However, be careful of zero or small \( y_{ijk} \) \( \Rightarrow \) there will be some doubt as to the accuracy of the chi-square approximation in goodness-of-fit test

- The chi-square approximation is better in comparing model than assessing goodness-of-fit

Analysis strategy: start with complex Poisson GLM (such as saturated model) and see how far the model can be reduced (by using deviance-based test to compare models).

- Binomial (multinomial) GLM approach for 3-way table

  When \( y_{ijk} \)'s are regarded as fixed, we can treat \( Y_{X_3} \) as a response and \( X_1, X_2 \) as covariates

  \( Q_1 \): what information gone? \( Q_2 \): what information still attainable?

  - **Ans** for \( Q_1 \): information about \( \pi_{ij} \)
  - **Ans** for \( Q_2 \): information about \( \pi_{k|ij} = \pi_{k|ij}^{++} \)

  \( Y_{X_3} = y_{ij1} \sim \text{binomial}(y_{ij+}, \pi_{k=1|ij}) \) if \( K=2 \)
  
  \[ Y_{X_3} = (y_{ij1}, \ldots, y_{ijK}) \sim \text{multinomial}(y_{ij+}, \pi_{k=1|ij}, \ldots, \pi_{k=K|ij}) \] if \( K > 2 \)

- **Q**: how is a binomial GLM connected to a Poisson GLM?

  - \( Y_{X_3} \sim 1 \Leftrightarrow S_2 \) (joint independent)
  
  \[ \pi_{ijk} = \pi_{ij} + \pi_{++k} \Leftrightarrow \pi_{k|ij} = \pi_{++k} \]

  - The binomial GLM implicitly assumes an association between \( X_1 \) and \( X_2 \) (**Q**: why?)
  
  - Poisson GLM allows us to drop the \( X_1 : X_2 \) term, but binomial GLM does not

  Using binomial GLM loses little when we are interested in the relationship between the response \( X_3 \) and the two predictors \( X_1, X_2 \), and not interested in the association between \( X_1 \) and \( X_2 \) (**Q**: why?)

  - \( Y_{X_3} \sim X_1 \Leftrightarrow X_2, X_3 \) are independent given \( X_1 \) (**Q**: \( Y_{X_3} \sim X_2 ? \))
  
  - \( Y_{X_3} \sim X_1 + X_2 \Leftrightarrow S_4 \) (uniform association)

  - **Q**: why is \( Y_{X_3} \sim X_1 + X_2 + X_1 : X_2 \) not associated with \( S_4 \)?

  - The saturated binomial GLM corresponds to a Poisson GLM for different association

- **Q**: Poisson or binomial GLM approach?

  - Binomial when one variable is clearly identified as the response
  
  - Poisson when relationship between 3 variables is more symmetric
• Correspondence analysis
  ➢ Cannot directly apply to 3-way table
  ➢ Can combine two of the factors, say $X_1$ and $X_2$, into a factor with $IJ$ levels and apply correspondence analysis on the 2-way table formed by the new factor and $X_3$
  ➢ Q: which two factors should be chosen to merge? Ans: pick up the two whose association is least interesting to us

• Simpson’s paradox
  ➢ example:

<table>
<thead>
<tr>
<th></th>
<th>smoker dead</th>
<th>alive</th>
<th>smoker dead</th>
<th>alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>14</td>
<td>95</td>
<td>109 (47)</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>7</td>
<td>114</td>
<td>127 (53)</td>
<td></td>
</tr>
<tr>
<td>age=35-44 younger</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>29</td>
<td>7</td>
<td>36 (22)</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>101</td>
<td>28</td>
<td>129 (78)</td>
<td></td>
</tr>
<tr>
<td>age=65-74 older</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>43</td>
<td>102</td>
<td>145 (37)</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>108</td>
<td>142</td>
<td>250 (63)</td>
<td></td>
</tr>
</tbody>
</table>

⇒ marginal total over age

<table>
<thead>
<tr>
<th></th>
<th>smoker dead</th>
<th>alive</th>
<th>smoker dead</th>
<th>alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>13</td>
<td>.87</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>.06</td>
<td>.94</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>age=35-44 younger</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>.81</td>
<td>.19</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>.78</td>
<td>.22</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>age=65-74 older</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>.30</td>
<td>.70</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>.43</td>
<td>.57</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

marginal association added over $X_3$ is different from the conditional association observed within each category of $X_3$

Q: Why it occurs? why $y_{ijk}$ give a contradictory result to $y_{ijk}'$?

Note that smoker are more concentrated in the younger age group and younger people are more likely to live longer

• Mantel-Haenszel test for $2 \times 2 \times K$ table
  ➢ Designed to test independence in $2 \times 2$ tables across $K$ categories

Note that the association of a $2 \times 2$ table can be completely characterized/measured by its odd-ratio $\Delta$

$\Delta = 1 \Leftrightarrow$ independence
$\Delta > 1 \Leftrightarrow$ positive association
$\Delta < 1 \Leftrightarrow$ negative association

made by Shao-Wei Cheng (NTHU)
Null and alternative hypotheses of Mantel-Haenszel test

- $H_0$: $\Delta_1 = \Delta_2 = \ldots = \Delta_K = 1$ (conditional independence)
- $H_1$: $\Delta_1 = \Delta_2 = \ldots = \Delta_K \neq 1$ (uniform association)

- It only makes sense to use the test when the odds ratios of the $K$ $2 \times 2$ tables do not vary greatly, i.e., the null of the test for uniform association, $\Delta_1 = \Delta_2 = \ldots = \Delta_K$, has been accepted.

