516 707 Categorical Data Analysis
for Health Research:
Comparing Marginal Proportions
and Measuring Agreement for Dependence Data

Assistant Professor Nikom Thanomseng
Dept. of Biostatistics & Demography
Faculty of Public Health, KKU
Web Page: http://home.kku.ac.th/nikom
Email: nikom@kku.ac.th

Outline
- Dependent Data
- Comparing Marginal proportion
  - McNemar Test
  - Exact method
  - Estimate Difference of proportion
  - matched odds ratio
- Comparing marginal of Square Contingency Tables
  - Stuart- Maxwell, Bhapkar
- Measuring Agreement Between Judges
  - Kappa & weight Kappa (Cohen’s Kappa)

Dependent Data
- Before-After or Pre and Post Test or Repeated Measured
  - Measurement on the same subject.
- Matching or Match Paired
  - each paired of subjects as much alike as possible
    with respect to extraneous variable.
  - matched case-control study
- Natural Pairs
  - Human Twins or siblings
  - litter mates in animal
Comparing Dependent Proportion

- How can we compare the probabilities of a “yes” outcome for the two group questions?

Let \( \pi_{ij} \) denote the probability of outcome \( i \) for question 1 and \( j \) for question 2.

- The probabilities of a “yes” outcome are \( \pi_{1+} \) for question 1
- The probabilities of a “yes” outcome are \( \pi_{+1} \) for question 2.

When these are identical, the probabilities of a “no” outcome are also identical.

There is then said to be marginal homogeneity. Since

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

marginal homogeneity in 2 × 2 tables is equivalent to \( \pi_{12} = \pi_{21} \)

\begin{tabular}{|c|c|c|c|}
\hline
Case & Question 1 & Question 2 & Total \hline
1 (yes) & \( n_{11} \) & \( n_{12} \) & \( n_{1+} \) \hline
2 (no) & \( n_{21} \) & \( n_{22} \) & \( n_{2+} \) \hline
\hline
\end{tabular}

Let \( \pi_i \) denote the probability of outcome \( i \) for question 1 and \( j \) for question 2.

- The probabilities of a “yes” outcome are \( \pi_i \) for question 1
- The probabilities of a “yes” outcome are \( \pi_j \) for question 2.

McNemar’s Test


**McNemar’s Test**

- It is applied to 2 × 2 contingency tables with a dichotomous trait, with matched pairs of subjects,
- to determine whether the row and column marginal frequencies are equal (“marginal homogeneity”).

McNemar’s Test for (2 × 2) tables

\[ H_0 : \pi_{1+} = \pi_{+1} \text{ or equivalently } \pi_{12} = \pi_{21} \]

Define \( n^* = n_{12} + n_{21} \). Consider the binomial variable with \( n^* \) trials that has it’s two possible outcomes \( n_{12} \) and \( n_{21} \) in the (2 × 2) table.
McNemar’s Test (cont.)

If H_0 is true, then
- Expect similar value n_{12} ~ n_{21}.
- Probability of n_{12} equals 1/2, and Probability of n_{21} equals 1/2.
  
For "small" n∗, just compute the exact probability (p-value).
For "large" n∗ (n∗ ≥ 10), use the normal approximation:
mean = 1/2(n∗) & standard deviation = √(1/2)(1/2)

- The standardized normal test statistic equals
  
- or z² to the chi-square distribution with df = 1.
- The chi-squared test for a comparison of two dependent proportions is called the McNemar test.

Estimating Differences of Proportions
- A confidence interval for the true difference of proportions is more informative than a significance test.
- Let \( p_{ij} = n_{ij}/n \) denote the sample cell proportions.
- The difference \( p_{i+} - p_{i-} \) between the sample marginal proportions estimates the true difference \( \pi_{i+} - \pi_{i-} \).
- The estimated variance of the sample difference equals
  
- square root is the SE for a confidence interval

- \((1-\alpha)100\%\) difference of proportions is.

