Chapter 11: Analysis of matched pairs

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis
Example: Prime minister approval (PMA) data. \( n_{++} = 1600 \) voting age British citizens were asked if they approved of the Prime Minister. The same people were asked again 6 months later. The 1600 are cross classified according to their two (binary) responses \((X, Y)\):

<table>
<thead>
<tr>
<th>First survey</th>
<th>Second survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approved</td>
<td>Approved   794</td>
</tr>
<tr>
<td>Disapproved</td>
<td>Approved   86</td>
</tr>
</tbody>
</table>

Here, each person is matched with his or her self. This is also called repeated measures data.

Here we see people tend to approve both times or disapprove both times more often than change their opinion. Question: of those that change their opinion, which direction do they tend to go? Hint: \( \frac{150}{150+794} \approx 0.16 \) & \( \frac{86}{86+570} \approx 0.13 \).
Let $\pi_{ab} = P(X = a, Y = b)$ and $n_{ab}$ be the number of such pairs.

<table>
<thead>
<tr>
<th>First survey</th>
<th>Second survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approve $X = 1$</td>
<td>Approve $Y = 1$</td>
</tr>
<tr>
<td>Disapprove $X = 2$</td>
<td>Disapprove $Y = 2$</td>
</tr>
<tr>
<td></td>
<td>$\pi_{21}$ &amp; $n_{21}$</td>
</tr>
</tbody>
</table>

We assume $(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{mult}\{n_+, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})\}$.

When $\pi_{1+} = \pi_{+1}$ then $P(X = 1) = P(Y = 1)$ and we have *marginal homogeneity*. This is of course equivalent to $P(X = 2) = P(Y = 2)$ by looking at complimentary events.

In the prime minister approval data, this would indicate that the proportion of people that approve at time zero is equal to the proportion that approve at 6 months. Does it imply that no one has changed their mind?

Let $p_{ab} = n_{ab}/n_+$ be the sample proportion in each cell.
How does the proportion change?

Define the difference $\delta = \pi_{+1} - \pi_{1+} = P(Y = 1) - P(X = 1)$. What does this measure for the prime minister approval data?

$\delta$ is estimated by

$$d = p_{+1} - p_{1+} = \frac{n_{11} + n_{21} - (n_{11} + n_{12})}{n_{++}}.$$ 

Considering the covariance for multinomial vector elements, we have a $(1 - \alpha)100\%$ CI for $\delta$ is

$$d \pm z_{\alpha/2} \hat{\sigma}(d),$$

where

$$\hat{\sigma}(d) = \sqrt{\frac{[(p_{12} + p_{21}) - (p_{12} - p_{21})^2]}{n}}.$$
11.1.2 Formal test of $H_0 : \delta = 0$

To test $H_0 : \delta = 0$, i.e. $H_0 : P(X = 1) = P(Y = 1)$, the Wald test statistic is $z_0 = d/\hat{\sigma}(d)$. The score test statistic is

$$z_0 = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}.$$ 

A $p$-value for testing $H_0 : \delta = 0$ is $P(|Z| > |z_0|)$; this latter test is *McNemar’s test*.

For the PMA data, a 95% CI for $\delta$ is $(-0.06, -0.02)$. The number of people approving of the prime minister has dropped by 2% to 6%. The McNemar (score) test statistic for testing $H_0 : P(X = 1) = P(Y = 1)$ is $z_0 = -4.17$ yielding a $p$-value of 0.00003.

Does this mean that between 2% and 6% of the people have changed their minds? (Answer: no).
11.1.4 Using correlation to increase precision

By having a person serve as their own control we increase the precision with which this difference is estimated (relative to two iid samples at an initial time and 6 months later). In some sense it is easier to measure how people's attitudes are changing by looking directly at changes within an individual instead of considering separate populations at time zero and 6 months later.

Note that

\[ n \ \text{var}(d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}). \]

When the response is positively correlated, \( \pi_{11}\pi_{22} > \pi_{12}\pi_{21} \) and the variance is smaller relative to two independent samples.
McNemar’s test statistic is not a function of diagonal elements, but the sample difference $d$ and $\hat{\sigma}(d)$ are. The diagonal elements contribute to how correlated $Y_{i1}$ and $Y_{i2}$ are, i.e. the tendency for people to not change their mind on the PM: $P(Y_{i1} = Y_{i2} = 1) = n_{11}/n_{++}$ and $P(Y_{i1} = Y_{i2} = 2) = n_{22}/n_{++}$. Of those that make a switch, the off-diagonal elements get at the direction and strength of the switch.

