Categorical Data Analysis
Lecture 14

Graduate School of Advanced Science and Engineering
Rei Monden

2025.1.23

Models for matched pairs

Models for matched pairs

Today we will see how to compare categorical responses for two groups with paired observations.

Such data are common, for example, in

- Longitudinal designs.
- Designs including two similar variables with the same response categories.

Comparing dependent proportions for binary matched pairs

Comparing dependent proportions for binary matched pairs

Y_1	Y_2		Total
	Yes	No	
Yes	n_{11}	n_{12}	$n_{1\cdot}$
No	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Observed counts per cell.

Y_1	Y_2	
	Yes	No
Yes	π_{11}	π_{12}
No	π_{21}	π_{22}

True probabilities per cell.

Let $\pi_{ij} = P(Y_1 = i, Y_2 = j)$ denote the (i, j)-th cell probability, for $i, j = 1, 2$.

There are two *marginal* probabilities of success:

- $P(Y_1 = 1) = \pi_{11} + \pi_{12}$
- $P(Y_2 = 1) = \pi_{11} + \pi_{21}$.

Comparing dependent proportions for binary matched pairs

Y_1	Y_2		Total
	Yes	No	
Yes	n_{11}	n_{12}	$n_{1\cdot}$
No	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Observed counts per cell.

Y_1	Y_2	
	Yes	No
Yes	π_{11}	π_{12}
No	π_{21}	π_{22}

True probabilities per cell.

There's marginal homogeneity when both probabilities of success coincide:

$$P(Y_1 = 1) = P(Y_2 = 1),$$

or equivalently when

$$\pi_{12} = \pi_{21}.$$

McNemar test comparing marginal proportions

The McNemar test is suitable for binary response, matched-pairs, data.

This test can be used to compare π_{12} and π_{21}.

$\mathcal{H}_0 : \pi_{12} = \pi_{21}$ (marginal homogeneity).
$\mathcal{H}_1 : \pi_{12} \neq \pi_{21}$ (no marginal homogeneity).

Under $\mathcal{H}_0$, we expect to observe $n_{12} \approx n_{21}$.

Denoting $n^* = n_{12} + n_{21}$, then under $\mathcal{H}_0$ we expect that

$$n_{12} \sim \text{Bin}(n^*, .5)$$

In words:
If $\mathcal{H}_0$ is true, we expect about half of the n^* observations to belong to each cell.

McNemar test comparing marginal proportions

$$\boxed{n_{12} \sim \text{Bin}(n^*, .5)}$$

We can now use the normal approximation to the binomial distribution:

If $X \sim Bin(n, p)$ *then* $X \underset{\text{approx.}}{\sim} \mathcal{N}(np, \sqrt{np(1-p)})$, *if* np *and* $n(1-p)$ *are at least 5.*

Since under $\mathcal{H}_0$ we have $p = .5$, then a minimum request to use the above approximation is to assume that $n^* > 10$.

Thus, assuming that $n^* > 10$, we have the McNemar test statistic:

$$\frac{n_{12} - .5n^*}{\sqrt{n^*(.5)(.5)}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0, 1).$$

McNemar test comparing marginal proportions – Example

Pay higher taxes (y_1)	Cut living standards (y_2)		Total
	Yes	No	
Yes	227	132	359
No	107	678	785
Total	334	810	1144

Opinions related to environment.

```
# Import data frame from file:
environ.df <- read.table("Envir_opinions.dat", header = TRUE)
environ.df
```

Output:

```
  person y1 y2
1      1  1  1
2      2  1  1
3      3  1  1
..............
```

McNemar test comparing marginal proportions – Example

```
# 2x2 contingency table:
environ.tab <- xtabs(~y1 + y2, data = environ.df)
environ.tab
```

Output:

```
   y2
y1    1   2
  1 227 132
  2 107 678
```

McNemar test comparing marginal proportions – Example

```
# McNemar test:
mcnemar.test(environ.tab,
             correct=FALSE # no continuity correction
             )
```

Output:

```
------------------------------------------------------------
  McNemar's chi-squared = 2.6151, df = 1, p-value = 0.1059
```

McNemar test comparing marginal proportions – Example

Pay higher taxes (y_1)	Cut living standards (y_2)		Total
	Yes	No	
Yes	227	132	359
No	107	678	785
Total	334	810	1144

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0,1)$$

```
------------------------------------------------------------
  McNemar's chi-squared = 2.6151, df = 1, p-value = 0.1059
```

Indeed,

- $z = \frac{132-107}{\sqrt{132+107}} = 1.6171$
- $p = P(|Z| > 1.6171) = .106.$

Similarly, nothing that if $X \sim \mathcal{N}(0,1)$ then $X^2 \sim \chi^2(1)$:

- $x = 1.6171^2 = 2.6151$
- $p = P(X^2_{df=1} > 2.6151) = .106.$

McNemar test comparing marginal proportions – Example

Pay higher taxes (y_1)	Cut living standards (y_2)		Total
	Yes	No	
Yes	227	132	359
No	107	678	785
Total	334	810	1144

Conclusion: $\chi^2(1) = 2.62$, $p = .106$.
We fail to reject the marginal homogeneity assumption.

