Models for Matched pairs

Today’s topics:

1. Introduction

2. Comparing proportions

3. Logit models

4. Loglinear models

5. Rater agreement

6. Preference data

Section skipped:
10.7
Introduction

Example:

<table>
<thead>
<tr>
<th></th>
<th>First Survey</th>
<th>Second Survey</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approve</td>
<td>Disapprove</td>
<td></td>
</tr>
<tr>
<td>First Survey</td>
<td>794</td>
<td>150</td>
<td>944</td>
</tr>
<tr>
<td></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>

The table is square. The row and columns categories are the same.

Let \( n_{ab} \) be the observed count and \( p_{ab} = n_{ab}/n \) be the sample proportion. We treat \( n_{ab} \) as a sample from a multinomial distribution with probabilities \( \pi_{ab} \) and sample size \( n \).

Then \( p_{a+} \) is the proportion in category \( a \) for the first observation and \( p_{+a} \) for the second observation. The goal is to compare these proportions, but with matched pairs these proportions are correlated and methods for independent samples are inappropriate.
Comparing proportions

For binary outcomes when $\pi_{1+} = \pi_{+1}$ then $\pi_{2+} = \pi_{+2}$ also, and there is marginal homogeneity.

Since

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

marginal homogeneity in a $2 \times 2$-table is equal to $\pi_{12} = \pi_{21}$. The table then shows symmetry.
Comparing proportions

For comparing proportions we can use

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

Let an estimate be

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

The 95% confidence interval can be obtained by

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

with

$$\hat{\sigma}^2(d) = \left[ p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21}) \right]$$

$$= \left[ (p_{12} + p_{21}) - (p_{12} - p_{21})^2 \right] / n$$
Comparing proportions

The square of the ratio

\[ \frac{d}{\hat{\sigma}(d)} \]

is a Wald statistic for testing the null-hypothesis of marginal homogeneity.

Under \( H_0 \) an alternative estimate of the estimated variance is

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

The score statistic is \( z_0 = d/\hat{\sigma}_0^2(d) \).

This score statistic simplifies to

\[
z_0 = \frac{n_{21} - n_{12}}{(n_{12} + n_{21})^{1/2}}
\]

This statistic is called McNemar’s statistic.
Subject specific tables

The data can be represented in a different format where for each subject we have a $2 \times 2$ table of survey against response.

As an example for 4 subjects:

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

A test of conditional independence in the $2 \times 2 \times n$ table provides a test of marginal homogeneity. The Cochran Mantel Haenszel statistic can be used, which is equivalent to McNemar’s test.

This table is called the subject-specific table whereas the $2 \times 2$ table presented earlier is the population averaged table.
Logit models for subject specific tables

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

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

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

with a dummy variable \(x_1 = 0\) and \(x_2 = 1\).

Alternatively

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

where the parameter \(\beta\) is a log odds ratio with the marginal distributions,

\[
\log \left[\frac{p_{+1}p_{+2}}{p_{+1}+p_{+2}}\right].
\]

These models are called marginal models. They focus on the marginal distributions of responses for the two observations.
Logit models for subject specific tables

Let \((Y_{i1}, Y_{i2})\) be the pair of observations for subject \(i, i = 1, \ldots, n.\)

The model

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

is a conditional model, since the effect \(\beta\) is defined conditional on the subject. The effect is subject-specific.

For the identity link, subject specific and population averaged effects are identical. For nonlinear models the effects differ.

For the logit link for example

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

The average of these over subjects does not have the form \(\exp(\alpha + \beta x_t)/(1 + \exp(\alpha + \beta x_t)).\)
Logit models for subject specific tables

Estimation of conditional models is generally difficult, since the number of parameters is dependent upon the number of observations.

This can be solved using

1. Conditional likelihood estimation: Condition on the sufficient statistics of the $\alpha_i$ and estimate the $\beta$. This can be done with $R$’s package clogit.

2. Random effects models: Instead of estimating each $\alpha_i$ we assume that the $\alpha_i$ are normally distributed with mean $\mu$ and variance $\sigma^2$. This can be estimated using multilevel software like MLwiN.
Loglinear models for square tables

An alternative analysis of square tables models the joint distribution directly using loglinear or logit models.

An example is the symmetry model. For an $I \times I$-table the joint distribution satisfies symmetry if

$$\pi_{ab} = \pi_{ba}, \text{ whenever } a \neq b$$

Given symmetry we have

$$\pi_{a+} = \sum_b \pi_{ab} = \sum_b \pi_{ba} = \pi_{+a}$$

for all $a$, i.e. we have marginal homogeneity.
Loglinear models for symmetry

For expected frequencies $\mu_{ab} = n\pi_{ab}$ the loglinear form is

$$\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \lambda_{ab}$$

where all $\lambda_{ab} = \lambda_{ba}$. This parametrization requires identification constraints. A simpler form is

$$\log \mu_{ab} = \lambda_{ab}$$

with all $\lambda_{ab} = \lambda_{ba}$.