Example
Diabetes and the occurrence of acute myocardial infarction (MI) matched for sex and age group (Coulehan JL et al., 1986)

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>DM</th>
<th>No DM</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>n_{1+} = 17</td>
<td>n_{1+} = 25</td>
<td>n_{1+} = 98</td>
</tr>
<tr>
<td>No DM</td>
<td>n_{2+} = 37</td>
<td>n_{2+} = 82</td>
<td>n_{2+} = 82</td>
</tr>
<tr>
<td>Total</td>
<td>n = 100</td>
<td>n = 153</td>
<td>n = 180</td>
</tr>
</tbody>
</table>

McNemar’s Test for (2 × 2) tables

\( H_0: \pi_{1+} = \pi_{1-} \) or equivalently \( \pi_{12} = \pi_{21} \) (marginal homogeneity)

\[ \chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(37 - 16)^2}{37 + 16} = 8.3207547 \]

95% CI = \( p_{1+} - p_{1-} \pm z_{0.025} \sqrt{\frac{(n_{1+} + n_{2+}) - (n_{12} - n_{21})^2}{n^2}} \)
### STATA Example

```
.mcci 9 37 16 82

<table>
<thead>
<tr>
<th>Controls</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>9</td>
<td>37</td>
<td>46</td>
</tr>
<tr>
<td>Unexposed</td>
<td>16</td>
<td>82</td>
<td>98</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>119</td>
<td>144</td>
</tr>
</tbody>
</table>

McNemar's chi2(1) = 8.32  Prob > chi2 = 0.0039
Exact McNemar significance probability = 0.0055

Proportion with factor
Cases .3194444
Controls .1736111 [95% Conf. Interval]
difference .1458333 .0427057 .2489609
rel. diff. .1764706 .0676581 .285283
ratio 1.84 1.208045 2.802546
odds ratio 2.3125 1.255127 4.452503 (exact)

Exact Method

\( n^* (n_{12} + n_{21}) < 10 \)  \( n^* \) = number of discordant pair

\( n^* \) = number of discordant pair

\( n_{12} \) = number of type A discordant pairs

\( n_{21} \) = number of type B discordant pairs

\[
\begin{array}{ccc}
\text{Case} & \text{Control} & \text{Total} \\
\hline
\text{Exposed} & \text{A} & n_{12} \\
\text{Unexposed} & \text{B} & n_{21} \\
\hline
\text{Total} & n & n \\
\end{array}
\]

McNemar's Test for Correlated Proportion : Exact Method

\( n^* (n_{12} + n_{21}) < 10 \)  \( n^* \) = number of discordant pair

(Rosner 2000)

\[
p = 2^n \sum_{k=0}^{n_{12}} \left( \begin{array}{c} n^* \\ k \end{array} \right) \left( \frac{1}{2} \right)^n ; \text{if } n_{12} < \frac{n^*}{2}
\]

\[
p = 2^n \sum_{k=n_{12}}^{n} \left( \begin{array}{c} n^* \\ k \end{array} \right) \left( \frac{1}{2} \right)^n ; \text{if } n_{12} \geq \frac{n^*}{2}
\]

\[
\binom{n^*}{k} = \frac{n^*!}{k!(n^*-k)!}
\]
\[ p = 2 \times \sum_{k=m}^{n^*} \left( \binom{n^*}{k} \frac{1}{2} \right)^k \text{ if } n_{12} > \frac{n^*}{2} \]

\[ p = 2 \times \left[ \frac{8}{7} \left( \frac{1}{2} \right)^8 + \frac{8}{7} \left( \frac{1}{2} \right)^8 \right] \]

\[ \left( \frac{8}{7} \right) = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} (1) \]

\[ p = 2 \left[ \frac{1}{2} \left( \frac{1}{2} \right)^3 + 8 \times \left( \frac{1}{2} \right)^2 \right] = 2 \left( \frac{0.3515625}{0.0703125} \right) = 0.0703125 \]

\[ m \leq c \leq 3719 \]

<table>
<thead>
<tr>
<th>Controls</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

**McNemar's chi2(1) = 4.50**  
**Prob > chi2 = 0.0339**  
**Exact McNemar significance probability = 0.0703**

<table>
<thead>
<tr>
<th>Cases</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

**Proportion with factor**

<table>
<thead>
<tr>
<th>Cases</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Controls</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cases</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>
\[ p = 2 \left( 8 \times \frac{1}{2} + 1 \times \frac{1}{2} \right) = 2 \times 0.03515625 = 0.0703125 \]

<table>
<thead>
<tr>
<th>Controls</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Unexposed</td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

McNemar's chi2(1) = 4.50  Prob > chi2 = 0.0339
Exact McNemar significance probability = 0.0703

Proportion with factor

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
<th>.2</th>
<th>Unexposed</th>
<th>.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.0039</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.0039</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matched Analysis: odds ratio
- estimate the strength association between two paired or dependent dichotomous variables.

\[ \theta = \frac{n_{12}}{n_{21}} \]

\[ (1-\alpha)100\% \text{ CI} = e^{\ln(\theta) \pm z_{\alpha/2} \times se[\ln(\theta)]} \]

\[ se[\ln(\theta)] = \sqrt{\frac{n_{12} + n_{21}}{n_{12}n_{21}}} \]

\[ n = \frac{K}{2} \]

\[ n_{12} > \frac{n}{2} \quad K = 7.8 \]

\[ n_{12} < \frac{n}{2} \quad K = 0.1 \]
STATA Example
```
. mcci 9 37 16 82
Controls               | Cases            | Exposed   Unexposed  | Total
-----------------+------------------------+------------
Exposed | 9 37 | 46
Unexposed | 16 82 | 98
-----------------+------------------------+------------
Total | 25 119 | 144