In other words, the sample marginal proportions

$$\widehat{P}(Y_1 = 1) = \frac{359}{1144} = .314 \quad \text{and} \quad \widehat{P}(Y_2 = 1) = \frac{334}{1144} = .292$$

are not statistically different from each other.

Estimating the difference between dependent proportions

$$\widehat{P}(Y_1 = 1) = \frac{359}{1144} = .314 \quad \text{and} \quad \widehat{P}(Y_2 = 1) = \frac{334}{1144} = .292$$

The estimated difference between the two marginal proportions is therefore

$$\widehat{P}(Y_1 = 1) - \widehat{P}(Y_2 = 1) = .314 - .292 = .022.$$

Let's use R to compute both the Wald and the score 95% CIs for this difference.

Estimating the difference between dependent proportions

```
environ.tab <- xtabs(~y1 + y2, data = environ.df)

library(PropCIs)

# Wald 95% CI:
diffpropci.Wald.mp(environ.tab[2, 1], # n21
                   environ.tab[1, 2], # n12
                   sum(environ.tab),  # n
                   0.95               # confidence level
                   )
```

Output:

```
95 percent confidence interval:
 -0.004602847  0.048309140
sample estimates:
[1] 0.02185315
```

Estimating the difference between dependent proportions

```
environ.tab <- xtabs(~y1 + y2, data = environ.df)

library(PropCIs)

# Score 95% CI:
scoreci.mp(environ.tab[2, 1], # n21
           environ.tab[1, 2], # n12
           sum(environ.tab),  # n
           0.95               # confidence level
)
```

Output:

```
95 percent confidence interval:
 -0.004661338  0.048494581
```

Analyzing rater agreement

Analyzing rater agreement

We continue with matched-pair data, but now under a different perspective:

> *Each matched pair now consists of ratings of two observers on a common object. The ratings are on a categorical scale.*

The main goal is to establish the strength of agreement between the two observers.

Analyzing rater agreement – Independence model

We can fit the loglinear model of independence to these data (recall Chapter 7):

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y.$$

It will indicate the degree of *agreement* expected if no *association* exists between the ratings.

Notice that agreement and association are not the same:

- Agreement: Scores from both raters coincide.
- Association: Scores from both raters covary (without necessarily coinciding).

E.g., if rater A is systematically more liberal than rater B, scores will have low *agreement* but high association.

Analyzing rater agreement – Independence model

The most interesting element of the loglinear model of independence in the current context is actually its cell standardized residuals:

> *Cell with positive std. residual $\Rightarrow$ the cell has more counts than expected under independence.*

Ideally, large positive standardized residuals occur on the main diagonal (agreement cells).

Analyzing rater agreement – Independence model

Pathologist X	Pathologist Y				Total
	1	2	3	4	
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

Diagnoses of carcinoma.

These are the ratings of two pathologists on the presence and extent of carcinoma in 118 slides.

The rating scale is ordered from 1 (least severe) through 4 (most severe).

Analyzing rater agreement – Independence model

Pathologist X	Pathologist Y				Total
	1	2	3	4	
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

Diagnoses of carcinoma.

Perfect agreement would imply that all 118 cases fell on the main diagonal.

Perfect disagreement would imply that 0 cases fell on the main diagonal.

Clearly here we have a midterm:

$$\text{proportion of agreement} = \frac{22 + 7 + 36 + 10}{118} = .636.$$

Analyzing rater agreement – Independence model

```
# Import data frame from file:
pathol.df <- read.table("Pathologists.dat", header = TRUE)
pathol.df
```

Output:

```
   X Y count
1  1 1    22
2  1 2     2
3  1 3     2
4  1 4     0
5  2 1     5
-------------
```

Analyzing rater agreement – Independence model

```
pathol.ind <- glm(count ~ factor(X) + factor(Y),
                  family = poisson(link = "log"),
                  data   = pathol.df)

summary(pathol.ind)
```

Output:

```
Residual deviance: 117.96  on  9  degrees of freedom
```

Conclusion: $\chi^2(9) = 118.0$, $p < .001$.
The independence model fits significantly worse than the saturated model.

