ST3241 Categorical Data Analysis I
Models For Matched Pairs

Dependent Proportions and Conditional Models
Example: Rating of Performance

- For a poll of a random sample of 1600 voting-age British citizens, 944 indicated approval of the Prime minister’s performance in the office.

- Six months later, of these same 1600 people, 800 indicated approval.
**Example**

<table>
<thead>
<tr>
<th>First Survey</th>
<th>Approve</th>
<th>Disapprove</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approve</td>
<td>794</td>
<td>150</td>
<td>944</td>
</tr>
<tr>
<td>Disapprove</td>
<td>86</td>
<td>570</td>
<td>656</td>
</tr>
<tr>
<td>Total</td>
<td>880</td>
<td>720</td>
<td>1600</td>
</tr>
</tbody>
</table>
Dependent Categorical Data

- To compare categorical responses for two samples when each sample has the same subjects or when a natural pairing exists between each subject in one sample and a subject from the other sample.

- The responses in the two samples are then statistically dependent.

- The pairs of observations are called *matched pairs*.

- A two-way table having the same categories for both classifications summarizes such data.
Comparing Dependent Proportions

- Let $n_{ij}$ = the number of subjects making response $i$ at the first survey and response $j$ at the second.

- In the example, the sample proportions approving are $\frac{944}{1600} = 0.59$ and $\frac{880}{1600} = 0.55$.

- These marginal proportions are correlated, and statistical analyses must recognize this.

- Let $\pi_{ij}$ = probability that a subject makes response $i$ at survey 1 and response $j$ at survey 2.
Dependent Proportions

• The probabilities of approval at the two surveys are $\pi_{1+}$ and $\pi_{+1}$, the first row and first column totals.

• When these are identical, the probabilities of disapproval are also identical, and there is marginal homogeneity.

• Note that,

$$\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$$

• Marginal homogeneity is equivalent to equality of off-maindiagonal probabilities; that is $\pi_{12} = \pi_{21}$.

• The table shows symmetry across the main diagonal.
Inference For Dependent Proportions

Use $\delta = \pi_{+1} - \pi_{1+}$

Let $d = p_{+1} - p_{1+} = p_{2+} - p_{+2}$.

• From the results on multinomial distributions,

$$cov(p_{+1}, p_{1+}) = (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/n$$

• Thus,

$$var(\sqrt{n}d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})$$

• For large samples, $d$ has approximately a normal distribution.

• A confidence interval for $\delta$ is then $d \pm z_{\alpha/2}\hat{\delta}(d)$
Inference For Dependent Proportions

• Here

\[ \hat{\delta}^2(d) = \frac{[p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})]}{n} = \frac{[(p_{12} + p_{21}) - (p_{12} - p_{21})^2]}{n} \] (1)

• The hypothesis of marginal homogeneity is \( H_0 : p_{i1+} = p_{i+1} \) (i.e. \( \delta = 0 \)).

• Wald test statistic is: \( z = d / \hat{\sigma}(d) \)
McNemar’s Test

• Under $H_0$, an alternative estimated variance is

$$\hat{\sigma}_0^2(d) = \frac{p_{12} + p_{21}}{n} = \frac{n_{12} + n_{21}}{n^2}$$

• The score test statistic is

$$z_0 = \frac{d}{\hat{\sigma}_0(d)} = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}$$

• The square of $z_0$ is a chi-squared distribution with $df = 1$.

• The test using it called McNemar’s test.

• It depends only on cases classified in different categories for the two observations.
Example:

- The sample proportions of approval of the prime minister’s performance are $p_{1+} = 0.59$ for the first survey and $p_{+1} = 0.55$ for the second.

- A 95% CI for $\pi_{+1} - \pi_{1+}$ is
  
  $$(0.55 - 0.59) \pm 1.96(0.0095) = (-0.06, -0.02).$$

- The approval rating appears to have dropped between 2 and 6%.