McNemar’s chi2(1) = 8.32  Prob > chi2 = 0.0039
Exact McNemar significance probability = 0.0055
```

Proportion with factor
```
Cases | .3194444
Controls | .1736111

[95% Conf. Interval]
difference | .1458333 .0427057 .2489609
ratio | 1.84 1.208045 2.802546
rel. diff. | .1764706 .0676581 .285283
odds ratio | 2.3125 1.255127 4.452503 (exact)
```

COMPARING MARGINS OF SQUARE CONTINGENCY TABLES
- Matched pairs analyses generalize to I > 2 outcome categories.
- Let (Y1, Y2) denote the observations for a randomly selected subject.
- A square I x I table \{nij\} shows counts of possible outcomes (i, j) for (Y1, Y2).

```
<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
<th>J</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>n11</td>
<td>n12</td>
<td>n1</td>
<td>n1</td>
</tr>
<tr>
<td>2</td>
<td>n21</td>
<td>n22</td>
<td>n2</td>
<td>n2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>J</td>
</tr>
<tr>
<td>total</td>
<td>n11</td>
<td>n12</td>
<td>n1</td>
<td>n1</td>
</tr>
</tbody>
</table>
```

COMPARING MARGINS OF SQUARE CONTINGENCY TABLES
Let \( \pi_i = P(Y1 = i, Y2 = j) \). Marginal homogeneity is
\( P(Y1 = i) = P(Y2 = i) \) for i = 1, ..., I