We may be interested in the how the odds of approving change over 6 months for a randomly selected individual from the population (conditional inference), or we may be interested in how the odds of approval change across the the two populations: everyone at time zero, and everyone at 6 months.
We can recast this as a *marginal* logit model

$$\text{logit } P(Y_{ij} = 1) = \mu + \beta' x_{ij},$$

where $x_{i1} = 0$ and $x_{i2} = 1$ are “before” and “after” covariates. For the PMA example, the covariates represent time.

In general, $x_{ij}$ are any covariates of interest, but the correlation between $Y_{i1}$ and $Y_{i2}$, $\alpha = \text{corr}(Y_{i1}, Y_{i2})$ must be accounted for in some way in estimating $\beta$. For the PDA example this correlation is quite high, the polychoric correlation is estimated to be $\hat{\rho} = 0.90$ with $\hat{\sigma}(\hat{\rho}) = 0.01$.

We will discuss marginal categorical models that account for such correlation, or *clustering*, fit via GEE in Chapter 12.

When fitting this type of model in GENMOD, $\hat{\beta} = -0.163$ and so $e^{\hat{\beta}} = 0.85$. $\hat{\text{corr}}(Y_{i1}, Y_{i2}) = \hat{\alpha} = 0.70$. 
Let \((Y_{i1}, Y_{i2})\) be a pair of ordered responses from the \(i^{th}\) subject, \(i = 1, \ldots, n\). Consider

\[
\logit P(Y_{ij} = 1) = \alpha_i + \beta x_j,
\]

where \(x_1 = 0\) and \(x_2 = 1\). Here, \(j = 1, 2\) can be thought of as time, with \(Y_{i1}\) denoting the first observation taken on subject \(i\) and \(Y_{i2}\) being the second. Then

\[
\frac{P(Y_{i1} = 1)}{P(Y_{i1} = 0)} = e^{\alpha_i} \quad \text{and} \quad \frac{P(Y_{i2} = 1)}{P(Y_{i2} = 0)} = e^{\alpha_i} e^{\beta}.
\]

And so

\[
\theta_{21} = \frac{P(Y_{i2} = 1)/P(Y_{i2} = 0)}{P(Y_{i1} = 1)/P(Y_{i1} = 0)} = e^\beta,
\]

which does not depend on the subject \(i\).
The $\alpha_1, \ldots, \alpha_n$ are subject-specific effects that correlate $Y_{i1}$ and $Y_{12}$. Large $\alpha_i$ indicates that both $Y_{i1} = 1$ and $Y_{i2} = 1$ are likely. Small $\alpha_i$ indicates that both $Y_{i1} = 0$ and $Y_{i2} = 0$ are likely.

The model assumes that given the $\alpha_1, \ldots, \alpha_n$, the responses are independent. That is, $Y_{i1} \perp Y_{i2} | \alpha_i$ across all $i = 1, \ldots, n$.

An estimate of $e^\beta$ provides a conditional odds ratio. For a given person, the odds of success are $e^\beta$ more likely at time $j = 2$ over time $j = 1$. It is conditional on the value of $\alpha_i$, i.e. the person.

When $\alpha_1 = \alpha_2 = \cdots = \alpha_n$ then there is no person-to-person variability in the response pair $(Y_{i1}, Y_{i2})$. The pairs $(Y_{i1}, Y_{i2})$ are then iid from the population.
The joint mass function for the \( n \) pairs \( \{(Y_{11}, Y_{12}), \ldots, (Y_{n1}, Y_{n2})\} \) is given by

\[
\prod_{i=1}^{n} \left( \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \right)^{y_{i1}} \left( \frac{1}{1 + e^{\alpha_i}} \right)^{1-y_{i1}} \left( \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \right)^{y_{i2}} \left( \frac{1}{1 + e^{\alpha_i + \beta}} \right)^{1-y_{i2}}.
\]