- For testing marginal homogeneity, the test statistic using the null variance is
  
  $$z_0 = \frac{86 - 150}{(86 + 150)^{1/2}} = -4.17$$

- It shows evidence of a drop in the approval rating.
Small Sample Tests

- The null hypothesis of marginal homogeneity for binary matched pairs is $H_0 : \pi_{12} = \pi_{21}$ or $\pi_{12}/(\pi_{12} + \pi_{21}) = 0.5$.
- For small samples, an exact test conditions on $n^* = n_{12} + n_{21}$.
- Under $H_0$, $n_{21}$ has a binomial($n^*, 0.5$) distribution.
- For our example, $n^* = 86 + 150 = 236$.
- For the alternative hypothesis, $H_a : \pi_{+1} < \pi_{1+}$, the p-value is the probability of at least 150 successes out of 236 trials, which equals 0.00002.
### Connection Between McNemar’s and CMH Test

<table>
<thead>
<tr>
<th>Subject</th>
<th>Survey</th>
<th>Approve</th>
<th>Disapprove</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>First</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>First</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>First</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Connection

- An alternative representation of binary responses for $n$ matched pairs presents the data in $n$ partial tables, one $2 \times 2$ table for each pair.

- For our example, we have $2 \times 2 \times 1600$ contingency table.

- For each subject, suppose that the probability of approval is identical in each survey. Then conditional independence exists between the opinion outcome and the survey time, controlling for the subject.
Discussion

• The probability of approval is also the same for each survey in the marginal table collapsed over the subjects.

• But this implies that the true probabilities satisfy marginal homogeneity. .. Thus, a test of conditional independence in the $2 \times 2 \times 1600$ table provides a test of marginal homogeneity.

• The CMH test statistic for this purpose is identical to the squared McNemar’s statistic.

• McNemar’s test is a special case of the CMH test applied to the binary responses of $n$ matched pairs displayed in n partial tables.
Conditional Logistic Regression For Binary Matched Pairs

- Let \((Y_1, Y_2)\) denote the pair of binary observations for a randomly selected subject.

- The difference \(\delta = P(Y_2 = 1)P(Y_1 = 1)\) between marginal probabilities occurs as a parameter in

\[
P(Y_t = 1) = \alpha + \delta x_t
\]

where \(x_1 = 0\) and \(x_2 = 1\).

- Then \(P(Y_1 = 1) = \alpha\) and \(P(Y_2 = 1) = a + \delta\).

- Alternatively, the logit link yields

\[
\text{logit}[P(Y_t = 1)] = \alpha + \beta x_t
\]

- The parameter \(\beta\) is a log odds ratio with the marginal distributions.
Marginal Models

• The previous models are marginal models.

• They focus on the marginal distributions of responses for the two observations.

• For instance, the ML estimate of $\beta$ in the previous logit model is the log odds ratio of marginal proportions

$$
\hat{\beta} = \log \frac{P_{+1}P_{2+}}{P_{+2}P_{1+}}
$$
Conditional Models

- By contrast, the subject specific tables implicitly allow probabilities to vary by subject.

- Let \((Y_{i1}, Y_{i2})\) denote the \(i\)-th pair of observations, \(i = 1, \ldots, n\).

- A model then has the form

\[
\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_t
\]

- This is called a *conditional* model, since the effect \(\beta\) is defined conditional on the subject.

- Its estimate describes conditional association for the three-way table stratified by the subject.

- The effect is *subject-specific*, since it is defined at the subject level.
A Logit Model with Subject-Specific Probabilities

- It permits subjects to have their own probability distribution.
- This model for $Y_{it}$, observation $t$ for subject $i$, is

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_t$$

where $x_1 = 0$ and $x_2 = 1$.
- It assumes a common effect $\beta$.
- For subject $i$,

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}, P(Y_{i2} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)},$$

- The parameter $\beta$ compares the response distributions.
Discussions

• A subject with a large positive $\alpha_i$ has high $P(Y_{it} = 1)$ for each $t$ and is likely to have success each time.

• The greater the variability in $\{\alpha_i\}$, the greater the overall positive association between responses.

• This is true for any $\beta$.

• No association occurs only when $\{\alpha_i\}$ are identical.

• The model takes into account for the dependence in matched pairs.