This is a minor problem.

What we care about in this context is to look at the standardized residuals.

Analyzing rater agreement – Independence model

```
ind.stdres <- rstandard(pathol.ind, type = "pearson")
matrix(ind.stdres, ncol = 4, byrow = TRUE)
```

Output:

```
       [,1]   [,2]   [,3]   [,4]
[1,]  8.487 -0.473 -5.951 -1.757
[2,] -0.502  3.201 -0.542 -1.757
[3,] -4.078 -1.215  5.509 -2.278
[4,] -3.300 -1.323  0.275  5.926
```

Conclusion:
The main diagonal displays large positive standardized residuals, which is *great* news:

> *Agreement for each category is greater than expected by chance (i.e., under independence).*

So we *do* have some evidence of agreement between both raters!

Analyzing rater agreement – Quasi-independence model

The quasi-independence model is better to assess rater agreement.

Here's the model:

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

where

$$I(i = j) = \begin{cases} 1, i = j \\ 0, \text{otherwise} \end{cases}.$$

The model adds one parameter, δ_i, per cell in the main diagonal.

$\delta_i > 0 \quad \Rightarrow \quad$ there are more agreements in cell (i, i) than expected under independence.

We can then fit this model and focus on these parameters.

Analyzing rater agreement – Quasi-independence model

```
# Import data frame from file:
pathol.df <- read.table("Pathologists.dat", header = TRUE)
pathol.df
```

Output:

```
  X Y count diag
1 1 1    22    1
2 1 2     2    0
3 1 3     2    0
4 1 4     0    0
5 2 1     5    0
-------------
```

Analyzing rater agreement – Quasi-independence model

```
pathol.qind <- glm(count ~ factor(X) + factor(Y) + factor(diag),
                   family = poisson(link = "log"),
                   data   = pathol.df)

summary(pathol.qind)
```

Output:

```
factor(diag)1    3.8611     0.7297   5.291 1.22e-07 ***
factor(diag)2    0.6042     0.6900   0.876  0.38119
factor(diag)3    1.9025     0.8367   2.274  0.02298 *
---------------------------------------------------------
Residual deviance:  13.178  on  5  degrees of freedom
```

Conclusion:

- This model still fits significantly worse than the saturated model, but much less so ($\chi^2(5) = 13.18$, $p = .02$).
- The first three parameters are indeed positive, and δ_1 and δ_3 is significantly different from 0. This is evidence towards rater agreement.

Note: In this case, δ_4 cannot be estimated due to constraints imposed by the zero-cells in the data.

Analyzing rater agreement – Quasi-independence model

Observed counts

Pathologist X	Pathologist Y				Total
	1	2	3	4	
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

Expected counts (quasi-independence model)

Pathologist X	Pathologist Y				Total
	1	2	3	4	
1	22	0.7	3.3	0	26
2	2.4	7	16.6	0	26
3	0.8	1.2	36	0	38
4	1.9	3.0	13.1	10	28
Total	27	12	69	10	118

The expected counts are computed as follows:

```
pathol.qind.pred <- predict(pathol.qind, type = "response")
matrix(pathol.qind.pred, ncol = 4, byrow = TRUE)
```

Analyzing rater agreement – Quasi-independence model

Observed counts

Pathologist X	Pathologist Y				Total
	1	2	3	4	
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

Expected counts (quasi-independence model)

Pathologist X	Pathologist Y				Total
	1	2	3	4	
1	22	0.7	3.3	0	26
2	2.4	7	16.6	0	26
3	0.8	1.2	36	0	38
4	1.9	3.0	13.1	10	28
Total	27	12	69	10	118

Thus the quasi-independence model:

- Fits the agreement conditions perfectly.
- The disagreement conditions are still modeled under the independence assumption.

Analyzing rater agreement – Cohen's kappa

One last note concerns the well-known Cohen's kappa coefficient of agreement.

This is a number instead of a model:

$$\kappa = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}.$$

κ compares the observed agreement to the expected agreement if the ratings were independent:

- $\kappa = 0$: Agreement as under independence.
- $\kappa = 1$: Perfect agreement.