The pairwise success totals \( S_i = y_{i1} + y_{i2} \in \{0, 1, 2\} \) are sufficient for \( \alpha_i \). We can compute (not obvious, see book)

\[
\begin{align*}
P(Y_{i1} = 0, Y_{i2} = 0 | S_i = 0) &= 1 \\
P(Y_{i1} = 1, Y_{i2} = 1 | S_i = 2) &= 1 \\
P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) &= \frac{e^{\beta}}{1 + e^{\beta}} \\
P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) &= \frac{1}{1 + e^{\beta}}
\end{align*}
\]
Conditional inference is based on conditioning on \( \{S_1, \ldots, S_n\} \). Let \( n_{12} = \sum_{i=1}^{n} I\{Y_{i1} = 1, Y_{i2} = 0\} \), \( n_{21} = \sum_{i=1}^{n} I\{Y_{i1} = 0, Y_{i2} = 1\} \), and \( n^* = n_{12} + n_{21} \) are the total number with \( S_i = 1 \). The conditional likelihood is

\[
\prod_{i:S_i=1} \left( \frac{e^\beta}{1 + e^\beta} \right)^{y_{i1}} \left( \frac{1}{1 + e^\beta} \right)^{y_{i2}} = \frac{[e^\beta]^{n_{21}}}{[1 + e^\beta]^{n^*}}.
\]

It pleasantly turns out (p. 421) that \( \hat{\beta} = \log(n_{21}/n_{12}) \) and

\[
\hat{\sigma}(\hat{\beta}) = \sqrt{1/n_{21} + 1/n_{12}}.
\]
We have $\hat{\beta} = \log(86/150) = -0.556$ and $\hat{\sigma}(\hat{\beta}) = 0.135$. So the odds of a randomly selected person saying the prime minister is doing a good job after 6 months is estimated to be $e^{-0.556} = 0.57$ times their initial odds.

An alternative approach to conditioning on sufficient statistics is to specify a full model and treat the $\alpha_i$ as subject-specific random effects. If we can think of subjects as being exchangeable, then a common assumption is

$$\alpha_1, \ldots, \alpha_n \overset{iid}{\sim} N(\mu, \sigma^2).$$

There are only three parameters ($\mu, \sigma, \beta$) in the likelihood (after averaging out the $\alpha_1, \ldots, \alpha_n$). Studies have shown that estimating $\beta$ is robust to the distributional assumption placed on $\alpha_1, \ldots, \alpha_n$. More to come in Chapter 13.
Recall the model

$$\text{logit}P(Y_{ij} = 1|\alpha_i) = \alpha_i + I\{j = 2\}\beta, \quad \alpha_1, \ldots, \alpha_n \overset{\text{iid}}{\sim} (\mu, \sigma^2).$$

Let’s look at more familiar models from STAT 705. Say we want to find how coffee affects myocardial bloodflow (ml/min/g). We can find a sample of $n$ individuals that have not had their morning coffee and measure their bloodflow $Y_{11}, \ldots, Y_{1n_1}$. Then we can find another independent sample of $n$ individuals that had their morning coffee and measure their bloodflow $Y_{21}, \ldots, Y_{2n_2n}$. Assume

$$Y_{11}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} N(\mu, \sigma_y^2) \text{ indep.} \quad Y_{21}, \ldots, Y_{2n_2n} \overset{\text{iid}}{\sim} N(\mu + \beta, \sigma_y^2).$$

Then

$$\hat{\beta} = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot} \sim N(\beta, 2\sigma_y^2).$$
Now instead, let’s block on the individual and record their bloodflow before morning coffee $Y_{1i}$ and after $Y_{2i}$. Assume the mixed effects model

$$Y_{ij} = \alpha_i + I\{j = 2\} \beta + e_{ij}, \quad \alpha_1, \ldots, \alpha_n \iid N(0, \sigma^2_\alpha) \text{ indep. } e_{ij} \iid N(0, \sigma^2).$$

This implies

$$Y_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} \iid N_2 \left( \begin{bmatrix} \mu \\ \mu + \beta \end{bmatrix}, \begin{bmatrix} \sigma^2 + \sigma^2_\alpha & \sigma^2_\alpha \\ \sigma^2_\alpha & \sigma^2 + \sigma^2_\alpha \end{bmatrix} \right).$$

One can then show

$$\hat{\beta} = \bar{Y}_{2\bullet} - \bar{Y}_{1\bullet} \sim N(\beta, 2\sigma^2).$$

Here, $\sigma^2_Y = \sigma^2 + \sigma^2_\alpha$. If $\sigma^2 << \sigma^2_Y$ then precision is improved drastically!
Generalize to repeated measures within a cluster

We can think of taking two or more observations within a cluster (an individual, matched covariates, etc.)