• Fitting it takes into account nonnegative association through the structure of the model.
**Conditional ML Inference**

- Large number of $\{\alpha_i\}$ causes difficulties with the fitting process.
- Conditional ML treats them as nuisance parameters and maximizes the likelihood function for a conditional distribution that eliminates them.
- The conditional ML estimator of $\beta$ in the subject specific model and its standard error are:
  \[
  \hat{\beta} = \log(n_{21}/n_{12}), \quad SE = \sqrt{1/n_{21} + 1/n_{12}}
  \]
- For details see, *Agresti (2002), Categorical Data Analysis*
Matched Case-Control Studies

- The two observations \((y_{i1}, y_{i2})\) in a matched pair need not refer to the same subject.
- Case-control studies that match a single control with each case yield matched-pairs data.
- For a binary response \(Y\), each case \((Y = 1)\) is matched with a control \((Y = 0)\) according to criteria that could affect the response.
- Subjects in matched pairs are measured on the predictor variables of interest, \(X\), and the \(XY\) association is analyzed.
Example: Diabetes For MI Case-Control Pairs

<table>
<thead>
<tr>
<th>MI Controls</th>
<th>Diabetes</th>
<th>No Diabetes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>9</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>No Diabetes</td>
<td>37</td>
<td>82</td>
<td>119</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>98</td>
<td>144</td>
</tr>
</tbody>
</table>

Case-control study matched 144 victims of MI according to age and gender with 144 people free of heart disease.

Subjects were asked whether they had ever been diagnosed as having diabetes ($x = 0$, no; $x = 1$, yes).
**Example**

- Consider the model:

\[
\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_{it}
\]

- The probabilities modeled refer to the distribution of \( Y \) given \( X \), but the retrospective study provides information only about the distribution of \( X \) given \( Y \).

- One can estimate the odds ratio \( \exp(\beta) \) which, however, relates to both conditional distributions.

- The conditional ML estimate of \( \exp(\beta) \) is simply

\[
\frac{n_{21}}{n_{12}} = \frac{37}{16} = 2.3
\]
Symmetry For Square Tables

• For an $I$ category response with an $I \times I$ table for matched pairs, the cell probabilities $\pi_{ij}$ satisfy marginal homogeneity is

$$\pi_{i+} = \pi_{+i}, \ i = 1, \ldots, I$$

• The probabilities in the square table satisfy symmetry is

$$\pi_{ij} = \pi_{ji}$$

for all pairs of cells.

• When $I > 2$, though, marginal homogeneity can occur without symmetry.
Symmetry As Logit and Loglinear Models

- The symmetry condition has the simple logit form
  \[ \log(\pi_{ij}/\pi_{ji}) = 0 \] for all \( i \) and \( j \).

- The *symmetry model* also has a loglinear model representation:
  \[ \log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij} \]
  where all \( \lambda_{ij} = \lambda_{ji} \).

- This is the special case of the saturated loglinear model with
  \( \lambda_{ij}^{XY} = \lambda_{ji}^{XY} \) and \( \lambda_i^X = \lambda_i^Y \)

- We have \( \log \mu_{ij} = \log \mu_{ji} \), so that \( \mu_{ij} = \mu_{ji} \)
Symmetry Models

• The ML fit of the *symmetry* model is

\[ \hat{\mu}_{ij} = \frac{n_{ij} n_{ji}}{2} \]

• The fit satisfies \( \hat{\mu}_{ij} = \hat{\mu}_{ji} \)

• It has \( \hat{\mu}_i = n_{ii} \), a perfect fit in the main diagonal.

• The residual \( df \) for chi-squared goodness-of-fit tests equal \( I(I-1)/2 \).

