Mathematical Biostatistics Boot Camp 2: Lecture 11, Matched Two by Two Tables

Brian Caffo

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

October 3, 2013

Table of contents

Matched pairs binary data

First survey	Second Survey: Approve	Second Survey: Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Controls	Cases: Exposed	Cases: Unexposed	Total
Exposed	27	29	56
Unexposed	3	4	7
Total	30	33	63

[1]

[1] Both data sets from Agresti, Categorical Data Analysis, second edition

Dependence

- Matched binary can arise from
  - Measuring a response at two occasions
  - Matching on case status in a retrospective study
  - Matching on exposure status in a prospective or cross-sectional study
- The pairs on binary observations are dependent, so our existing methods do not apply
- We will discuss the process of making conclusions about the marginal probabilities and odds

Notation

	time 2		
time 1	Yes	No	Total
Yes	n_{11}	n_{12}	n_{1+}
no	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

	time 2		
time 1	Yes	No	Total
Yes	π_{11}	π_{12}	π_{1+}
no	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

- We assume that the $(n_{11}, n_{12}, n_{21}, n_{22})$ are multinomial with n trials and probabilities $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$
- π_{1+} and π_{+1} are the marginal probabilities of a yes response at the two occasions
- $\pi_{1+} = P(\text{Yes} \mid \text{Time } 1)$
- $\pi_{+1} = P(\text{Yes} \mid \text{Time } 2)$

Marginal homogeneity

- Marginal homogeneity is the hypothesis $H_0 : \pi_{1+} = \pi_{+1}$
- Marginal homogeneity is equivalent to symmetry $H_0 : \pi_{12} = \pi_{21}$
- The obvious estimate of $\pi_{12} - \pi_{21}$ is $n_{12}/n - n_{21}/n$
- Under H_0 a consistent estimate of the variance is $(n_{12} + n_{21})/n^2$
- Therefore

$$\frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

follows an asymptotic χ^2 distribution with 1 degree of freedom

McNemar's test

- The test from the previous page is called McNemar's test
- Notice that only the discordant cells enter into the test
    - n_{12} and n_{21} carry the relevant information about whether or not π_{1+} and π_{+1} differ
    - n_{11} and n_{22} contribute information to estimating the magnitude of this difference

Example

- Test statistic $\frac{(80-150)^2}{86+150} = 17.36$
- P-value $= 3 \times 10^{-5}$
- Hence we reject the null hypothesis and conclude that there is evidence to suggest a change in opinion between the two polls
- In R

```
mcnemar.test(matrix(c(794, 86, 150, 570), 2),
                   correct = FALSE)
```

The `correct` option applies a continuity correction

Estimation

- Let $\hat\pi_{ij} = n_{ij}/n$ be the sample proportions
- $d = \hat\pi_{1+} - \hat\pi_{+1} = (n_{12} - n_{21})/n$ estimates the difference in the marginal proportions
- The variance of d is

$$\sigma_d^2 = \{\pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})\}/n$$

- $\frac{d - (\pi_{1+} - \pi_{+1})}{\hat\sigma_d}$ follows an asymptotic normal distribution
- Compare σ_d^2 with what we would use if the proportions were independent

Example

- $d = 944/1600 - 880/1600 = .59 - .55 = .04$
- $\hat\pi_{11} = .50$, $\hat\pi_{12} = .09$, $\hat\pi_{21} = .05$, $\hat\pi_{22} = .36$
- $\hat\sigma_d^2 = \{.59(1 - .59) + .55(1 - .55) - 2(.50 \times .36 - .09 \times .05)\}/1600$
- $\hat\sigma_d = .0095$
- 95% CI - $.04 \pm 1.96 \times .0095 = [.06, .02]$
- Note ignoring the dependence yields $\hat\sigma_d = .0175$

Relationship with CMH test

- Each subject's (or matched pair's) responses can be represented as one of four tables.