Let \((Y_{i1}, Y_{i2})\) be a pair of correlated binary observations from within the same cluster. The data look like

<table>
<thead>
<tr>
<th>(Y_{i1})</th>
<th>(Y_{i2})</th>
<th>(x_{i1})</th>
<th>(x_{i2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_{11})</td>
<td>(Y_{12})</td>
<td>(x_{11})</td>
<td>(x_{12})</td>
</tr>
<tr>
<td>(Y_{21})</td>
<td>(Y_{22})</td>
<td>(x_{21})</td>
<td>(x_{22})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>(Y_{n1})</td>
<td>(Y_{n2})</td>
<td>(x_{n1})</td>
<td>(x_{n2})</td>
</tr>
</tbody>
</table>

The logit model specifies

\[
\text{logit } P(Y_{ij} = 1) = \alpha_i + x'_{ij}\beta,
\]

where \(i = 1, \ldots, n\) is a pair number and \(j = 1, 2\) denotes the observation within a cluster.
As before, we condition on the sufficient statistics for $\beta$, namely $S_i = Y_{i1} + Y_{i2}$. We have

\[
P(Y_{i1} = Y_{i2} = 0 | S_i = 0) = 1
\]

\[
P(Y_{i1} = Y_{i2} = 1 | S_i = 2) = 1
\]

\[
P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \frac{\exp(x'_{i2}\beta)}{\exp(x'_{i1}\beta) + \exp(x'_{i2}\beta)}
\]

\[
P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) = \frac{\exp(x'_{i1}\beta)}{\exp(x'_{i1}\beta) + \exp(x'_{i2}\beta)}.
\]

The conditional likelihood is formed as before in the simpler case and inference obtained in PROC LOGISTIC using the STRATA statement.

Let’s examine the PMA data using thinking of $(Y_{i1}, Y_{i2})$ as repeated measurements within an individual with corresponding covariates $x_{i1} = 0$ and $x_{i2} = 1$ denoting time.
```
data Data1;
  do ID=1 to 794; approve=1; time=0; output; approve=1; time=1; output; end;
  do ID=795 to 944; approve=1; time=0; output; approve=0; time=1; output; end;
  do ID=945 to 1030; approve=0; time=0; output; approve=1; time=1; output; end;
  do ID=1031 to 1600; approve=0; time=0; output; approve=0; time=1; output; end;
proc logistic data=Data1; strata ID; model approve(event='1')=time;
```

The LOGISTIC Procedure

Conditional Analysis

Testing Global Null Hypothesis: \( \beta = 0 \)

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>17.5752</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score</td>
<td>17.3559</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald</td>
<td>16.9152</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>-0.5563</td>
<td>0.1353</td>
<td>16.9152</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.573</td>
<td>0.440 0.747</td>
</tr>
</tbody>
</table>
11.2.5 Matched case-control studies

Let \((Y_{i1} = 0, Y_{i2} = 1)\) be a pair of binary observations from two different subjects matched on criteria that could affect the outcome. The data look like

<table>
<thead>
<tr>
<th>Control (Y_{i1})</th>
<th>Case (Y_{i2})</th>
<th>Case (x_{i1})</th>
<th>Control (x_{i2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>(x_{11})</td>
<td>(x_{12})</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>(x_{21})</td>
<td>(x_{22})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>(x_{n1})</td>
<td>(x_{n2})</td>
</tr>
</tbody>
</table>

The logit model specifies

\[
\text{logit } P(Y_{ij} = 1) = \alpha_i + \mathbf{x}_{ij}'\mathbf{\beta},
\]

where \(i = 1, \ldots, n\) is a pair number and \(j = 1, 2\) denotes case or control.
\textit{By construction} we have all $S_i = y_{i1} + y_{i2} = 1$ and analogous to our conditional approach for a pair of binary responses within an individual, we have

$$P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \frac{e^{x_{i2}' \beta}}{e^{x_{i1}' \beta} + e^{x_{i2}' \beta}},$$

which does not depend on $\alpha_i$, and the conditional likelihood for $\beta$ is formed by taking the product over $i = 1, \ldots, n$.

Even though the number of cases and the number of controls are fixed at $n$, the logit link allows us to determine the effect of covariates on the \textit{odds} of being a case versus a control. That is the odds of being a case instead of a control is increased by $e^{\beta_j}$ when $x_j$ is increased by unity.
**Example**: \( n_{++} = 144 \) pairs of Navajo Indians, one having myocardial infarction (MI) and the other free of heart disease, were matched on age and gender yielding 288 Navajo total. It is of interest to determine how the presence of diabetes affects the odds of MI. Here’s the cross-classification of the pairs:

<table>
<thead>
<tr>
<th>MI controls</th>
<th>MI cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diabetes</td>
</tr>
<tr>
<td>Diabetes</td>
<td>9</td>
</tr>
<tr>
<td>No diabetes</td>
<td>37</td>
</tr>
</tbody>
</table>

