Two-way Contingency Table

- Two cross-classified categorical variables $X_1$ and $X_2$
  - $X_1$ has $I$ categories, denoted by $i = 1, 2, \ldots, I$
  - $X_2$ has $J$ categories, denoted by $j = 1, 2, \ldots, J$

- Classifications of subjects in some population on $X_1$ and $X_2$ have $IJ$ possible combinations.

Define the population parameters:

- $\pi_{ij} = \text{the proportion of the subjects in the population with } X_1 = i \text{ and } X_2 = j$
  - arrange $\pi_{ij}$'s in the cells of a rectangular table having $I$ rows for categories of $X_1$ and $J$ columns for categories of $X_2$ to display the population distribution
  
  $\pi_{i+} = \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij} \Rightarrow \text{marginal proportion}$
  
  $\pi_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = \sum_{i=1}^{I} \pi_{i+} = \sum_{j=1}^{J} \pi_{+j} = 1$

- $\pi_{i|j} = \frac{\pi_{ij}}{\pi_{+j}} \text{ and } \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \Rightarrow \text{conditional proportion}$
  
  $\Rightarrow \sum_{i=1}^{I} \pi_{i|j} = 1, \forall j \text{ and } \sum_{j=1}^{J} \pi_{j|i} = 1, \forall i$

- **Q:** For the population, what questions might be of interest?
  
  $\pi_{1+} = \cdots = \pi_{I+}$? or $\pi_{+1} = \cdots = \pi_{+J}$?

- Are $X_1$ and $X_2$ observed from a randomly sampled subject independent, i.e., does $X_1$ affect $X_2$ and vice versa?
  
  If $X_1$ and $X_2$ are independent, then

  - $\pi_{ij} = P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j) = \pi_{i+} \pi_{+j}$
  - $\pi_{i|j} = P(X_1 = i | X_2 = j) = P(X_1 = i) = \pi_{i+}, \forall j$
  - $\pi_{j|i} = P(X_2 = j | X_1 = i) = P(X_2 = j) = \pi_{+j}, \forall i$
  - $\pi_{11}; \cdots; \pi_{1J} = \pi_{21}; \cdots; \pi_{2J} = \cdots = \pi_{I1}; \cdots; \pi_{IJ}$
  - $\pi_{11}; \cdots; \pi_{I1} = \pi_{12}; \cdots; \pi_{1J} = \cdots = \pi_{I1}; \cdots; \pi_{IJ}$

  - For $2 \times 2$ table, odd ratio
    
    $\Delta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\pi_{11} \times \pi_{22}}{\pi_{12} \times \pi_{21}} = 1$
For $I \times J$ table and any $1 \leq i < I$ and $1 \leq j < J$,

$$\frac{\Delta_{ij}}{\sum_{i,j} \pi_{ij} \times \pi_{I,J}} = 1$$

• For a sample drawn from the population, let
  $y_{ij}$ = total number of subjects in the sample with $X_1 = i$ and $X_2 = j$.
  marginal totals (row totals or column totals)
  $y_{i+} = \sum_{j=1}^{J} y_{ij}$ and $y_{+j} = \sum_{i=1}^{I} y_{ij}$
  grand total $y_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} = \sum_{i=1}^{I} y_{i+} = \sum_{j=1}^{J} y_{+j}$

• When the cells of the rectangular table contain $y_{ij}$'s, it is called a $I \times J$ contingency table.

• The above treatments for $\pi$'s and $y$'s can be generalized to more than two categorical variables.

• Q: how to model the data (i.e., what's the joint distribution of $y_{ij}$'s)?
  The statistical modeling of the data depends on the sampling schemes.

Consider an example of wafer data:

Consider the sampling schemes:

1. Observe the manufacturing process for a certain period of time
2. Decide to sample 450 wafers
3. Decide to sample 400 wafers without particles and 50 wafers with particles
4. Scheme 3 and the 450 wafers must also include, by design, 334 good wafers and 116 bad ones

- Note 1: the first three schemes are all plausible
- Note 2: scheme 4 seems less likely in this example; such a scheme is more attractive when one level of each variable is relatively rare and we choose to over-sample both levels to ensure some representation

- Scheme 1
  Model: $y$: fixed; $Y$: random; red square: free

<table>
<thead>
<tr>
<th>Quantity</th>
<th>No Particles</th>
<th>Particles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>320</td>
<td>14</td>
</tr>
<tr>
<td>Bad</td>
<td>80</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>50</td>
</tr>
</tbody>
</table>
Suppose the data (from an \( I \times J \) table) is fitted with a Poisson GLM with log link

- When \( \pi_{ij} = \pi_i + \pi_j \) (\( X_1 \) and \( X_2 \) independent),
  \[
  \eta_{ij} = \log(\mu_{ij}) = \log(t \pi_{ij}) = \log(t \pi_i + \pi_j) = \log(t) + \log(\pi_i) + \log(\pi_j)
  \]
  \[ \Rightarrow \text{corresponds to a main-effect model, i.e., } Y_{ij} \sim X_1 + X_2 = S \]

