Categorical Data Analysis Note
Edgard Maboudou

Chapter 8: Models for Matched Pairs

Example

- 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, 880 indicated approval.
- n individuals: response at occasion 1.
- n individuals: response at occasion 2.

		Occasion two		
		1	2	
Occasion one	1	n11	n12	n1+
	2	n21	n22	n2+
		n+1	n+2	n

		Response two		
		1	2	
Response one	1	p11	p12	p1+
	2	p21	p22	p2+
		p+1	p+2	1

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.
- 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 $944/1600 = 0.59$ and $880/1600 = 0.55$.
- These marginal proportions are correlated, and statistical analyses must recognize this.

- Let π_{ij} = probability that a subject makes response i at survey 1 and response j at survey 2.
- The probabilities of approval at the two surveys are π_{1+} and π_{+1}, the first row and first column totals.
- Are π_{1+} and π_{+1} the same?

- When these are identical, the probabilities of disapproval are also identical, and there is *marginal homogeneity*.
- Marginal homogeneity is equivalent to equality of off-main diagonal probabilities; that is $\pi_{12} = \pi_{21}$.
- The table shows *symmetry* across the main diagonal.
- $p_{1+} = \frac{n_{1+}}{n}$ and $p_{+1} = \frac{n_{+1}}{n}$ are not independent since we have dependent samples.
- Thus, we cannot use earlier methods.
- For example, CI

$$p_{1+} - p_{+1} \pm z_{\alpha/2}\sqrt{\frac{p_{1+}(1-p_{1+})}{n} + \frac{p_{+1}(1-p_{+1})}{n}}.$$

  is not appropriate

Dependent Categorical Data

- Use $\delta = \pi_{1+} - \pi_{+1}$
- Let $d = p_{1+} - p_{+1}$
- From the results on multinomial distributions,

- Thus,

- For large samples, d has approximately a normal distribution.
- A confidence interval for δ is then $d \pm z_{\alpha/2}\hat{\sigma(d)}$
- Here

- The hypothesis of marginal homogeneity is H_0: $\pi_{1+} = \pi_{+1}$ (i.e. $\delta = 0$)
- Wald test statistic is:

McNemar's Test

For large n

- Formal test for H_0: $\pi_{1+} = \pi_{+1}$ (test for marginal homogeneity).
- Under H_0, an alternative estimated variance is

$$\hat{\sigma}^2 = \frac{p_{12} + p_{21}}{n} = \frac{n_{12} + n_{21}}{n^2}.$$

- $p_{1+} - p_{+1}$ is approximately $N(0, \sigma^2)$ if H_0 is true, where σ^2 is estimated by $\hat{\sigma}^2$ above
- The score test statistic is

- The square of z_0 is a chi-squared distrbution with $df = 1$
- The test used here is called **McNemar's test**.
- It depends only on cases classified in different categories for the two observations.
- Large sample if $n_* = n_{12} + n_{21} > 10$

Small Sample Test

- The null hypothesis of marginal homogeneity for binary matched pairs is H_0: $\pi_{12} = \pi_{21}$ or $\frac{\pi_{12}}{\pi_{12}+\pi_{21}} = 0.5$.
- 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, i.e.
- Given n_*, $n_{12} \sim Bin\left(n_*, \frac{\pi_{12}}{\pi_{12}+\pi_{21}} = 0.5\right)$.
- So the exact test computes P-value of observed n_{12} from binomial distribution.

Example: Each of 50 individuals were asked if they approved of some political issue. One year later, the same survey is repeated again.

		Time 2		
		App	Dis	
Time	App	18	8	26
1	Dis	4	20	24
		22	28	50

Has the proportion approving changed?

(Example continued)

In R, we can use the command mcnemar.test as follow:

```
> table10.1<-matrix(c(18,8,4,20),byrow=T,ncol=2);table10.1
     [,1] [,2]
[1,]   18    8
[2,]    4   20
> mcnemar.test(table10.1,correct=F)

        McNemar's Chi-squared test

data:  table10.1
McNemar's chi-squared = 1.3333, df = 1, p-value = 0.2482

> mcnemar.test(table10.1,correct=T)

        McNemar's Chi-squared test with continuity correction

data:  table10.1
McNemar's chi-squared = 0.75, df = 1, p-value = 0.3865
```

Also, we can use small sample test.

Example: 95% CI for $\pi_{1+} - \pi_{+1}$ from previous example.
Sample Proportions

In R,

```
> t10.1_prop<-prop.table(table10.1)## to get the table of sample proportions
> margin.table(t10.1_prop,1) #marginal sum for rows
[1] 0.52 0.48
> margin.table(t10.1_prop,2) #marginal sum for columns
[1] 0.44 0.56
> prop.diff<-margin.table(t10.1_prop,2)[1]-margin.table(t10.1_prop,1)[1]
> off.diag<-diag(t10.1_prop[1:2,2:1]);off.diag
[1] 0.16 0.08
> ci<-prop.diff+c(-1,1)*qnorm(.975)*sqrt((sum(off.diag)-diff(off.diag)^2)/sum(ta
> ci
[1] -0.21396752  0.05396752
```

Symmetry and Quasi-symmetry Models

- For a 2×2 table, marginal homogeneity means $\pi_{1+} = \pi_{+1}$, which is equivalent to $\pi_{12} = \pi_{21}$.
- That is symmetry from the table: $\pi_{ij} = \pi_{ji}$ for all i, j.
- Symmetry implies Marginal homogeneity, since $\pi_{i+} = \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$.
- Yet, the converse is not true. Marginal homogeneity does not imply symmetry except when $I = 2$.