The data are conditionally analyzed using the STRATA subcommand in PROC LOGISTIC.

```plaintext
data Data1;
do ID=1 to 9; case=1; diab=1; output; case=0; diab=1; output; end;
do ID=10 to 25; case=1; diab=0; output; case=0; diab=1; output; end;
do ID=26 to 62; case=1; diab=1; output; case=0; diab=0; output; end;
do ID=63 to 144; case=1; diab=0; output; case=0; diab=0; output; end;
proc logistic data=Data1;
   strata ID;
   model case(event='1')=diab;
```
We estimate that the odds of MI increase by 2.3 when diabetes is present, with a 95% CI of (1.3, 4.2). Diabetes significantly affects the outcome MI.
Endometrial cancer example

The following data is from Breslow and Day (1980) and is analyzed in the SAS documentation. There’s 63 matched pairs, consisting of one case of endometrial cancer (Outcome=1) and a control without cancer (Outcome=0). The case and corresponding control have the same ID, specified in the `strata` subcommand. Two prognostic factors are included: Gall (= 1 for gall bladder disease) and Hyper (= 1 for hypertension). The goal of the case-control analysis is to determine the relative risk of endometrial cancer for gall bladder disease, controlling for the effect of hypertension.

data d1;
  do ID=1 to 63; do Outcome = 1 to 0 by -1; input Gall Hyper @@; output; end; end;
datalines;
0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 1 0 0 1 0 1 0 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 1 1 0 0 1 1 0 1 0 1 0 0 1 0 1 0 0 0 1 1 0 0 1 1 0 0 0 1 0 1 0 0
0 0 1 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 1 0 0
0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 1 1 0 0 0 1 0 1 0 1 0 0
0 0 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 1 0 1
0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0
1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0
;
proc logistic data=d1; strata ID;
  model outcome(event='1')= Gall Hyper; run;
Adjusting for hypertension, the odds of developing endometrial cancer are about 2.6 times as great (and almost significant!) for those with gall bladder disease. How about the relative risk?
Generalization: more than a pair of binary outcomes, \( j = 1, 2, \ldots, J_i \). For example, repeated measures on subject \( i \), or \( J_i \) rats from litter \( i \).

Section 11.1 presented marginal inference, \( \delta = P(Y = 1) - P(X = 1) \). Answers how does probability marginally change, averaged over everyone in population.

Section 11.2 deals with a conditional interpretation. \( \theta_{21} \) was how odds of success change over time \( j = 2 \) versus \( j = 1 \) for any randomly sampled individual in the population.

In matched case-control study, we use the \( \alpha_i \) to induce correlation in responses \( (Y_{i1}, Y_{i2}) \) within two like individuals.

For sparse data, one can include an additional EXACT subcommand in PROC LOGISTIC to get exact tests and odds ratio estimates, e.g. exact diab / estimate=both;
The conditional odds ratio 0.57 is smaller than the population averaged odds ratio 0.85. Is this reasonable? Yes. Many people either like or dislike the PM. If one’s $\alpha_i << 0$ then this person strongly dislikes the PM regardless of $\beta$. After 6 months, this person perhaps dislikes the PM a bit less, but the probability in either case is likely to be small.

Conditional inference holds for an individual with repeated measures, or individuals in a matched (blocked!) set. Because the conditional approach essentially blocks on like variables (measurements within an individual; outcomes matched on gender, age, cholesterol, etc.) it accounts for, and can reduce variability associated with estimating the effect of interest. The marginal inference holds for the population as a whole, averaged over the blocking effects.

Which inference is preferred? It depends on the question!
11.4 Testing for symmetry in a square $I \times I$ table

Consider an $I \times I$ table which cross-classifies $(X, Y)$ on the same outcomes.

<table>
<thead>
<tr>
<th></th>
<th>$Y = 1$</th>
<th>$Y = 2$</th>
<th>$\cdots$</th>
<th>$Y = I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 1$</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$\cdots$</td>
<td>$\pi_{1I}$</td>
</tr>
<tr>
<td>$X = 2$</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$\cdots$</td>
<td>$\pi_{2I}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$X = I$</td>
<td>$\pi_{I1}$</td>
<td>$\pi_{I2}$</td>
<td>$\cdots$</td>
<td>$\pi_{II}$</td>
</tr>
</tbody>
</table>

**Marginal homogeneity** happens when $P(X = i) = P(Y = i)$ ($\pi_{+i} = \pi_{+i}$) for $i = 1, \ldots, I$. This is important, for example, when determining if classifiers (like X-ray readers) tend to classify in roughly the same proportions. If not, perhaps one reader tends to diagnose a disease more often than another reader.