- When \( \pi_{ij} = \pi_i + \pi_j \) and \( \pi_1 = \cdots = \pi_I = \pi_1 = \cdots = \pi_J \)
  \[ \eta_{ij} = \log(t) + \log(\pi_j) \text{ (or } \eta_{ij} = \log(t) + \log(\pi_i) \text{)} \]
  \[ \Rightarrow \text{corresponds to the model } Y_{ij} \sim X_2 \text{ (or } Y_{ij} \sim X_1 \text{)} \]

- When \( \pi_{ij} \neq \pi_i + \pi_j \) (\( X_1 \) and \( X_2 \) not independent)
  \[ \Rightarrow \text{add interaction } X_1 : X_2 \]
  \[ \Rightarrow \text{may consider } Y_{ij} \sim X_1 + X_2 + X_1 : X_2 = L \text{ (saturated model)} \]

**Q:** what type of \( \pi \)'s corresponds to the following models?

- \( Y_{ij} \sim 1 \)
- \( Y_{ij} \sim X_1 + X_2 + X_1 : X_2 \)
- \( Y_{ij} \sim X_1 + X_2 \)

Recall. For a Poisson GLM with log link, \( X^T Y = X^T \hat{\mu} \)

For models without interactions,

\[ \Rightarrow X^T Y \text{ is only related to marginal totals} \]

\[ \Rightarrow \text{the fitted values } \hat{\mu} \text{ is a function of marginal totals} \]

\[ \Rightarrow \text{for example, for main-effect model } Y_{ij} \sim X_1 + X_2 \]

\[ \hat{\mu}_{ij} = \hat{\pi}_{++} \hat{\pi}_{i+} \hat{\pi}_{+j} = \frac{Y_{+} Y_{+j}}{Y_{++}} \]

\[ \Rightarrow \text{To test whether } \pi_{ij} = \pi_i + \pi_j (H_0) \Rightarrow H_0: S \text{ vs. } H_1: L \backslash S \]

- Deviance based:
  \[ D_S - D_L \overset{a}{\sim} \chi^2_{(I-1)(J-1)} \]
Pearson’s $X^2$ (goodness-of-fit measure) under $S$:
\[
X^2_S = \sum_{i,j} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2(I-1)(J-1)
\]

- **Yate’s continuity correction**:
  - Subtract 0.5 from $Y_{ij} - \hat{\mu}_{ij}$ when it is positive
  - Add 0.5 to $Y_{ij} - \hat{\mu}_{ij}$ when it is negative
  this give superior results for small samples

- **To test $H_0$:** $\pi_{1+} = \cdots = \pi_{I+}$ (or $\pi_{+1} = \cdots = \pi_{+J}$),
  compare models $S^*$ and $L^*$ ($H_0$: $S^*$ vs. $H_1$: $L^* \setminus S^*$), where

  $\begin{align*}
  S^*: & Y_{ij} \sim X_1 + X_2 \quad \text{and} \quad L^*: Y_{ij} \sim X_1 + X_2 + X_1 \cdot X_2 \\
  S^*: & Y_{ij} \sim X_1 \cdot X_2 \quad \text{and} \quad L^*: Y_{ij} \sim X_1 + X_1 \cdot X_2 \\
  S^*: & Y_{ij} \sim 1 \quad \text{and} \quad L^*: Y_{ij} \sim X_1
  \end{align*}$

- **Deviance-based test**: $D_{S^*} - D_{L^*} \overset{a}{\sim} \chi^2_{df_{S^*} - df_{L^*}}$

  - Can be generalized to $X_1$ with $I$ levels and $X_2$ with $J$ levels

  - **Scheme 2**:

    - **Model**: for a random sample, we can assume
      \[
      (Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{multinomial}(y_{++}, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})
      \]
      where $\pi_{ij}$ ($i=1, 2$; $j=1, 2$) is linked to $X_1$ and $X_2$ according to the model we choose

    - **Connection between Poisson and multinomial**:
      Let $Y_i \sim \text{Poisson}(\lambda_i)$, $i=1,\ldots,k$, and independent,
      \[
      (Y_1, \ldots, Y_k | \Sigma Y_i = n) \sim \text{multinomial}(n, \lambda_1/\Sigma \lambda_i, \ldots, \lambda_k/\Sigma \lambda_i)
      \]
      - the parameter $t$ (value of size variable) in Poisson is removed, but $\pi_{ij}$'s are not affected
      - would expect there is a lot of similarity between the inferences for Poisson and multinomial models

    - **Log-likelihood of the multinomial**:
      \[
      \log(L) \propto \sum_{i,j} Y_{ij} \log(\pi_{ij})
      \]
      (cf., log-likelihood for Poisson $\propto \sum_{i,j} Y_{ij} \log(\mu_{ij}) - \mu_{ij}$)
The inferences in the multinomial model would coincide with that in Poisson model, i.e.,
- same estimates (MLE)
- same test statistics and p-values

The Poisson model is easier to execute in R, so we can fit a Poisson GLM for data from a multinomial sampling scheme.

Can be generalized to $I \times J$ table in the same manner.