```
<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
<th>J</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( \pi_{11} )</td>
<td>( \pi_{12} )</td>
<td>( \pi_1 )</td>
<td>( \pi_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( \pi_{21} )</td>
<td>( \pi_{22} )</td>
<td>( \pi_2 )</td>
<td>( \pi_2 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>J</td>
</tr>
<tr>
<td>total</td>
<td>( \pi_{11} )</td>
<td>( \pi_{12} )</td>
<td>( \pi_1 )</td>
<td>( \pi_1 )</td>
</tr>
</tbody>
</table>
```
- $H_0$: marginal homogeneity by exploiting the large-sample normality of marginal proportions. Let $d_i = p_{ii} - p_{ii}$
- compare the marginal proportions in column $i$ and row $i$.
- Let $d$ be a vector of the first $l-1$ differences.
- It is redundant to include $d_l$, since $\sum d_i = 0$
- the estimated covariance matrix of $d$ is

$$V = (l-1)(l-1) \text{ matrix}$$
$$V_{ii} = (n_{i.} + n_{.i} - 2n_{ii}) ; \quad s_{ij} = -(n_{ij} + n_{ji})$$

$d_i = (l-1) \text{ matrix, or } d_i = p_{ii} - p_{.i}$ or $d_{ij} = n_{i.} - n_{.i}$

---

**Stuart-Maxwell**

$$\chi^2_{\text{Stuart-Maxwell}} = d^t V d$$

$d_i = \text{matrix (}i-1); d_i = n_{i.} - n_{.i}$

$v = \text{matrix (}i-1)(j-1)$

$$v_{i.} = (n_{i.} + n_{.i} - 2n_{ii}) ; \quad s_{ij} = -(n_{ij} + n_{ji})$$

When $k = 3$ Derived by Fleiss & Everitt (1971)

$$\chi^2_{\text{Stuart-Maxwell}} = \frac{\sum (n_{ij}^2 - n_{.i}^2 - n_{.j}^2 + n_{ii}^2)}{2n_{ij}n_{ji}}$$

Let $d_i = n_{i.} - n_{.i}; \quad s_{ij} = \frac{n_{ij} + n_{ji}}{2}$

---

**Square Table for Comparing marginal of Schizophrenia by diagnostics A & B (Fleiss, 1981)**

<table>
<thead>
<tr>
<th>Schizophrenia</th>
<th>Affective</th>
<th>Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia</td>
<td>35</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>Affective</td>
<td>15</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Others</td>
<td>10</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

Let $d_1 = n_{1.} - n_{.1} = 40 - 60$

$$\pi_{23} = \frac{n_{23} + n_{33}}{2} = \frac{5 + 5}{2}$$
\[ \chi^2_{\text{Stuart – Maxwell}} = d'vd \]

\[ d_i = \text{matrix (i-1)}, \quad d_{ij} = n_i - n_j \]

\[ V = \text{matrix (i-1)(j-1)} \]

\[ V_{ij} = (n_i + n_j - 2n_{ij}) \quad s_{ij} = -(n_{ij} + n_{ji}) \]

When \( k=3 \) Derived by Fleiss & Everitt (1971)

\[ d = \begin{bmatrix} 40 & -20 \\ 40 & 10 \end{bmatrix} \]

\[ V = \begin{bmatrix} 60 + 40 - 2(35) & -(5+15) \\ -(5+15) & 30 - 20 \end{bmatrix} \]

\[ = \begin{bmatrix} 30 - 20 \\ -20 & 30 \end{bmatrix} \]

\[ \chi^2 = d' V^2 d = \begin{bmatrix} -20 \end{bmatrix} \begin{bmatrix} 30 - 20 \\ -20 & 30 \end{bmatrix} \begin{bmatrix} -20 \end{bmatrix} \]

\[ = 14.00 \]

\[ \begin{array}{c|cc|c|c|c} \hline & \text{Schizophrenia} & \text{Affective} & \text{Others} & \text{Total} \\ \hline \text{Schizophrenia} & 35 & 5 & 0 & 40 \\ \text{Affective} & 15 & 20 & 5 & 40 \\ \text{Others} & 10 & 5 & 5 & 20 \\ \hline \text{Total} & 60 & 30 & 10 & 100 \\ \hline \end{array} \]

\[ \chi^2_{\text{Stuart-Maxwell}} = \frac{n_1 d_1^2 + n_2 d_2^2 + n_3 d_3^2}{1213 1322 1223} \]

\[ = \frac{5+5}{2} + \frac{0+10}{2} + \frac{5+15}{2} \]

\[ = 14.00 \]

\[ \begin{array}{c c c c c c} \hline \text{Stuart-Maxwell} & \text{Schr.} & \text{Afl.} & \text{Oth.} & \text{Total} \\ \hline \text{Schr.} & 15 & 5 & 0 & 40 \\ \text{Afl.} & 15 & 20 & 5 & 40 \\ \text{Oth.} & 10 & 5 & 5 & 20 \\ \hline \text{Total} & 40 & 30 & 10 & 100 \\ \hline \end{array} \]
<table>
<thead>
<tr>
<th></th>
<th>col1</th>
<th>col2</th>
<th>col3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>row1</td>
<td>35</td>
<td>5</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>row2</td>
<td>15</td>
<td>20</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>row3</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>30</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \frac{(d'V^2d)^{-1}d}{W} \]

- The Bhapkar & Stuart-Maxwell tests are asymptotically equivalent (Keefe, 1982)

\[
d_i = \text{matrix (i-1)} , d_{ij} = n_{i} - n_{j} \\
V = \text{matrix (i-1)(j-1)} \\
V_{i} = (n_{i} + n_{j} - 2n_{ij}) ; s_{ij} = - (n_{ij} + n_{ji}) \\
W = nd'V^2d , \quad \text{Agresti (2002)}
\]

\[
X^2_{\text{Bhapka}} = d'(V - n^{-1}dd')^{-1}d
\]