Symmetry, a stronger assumption, implies marginal homogeneity.
An $I \times I$ table is symmetric if $P(X = i, Y = j) = P(X = j, Y = i)$ ($\pi_{ij} = \pi_{ji}$).

This simply reduces the number of parameters from $I^2$ (subject to summing to one) to $I(I + 1)/2$ (subject to summing to one). For example, in a $3 \times 3$ table this forces

<table>
<thead>
<tr>
<th></th>
<th>$Y = 1$</th>
<th>$Y = 2$</th>
<th>$Y = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 1$</td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$\pi_3$</td>
</tr>
<tr>
<td>$X = 2$</td>
<td>$\pi_2$</td>
<td>$\pi_4$</td>
<td>$\pi_5$</td>
</tr>
<tr>
<td>$X = 3$</td>
<td>$\pi_3$</td>
<td>$\pi_5$</td>
<td>$\pi_6$</td>
</tr>
</tbody>
</table>

subject to $\pi_1 + \pi_4 + \pi_6 + 2\pi_2 + 2\pi_3 + 2\pi_5 = 1$.

The symmetric model is easily fit by specifying the cell probabilities by hand in GENMOD. A test of the symmetric model versus the saturated model is a test of $H_0 : \pi_{ij} = \pi_{ji}$ and can be carried out by looking at the Deviance statistic (yielding a LRT).
Recent example

The following table is from Yule (1900)

<table>
<thead>
<tr>
<th></th>
<th>Husband</th>
<th>Wife</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tall</td>
<td>Medium</td>
</tr>
<tr>
<td>Tall</td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td>Medium</td>
<td>20</td>
<td>51</td>
</tr>
<tr>
<td>Short</td>
<td>12</td>
<td>25</td>
</tr>
</tbody>
</table>

Let \((X, Y)\) be the heights of the (Husband, Wife). The table is symmetric if \(P(X = i, Y = j) = P(X = j, Y = i)\). For example, symmetry forces the same proportion of pairings of (Husband, Wife) = (Tall, Short) and (Husband, Wife) = (Short, Tall). This assumes the following structure

<table>
<thead>
<tr>
<th></th>
<th>Husband</th>
<th>Wife</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tall</td>
<td>Medium</td>
</tr>
<tr>
<td>Tall</td>
<td>(\pi_1)</td>
<td>(\pi_2)</td>
</tr>
<tr>
<td>Medium</td>
<td>(\pi_2)</td>
<td>(\pi_4)</td>
</tr>
<tr>
<td>Short</td>
<td>(\pi_3)</td>
<td>(\pi_5)</td>
</tr>
</tbody>
</table>

subject to \(\pi_1 + 2\pi_2 + 2\pi_3 + \pi_4 + 2\pi_5 + \pi_6 = 1\).
data hw;
  input h w symm count @@;
datalines;
  1 1 1 18 1 2 2 28 1 3 3 14
  2 1 2 20 2 2 4 51 2 3 5 28
  3 1 3 12 3 2 5 25 3 3 6 9
;
proc genmod; class symm;
  model count=symm / link=log dist=poi;

The GENMOD output gives us

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>3</td>
<td>1.6635</td>
<td>0.5545</td>
</tr>
</tbody>
</table>

A test of symmetry versus the saturated model gives a $p$-value of $P(\chi^2_3 > 1.66) = 0.65$. We accept that the symmetric model fits. Symmetry implies marginal homogeneity, $P(X = i) = P(Y = i)$. Husbands and wives are tall, medium, or short in the same proportions.

Furthermore, for example, short wives and tall husbands occur with the same probability as tall wives with short husbands.
Consider an $I \times I$ table where $X$ and $Y$ are cross-classified on the same scale. Below are $n = 118$ slides classified for carcinoma of the uterine cervix by two pathologists as (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma \textit{in situ}, or (4) squamous or invasive carcinoma.

<table>
<thead>
<tr>
<th>Pathologist A</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>Total</td>
<td>27</td>
<td>12</td>
<td>69</td>
<td>10</td>
<td>118</td>
</tr>
</tbody>
</table>

If $A$ and $B$ were the same person then $\pi_{ij} = 0$ when $i \neq j$, i.e. there’d only be nonzero diagonal elements. Nonzero off-diagonal elements reflect disagreement and the further off the diagonal they are, the more severe the disagreement.
Marginal homogeneity

For example there are two slides classified by $B$ as carcinoma \textit{in situ} (not metastasized beyond the original site) that $A$ classified as negative.