- Scheme 3:

  - Model: for a random sample, can assume
    \[
    Y_{1j} \sim \text{binomial}(y_{+j}, \pi_{i=1|j} = \frac{\pi_{1j}}{\pi_{+j}}), \quad j = 1, 2
    \]
    where $\pi_{i=1|j}$ is linked to the covariate $X_2 (=j)$ only according to the model we choose.

  - $Q$: compared to schemes 1 and 2, what information has been gone/questionable in this scheme?

  - Suppose fit the data with a Binomial GLM with logit link:
    \[
    Y_{1j} \sim \text{binomial}(y_{+j}, \pi_{i=1|j} = \frac{\pi_{1j}}{\pi_{+j}}), \quad j = 1, 2
    \]
    \[
    \Rightarrow \text{corresponds to a constant-effect model, i.e., } Y_{1j} \sim 1 \equiv S
    \]
    \[
    \Rightarrow Y_{1j} \sim X_2 \equiv L \text{ (saturated model)}
    \]
    \[
    \Rightarrow \text{To test whether } \pi_{ij} = \pi_{i+} \pi_{+j} \text{ (} X_1 \text{ and } X_2 \text{ independent)}
    \]
    \[
    \Rightarrow \text{Deviance based: } D_S - D_L = D_S
    \]
    \[
    \Rightarrow \text{Pearson } X^2 \text{ under } S
    \]
    \[
    \Rightarrow \text{Can be generalized to } X_2 \text{ with } J > 2 \text{ levels in the same manner}
    \]
    \[
    \Rightarrow \text{For the case that } X_1 \text{ has } I > 2 \text{ levels}
    \]
    \[
    (Y_{11}, \ldots, Y_{1J}) \sim \text{multinomial}(y_{+j}, \pi_{i=1|j}, \ldots, \pi_{i=I|j}), \quad j = 1, \ldots, J
    \]
    \[
    \Rightarrow \text{called product multinomial model (cf., unrestricted multinomial model in scheme 2)}
    \]
• Scheme 4:

Model: if \( \pi_{ij} = \pi_{i+} \pi_{+j} (H_0) \), for a random sample,

\[
Y_{11} \sim \text{hypergeometric}(y_{1+}, y_{+1}, y_{+2}), \text{ i.e.,} \]

\[
P(Y_{11} = y_{11}) = \frac{\binom{y_{1+}}{y_{11}} \binom{y_{+2}}{y_{12}}}{{\binom{y_{+1}}{y_{11}}}}
\]

\[
y_{11} \leq \min\{y_{+1}, y_{+2}\} = \frac{y_{11}! \cdot y_{12}! \cdot y_{21}! \cdot y_{22}! \cdot y_{++}!}{y_{11}! \cdot y_{12}! \cdot y_{21}! \cdot y_{22}! \cdot y_{11}!}
\]

- Under Scheme 3 and \( H_0 \), the joint pmf of \((Y_{11}, Y_{12}, Y_{21}, Y_{22})\) is:

\[
\frac{\binom{y_{1+}}{y_{11}} \pi_{1+} y_{11} \pi_{21} y_{21} \times \binom{y_{+2}}{y_{12}} \pi_{1+} y_{12} \pi_{22} y_{22}}{\binom{y_{+1}}{y_{11}} \pi_{1+} y_{11} + \pi_{21} y_{21} + \pi_{22} y_{22}}
\]

- Under Scheme 3 and \( H_0 \), the sufficient statistics of \( \pi_{1+} \) and \( \pi_{2+} \) are \( Y_{1+} \) and \( Y_{2+} \), respectively, and their joint pmf is:

When \( \pi_{ij} \neq \pi_{i+} \pi_{+j} (H_1) \), the probability a black ball is drawn is different from the probability a white ball is drawn.

To test whether \( \pi_{ij} = \pi_{i+} \pi_{+j} \) (Fisher’s exact test)

- Because \( Y_{11} \) can only take a limited number of values, can compute the probability of all these outcomes under \( H_0 \) ⇒ can compute the total probability (p-value) of all outcomes that are more extreme than the one observed

- **Q**: what outcomes are more extreme? Some options:
  - The outcomes with probability \( \leq P(Y_{11} = y_{11}) \)
  - Outcomes \( y_{11}' \) s.t. \( |y_{11}' - E(Y_{11})| \geq |y_{11} - E(Y_{11})| \)
  - Others (see Agresti, 2013, 3.5)

Generalization to \( I \times J \) table for testing \( H_0: \pi_{ij} = \pi_{i+} \pi_{+j} \)

⇒ use multiple hypergeometric as null distribution, whose probability mass function is:

\[
\left(\prod_i y_{i+!}\right) \left(\prod_j y_{+j!}\right) / (y_{++!} \times \prod_{ij} y_{ij!})
\]

Some notes:

- The situation that both marginal totals are fixed is rather less common in practical sampling applications
It can arise when classifying objects into one of 2 types when the true proportions of each type are known
- Example: the lady tasting tea

- It suggests a more accurate test for independence
  - Fisher’s exact test is attractive because the null distribution for deviance-based and Pearson’s $X^2$ test statistics is only approximately $\chi^2$ distributed.
  - For tables with small counts, this $\chi^2$ approximation is suspicious, which makes the exact method valuable.
  - Fisher’s exact test becomes more difficult to compute for larger tables. However, the $\chi^2$ approximation will tend to be accurate for larger tables.