• The adjusted residuals equal

\[ r_{ij} = \frac{n_{ij} - n_{ji}}{\sqrt{n_{ij} + n_{ji}}} \]

• Only one residual for each pair of categories is non-redundant, since

\[ r_{ji} = -r_{ij} \]
## Coffee Market Share Example

<table>
<thead>
<tr>
<th>First purchase</th>
<th>High Point</th>
<th>Taster’s</th>
<th>Sanka</th>
<th>Nescafe</th>
<th>Brim</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Point</td>
<td>93</td>
<td>17</td>
<td>44</td>
<td>7</td>
<td>10</td>
<td>171</td>
</tr>
<tr>
<td>Taster’s</td>
<td>9</td>
<td>46</td>
<td>11</td>
<td>0</td>
<td>9</td>
<td>75</td>
</tr>
<tr>
<td>Sanka</td>
<td>17</td>
<td>11</td>
<td>155</td>
<td>9</td>
<td>12</td>
<td>204</td>
</tr>
<tr>
<td>Nescafe</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>15</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>Brim</td>
<td>10</td>
<td>4</td>
<td>12</td>
<td>2</td>
<td>27</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>82</td>
<td>231</td>
<td>33</td>
<td>60</td>
<td>541</td>
</tr>
</tbody>
</table>
### Coffee Market Share Example

<table>
<thead>
<tr>
<th>First purchase</th>
<th>High Point</th>
<th>Taster’s</th>
<th>Sanka</th>
<th>Nescafe</th>
<th>Brim</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Point</td>
<td>93</td>
<td>17</td>
<td>44</td>
<td>7</td>
<td>10</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>(93)</td>
<td>(13.0)</td>
<td>(30.5)</td>
<td>(6.5)</td>
<td>(10.0)</td>
<td></td>
</tr>
<tr>
<td>Taster’s</td>
<td>9</td>
<td>46</td>
<td>11</td>
<td>0</td>
<td>9</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>(13.0)</td>
<td>(46)</td>
<td>(11.0)</td>
<td>(2.0)</td>
<td>(6.5)</td>
<td></td>
</tr>
<tr>
<td>Sanka</td>
<td>17</td>
<td>11</td>
<td>155</td>
<td>9</td>
<td>12</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>(30.5)</td>
<td>(11.0)</td>
<td>(155)</td>
<td>(9.0)</td>
<td>(12.0)</td>
<td></td>
</tr>
<tr>
<td>Nescafe</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>15</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>(6.5)</td>
<td>(2.0)</td>
<td>(9.0)</td>
<td>(15)</td>
<td>(2.0)</td>
<td></td>
</tr>
<tr>
<td>Brim</td>
<td>10</td>
<td>4</td>
<td>12</td>
<td>2</td>
<td>27</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>(10.0)</td>
<td>(6.5)</td>
<td>(12.0)</td>
<td>(2.0)</td>
<td>(27.0)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>82</td>
<td>231</td>
<td>33</td>
<td>60</td>
<td>541</td>
</tr>
</tbody>
</table>
Coffee Market Share Example

- The symmetry model fitted to these data has $G^2 = 22.4729$ and $\chi^2 = 20.4123$, with $df = 10$.
- The lack of fit results primarily from the discrepancy between $n_{13}$ and $n_{31}$.
- For that pair, the adjusted residual equals $(44 - 17)/(44 + 17)^{1/2} = 3.5$
- Consumers of High Point changed to Sanka more often than the reverse.
- Otherwise the symmetry model fits most of the table well
Quasi-Symmetry

• The symmetry model is so simple that it rarely fits well.
• One can accommodate marginal homogeneity by permitting the loglinear main-effects terms to differ.
• The resulting model, called the \textit{quasi} – \textit{symmetry model}, is

\[ \log \mu_{ij} = \lambda + \lambda_{i}^{X} + \lambda_{j}^{Y} + \lambda_{ij} \]

where \( \lambda_{ij} = \lambda_{ji} \) for all \( i \) and \( j \).
Quasi-Symmetry Model

- The fitted marginal totals equal the observed totals $\hat{\mu}_{i+} = n_{i+}$ and $\hat{\mu}_{+i} = n_{+i}, i = 1, \cdots, I$

- The symmetry model is the special case.

- The independence model is the special case in which all $\lambda_{ij} = 0$.

- This model is useful partly because it contains these two models as special cases.