Perfect agreement occurs when $\pi_{11} + \pi_{22} + \pi_{33} + \pi_{44} = 1$. The strength of agreement has to do with how close this is to one.

Marginal homogeneity occurs when the two classifiers agree on the proportion of each classification in the population, but not necessarily the classifications themselves. If marginal homogeneity is not satisfied, then one classifier tends to classify a fixed category more often than the other.
Classifiers are independent if
\[ P(X = i, Y = j) = P(X = i)P(Y = j), \]
and in this case agreement for category \( i \) happens with probability
\[ P(X = i, Y = i) = P(X = i)P(Y = i) = \pi_i + \pi_i. \]
The kappa statistic looks at the difference between the probability of
agreement \( \sum_{i=1}^{I} \pi_{ii} \) and agreement due to "chance" \( \sum_{i=1}^{I} \pi_i + \pi_i \),
normalized by the largest this can be when \( \sum_{i=1}^{I} \pi_{ii} = 1 \):
\[
\kappa = \frac{\sum_{i=1}^{I} \pi_{ii} - \pi_i + \pi_i}{1 - \sum_{i=1}^{I} \pi_i + \pi_i},
\]
and is estimated by simply replacing \( \pi_{ij} \) by \( \hat{\pi}_{ij} = \frac{n_{ij}}{n_{++}} \).
data table;
  input A B count @@;
datalines;
1 1 22 1 2 2 1 3 2 1 4 0
2 1 5 2 2 7 2 3 14 2 4 0
3 1 0 3 2 2 3 3 36 3 4 0
4 1 0 4 2 1 4 3 17 4 4 10
;
proc freq order=data; weight count; tables A*B / plcorr agree;

The FREQ Procedure

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.9332</td>
<td>0.0340</td>
</tr>
<tr>
<td>Polychoric Correlation</td>
<td>0.9029</td>
<td>0.0307</td>
</tr>
</tbody>
</table>

Test of Symmetry

<table>
<thead>
<tr>
<th>Statistic (S)</th>
<th>Value</th>
<th>ASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr &gt; S</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Kappa Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Kappa</td>
<td>0.4930</td>
<td>0.0567</td>
<td>0.3818</td>
</tr>
<tr>
<td>Weighted Kappa</td>
<td>0.6488</td>
<td>0.0477</td>
<td>0.5554</td>
</tr>
</tbody>
</table>

Sample Size = 118
There’s a test for symmetry! The statistic is the same as the Pearson GOF test for the symmetric log-linear model, i.e. a score test for testing $H_0 : \pi_{ij} = \pi_{ji}$. What do we conclude?

How about $\hat{\gamma} = 0.93$ and $\hat{\rho} = 0.90$, both highly significant? What does that tell us?

Finally, $\hat{\kappa} = 0.49$ with 95% CI about (0.4, 0.6). The difference between observed agreement and that expected purely by chance is between 0.4 and 0.6, moderately strong agreement.

The weighted kappa statistic is valid for an ordinal response and weights differences in classifications according to how “severe” the discrepancy. See p. 435.

$\kappa$ is one number summarizing agreement. It may be much more interesting to quantify where or why disagreement occurs via models.
Recall that McNemar’s test tests $H_0 : P(X = 1) = P(Y = 1)$ for a $2 \times 2$ table. This is output from PROC FREQ in SAS using AGREE.

Often, when comparing raters, we have more than 2 categories. A general test of marginal homogeneity tests $H_0 : P(X = i) = P(Y = i)$ for $i = 1, \ldots, I$. mh is a small program written by John Uebersax to perform overall tests of marginal homogeneity, among other things.

MH Program: Marginal Homogeneity Tests for N x N Tables
Version 1.2 - John Uebersax
2008-04-24 2:19 PM

***INPUT***
Diagnoses of Carcinoma (Agresi Table 10.8)
4 categories
Path A is row variable
Path B is column variable
ordered categories
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>17</td>
<td>10</td>
</tr>
</tbody>
</table>

Total number of cases: 118
### ***BASIC TESTS***

Four-fold tables tested

<table>
<thead>
<tr>
<th></th>
<th>Path A</th>
<th>Path B</th>
<th>Path A</th>
<th>Path B</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>4</td>
<td>5</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>5</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>2</td>
<td>33</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>0</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

McNemar Tests for Each Category

<table>
<thead>
<tr>
<th>Level (k)</th>
<th>Frequency</th>
<th>Proportion (Base Rate)</th>
<th>Chi-squared(a)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Path A</td>
<td>Path B</td>
<td>Path A</td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>27</td>
<td>0.220</td>
<td>0.229</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>12</td>
<td>0.220</td>
<td>0.102</td>
</tr>
<tr>
<td>3</td>
<td>38</td>
<td>69</td>
<td>0.322</td>
<td>0.585</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>10</td>
<td>0.237</td>
<td>0.085</td>
</tr>
</tbody>
</table>

(a) or exact test

* p < Bonferroni-adjusted significance criterion of 0.017.