- Reading: Faraway (2006, 1st ed.), 4.1, 4.2

**Correspondence Analysis (CA)**

- Q: when independence of a 2-way contingency table is rejected, how to know where the dependence is coming from?

  ➢ Interaction terms in a Poisson GLM contain dependence information; however, interpretation of them could be difficult.

**CA: a visual residual analysis for contingency table**

- Singular value decomposition

  ➢ $R=\begin{bmatrix} R_{ij} \end{bmatrix}$: an $r \times c$ matrix. W.l.o.g, assume $r \geq c$ and $\text{rank}(R)=c$, then

  \[
  R = U_{r \times c} D_{c \times c} V_{c \times c}^T = \sum_k d_k U_k V_k^T, \text{ i.e.,}
  \]

  \[
  R_{ij} = \sum_{k=1}^c U_{ik} d_k V_{jk}, \text{ where}
  \]

  - $U=\begin{bmatrix} U_{ij} \end{bmatrix}=\begin{bmatrix} U_1, \ldots, U_c \end{bmatrix}$: an $r \times c$ column orthonormal matrix, i.e., $U^TU=I_{r \times c}$; its columns are called left singular vectors
  
  - $V=\begin{bmatrix} V_{ij} \end{bmatrix}=\begin{bmatrix} V_1, \ldots, V_c \end{bmatrix}$: a $c \times c$ column orthonormal matrix, i.e., $V^TV=I_{c \times c}$; its columns called right singular vectors
  
  - $D=\text{diag}(d_1, \ldots, d_c)$, $d_1 \geq \ldots \geq d_c > 0$, called singular values

  ➢ Some properties

  - Columns of $U_{r \times c}$ are eigenvectors of $(RR^T)_{r \times r}$
  - Columns of $V_{c \times c}$ are eigenvectors of $(R^TR)_{c \times c}$
  - $\{d_1^2, \ldots, d_c^2\}$ are eigenvalues of $RR^T$ and $R^TR$

- Procedure of correspondence analysis on Pearson residuals
a) Fit a GLM corresponding to independence on the contingency table and compute its Pearson residuals, \( r_P \)'s (\( Q: \) what information contained in the \( r_P \)'s?)
b) Write \( r_P \)'s in the matrix form \([R_{ij}] = R_{r \times c}\) as in contingency table
c) Perform the singular value decomposition on \( R = UDV^T \)
d) It is common for the first few singular values of \( R \) to be much larger than the rest. \textit{Suppose that the first 2 dominate}. Then,
\[
R_{ij} \approx U_{i1} d_1 V_{j1}^T + U_{i2} d_2 V_{j2}^T = (U_{i1} \sqrt{d_1})(V_{j1} \sqrt{d_1}) + (U_{i2} \sqrt{d_2})(V_{j2} \sqrt{d_2}) \equiv U_{i1}' V_{j1}' + U_{i2}' V_{j2}'
\]

- Some notes:
  - \( Q: \) what should we look for in a correspondence plot?
  - \( Q: \) what does a large positive \( R_{ij} \) mean? a large negative \( R_{ij} \)?
  - \( \sum_k d_k^2 = \text{Pearson's } X^2 \), because \( \sum_{ij} r_P^2 = \text{trace}(R^T R) = \sum_k d_k^2 \) (\( Q: \) What does a large \( \sum_k d_k^2 \) indicate?)
  - \( Q: \) what should we look for in a correspondence plot?
  - \( \text{Large values in } |U_k'| \) (and \( |V_k'| \))
    - the profiles of the rows (or the columns) correspond to the large values are different from the marginal dist.
- e.g.: BLOND hair ⇒ the distribution of eye colors within this group is not typical
- e.g.: BROWN hair ⇒ the distribution of eye colors within this group close to the marginal distribution of columns

Row and column levels close together and far from the origin
- a large positive $R_{ij}$ would be associated with the combination
- e.g.: BLOND hair ↔ blue eye ⇒ strong association

Row and column levels situate apart on either side of the origin
- a large negative $R_{ij}$ would be associated with the combination
- e.g.: BLOND hair ↔ brown eye ⇒ relatively fewer people

Points of two row (or two column) levels are close together
- The two rows/columns have a similar pattern of association
  ⇒ might consider to combine the two categories
- e.g.: hazel eye ↔ green eye ⇒ similar hair color distribution

- Other versions of CA: see Venables and Ripley (2002, corresp in the MASS package of R), or Blasius and Greenacre (1998)

- Reading: Faraway (2006, 1st ed.), 4.2

**Matched Pairs Design (MPD)**

- **Design**
  - A block factor: $y_{++}$ levels, each level represents a block, each block of size 2, i.e., 2 experimental units (EUs) in one block
  - A treatment factor: 2 levels A and B, randomly assigned to the 2 EUs in each blocks
  - A response variable: categorical
  - 2 formats of representing data
  - Comparison 1: MPD ↔ MCCD
  - Comparison 2: MPD ↔ Paired sample $t$-test