- Fitting the quasi – symmetry model requires iterative procedures for loglinear models.
## Coffee Market Share Example

<table>
<thead>
<tr>
<th>First Purchase</th>
<th>High Point</th>
<th>Taster’s</th>
<th>Sanka</th>
<th>Nescafe</th>
<th>Brim</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Point</td>
<td>93</td>
<td>17</td>
<td>44</td>
<td>7</td>
<td>10</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>(93)</td>
<td>(16.8)</td>
<td>(40.9)</td>
<td>(7.4)</td>
<td>(12.9)</td>
<td></td>
</tr>
<tr>
<td>Taster’s</td>
<td>9</td>
<td>46</td>
<td>11</td>
<td>0</td>
<td>9</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>(9.2)</td>
<td>(46)</td>
<td>(11.6)</td>
<td>(1.7)</td>
<td>(6.5)</td>
<td></td>
</tr>
<tr>
<td>Sanka</td>
<td>17</td>
<td>11</td>
<td>155</td>
<td>9</td>
<td>12</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>(20.1)</td>
<td>(10.4)</td>
<td>(155)</td>
<td>(7.2)</td>
<td>(11.3)</td>
<td></td>
</tr>
<tr>
<td>Nescafe</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>15</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>(5.6)</td>
<td>(2.3)</td>
<td>(10.8)</td>
<td>(15)</td>
<td>(2.3)</td>
<td></td>
</tr>
<tr>
<td>Brim</td>
<td>10</td>
<td>4</td>
<td>12</td>
<td>2</td>
<td>27</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>(7.1)</td>
<td>(6.5)</td>
<td>(12.7)</td>
<td>(1.7)</td>
<td>(27.0)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>82</td>
<td>231</td>
<td>33</td>
<td>60</td>
<td>541</td>
</tr>
</tbody>
</table>
Coffee Market Share Example

• The quasi-symmetry model has $G^2 = 9.9740$ and $\chi^2 = 8.5303$ with $df = 6$.

• Permitting the marginal distributions to differ yields a better fit than the symmetry model provides.

• It is plausible that, adjusting for market shares, brand switching has a symmetry pattern.
An Ordinal Quasi-Symmetry Model

- Consider the case with ordinal responses.
- Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote ordered scores for both the row and column categories.
- The *ordinal quasi-symmetry model* has form
  \[ \log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \beta u_j + \lambda_{ij} \]
  where $\lambda_{ij} = \lambda_{ji}$ for all $i$ and $j$.
- It is the special case of the *quasi-symmetry model* in which
  \[ \lambda^Y_j - \lambda^X_j = \beta u_j \]
- The difference in main effect terms has a linear trend across the response categories.
- The symmetry model is the special case $\beta = 0$. 
Discussions

- For this model, the fitted marginal counts need not equal the observed marginal counts, but they do have the same means.

- When $\beta > 0$, the $\pi_{1+} > \pi_{1+}$, $\pi_{1+} + \pi_{2+} > \pi_{1+} + \pi_{2+}$ and so forth.

- That is, responses are more likely to be at the low end of the ordinal scale for the row variable than for column variable.

- When $\beta < 0$, the mean response is higher for the row classification.

- The ordinal quasi-symmetry model is equivalent to the model of logit form

$$\text{logit}(\pi_{ij}/\pi_{ji}) = \beta(u_j - u_i)$$
Testing Marginal Homogeneity

• For quasi-symmetry model, marginal homogeneity is the special case in which the two sets of main effect parameters are identical; that is, all $\lambda_i^X = \lambda_i^Y$.

• But this is simply the symmetry model.

• A test of marginal homogeneity tests the null hypothesis that the symmetry ($S$) model holds against the alternative hypothesis of quasi-symmetry ($QS$).