Analyzing rater agreement – Cohen's kappa

```
library(psych)

pathol.mat <- matrix(pathol.df$count,
                     ncol  = 4,
                     byrow = TRUE)
pathol.mat
```

Output:

```
     [,1] [,2] [,3] [,4]
[1,]   22    2    2    0
[2,]    5    7   14    0
[3,]    0    2   36    0
[4,]    0    1   17   10
```

Analyzing rater agreement – Cohen's kappa

```
cohen.kappa(pathol.mat)
```

Output:

```
Cohen Kappa and Weighted Kappa correlation coefficients and CIs
                 lower estimate upper
unweighted kappa  0.38     0.49  0.60
-----------------------------------------
```

Conclusion:

The difference between the observed agreement to the expected agreement if the ratings were independent is about 50% of the maximum possible difference.

Analyzing rater agreement – Cohen's kappa

The loglinear modeling approach is better than Cohen's κ, because κ is too sensitive to the marginal distributions of the contingency table.

Exercise 14-1

Apply the McNemar test to the data on smoking and birth weight (`birthweight.csv`). For both the 'low birth weight' and the 'normal birth weight' variables, the codes are 0 = nonsmoker, 1 = smoker.

Provide all the R code and interpret the results (use the significance level 5%).

Exercise 14-2

Subjects were asked whether they believe in heaven and whether they believe in hell (yes or no answer). The results are stored in the file `Heaven.csv`. For both variables, 0 = No, 1 = Yes.

a. Test the hypothesis that the population proportions answering *yes* were identical for heaven and hell. Use significance level 5%.
b. Find a 90% confidence interval for the difference between the population proportions using both the Wald test and Score test. Interpret.

Exercise 14-3

Two neurologists A and B gave diagnoses of multiple sclerosis. The results are stored in the `Neurologists.dat`. The diagnoses are categorized into 4 groups:

- 1 = Certain multiple sclerosis
- 2 = Probable multiple sclerosis
- 3 = Possible multiple sclerosis
- 4 = Doubtful, unlikely or definitely not multiple sclerosis

a. Use the independence model and residuals to study the pattern of agreement. Interpret.
b. Use the quasi-independence model to study the pattern and strength of agreement between the neurologists. Interpret the results.
c. Use kappa to describe agreement. Interpret.

Exercise 14-1 (日本語)

Moodle の Example data フォルダー内にある喫煙と出生体重に関するデータ（birthweight.csv）に McNemar 検定を適用せよ．「低出生体重」と「正常出生体重」の両変数において, 0 = 非喫煙者, 1 = 喫煙者を表している．

分析に用いた R コードをレポートに表示し, 結果を解釈せよ（有意水準は 5% とする）．

Exercise 14-2 (日本語)

実験参加者に対して，「天国を信じるか？(0= いいえ, 1= はい)」「地獄を信じるか？(0= いいえ, 1= はい)」という質問を集計した結果が Moodle の Example data フォルダー内の `Heaven.csv` に保存されている．

a. 天国と地獄に「はい」と回答した母集団の割合が同一であるという仮説を検定せよ．ただし，分析に用いた R コードをレポートに表示し，有意水準は 5% とする．
b. 母集団の割合の差に対する 90% 信頼区間を Wald 検定およびスコア検定の両方を用いて求め，結果を解釈せよ．ただし，分析に用いた R コードをレポートに表示すること．

Exercise 14-3 (日本語)

2 人の神経内科医 A と B が多発性硬化症の診断を行いました。その結果が `Neurologists.dat` に保存されています。診断は以下の 4 つのグループに分類されています:

- 1 = 確定的な多発性硬化症
- 2 = 可能性の高い多発性硬化症
- 3 = 可能性のある多発性硬化症
- 4 = 疑わしい、多発性硬化症である可能性は低い、または確実に多発性硬化症ではない

a. 独立モデルと残差を使用して, 一致のパターンを調べよ. 結果を解釈せよ.
b. 準独立モデルを使用して, 神経科医間の一致のパターンおよび強度を調べよ. 結果を解釈せよ.
c. カッパ係数を使用して一致の度合いを表し, その結果を解釈せよ.

Final Exam

- The final exam will be given on Thursday, Jauary 30th 8:45 – 12:00 at **EDU** K201.
- The exam will consist entirely of multiple-choice questions, and you will need a black pen to fill out the answer sheet. Please ensure you bring your own black pen to the exam.
- The questions will cover the material discussed in the lectures up to and including Lecture 14.

最終筆記試験

- 最終筆記試験は第 1 回の講義でアナウンスした通り、2025 年 1 月 30 日 (木)8:45 – 12:00に教 K201 で行います。
- 試験はすべて選択式の問題で構成され、回答用紙を記入するために黒のペンが必要です。必ずご自身で黒のペンを持参してください。
- テスト範囲は、第 14 回講義までに扱った内容です。