- Data for contingency table: observe one type of categorical measure on two matched objects (EUs)
  - In contrast, in the typical 2-way contingency table, observe two (different) types of categorical measures ($X_1$ and $X_2$) on one object
  - e.g., left ($X_1$) and right ($X_2$) eye performance of a person
• Contingency table for matched pair data is a square matrix and
  ➢ no marginal totals are fixed in advance
  ➢ grand total $Y_{++}$ could be random or fixed
• Q: what questions are of interest for matched pair data?
  ➢ row and column marginals are homogeneous, i.e., $\pi_{ij} = \pi_{ji}$?
  ➢ $[\pi_{ij}]_{I \times I}$ is a symmetric matrix, i.e., $\pi_{ij} = \pi_{ji}$?
    - symmetry implies marginal homogeneity (MH), but, the reverse statement not necessarily true (except for $2 \times 2$ table)
    - Q: how to interpret symmetry?
  ➢ When row and column marginal totals are quite different, might be interested in whether
    $\pi_{ij} = \pi_{i+} \pi_{+j} \gamma_{ij}$, where $\gamma_{ij} = \gamma_{ji}$?
    - It is called quasi-symmetry (QS)
    - MH + QS $\iff$ symmetry
  ➢ $X_1$ and $X_2$ are independent, i.e., $\pi_{ij} = \pi_{i+} \pi_{+j}$ for all $i$ and $j$?
    - symmetry implies marginal homogeneity (MH), but, the reverse statement not necessarily true (except for $2 \times 2$ table)
  ➢ Q: how to interpret symmetry?

• Tests for these hypotheses based on log-linear model, e.g.,
  $Y = (y_{11}, y_{21}, y_{31}, y_{12}, y_{22}, y_{32}, y_{13}, y_{23}, y_{33})^T$
  Test for symmetry ($H_0$) hypothesis:
    - Generate a vector with $I^2$ components for an $(I(I+1)/2)$-level nominal factor with the structure:
      $\text{sym-factor} = (l_1, l_2, l_3, l_4, l_5, l_6)^T$
      $\text{sym-factor} = S_{sym}$
    - Deviance-based/Pearson $X^2$ goodness-of-fit test for $S_{sym}$
  Test for QS ($H_0$) hypothesis
    - $\log(\pi_{ij}) = \log(\pi_{i+} \pi_{+j} \gamma_{ij}) = \log(\pi_{i+}) + \log(\pi_{+j}) + \log(\gamma_{ij})$
\[ Y \sim X_1 + X_2 + \text{sym-factor} = S_{qsym} \]

- Deviance-based/Pearson \( X^2 \) goodness-of-fit test for \( S_{qsym} \)

Test for MH (\( H_0 \)) hypothesis

- No log-linear models that directly correspond to MH
- An indirect test using log-linear models when \( S_{qsym} \) already holds
  - Deviance-based test for \( H_0: S_{sym} \) vs. \( H_1: S_{qsym} \backslash S_{sym} \)
- Other approaches, see Agresti (2013), 11.3

Test for QI (\( H_0 \)) hypothesis

- Approach 1
  - Omit the diagonal data, i.e., let
    \[ Y' = (y_{21}, y_{31}, y_{12}, y_{32}, y_{13}, y_{23})^T \]
  - \( Y' \sim X_1 + X_2 = S_{qindep1} \)
  - Deviance-based/Pearson \( X^2 \) goodness-of-fit test for \( S_{qindep1} \)

- Approach 2
  - Generate a vector with \( I^2 \) components for an \((I+1)\)-level nominal factor with the structure:
    \[ \text{QI-factor} = (l_1, l_0, l_0, l_0, l_2, l_0, l_0, l_0, l_3)^T \]
  - \( Y \sim X_1 + X_2 + \text{QI-factor} = S_{qindep2} \)
  - Deviance-based/Pearson \( X^2 \) goodness-of-fit test for \( S_{qindep2} \)

- Reading: Faraway (2006, 1st ed.), 4.3

Three-Way Contingency Table

- The \( \pi \)'s and \( Y \)'s are defined in the same manner as in the 2-way table
- Poisson GLM approach to investigate how \( X_1, X_2, X_3 \) interact

- Mutual independence (\( X_1, X_2, X_3 \) are independent)
  - \( \pi_{ijk} = \pi_{i+} \pi_{+j} \pi_{++k} \)
  - \( \log(\pi_{ijk}) = \log(\pi_{i+} \pi_{+j} \pi_{++k}) = \log(\pi_{i+}) + \log(\pi_{+j}) + \log(\pi_{++k}) \)
The estimates of parameters in this model correspond only to the marginal totals \( y_{i+}, y_{+j}, \) and \( y_{++} \).

