Mathematical Biostatistics Boot Camp 2: Lecture 9, Simpson's Paradox and Confounding

Brian Caffo

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

September 30, 2013

Table of contents

Simpson's (perceived) paradox

		Death penalty		
Victim	Defendant	yes	no	% yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
	White	53	430	11.0
	Black	15	176	7.9
White		64	451	12.4
Black		4	155	2.5

[1]

[1] From Agresti, Categorical Data Analysis, second edition

Discussion

- Marginally, white defendants received the death penalty a greater percentage of time than black defendants
- Across white and black victims, black defendant's received the death penalty a greater percentage of time than white defendants
- Simpson's paradox refers to the fact that marginal and conditional associations can be opposing
- The death penalty was enacted more often for the murder of a white victim than a black victim. Whites tend to kill whites, hence the larger marginal association.

Example

- Wikipedia's entry on Simpson's paradox gives an example comparing two player's batting averages

	First Half		Second Half		Whole Season	
Player 1	$4/10$	$(.40)$	$25/100$	$(.25)$	$29/110$	$(.26)$
Plater 2	$35/100$	$(.35)$	$2/10$	$(.20)$	$37/110$	$(.34)$

- Player 1 has a better batting average than Player 2 in both the first and second half of the season, yet has a worse batting average overall
- Consider the number of at-bats

Berkeley admissions data

- The Berkeley admissions data is a well known data set regarding Simpsons paradox

```
?UCBAdmissions
data(UCBAdmissions)
     apply(UCBAdmissions, c(1, 2), sum)
          Gender
Admit       Male Female
  Admitted 1198    557
  Rejected 1493   1278
            .445   .304 <- Acceptance rate
```

Acceptance rate by department

```
> apply(UCBAdmissions, 3,
        function(x) c(x[1] / sum(x[1 : 2]),
                       x[3] / sum(x[3 : 4])
                      )
        )
Dept   M    F
   A 0.62 0.82
   B 0.63 0.68
   C 0.37 0.34
   D 0.33 0.35
   E 0.28 0.24
   F 0.06 0.07
```

Why? The application rates by department

```
> apply(UCBAdmissions, c(2, 3), sum)
        Dept
Gender     A   B   C   D   E   F
  Male   825 560 325 417 191 373
  Female 108  25 593 375 393 341
```

Discussion

- Mathematically, Simpson's pardox is not paradoxical

$$\begin{aligned} a/b &< c/d \\ e/f &< g/h \\ (a+e)/(b+f) &> (c+g)/(d+h) \end{aligned}$$

- More statistically, it says that the apparent relationship between two variables can change in the light or absence of a third

Confounding

- Variables that are correlated with both the explanatory and response variables can distort the estimated effect
  - Victim's race was correlated with defendant's race and death penalty
- One strategy to adjust for confounding variables is to **stratify** by the confounder and then combine the strata-specific estimates
  - Requires appropriately weighting the strata-specific estimates
- Unnecessary stratification reduces precision

Aside: weighting

- Suppose that you have two unbiased scales, one with variance 1 lb and and one with variance 9 lbs
- Confronted with weights from both scales, would you give both measurements equal creedance?
- Suppose that $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu, \sigma_2^2)$ where σ_1 and σ_2 are both known
- log-likelihood for μ

$$-(x_1 - \mu)^2/2\sigma_1^2 - (x_2 - \mu)^2/2\sigma_2^2$$

Continued

- Derivative wrt μ set equal to 0

$$(x_1 - \mu)/\sigma_1^2 + (x_2 - \mu)/\sigma_2^2 = 0$$

- Answer

$$\frac{x_1 r_1 + x_2 r_2}{r_1 + r_2} = x_1 p + x_2 (1 - p)$$

  where $r_i = 1/\sigma_i^2$ and $p = r_1/(r_1 + r_2)$
- Note, if X_1 has very low variance, its term dominates the estimate of μ
- General principle: instead of averaging over several unbiased estimates, take an average weighted according to inverse variances
- For our example $\sigma_1^2 = 1$, $\sigma_2^2 = 9$ so $p = .9$

Mantel/Haenszel estimator

- Let n_{ijk} be entry i, j of table k
- The k^{th} sample odds ratio is $\hat{\theta}_k = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$
- The Mantel Haenszel estimator is of the form $\hat{\theta} = \frac{\sum_k r_k \hat{\theta}_k}{\sum_k r_k}$
- The weights are $r_k = \frac{n_{12k}n_{21k}}{n_{++k}}$
- The estimator simplifies to $\hat{\theta}_{MH} = \frac{\sum_k n_{11k}n_{22k}/n_{++k}}{\sum_k n_{12k}n_{21k}/n_{++k}}$
- SE of the log is given in Agresti (page 235) or Rosner (page 656)

	Center 1		Center 2		Center 3		Center 4		Center 5		Center 6		Center 7		Center 8	
	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F
T	11	25	16	4	14	5	2	14	6	11	1	10	1	4	4	2
C	10	27	22	10	7	12	1	16	0	12	0	10	1	8	6	1
n	73		52		38		33		29		21		14		13	

S - Success, F - failure

T - Active Drug, C - placebo[2]

$$\hat\theta_{MH} = \frac{(11\times 27)/73 + (16\times 10)/25 + \ldots + (4\times 1)/13}{(10\times 25)/73 + (4\times 22)/25 + \ldots + (6\times 2)/13} = 2.13$$

Also $\log\hat\theta_{MH} = .758$ and $\hat{SE}_{\log\hat\theta_{MH}} = .303$

[2] Data from Agresti, Categorical Data Analysis, second edition

CMH test

- $H_0 : \theta_1 = \ldots = \theta_k = 1$ versus $H_a : \theta_1 = \ldots = \theta_k \neq 1$
- The CHM test applies to other alternatives, but is most powerful for the H_a given above
- Same as testing conditional independence of the response and exposure given the stratifying variable
- CMH conditioned on the rows and columns for each of the k contingency tables resulting in k hypergeometric distributions and leaving only the n_{11k} cells free

CMH test cont'd

- Under the conditioning and under the null hypothesis
  - $E(n_{11k}) = n_{1+k} n_{+1k} / n_{++k}$
  - $\mathrm{Var}(n_{11k}) = n_{1+k} n_{2+k} n_{+1k} n_{+2k} / n_{++k}^2 (n_{++k} - 1)$
- The CMH test statistic is

$$\frac{[\sum_k \{n_{11k} - E(n_{11k})\}]^2}{\sum_k \mathrm{Var}(n_{11k})}$$

- For large sample sizes and under H_0, this test statistic is $\chi^2(1)$ (regardless of how many tables you are summing up)

In R

```
dat <- array(c(11, 10, 25, 27,  16, 22,  4, 10,
               14,  7,  5, 12,   2,  1, 14, 16,
                6,  0, 11, 12,   1,  0, 10, 10,
                1,  1,  4,  8,   4,  6,  2,  1),
             c(2, 2, 8))
mantelhaen.test(dat, correct = FALSE)
```

Results: $CMH_{TS} = 6.38$
P-value: .012
Test presents evidence to suggest that the treatment and response are not conditionally independent given center

Some final notes on CMH

- It's possible to perform an analogous test in a random effects logit model that benefits from a complete model specification
- It's also possible to test heterogeneity of the strata-specific odds ratios
- Exact tests (guarantee the type I error rate) are also possible `exact = TRUE` in R