for this set of alpha's
plot(sort(devs), quantiles, xlab="order statistics",
ylab="quantiles")
# draw on 45 degree line
abline(0,1, lwd=2, col="red")
The Q-Q plot

Close but not the same!