The coding we use will determine exactly how the parameters relate to the margin totals, e.g., let \( \beta \) be a main effect of \( X_1 \) that codes \( i_1 \) and \( i_2 \) categories as 0 (reference) and 1.

\[ \frac{e^\beta}{1+e^\beta} = \frac{\hat{\pi}_{i_2++}}{\hat{\pi}_{i_1++} + \hat{\pi}_{i_2++}} \]
\[ = \frac{y_{i_2++}}{y_{i_1++} + y_{i_2++}} \]

Insignificant factor, say \( X_1 \Rightarrow \pi_{1++} = \pi_{2++} = \ldots = \pi_{I++} \)

\( Y \sim X_1 + X_2 + X_3 \equiv S_1 \)

\( X_1, X_2 \) are independent given \( X_3 \)

\[ \log(\pi_{ijk}) = \log(\pi_{ij+} \pi_{++k}) = \log(\pi_{ij+}) + \log(\pi_{++k}) \]

\( Y \sim X_1 + X_2 + X_3 \equiv S_2 (\supset S_1) \)

Join independence (\( \{X_1, X_2\} \) and \( X_3 \) are independent)

\[ \pi_{ijk} = \pi_{ij+} \pi_{++k} \iff \pi_{ijk} = \pi_{ij+} \pi_{jk}/\pi_{++k} \]

\( Y \sim X_1 + X_1;X_2 + X_3 \equiv S_3 \)(\( \supset S_1 \))

Conditional independence (\( X_1, X_2 \) are independent given \( X_3 \))

\[ \log(\pi_{ijk}) = \log(\pi_{i+k} \pi_{+jk}/\pi_{++k}) = \log(\pi_{i+k}) + \log(\pi_{+jk}) - \log(\pi_{++k}) \]

Note that \( S_3 \not\supset S_2 \), but \( X_2 \) is jointly independent of \( \{X_1, X_3\} \) implies that \( X_1, X_2 \) are independent given \( X_3 \)

Q: can this conditional independence imply independence between \( X_1 \) and \( X_2 \), i.e., \( \pi_{ij+} = \pi_{i+j+} \pi_{j+} \) ? (Ans: No. Check singular value decomposition in LNp.5-15)
Uniform association (UA)

- Consider a model with all two-factor interactions

$$ Y \sim X_1 + X_2 + X_3 + X_1: X_2 + X_1: X_3 + X_2: X_3 \equiv S_4 (\supseteq S_3) $$

- $S_4$ has no simple interpretation in terms of independence

- $S_4$ asserts that for every level of one variable, say $X_3$, we have the same association between $X_1$ and $X_2$

$$ Y \sim \underbrace{X_1 + X_2 + X_3}_{\text{model}} + \underbrace{X_1: X_2 + X_1: X_3 + X_2: X_3}_{\text{interaction}} $$

- For each level of $X_3$, the reduced models of $S_4$ have different coefficients for the main effects of $X_1$ and $X_2$, but have the same coefficients for the interaction $X_1: X_2$

- In particular, $I=J=2$, same fitted odds-ratio between $X_1$ and $X_2$ for each category of $X_3$. Note that

$$ \text{fitted odd-ratio} = \frac{\hat{y}_{11k} \hat{y}_{22k}}{\hat{y}_{12k} \hat{y}_{21k}} = \frac{\hat{\pi}_{11k} \hat{\pi}_{22k}}{\hat{\pi}_{12k} \hat{\pi}_{21k}} = e^{\hat{\beta}_{12k}} $$

where $\hat{\beta}_{12k}$ is the coefficient of the $X_1: X_2$ term (under a coding $\propto \{+1, -1\}$) in the reduced model of $X_3=k$.

- $S_4$ is not saturated $\Rightarrow$ some degrees of freedoms left for goodness-of-fit test

- A saturated model corresponds to a 3-way table with different association between, say $X_1$ and $X_2$, across $K$ levels of $X_3$ whereas $Y \sim 1$ corresponds to a 3-way table with constant $\pi$

Q: What does uniform association mean? How to interpret the association? How does it connect with interaction terms?
Q: how to examine whether $X_1, X_2, X_3$ in a 3-way table are mutually independent ($S_1$), jointly independent ($S_2$), conditionally independent ($S_3$), or uniformly associated ($S_4$), individually?

- **Ans:** Perform deviance-based/Pearson’s $X^2$ goodness-of-fit (GoF) tests for $S_1, S_2, S_3, S_4$ (as $H_0$), respectively.
- However, be careful of zero or small $y_{ijk}$ (rule of thumb: 20% of cells less than 5) in the table ⇒ there will be some doubt about the accuracy of chi-square approximation in GoF test.
- The chi-square approximation is better in comparing models than assessing GoF.

Analysis strategy: start with complex Poisson GLM (e.g., saturated one) and see how far the model can be reduced (e.g., using model selection or sequential deviance-based tests as in ANOVA to compare models).

- Binomial (or multinomial) GLM approach for 3-way table:
  - If $y_{ij}$’s regarded as fixed, can treat $Y_{X_3}$ as response and $X_1, X_2$ as covariates.

Q1: what information been gone? Q2: what still attainable?