• The LR test compares the $G^2$ goodness-of-fit statistics,

\[ G^2(S|QS) = G^2(S) - G^2(QS) \]

• For $I \times I$ tables, the test has $df = I - 1$. 
Example: Diagnoses of Carcinoma

<table>
<thead>
<tr>
<th>Pathologist X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>36</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>27</td>
<td>12</td>
<td>69</td>
<td>10</td>
<td>118</td>
</tr>
</tbody>
</table>
**Example: Diagnoses of Carcinoma (Independence Model Fit)**

<table>
<thead>
<tr>
<th>Pathologist X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>(8.5)</td>
<td>(-0.5)</td>
<td>(-5.9)</td>
<td>(-1.8)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>(-0.5)</td>
<td>(3.2)</td>
<td>(-0.5)</td>
<td>(-1.8)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>36</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>(-4.1)</td>
<td>(-1.2)</td>
<td>(5.5)</td>
<td>(-2.3)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>(-3.3)</td>
<td>(-1.3)</td>
<td>(0.3)</td>
<td>(5.9)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>12</td>
<td>69</td>
<td>10</td>
<td>118</td>
</tr>
</tbody>
</table>
Analyzing Rater Agreement

• Let $\pi_{ij} = P(X = i, Y = j)$ denote the probability that observer $X$ classifies a slide in category $i$ and observer $Y$ classifies it in category $j$.

• Their ratings of a particular subject *agree* if they classify the subject in the same category.

• $\sum_i \pi_{ii}$ is the total probability of *agreement*.

• We distinguish between *agreement* and *association*.

• Strong agreement requires strong association, but strong association can exist without strong agreement.
Quasi Independence Model

- One useful generalization of independence model is the quasi-independence model

\[ \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta_i I(i = j) \]

- The indicator \( I(i = j) \) equals 1 when \( i = j \) and equals 0 when \( i \neq j \).
- It treats the main diagonal differently from the rest of the table.
- The ML fit in those cells is perfect. .. For the remaining cells the independence model still applies.
Examples

• For the diagnoses of carcinoma example, the quasi-independence model has $G^2 = 13.2$ and $\chi^2 = 11.5$ with $df = 5$.

• It fits much better than the independence model, though some lack of fit remains.

• For the Coffee brand example, the quasi-independence model has $G^2 = 13.8$ with $df = 11$.

• This is a dramatic improvement over independence, which has $G^2 = 346.4$ with $df = 16$. 
Some SAS Codes (Read The Data)

data coffee;
    input purchase1 purchase2 symm qi count @@;
datalines;
1 1 1 1 93 1 2 2 6 17 1 3 3 6 44 1 4 4 6 7 1 5 5 6 10  
2 1 2 6 9 2 2 6 2 46 2 3 7 6 11 2 4 8 6 0 2 5 9 6 9 
3 1 3 6 17 3 2 7 6 11 3 3 10 3 155 3 4 11 6 9 3 5 12 6 12 
4 1 4 6 6 4 2 8 6 4 4 3 11 6 9 4 4 13 4 15 4 5 14 6 2 
5 1 5 6 10 5 2 9 6 4 5 3 12 6 12 5 4 14 6 2 5 5 15 5 27
SAS Codes Symmetry Model

```sas
proc genmod data=coffee order=data;
   class symm; model count = symm / d=poi link=log;
   output out=temp pred=pred;
run;
proc freq data=temp;
   weight pred;
   tables purchase1*purchase2/nopercent norow
      nocol;
run;
```
### Partial Output

<table>
<thead>
<tr>
<th>purchase1</th>
<th>purchase2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>93</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>30.5</td>
</tr>
<tr>
<td>4</td>
<td>6.5</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>153</td>
</tr>
</tbody>
</table>
SAS Codes (Quasi-Symmetry Model)

proc genmod data=coffee;
   class purchase1 purchase2 symm;
   model count = purchase1 purchase2 symm
      /dist=poi link=log;
   output out=temp pred=pred;
run;
proc freq data=temp;
   weight pred;
   tables purchase1*purchase2/nopercent norow nocol;
run;
SAS Codes (Quasi Independence Model)

proc genmod data=coffee;
  class purchase1 purchase2 qi;
  model count = purchase1 purchase2 qi /dist=poi
    link=log;
  output out=temp pred=pred;
run;
proc freq data=temp;
  weight pred;
  tables purchase1*purchase2/nopercent norow
  nocol;
run;