	Response	
Time	Yes	No
First	1	0
Second	1	0

	Response	
Time	Yes	No
First	1	0
Second	0	1

	Response	
Time	Yes	No
First	0	1
Second	1	0

	Response	
Time	Yes	No
First	0	1
Second	0	1

Result

- McNemar's test is equivalent to the CMH test where subject is the stratifying variable and each 2×2 table is the observed zero-one table for that subject
- This representation is only useful for conceptual purposes

Exact version

- Consider the cells n_{12} and n_{21}
- Under H_0, $\pi_{12}/(\pi_{12} + \pi_{21}) = .5$
- Therefore, under H_0, $n_{21} \mid n_{21} + n_{12}$ is binomial with success probability .5 and $n_{21} + n_{12}$ trials
- We can use this result to come up with an exact P-value for matched pairs data

- Consider the approval rating data
- $H_0 : \pi_{21} = \pi_{12}$ versus $H_a : \pi_{21} < \pi_{12}$ $(\pi_{+1} < \pi_{1+})$
- $P(X \leq 86 \mid 86 + 150) = .000$ where X is binomial with 236 trials and success probability $p = .5$
- For two sided tests, double the smaller of the two one-sided tests

Estimating the marginal odds ratio

- The marginal odds ratio is

$$\frac{\pi_{1+}/\pi_{2+}}{\pi_{+1}/\pi_{+2}} = \frac{\pi_{1+}\pi_{+2}}{\pi_{+1}\pi_{2+}}$$

- The maximum likelihood estimate of the margina *log* odds ratio is

$$\hat{\theta} = \log\{\hat{\pi}_{1+}\hat{\pi}_{+2}/\hat{\pi}_{+1}\hat{\pi}_{2+}\}$$

- The asymptotic variance of this estimator is

$$\begin{aligned} & \{(\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} \\ - & \; 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})\}/n \end{aligned}$$

Example

- In the approval rating example the marginal OR compares the odds of approval at time 1 to that at time 2
- $\hat\theta = \log(944 \times 720/880 \times 656) = .16$
- Estimated standard error $= .039$
- CI for the log odds ratio $= .16 \pm 1.96 \times .039 = [.084, .236]$

Conditional versus marginal odds

First	Second Survey		
survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Conditional versus marginal odds

- n_{ij} cell counts
- n total sample size
- π_{ij} the multinomial probabilities
- The ML estimate of the marginal *log* odds ratio is

$$\hat\theta = \log\{\hat\pi_{1+}\hat\pi_{+2}/\hat\pi_{+1}\hat\pi_{2+}\}$$

- The asymptotic variance of this estimator is

$$\begin{aligned} &\{(\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} \\ - &\ 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})\}/n \end{aligned}$$

Conditional ML

- Consider the following model

$$
\begin{aligned}
\mathrm{logit}\{P(\text{Person } i \text{ says Yes at Time 1})\} &= \alpha + U_i \\
\mathrm{logit}\{P(\text{Person } i \text{ says Yes at Time 2})\} &= \alpha + \gamma + U_i
\end{aligned}
$$

- Each U_i contains person-specific effects. A person with a large U_i is likely to answer Yes at both occasions.
- γ is the **log odds ratio** comparing a response of Yes at Time 1 to a response of Yes at Time 2.
- γ is **subject specific effect**. If you subtract the log odds of a yes response for two different people, the U_i terms would not cancel

Conditional ML cont'd

- One way to eliminate the U_i and get a good estimate of γ is to condition on the total number of Yes responses for each person
  - If they answered Yes or No on both occasions then you know both responses
  - Therefore, only discordant pairs have any relevant information after conditioning
- The conditional ML estimate for γ and its SE turn out to be

$$\log\{n_{21}/n_{12}\} \qquad \sqrt{1/n_{21} + 1/n_{12}}$$

Distinctions in interpretations

- The marginal ML has a marginal interpretation. The effect is averaged over all of the values of U_i.
- The conditional ML estimate has a subject specific interpretation.
- Marginal interpretations are more useful for policy type statements. Policy makers tend to be interested in how factors influence populations.
- Subject specific interpretations are more useful in clinical applications. Physicians are interested in how factors influence individuals.