Procedure of the test

- Suppose that the marginal totals for each $2 \times 2$ table carry no information, e.g., marginal totals are fixed in advance.
  - Under $H_0$, can assume a hypergeometric distribution in each table $\Rightarrow y_{0k}$ is sufficient for each table.
- Test statistic combine information, i.e., $y_{11k}$, from $K$ tables:
  
  \[
  (\sum_k y_{11k} - \sum_k E(y_{11k})) - \frac{1}{2}\frac{1}{\sum_k \text{var}(y_{11k})} \sim \chi^2
  \]

  where $E(y_{11k})$ and $\text{var}(y_{11k})$ are calculated under the $H_0$.

  - Can calculate an exact $p$-value for smaller dataset $\Rightarrow$ useful when data is sparse under which the $\chi^2$ approximations in tests based on GLM are questionable.
  - A version without the $1/2$ continuity correction is called the Cochran-Mantel-Haenszel statistic.

Reading: Faraway, 4.4

**Ordinal Variables**

- Some variables have a nature ordering between categories.
  - E.g., education: HS, BA, MA; political ideology: VL, SL, M, SC, VC
  - For ordinal variables, can use the methods for nominal variable.
    - But, more information can be extracted by taking advantage of the ordinal structure.
    - The ordinal structure matters only when # of categories $> 2$.
  - Treatments for ordinal response (in future lecture) and ordinal predictors are different.

- Treatment for ordinal predictors: assign each category a score.
  - It transforms an ordinal variable into a continuous variable.
  - The choice of scores requires some judgment.
    - If no particular preference, even spacing allows for the simplest interpretation.
    - For interval scales, midpoints of the intervals are often used.
  - Should check whether the inference is robust to different assignments of scores.
    - If qualitative conclusions are changed, this is an indication that you cannot make any strong finding.
• Linear-by-linear association Poisson GLM for 2-way tables:
  ➢ Consider a 2-way table with ordinal row and column variables:
     Assign scores $u_1 \leq u_2 \leq \ldots \leq u_I$ to rows, denoted by $u(X_1)$
     Assign scores $v_1 \leq v_2 \leq \ldots \leq v_J$ to columns, denoted by $v(X_2)$
  ➢ Model:
    $$\log(\mu_{ij}) = \log(t\pi_{ij}) = \log(t) + \log(\pi_{i+}) + \log(\pi_{+j}) + \gamma_i v_j$$
    where $u_i v_j$’s are known scores, $\gamma$ is an unknown parameter.

Some notes about $\gamma$:
- the values of $\gamma$ represents the amount of association and can be positive or negative
- $\gamma=0$ means independence
- Interpretation of $\gamma$: $\log \left( \frac{\mu_{ij+1,i+1}}{\mu_{ij,i+1} \mu_{ij+1,j+1}} \right) = \log(\pi_{i+1,j+1}) - \log(\pi_{i,j+1}) - \log(\pi_{i+1,j}) + \gamma_i v_j$
  - for evenly spaced scores, these log-odds-ratios are equal
    called uniform association in Goodman (1979)
- Latent (continuous) variable motivation for $\gamma$:
  - Assume $\pi_{ij}$’s are obtained by putting a grid on a bi-variate Normal for latent variables and $u_i$’s and $v_j$’s are cutpoints
  - $\gamma$ can then be identified with the correlation coefficient $\rho$ of the latent variables (c.f., positive and negative $\rho$)

➢ Q: For the test of independence, what is the benefit of $S_{O \times O}$ over the nominal approach, i.e., fitting a nominal-by-nominal model $Y \sim X_1 + X_2 + u(X_1) v(X_2)$? As shown in a lab example,
  - In the $N \times N$ approach, interaction effects reduce a deviance of 40.743 on 36 degrees of freedom, but
  - What if there exists an interaction effect which can reduce a deviance of 10.175 on one degree of freedom, and other 35 interaction effects for the rest deviance, i.e., 30.568

• Ordinal-by-nominal model (or nominal-by-ordinal model)
  ➢ column (or row) variable treated as a nominal variable
  ➢ It is called column (or row) effects model because the columns (or row) are not assigned scores; instead, their effects are estimated
  - An alternative viewpoint for ordinal column variable: column (or row) scores are regarded as parameters
  ➢ Column effects model:
    $$\log(\mu_{ij}) = \log(t\pi_{ij}) = \log(t) + \log(\pi_{i+}) + \log(\pi_{+j}) + \gamma_i v_j$$
    where $u_i$’s, $i=1,\ldots, I$, are known scores and $\gamma_j$’s, $j=1,\ldots, J-1$, are unknown parameters.
    $$Y \sim X_1 + X_2 + u(X_1): X_2 \equiv S_{O \times N} (\supset S_{O \times O})$$
Some notes about $\gamma_j$’s (which are called the column effects):

- Equality of the $\gamma_j$’s corresponds to the hypothesis of independence between $X_1$ and $X_2$. $U_i \cdot \gamma_j = U_i \cdot \gamma_j \Rightarrow$ main effect.

- For ordinal column variable, if the model $S_{O \times O}$ were a good fit, we would expect the estimates of the $\gamma_j$’s in $S_{O \times N}$ to be roughly proportional to $v_j$’s (e.g., for evenly spaced $v_j$’s, estimates of $\gamma_j$’s should follow a linear trend).

- We can use the estimates of $\gamma_j$’s in $S_{O \times N}$ to examine whether the chosen scores for columns in $S_{O \times O}$ (i.e., $v_j$’s) are appropriate and to possibly suggest better scores (see an example in lab).

- Row effect model is effectively the same model except with the roles of the variables reversed.

Advantages of using scores for ordinal variables:

- It can be helpful in reducing the complexity of models for categorical data with ordinal variables.
- It is especially useful in higher dimensional table where a reduction in the number of parameters is particularly welcome.
- It can also sharpen our ability to detect associations.

Reading: Faraway, 4.5