- **Ans** for Q1: information about $\pi_{ij}$
- **Ans** for Q2: information about $\pi_{kij}$

**Statistical Modeling**

- $Y_{X_3} = Y_{ij1} \sim \text{binomial}(y_{ij}, \pi_{k=1|ij})$ when $K=2$
- $Y_{X_3} = (Y_{ij1}, \ldots, Y_{ijK}) \sim \text{multinomial}(y_{ij}, \pi_{k=1|ij}, \ldots, \pi_{k=K|ij})$ when $K > 2$

Q: how is a binomial GLM connected to a Poisson GLM in 3-way tables?

- $Y_{X_3} \sim 1 \iff S_2$ (joint independence)
  - $\pi_{ijk} = \pi_{ij} \times \pi_{++k} \iff \pi_{k|ij} = \pi_{++k}$
  - 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 covariates $X_1$, $X_2$, and not interested in the association between $X_1$ and $X_2$.

- **Q:** how about $Y_{X_3} \sim 1 + X_1$?

$Y_{X_3} \sim 1 + X_1 \Leftrightarrow S_4$ (uniform association)

The saturated binomial GLM, $Y_{X_3} \sim 1 + X_1 + X_2 + X_1 \cdot X_2$, corresponds to a Poisson GLM for different association.

- **Q:** Poisson or binomial GLM approach? Which to use?
  - Binomial if one variable is clearly identified as the response
  - Poisson if 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 $I \times J$ 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

#### Example:

<table>
<thead>
<tr>
<th>$X_1 (i)$: age</th>
<th>smoker</th>
<th>dead</th>
<th>alive</th>
<th>$X_2 (j)$: smoker</th>
<th>dead</th>
<th>alive</th>
<th>$X_3 (k)$: dead or alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>age=35-44</td>
<td>yes</td>
<td>14</td>
<td>95</td>
<td>no</td>
<td>7</td>
<td>114</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>108</td>
<td>142</td>
<td>250</td>
<td>.43</td>
<td>.57</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>marginal total over age</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
</tr>
<tr>
<td>no</td>
</tr>
<tr>
<td>smoker</td>
</tr>
<tr>
<td>dead</td>
</tr>
<tr>
<td>alive</td>
</tr>
<tr>
<td>.30</td>
</tr>
<tr>
<td>.43</td>
</tr>
<tr>
<td>.70</td>
</tr>
<tr>
<td>.57</td>
</tr>
</tbody>
</table>
Marginal association added over $X_1$ is different from the conditional association observed within each category of $X_1$.

Q: Why it occurs? Why the table of $y_{jk}$ gives a contradictory result to the tables of $y_{jk|i}$?

If $\frac{y_{11\cdot}}{y_{11\cdot} + y_{12\cdot}} = \frac{y_{12\cdot}}{y_{12\cdot} + y_{12\cdot}} \leq \frac{y_{12\cdot}}{y_{12\cdot} + y_{12\cdot}} = \frac{y_{12\cdot}}{y_{12\cdot} + y_{12\cdot}}$ and $\frac{y_{21\cdot}}{y_{21\cdot} + y_{22\cdot}} = \frac{y_{22\cdot}}{y_{22\cdot} + y_{22\cdot}} \leq \frac{y_{22\cdot}}{y_{22\cdot} + y_{22\cdot}} = \frac{y_{22\cdot}}{y_{22\cdot} + y_{22\cdot}}$.

Note. $\frac{y_{1j\cdot} + y_{2j\cdot}}{y_{1j\cdot} + y_{2j\cdot} + y_{2j\cdot} + y_{2j\cdot}} = \frac{y_{1j\cdot}}{y_{1j\cdot} + y_{2j\cdot} + y_{2j\cdot} + y_{2j\cdot}} + \frac{y_{2j\cdot}}{y_{2j\cdot} + (1 - \frac{y_{1j\cdot} + y_{2j\cdot}}{y_{1j\cdot} + y_{2j\cdot} + y_{2j\cdot} + y_{2j\cdot}})}$

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

- Mantel-Haenszel (MH) test for $2\times2\times K$ table

Designed to test independence in $2\times2$ tables across $K$ categories.

Recall. Association of a $2\times2$ table can be completely characterized/measured by its odds-ratio $\Delta$.

- $\Delta = 1 \iff$ independence
- $\Delta > 1 \iff$ positive association
- $\Delta < 1 \iff$ negative association

Null and alternative hypotheses of MH test

- $H_0: \Delta_1 = \Delta_2 = \ldots = \Delta_K = 1$ (conditional independence)
- $H_1^*: \Delta_k \neq 1$ (different association) or $H_1: \Delta_1 = \Delta_2 = \ldots = \Delta_K \neq 1$ (uniform association)

The test works better when the odds ratios of the $K$ $2\times2$ tables do not vary greatly, e.g., the null of the GoF test for uniform association, $\Delta_1 = \Delta_2 = \ldots = \Delta_K$, does not rejected.

