Notations on a $2 \times 2$ Table

- Variables $X$ and $Y$ take two values each.
- Denote $P[X = i, Y = j] = \pi_{ij}$, $i = 1, 2$, $j = 1, 2$.
- Denote $\pi_{i+} = \sum_{i=1}^{I} \pi_{ij}$ and $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$.
- Denote $P[X = i] = \pi_{i+}$ and $P[Y = j] = \pi_{+j}$.
- Let $n_{ij}$ be the observed frequency for $X = i$, $Y = j$. Then, the MLE are

$$\hat{\pi}_{ij} = p_{ij} = \frac{n_{ij}}{n_{++}}$$

and

$$\hat{\pi}_{i+} = p_{i+} = \frac{n_{i+}}{n_{++}}, \quad \hat{\pi}_{+j} = p_{+j} = \frac{n_{+j}}{n_{++}}$$

where $n_{i+} = \sum_{j} n_{ij}$, $n_{+j} = \sum_{i} n_{ij}$, and $n_{++} = \sum_{i} \sum_{j} n_{ij}$.
Table 1: Two-Way Contingency Probability Table with $I = J = 2$

<table>
<thead>
<tr>
<th>Row</th>
<th>Column 1</th>
<th>Column 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$\pi_{1+}$</td>
</tr>
<tr>
<td>2</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$\pi_{2+}$</td>
</tr>
<tr>
<td>Total</td>
<td>$\pi_{+1}$</td>
<td>$\pi_{+2}$</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 2: Two-Way Contingency observation Table with $I = J = 2$

<table>
<thead>
<tr>
<th></th>
<th>Column</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Row</td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{1+}$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_{2+}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
<td>$n_{++}$</td>
</tr>
</tbody>
</table>
• The odds ratio is defined by

\[ \theta = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}. \]

• It can be estimated by

\[ \hat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}}. \]

• Interpretation: odds ratio indicates the increase of the risks from the first row to the second row.
Interpretation

- Interpretation: odds ratio indicates the increase of the risks from the first row to the second row.
- If $\theta = 1$, $X$ and $Y$ are independent.
Test and CI

- The asymptotic variance

\[ \hat{\sigma}^2_{\log(\hat{\theta})} = V[\log(\hat{\theta})] = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \]

- A \( z \)-test rejects \( H_0 \) of independence if

\[ \left| \frac{\log(\hat{\theta})}{\hat{\sigma}_{\log(\hat{\theta})}} \right| > z_{\alpha/2}. \]

- A \( (1 - \alpha)100\% \) CI for \( \log(\theta) \) is

\[ \log(\hat{\theta}) \pm z_{\alpha/2} \hat{\sigma}_{\log(\hat{\theta})}. \]

- A \( (1 - \alpha)100\% \) CI for \( \theta \) is the corresponding transformation as

\[ \hat{\theta}e^{\pm \frac{z_{\alpha/2}}{2}\hat{\sigma}_{\log(\hat{\theta})}}. \]
Exponential Family

- Suppose \( Y_1, \cdots, Y_n \) are independent random variables.

- Let \( f(y_i; \theta_i, \phi) \) be PMF or PDF of \( Y_i \), where \( \phi \) is a scale parameter.

- If we can write

\[
f(y_i; \theta_i, \phi) = \exp \left[ \frac{y_i \theta - b(\theta)}{a(\phi)} + c(y, \phi) \right],
\]

then we call the PMF or the PDF \( f(y_i; \theta_i) \) is an exponential family.
**Binomial Distribution**

Assume $X \sim Bin(n, p)$. Then, $E(X) = np$. The PMF is

$$\binom{n}{x} p^x (1 - p)^{n-x}$$

$$= \exp\{x \log \frac{p}{1 - p} + n \log(1 - p) - \log \binom{n}{x}\}.$$ 

So $y = x$, $\theta = \log[p/(1 - p)]$, $b(\theta) = n \log(1 + e^\theta)$, $\phi = 1$, $a(\phi) = 1$, $c(y, \phi) = -\log \binom{n}{x}$. 
Assume $X \sim \text{Poisson}(\lambda)$. Then, $E(X) = \lambda$. The PMF is

$$
\frac{\lambda^x}{x!} e^{-\lambda}
$$

$$
= \exp\{x \log(\lambda) - \lambda - \log(x!}\}.
$$

So $y = x$, $\theta = \log(\lambda)$, $b(\theta) = e^\theta$, $\phi = 1$, $a(\phi) = 1$, $c(y, \phi) = -\log(x!)$. 


