Lecture 24

Brian Caffo

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

December 19, 2007
Table of contents

1. Table of contents
2. Outline
3. Case-control methods
4. Rare disease assumption
5. Exact inference for the odds ratio
Outline

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

<table>
<thead>
<tr>
<th>Smoker</th>
<th>Lung cancer</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>688</td>
<td>650</td>
<td>1338</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>59</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>709</td>
<td>709</td>
<td>1418</td>
</tr>
</tbody>
</table>

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

\[
\frac{\text{Odds}(\text{case} \mid \text{smoker})}{\text{Odds}(\text{case} \mid \text{smoker}^C)} = \frac{P(C \mid S)/P(\bar{C} \mid S)}{P(C \mid \bar{S})/P(\bar{C} \mid \bar{S})}
\]

\[
= \frac{P(C, S)/P(\bar{C}, S)}{P(C, \bar{S})/P(\bar{C}, \bar{S})}
\]

\[
= \frac{P(C, S)P(\bar{C}, \bar{S})}{P(C, \bar{S})P(\bar{C}, S)}
\]

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 $RR$
\[ OR = \frac{P(S \mid C)/P(\bar{S} \mid C)}{P(S \mid \bar{C})/P(\bar{S} \mid \bar{C})} \]

\[ = \frac{P(C \mid S)/P(\bar{C} \mid S)}{P(C \mid \bar{S})/P(\bar{C} \mid \bar{S})} \]

\[ = \frac{P(C \mid S)}{P(C \mid \bar{S})} \times \frac{P(\bar{C} \mid \bar{S})}{P(\bar{C} \mid S)} \]

\[ = RR \times \frac{1 - P(C \mid \bar{S})}{1 - P(C \mid S)} \]

- \( OR \) approximate \( RR \) if \( P(C \mid \bar{S}) \) and \( P(C \mid S) \) are small (or if they are nearly equal)
Cross-sectional data

- \( P(\hat{D}) = \frac{10}{1010} \approx 0.01 \)
- \( \hat{OR} = \frac{9 \times 999}{1 \times 1} = 8991 \)
- \( \hat{RR} = \frac{9/10}{1/1000} = 900 \)
- \( D \) 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

<table>
<thead>
<tr>
<th></th>
<th>Lung cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>688</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>688</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1338</td>
</tr>
<tr>
<td>No</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>1418</td>
</tr>
</tbody>
</table>

1

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

\footnote{Data from Agresti, Categorical Data Analysis, second edition}
Exact inference for the OR

<table>
<thead>
<tr>
<th>Smoker</th>
<th>Cases</th>
<th>Controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>688</td>
<td>650</td>
<td>1338</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>59</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>709</td>
<td>709</td>
<td>1418</td>
</tr>
</tbody>
</table>

- $X$ 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
\begin{itemize}
  \item \text{logit}(p) = \log\left\{ \frac{p}{1 - p} \right\} \text{ is the log-odds}
  \item Differences in logits are log-odds ratios
  \item \text{logit}\{P(\text{Smoker} \mid \text{Case})\} = \delta
    \begin{itemize}
      \item \( P(\text{Smoker} \mid \text{Case}) = \frac{e^\delta}{1 + e^\delta} \)
    \end{itemize}
  \item \text{logit}\{P(\text{Smoker} \mid \text{Control})\} = \delta + \theta
    \begin{itemize}
      \item \( P(\text{Smoker} \mid \text{Control}) = \frac{e^{\delta+\theta}}{1 + e^{\delta+\theta}} \)
    \end{itemize}
  \item \( \theta \) is the log-odds ratio
  \item \( \delta \) is the nuisance parameter
\end{itemize}
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 $e^{\delta+\theta}/(1 + e^{\delta+\theta})$

$$P(X = x) = \binom{n_1}{x} \left\{\frac{e^\delta}{1 + e^\delta}\right\}^x \left\{\frac{1}{1 + e^\delta}\right\}^{n_1-x}$$

$$= \binom{n_1}{x} e^{x\delta} \left\{\frac{1}{1 + e^\delta}\right\}^{n_1}$$
\[ P(X = x) = \binom{n_1}{x} e^{x\delta} \left\{ \frac{1}{1 + e^\delta} \right\}^{n_1} \]

\[ P(Y = z - x) = \binom{n_2}{z - x} e^{(z-x)\delta + (z-x)\theta} \left\{ \frac{1}{1 + e^{\delta+\theta}} \right\}^{n_2} \]

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

\[ P(X = x \mid X + Y = z) = \frac{P(X = x)P(Y = z - x)}{\sum_u P(X = u)P(Y = z - u)} \]
Non-central hypergeometric distribution

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

- \(\theta\) 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 \(\theta = 0\)