The likelihood equations are

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \text{ and}$$
$$\hat{\mu}_{aa} = n_{aa}$$

The expected frequencies are

$$\hat{\mu}_{ab} = \frac{n_{ab} + n_{ba}}{2}$$

with $df = I(I - 1)/2$. 

Models for Matched pairs 11.11
**Quasi-symmetry**

The symmetry model often fails, because the marginal distributions differ. To accommodate for marginal heterogeneity the main effects may differ for the row and column categories, i.e.

\[
\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}
\]

where all \(\lambda_{ab} = \lambda_{ba}\). This is the quasi-symmetry model.

The likelihood equations are

\[
\begin{align*}
\hat{\mu}_{ab} + \hat{\mu}_{ba} &= n_{ab} + n_{ba} \\
\hat{\mu}_{a+} &= n_{a+} \\
\hat{\mu}_{+b} &= n_{+b}
\end{align*}
\]

Iterative methods are needed for solving these. For example using the IPF algorithm. The degrees of freedom are \(df = (I - 1)(I - 2)/2\).

The association symmetry implies that odds ratios on one side of the diagonal are equal to odds ratios on the other side.
Quasi-independence

In square tables there is often dependence because of the main diagonal. (The counts on the diagonal are larger than predicted by the independence model).

Conditional on the event that a pair falls of the diagonal, the relationship may have a simple structure (independence).

The quasi-independence model states that the two variables are independent given that the row and column outcomes differ. In loglinear form this is

\[ \log \mu_{ab} = \lambda + \lambda^X_a + \lambda^Y_b + \delta_a I(a = b) \]

where \( I(\cdot) \) is the indicator function

\[ I(a = b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \]
The likelihood equations are

\[
\hat{\mu}_{aa} = n_{aa} \\
\hat{\mu}_{a+} = n_{a+} \\
\hat{\mu}_{+b} = n_{+b}
\]

A perfect fit occurs along the diagonal. Iterative methods are needed for solving these. For example using the IPF algorithm. The degrees of freedom are \( df = (I - 1)^2 - I \).

The quasi-independence model applies to tables with \( I \geq 3 \).
Marginal Homogeneity

Marginal homogeneity is not a loglinear model. However, when there is quasi-symmetry and there is marginal homogeneity, the symmetry model results.

Therefore the likelihood ratio statistic

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

is a test of marginal homogeneity, with df = \( I - 1 \). It assumes that the Quasi-symmetry model holds.
Agreement between observers

Agreement and association are different facets of a joint distribution. Strong agreement requires strong association, but there may exist strong association without agreement.

The baseline model for agreement is the independence model, its residuals show patterns of agreement and disagreement.

The quasi-independence model can be used as a more complex model of agreement. Agreement is then represented by the $\delta_a$’s, i.e. larger values represent stronger agreement.

An alternative approach tries to summarize agreement in a single measure. **Cohen’s kappa** is such a measure:

$$\kappa = \frac{\sum_a \pi_{aa} - \sum_a \pi_a + \pi_a + a}{1 - \sum_a \pi_a + \pi_a + a}$$

$\kappa$ equals zero when the agreement merely equals that expected under independence. It equals 1 when perfect agreement occurs.

The sample value $\hat{\kappa}$ has a large-sample normal distribution. Using its estimated variance confidence intervals can be obtained which are more useful than the test $\kappa = 0$. 

Models for Matched pairs 11.16
Square tables may arise from pairwise preference statements. In these kind of data $n_{ab} + n_{ba}$ is constant, i.e. the number of judges.

Let $\pi_{ab}$ be the probability that $a$ is preferred over $b$. The Bradley-Terry model is a logit model for these preferences, i.e.

$$\log \frac{\pi_{ab}}{\pi_{ba}} = \beta_a - \beta_b$$

Identifiability restrictions are required.

The Bradley-Terry model is a logit formulation of quasi-symmetry.

$$\log \frac{\mu_{ab}}{\mu_{ba}} = (\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}) - (\lambda + \lambda_b^X + \lambda_a^Y + \lambda_{ba})$$

$$= (\lambda_a^X - \lambda_a^Y) - (\lambda_b^X - \lambda_b^Y)$$

$$= \beta_a - \beta_b$$