Gamma Distribution

Assume $X \sim \Gamma(\alpha, \beta)$, $\beta$ is known. Then, $E(X) = \alpha/\beta$. Then PMF is

$$\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}$$

$$= \exp\{\alpha \log x + \alpha \log(\beta) - \log(\Gamma(\alpha)) - \log(x) - \beta x\}.$$  

Assume $\alpha$ is known. Then $\theta = -\beta$, $y = x$, $b(\theta) = -\alpha \log(\beta)$, $\phi = 1$ and $a(\phi) = 1.$
Negative Binomial Distribution

Assume \( X \sim NB(k, p) \). The PDF is

\[
\binom{x - 1}{k - 1} p^k (1 - p)^{x-k} = \exp\{x \log(1 - p) + k \log \frac{p}{1 - p} - \log \left(\frac{x - 1}{k - 1}\right)\},
\]

for \( x = k, k + 1, \ldots \). So \( y = x, \theta = \log(1 - p) \), 
\( b(\theta) = -k \log[(1 - e^\theta)/e^\theta], \phi = 1, a(\theta) = 1 \) and 
\( c(y, \phi) = - \log \left(\frac{x - 1}{k - 1}\right) \). Then,

\[
E(X) = b'(\theta) = \frac{k}{1 - e^\theta} = \frac{k}{p}
\]

and

\[
V(X) = b''(\theta) = \frac{ke^\theta}{(1 - e^\theta)^2} = \frac{k(1 - p)}{p^2}.
\]
The definition of Generalized Linear Model (GLM) is based on exponential family. There are three components in GLM. They are

- **Random component.** Assume the distributions of the sample. Such as normal, binomial, Poisson and etc.

- **Systematic component.** Describe the form of predictor (independent) variables. Such as

  \[ \eta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{ip} x_{ip}. \]

- **Link function.** Connect the unknown parameters to model. Such as

  \[ g[\mu(\theta)] = \eta \]

for some \( g(\cdot) \), where \( \mu(\theta) \) is the expected value.
Canonical Link

It uses $\theta = \eta$ is the canonical link.

- Normal: identity link $g(\mu) = \mu$.
- Binomial: logistic link $g(\mu) = \log \frac{\mu}{1-\mu}$.
- Poisson: log link $g(\mu) = \log(\mu)$.
- Gamma: negative inverse link $g(\mu) = -1/\mu$.
- Negative binomial: $g(\mu) = \log[\mu/k(1 + \mu/k)]$. 
• The most important cases are binomial and Poisson.

• Canonical link is just one of the link functions.

• Estimation is based on maximum likelihood.
There are three link functions for binomial.

- Logistic link.
  \[
  \log \frac{p_i}{1 - p_i} = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j
  \]
called logistic linear model or logistic regression.

- Inverse CDF link.
  \[
  F^{-1}(p_i) = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j.
  \]
If \( F = \Phi \), it is the probit link, called probit model.

- Complementary loglog link.
  \[
  \log[- \log(1 - p_i)] = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j.
  \]
Logistic Regression

Consider the simplest case. That is

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i.$$ 

Suppose $\hat{\beta}$ and $\hat{\beta}_1$ are the MLEs.

- Odds ratio: as $x$ increases $a$ units, the estimate of odds ratio is $e^{a\hat{\beta}_1}$.
- The significance of the odds ratio can be directly read by the $p$-value of $\beta_1$.
- Confidence interval can also be derived respectively.
### Table 3: Blood Pressure and Heart Disease

<table>
<thead>
<tr>
<th>Blood Pressure</th>
<th>Heart Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>&lt; 117</td>
<td>3</td>
</tr>
<tr>
<td>117 – 126</td>
<td>17</td>
</tr>
<tr>
<td>127 – 136</td>
<td>12</td>
</tr>
<tr>
<td>137 – 146</td>
<td>16</td>
</tr>
<tr>
<td>147 – 156</td>
<td>12</td>
</tr>
<tr>
<td>157 – 166</td>
<td>8</td>
</tr>
<tr>
<td>167 – 186</td>
<td>16</td>
</tr>
<tr>
<td>&gt; 186</td>
<td>8</td>
</tr>
</tbody>
</table>
Goodness of Fit

Let $\hat{n}_{ij}$ be the predicted counts of the model.

- Pearson $\chi^2$ is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}.$$ 

- Loglikelihood ratio $\chi^2$ is

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log(n_{ij}/\hat{n}_{ij}).$$