Lecture 24

Brian Caffo

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

December 19, 2007

Table of contents

Outline

1. Odds ratios for retrospective studies
2. Odds ratios approximating the prospective RR
3. Exact inference for the odds ratio

Case-control methods

	Lung cancer		
Smoker	Cases	Controls	Total
Yes	688	650	1338
No	21	59	80
	709	709	1418

- Case status obtained from records
- Cannot estimate $P(\text{Case} \mid \text{Smoker})$
- Can estimate $P(\text{Smoker} \mid \text{Case})$

Continued

- Can estimate odds ratio b/c

$$\frac{Odds(\text{case} \mid \text{smoker})}{Odds(\text{case} \mid \text{smoker}^c)} = \frac{Odds(\text{smoker} \mid \text{case})}{Odds(\text{smoker} \mid \text{case}^c)}$$

Proof

C - case, S - smoker

$$
\begin{aligned}
& \frac{Odds(\text{case} \mid \text{smoker})}{Odds(\text{case} \mid \text{smoker}^c)} \\
= & \frac{P(C \mid S)/P(\bar{C} \mid S)}{P(C \mid \bar{S})/P(\bar{C} \mid \bar{S})} \\
= & \frac{P(C, S)/P(\bar{C}, S)}{P(C, \bar{S})/P(\bar{C}, \bar{S})} \\
= & \frac{P(C, S)P(\bar{C}, \bar{S})}{P(C, \bar{S})P(\bar{C}, S)}
\end{aligned}
$$

Exchange C and S and the result is obtained

Notes

- Sample OR is $\frac{n_{11}n_{22}}{n_{12}n_{21}}$
- Sample OR is unchanged if a row or column is multiplied by a constant
- Invariant to transposing
- Is related to RR

Notes continued

$$
\begin{aligned}
OR &= \frac{P(S \mid C)/P(\bar S \mid C)}{P(S \mid \bar C)/P(\bar S \mid \bar C)} \\
&= \frac{P(C \mid S)/P(\bar C \mid S)}{P(C \mid \bar S)/P(\bar C \mid \bar S)} \\
&= \frac{P(C \mid S)}{P(C \mid \bar S)} \frac{P(\bar C \mid \bar S)}{P(\bar C \mid S)} \\
&= RR \times \frac{1 - P(C \mid \bar S)}{1 - P(C \mid S)}
\end{aligned}
$$

- OR approximate RR if $P(C \mid \bar S)$ and $P(C \mid S)$ are small (or if they are nearly equal)

Rare disease assumption

	Disease		
Exposure	Yes	No	Total
Yes	9	1	10
No	1	999	1000
	10	1000	1010

- Cross-sectional data
- $P(\hat{D}) = 10/1010 \approx .01$
- $\hat{OR} = (9 \times 999)/(1 \times 1) = 8991$
- $\hat{RR} = (9/10)/(1/1000) = 900$
- D is rare in the sample
- D is not rare among the exposed

Notes

- $OR = 1$ implies no association
- $OR > 1$ positive association
- $OR < 1$ negative association
- For retrospective CC studies, OR can be interpreted prospectively
- For diseases that are rare among the cases and controls, the OR approximates the RR
- Delta method SE for log OR is

$$\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Example

	Lung cancer		
Smoker	Cases	Controls	Total
Yes	688	650	1338
No	21	59	80
	709	709	1418

[1]

- $\hat{OR} = \frac{688 \times 59}{21 \times 650} = 3.0$
- $\hat{SE}_{\log \hat{OR}} = \sqrt{\frac{1}{688} + \frac{1}{650} + \frac{1}{21} + \frac{1}{59}} = .26$
- $CI = \log(3.0) \pm 1.96 \times .26 = [.59, 1.61]$
- The estimated odds of lung cancer for smokers are 3 times that of the odds for non-smokers with an interval of $[\exp(.59), \exp(1.61)] = [1.80, 5.00]$

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

Exact inference for the OR

	Lung cancer		
Smoker	Cases	Controls	Total
Yes	688	650	1338
No	21	59	80
	709	709	1418

- X the number of smokers for the cases
- Y the number of smokers for the controls
- Calculate an exact CI for the odds ratio
- Have to eliminate a nuisance parameter

Notation

- $\mathrm{logit}(p) = \log\{p/(1-p)\}$ is the **log-odds**
- Differences in $\mathrm{logit}s$ are log-odds *ratios*
- $\mathrm{logit}\{P(\text{Smoker} \mid \text{Case})\} = \delta$
    - $P(\text{Smoker} \mid \text{Case}) = e^{\delta}/(1+e^{\delta})$
- $\mathrm{logit}\{P(\text{Smoker} \mid \text{Control})\} = \delta + \theta$
    - $P(\text{Smoker} \mid \text{Control}) = e^{\delta+\theta}/(1+e^{\delta+\theta})$
- θ is the log-odds ratio
- δ is the nuisance parameter

Notation

- X is binomial with n_1 trials and success probability $e^\delta/(1+e^\delta)$
- Y is binomial with n_2 trials and success probability $e^{\delta+\theta}/(1+e^{\delta+\theta})$

$$
\begin{aligned}
P(X = x) &= \left(\begin{array}{c} n_1 \\ x \end{array}\right) \left\{\frac{e^\delta}{1+e^\delta}\right\}^x \left\{\frac{1}{1+e^\delta}\right\}^{n_1 - x} \\
&= \left(\begin{array}{c} n_1 \\ x \end{array}\right) e^{x\delta} \left\{\frac{1}{1+e^\delta}\right\}^{n_1}
\end{aligned}
$$

$$P(X = x) = \left(\begin{array}{c} n_1 \\ x \end{array} \right) e^{x\delta} \left\{ \frac{1}{1 + e^{\delta}} \right\}^{n_1}$$

$$P(Y = z - x) = \left(\begin{array}{c} n_2 \\ z - x \end{array} \right) e^{(z-x)\delta + (z-x)\theta} \left\{ \frac{1}{1 + e^{\delta + \theta}} \right\}^{n_2}$$

$$P(X + Y = z) = \sum_u P(X = u) P(Y = z - u)$$

$$P(X = x \mid X + Y = z) = \frac{P(X = x)P(Y = z - x)}{\sum_u P(X = u)P(Y = z - u)}$$

Non-central hypergeometric distribution

$$P(X = x \mid X + Y = z; \theta) = \frac{\binom{n_1}{x}\binom{n_2}{z-x} e^{x\theta}}{\sum_u \binom{n_1}{u}\binom{n_2}{z-u} e^{u\theta}}$$

- θ is the log odds ratio
- This distribution is used to calculate exact hypothesis tests for $H_0 : \theta = \theta_0$
- Inverting exact tests yields exact confidence intervals for the odds ratio
- Simplifies to the hypergeometric distribution for $\theta = 0$