Symmetry Model for Square Tables

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

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

- The probabilities in the square table satisfy *symmetry*:

$$\pi_{ij} = \pi_{ji} \quad \text{for all} \quad i, j$$

- When $I > 2$, though, marginal homogeneity can occur without symmetry.
- 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:

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij}, \quad \text{where} \quad \lambda_{ij} = \lambda_{ji}$$

- This is the special case of the saturated loglinear model with $\lambda_{ij}^{XY} = \lambda_{ji}^{XY}$ and $\lambda_i^X = \lambda_i^Y$
- The ML fit of the *symmetry* model is

$$\hat{\mu}_{ij} = \frac{n_{ij} + n_{ji}}{2}.$$

- the model has expected frequency

$$\hat{\mu}_{ij} = \frac{n_{ij} + n_{ji}}{2}.$$

- We have $\log \mu_{ij} = \log \mu_{ji}$ so that $\mu_{ij} = \mu_{ji}$
- 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 equals $I(I-1)/2$.
- The adjusted residuals equal

- Only one residual for each pair of categories is non-redundant, since $r_{ij} = -r_{ji}$

Quasi-symmetry Model

- 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.
- It is like the previous model but allows marginal heterogeneity by rewriting $\lambda_i^X \neq \lambda_j^Y$.
- The resulting model, called the **quasi-symmetry model**, is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}, \quad \text{where} \quad \lambda_{ij} = \lambda_{ji}$$

- Symmetry and quasi-symmetry models have symmetric association.

$$\log \theta_{ij} = \log \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i+1,j}\mu_{i,j+1}}$$

- The fitted marginal totals equal the observed totals $\hat{\mu}_{i+} = n_{i+}$ and $\hat{\mu}_{+i} = n_{+i}$, $i = 1, \ldots, I$
- 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 log-linear models.

Ex. Migration data from the US census compares region of residence in 1985 with that of 1980 for 55,981 people. (Match pairs)

		1985			
		NE	NW	S	W
1980	NE	11607	100	366	124
	NW	87	13677	515	302
	S	172	225	17819	270
	W	63	176	286	10192

For the symmetry model, the R output is

```
> residence80<-c("NE","NW","S","W")
> residence80<-factor(residence80,levels=residence80)
> residence85<-residence80
> table<-expand.grid(res80=residence80,res85=residence85)
> table$count<-c(11607,100,366,124,87,13677,515,302,172,225,17819,270,63,176,286
> table$sym<-paste(pmin(as.numeric(table$res80),as.numeric(table$res85)),
+ pmax(as.numeric(table$res80),as.numeric(table$res85)),sep=",")
> table$sym<-factor(table$sym,levels=rev(table$sym))
> (fit.sym<-glm(count~sym,family=poisson(log),data=table))

Call:  glm(formula = count ~ sym, family = poisson(log), data = table)

Coefficients:
(Intercept)       sym3,4       sym2,4       sym1,4       sym3,3       sym2,3
     9.2294      -3.6017      -3.7529      -4.6914       0.5587      -3.3159
     sym2,2       sym1,2       sym1,1
     0.2941      -4.6914       0.1300
```

```
Degrees of Freedom: 15 Total (i.e. Null);  6 Residual
Null Deviance:      131000
Residual Deviance: 243.6        AIC: 393.7
```

For the quasi-symmetry model, the R output is

```
> options(contrast=c("contr.treatment","contr.poly"))
> table$res80a<-factor(table$res80,levels=rev(residence80))
> table$res85a<-factor(table$res85,levels=rev(residence80))
> (fit.qsym<-glm(count~sym+res80,family=poisson(log),data=table))

Call:  glm(formula = count ~ sym + res80, family = poisson(log), data = table)

Coefficients:
(Intercept)       sym3,4       sym2,4       sym1,4       sym3,3       sym2,3
    8.55762     -3.66438     -3.48922     -4.41090      0.43707     -3.13298
     sym2,2       sym1,2       sym1,1      res80NW       res80S       res80W
    0.91689     -4.04444      0.80174      0.04896      0.79333      0.67174

Degrees of Freedom: 15 Total (i.e. Null);  3 Residual
Null Deviance:      131000
Residual Deviance: 2.986        AIC: 159.2
> 1-pchisq(fit.qsym$deviance,df=fit.qsym$df.residual)
[1] 0.3937946
```

$\hat{\theta}_{13}$: the odds of living S as opposed to W in 1985 are higher by a factor of 1.86 for those in 1980 in NE compared to those in NW.

$\hat{\theta}_{31}$: the odds of living in the NE as opposed to NW in 1985 is higher by a factor of 1.86 for those in the S in 1980 compared to those in the W.

```
> exp<-fit.qsym$fitted.values
> obs<-table$count
> cbind(obs,exp)
      obs          exp
1   11607 11607.00000
2     100    95.78862
3     366   370.43747
4     124   123.77391
5      87    91.21138
6   13677 13677.00000
7     515   501.68254
8     302   311.10608
9     172   167.56253
10    225   238.31746
11  17819 17819.00000
12    270   261.12001
13     63    63.22609
14    176   166.89392
15    286   294.87999
16  10192 10192.00000
```

In the final:
– bring your own tables, calculator, two cheat sheets front-and-back.
– 80% of the exam will be on the last materials.
– 4 parts:
I. Fit-model with logit from a loglinear.
II. Model with ordinal associations.
III. McNemar Test.
part 1: do the test
part 2: confidence interval.
IV. Symmetry and Quasi-symmetry Models MAY BE there.