Procedure of the MH test

- Suppose the marginal totals of each $2\times2$ table carry no information (e.g., fixed in advance) or are conditioned.
  - Under $H_0$, can assume a hyper-geometric distribution for $y_{11k}$ in each $2\times2$ table
  - $y_{11k}$ is sufficient for testing independence of $k$th table.
MH statistic combine information of $y_{11k}$'s from $K$ tables:

$$\left( \sum_{k} \left[ y_{11k} - E(y_{11k}) \right] - 1/2 \right)^2 \frac{1}{\sum_{k} Var(y_{11k})} \sim \chi^2_k$$

where $E(y_{11k})$ and $Var(y_{11k})$ are calculated under the $H_0$

- can calculate an exact p-value for smaller dataset using hypergeometric distribution
  - useful when data is sparse, under which the $\chi^2$ approximations based on asymptotic thm is questionable

- MH test is sometimes called Cochran-Mantel-Haenszel test because a version without the 1/2 is published earlier by Cochran (1954).

- Reading: Faraway (2006, 1st ed.), 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

- The ordinal structure not matter when # of categories = 2

- For ordinal variables, can use the methods for nominal variable
  - But, more information can be extracted by taking advantage of the ordinal structure

- Treatments for ordinal response (future lecture) and ordinal covariates are different

- Treatment for ordinal predictors: assign each category a score
  - It kind of turns 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 based on scores

- Poisson GLM with linear-by-linear association for 2-way tables:
  - Consider table with ordinal row ($X_1$) and column ($X_2$) variables
Some notes about $\gamma$:

- 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 $v(X_2)$.

**Linear-by-linear association model:**

$$\eta_{i,j} = \log(\mu_{i,j}) = \log(t \pi_{i,j}) = \log(t) + \log(\pi_{i+}) + \log(\pi_{+j}) + \gamma \times u_i \times v_j$$

where $u_i$'s, $v_j$'s are known scores, and $\gamma$ is an unknown parameter.

$$Y \sim X_1 + X_2 + u(X_1)v(X_2) \equiv S_{O \times O}$$

Some notes about $\gamma$:

- Values of $\gamma$ represent the amount of association.
- $\gamma = 0 \iff$ independence.
- Positive and negative $\gamma$.

**Interpretation of $\gamma$ by log-odds-ratio:**

$$\log\left(\frac{\pi_{i,j} \pi_{i+1,j+1}}{\pi_{i,j+1} \pi_{i+1,j}}\right) = \log\left(\frac{\mu_{i,j} \mu_{i+1,j+1}}{\mu_{i,j+1} \mu_{i+1,j}}\right)$$

$$= (\eta_{i,j} + \eta_{i+1,j+1}) - (\eta_{i,j+1} + \eta_{i+1,j}) = \gamma (u_{i+1} - u_i)(v_{j+1} - v_j)$$

For evenly spaced scores, these log-odds-ratios are equal $\Rightarrow$ called uniform association in Goodman (1979).

**Latent (continuous) variable Z motivation for $\gamma$:**

- Assume $\pi_{i,j}$'s are obtained by putting a grid on an approximately bi-variate Normal ($Z_1, Z_2$) 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 (cf., positive and negative $\rho$).

**Q:** For the tests of independence or goodness-of-fit, what is the benefit of using $S_{O \times O}$ over the nominal approach, i.e., fitting a nominal-by-nominal model $S_{N \times N}$: $Y \sim X_1 + X_2 + X_1 : X_2$?

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
- The $O \times O$ interaction effect reduces a deviance of 10.175 on one degree of freedom, i.e., the other 35 interaction effects only reduce a deviance of 30.568.
- Ordinal-by-nominal model (or nominal-by-ordinal model)

- Rows (or columns) assigned scores, but column (or row) variable treated as a nominal variable
- called column (or row) effects model because the columns (or rows) are not assigned scores; instead, their effects are estimated
  - alternative viewpoint: the scores of the ordinal columns (or rows) regarded as parameters

- Column effects model:
  \[ n_{ij} = \log(\mu_{ij}) = \log(t \pi_{ij}) = \log(t) + \log(\pi_{i+}) + \log(\pi_{+j}) + u_i \times \gamma_j \]
  where \( u_i \)'s, \( i=1,\ldots, I \), are known scores, and \( \gamma_j \)'s, \( j=1,\ldots, J \), are unknown parameters (over-parameterized; only requires \( J-1 \) parameters),

- \( Y \sim X_1 + X_2 + u(X_1) \times X_2 \equiv S_{O \times N} (\supset S_{O \times O}) \)

- Some notes about \( \gamma_j \)'s, called the column effects:
  - Equality of the \( \gamma_j \)'s (then, \( u_i \times \gamma_j = u_i \times \gamma \)) corresponds to the hypothesis of independence between \( X_1 \) and \( X_2 \)
  - 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 \( S_{O \times N} (1) \) to examine whether the chosen scores for columns in \( S_{O \times O} \) (i.e., \( v_j \)'s) are appropriate, or (2) to possibly suggest better scores (see an example in lab)

- Some advantages of using scores for ordinal variables
  - helpful in reducing the complexity of models for categorical data with ordinal variables
  - especially useful in higher dimensional table where a reduction in the # of parameters is particularly welcome
  - can also sharpen our ability to detect associations