### Tests of Overall Marginal Homogeneity

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-squared</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhapkar chi-squared</td>
<td>38.528</td>
<td>3</td>
<td>0.0000</td>
</tr>
<tr>
<td>Stuart-Maxwell chi-squared</td>
<td>29.045</td>
<td>3</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

### Bowker Symmetry Test

<table>
<thead>
<tr>
<th>Chi-squared</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.286</td>
<td>6</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
***TESTS FOR ORDERED-CATEGORY DATA***

McNemar Test of Overall Bias
or Direction of Change

--------------------------------------------
Cases where Path A level is higher: 25
Cases where Path B level is higher: 18

Chi-squared = 1.140 df = 1 p = 0.2858

Four-fold tables tested (for thresholds tests)

<table>
<thead>
<tr>
<th></th>
<th>22</th>
<th>4</th>
<th>5</th>
<th>87</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>16</td>
<td>3</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0</td>
<td>18</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Tests of Equal Category Thresholds

---------------------------------------------------------------------
Proportion
of cases
below
level k

<table>
<thead>
<tr>
<th>Level (k)</th>
<th>Threshold(a)</th>
<th>Chi-squared(b)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.771</td>
<td>-0.743</td>
<td>exact test 1.0000</td>
</tr>
<tr>
<td>3</td>
<td>-0.149</td>
<td>-0.439</td>
<td>8.895 0.0029*</td>
</tr>
<tr>
<td>4</td>
<td>0.715</td>
<td>1.374</td>
<td>18.000 0.0000*</td>
</tr>
</tbody>
</table>

(a) for probit model
(b) or exact test

* p < Bonferroni-adjusted significance criterion of 0.017.
***GRAPHIC OUTPUT***
Marginal Distributions of Categories
for Path A (**) and Path B (==)

0.585 +

| ==
| ==
| ==
| ==
| == ** ==
| == ** ==
| ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** == ** =

Notes: x-axis is category number or level.
y-axis is proportion of cases.

Proportion of cases below each level

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path A</td>
<td>Path B</td>
<td>Scale</td>
<td></td>
</tr>
</tbody>
</table>
The Bhapkar test (p. 424, 11.3.1; more powerful than Stuart-Maxwell) for marginal homogeneity is highly significant with $p = 0.0000$. We reject marginal homogeneity. The graphical output indicates that both pathologists tend to classify ‘negative’ in roughly the same proportion, but that $B$ classifies ‘carcinoma in situ’ more often than $A$, whereas $A$ classifies ‘atypical squamous hyperplasia’ and ‘squamous or invasive carcinoma’ more often than $B$.

There is also an individual test for each category. $H_0 : P(X = i) = P(Y = i)$ is rejected for $i = 2, 3, 4$ but not $i = 1$. 
We are interested in whether one rater tends to classify slides ‘higher’ or ‘lower’ than the other. Off-diagonal elements above the diagonal are when B classifies higher than A; elements below the diagonal are when B classifies lower than A. The McNemar test of overall bias is not significant, indicating that one rater does not tend to rate higher or lower than the other.

The test for symmetry has the same test statistic and $p$-value as from SAS.

The program is easy to run on a Windows-based PC and free. There is a users guide and sample input and output files. Web location: http://www.john-uebersax.com/stat/mh.htm.
> library(coin)
Loading required package: survival
Loading required package: splines
> rate=c("N","ASH","CIS","SIC")
> ratings=as.table(matrix(c(22,5,0,0,2,7,2,1,2,14,36,17,0,0,0,10),nrow=4,
+ dimnames=list(PathA=rate,PathB=rate)))
> ratings

PathB
PathA  N  ASH  CIS  SIC
  N   22  2   2   0
  ASH  5  7  14   0
  CIS  0  2  36   0
  SIC  0  1  17  10
> mh_test(ratings)

Asymptotic Marginal-Homogeneity Test

data:  response by
groups (PathA, PathB)
stratified by block
chi-squared = 29.0447, df = 3, p-value = 2.192e-06