Modelling Association

Sudipto Banerjee
sudiptob@biostat.umn.edu

University of Minnesota
Assoc. in Observational Studies

- Natural sciences can study causality through designed experiments.
- More difficult in observational studies.
- Prospective study: we observe or choose a sample according to certain explanatory variables and follow up subjects for outcome. Enhanced validity through matching at the start of study for confounders.
- Retrospective study: We choose a sample according to the outcome of interest and then investigate what values of potential explanatory variables had previously existed in the patient’s history.
- Evidence is often summarized in the form of a $2 \times 2$ table, cross-classifying two binary variables.
Prospective study analysis

In the follow-up study we compare the relative risk of disease between an exposed group \((A)\) and a non-exposed group \((B)\).

These form two populations of size \(n_A\) and \(n_B\), with possibly differing rates of incidence of the outcome, \(p_A\) and \(p_B\). This is described by two binomial populations

\[ r_A \sim Bin(n_A), \quad r_B \sim Bin(n_B, p_B) \]

Hypothesis: \(p_A = p_B = p\).

Rothman’s Example (Program 3.1): There are \(n_A = 20\) exposed and \(n_B = 100\) non-exposed. Also, \(r_A = 10\) and \(r_B = 45\). The maximum likelihood estimates are \(\hat{p}_A = 0.5\) and \(\hat{p}_B = 0.45\); so the corresponding estimates of the risk difference, risk ratio and odds ratio are:

\[
\begin{align*}
\hat{p}_A - \hat{p}_B &= 0.05 \\
\frac{\hat{p}_A}{\hat{p}_B} &= 1.11 \\
\frac{\hat{p}_A(1 - \hat{p}_B)}{\hat{p}_B(1 - \hat{p}_A)} &= 1.22.
\end{align*}
\]
A second population

For a second data set with the same number of exposed and non-exposed subjects, the risk difference and risk ratio are lower than for the first data set (namely 0.04 and 1.05) but the odds ratio is higher, namely 1.26.

If we set noninformative beta priors on the probabilities $p_A$ and $p_B$, the sampling model shows the wide uncertainty attached to these binomial probabilities, especially in view of the small number of exposed cases. See Program 3.1 Follow Up Study Example. Note: The two data sets are combined into a single likelihood.

Below we present the results for the differences between the two populations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>difference in $\log(OR)$</td>
<td>-0.085 (-1.732, 1.425)</td>
</tr>
<tr>
<td>difference in $\log(RR)$</td>
<td>0.055 (-0.548, 0.567)</td>
</tr>
<tr>
<td>difference in $\log(RD)$</td>
<td>0.058 (-0.548, 0.567)</td>
</tr>
</tbody>
</table>
Simulating controls

- Bayesian methods facilitate incorporation of existing knowledge to situations where only data on cases may be available.
- Goal: accumulate a set of cases and investigate whether their exposure to a suspected causal factor is unusual. The control group is used to derive the posterior distribution of exposure.
- Zelen and Parker (1986): extensive information on exposure levels in the population may be available.
- Let $Y = 1$ for cases and $Y = 0$ for controls. Let $X = 1$ if the individual is exposed to the causal agent and 0 otherwise.
- Then:
  \[
  f(X \mid Y) = \frac{\exp(\alpha X + \beta Y X)}{1 + \exp(\alpha + \beta Y)}
  \]
- $X$ and $Y$ are independent only if $\beta = 0$. 
Typical case control data:

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$s$</td>
<td>$c - s$</td>
<td>$c$</td>
</tr>
<tr>
<td>Controls</td>
<td>$r$</td>
<td>$m - r$</td>
<td>$m$</td>
</tr>
<tr>
<td>Total</td>
<td>$e$</td>
<td>$n$</td>
<td>$t$</td>
</tr>
</tbody>
</table>

Likelihood:

$$L(s, r, c, m \mid \alpha, \beta) = \frac{\exp((\alpha + \beta)s + \alpha r)}{(1 + \exp(\alpha))^m(1 + \exp(\alpha + \beta))^c}.$$ 

A conjugate prior has the same form as the likelihood but with “prior data” $s'$, $r'$, $c'$ and $m'$. The posterior is:

$$P(\alpha, \beta \mid s, r, c, m) \propto \frac{\exp((\alpha + \beta)S + \alpha R)}{(1 + \exp(\alpha))^M(1 + \exp(\alpha + \beta))^C},$$

where $S = s + s'$, $R = r + r'$, $C = c + c'$ and $M = m + m'$. Also, $T = C + M$. 

contd.
Zelen and Parker: Derive prior data for controls, namely $r'$ exposed among $m'$ individuals without the disease or condition.

Simulated control data is the only control data. So $m = r = 0$ and $M = m'$, $R = r'$. These are set solely based upon the population exposure.

Example: if 30% of the nation's female population are smokers, then $r'/m' = 0.3$. Suppose the probability that this fraction exceeds 35% is put at 0.05. Then, using the normal approximation

$$\log \frac{r'/m'}{1 - r'/m'} + 1.64\sigma = \log \frac{0.35}{1 - 0.35},$$

where $\sigma$ is given by:

$$\sigma = \sqrt{\frac{1}{r'} + \frac{1}{(m' - r')}}.$$ 

Example: A BUGS program, modified from Congdon to specify the initial values, is given in the course web-site. It analyzes an example with just eight young women, of whom seven had contracted adenocarcinoma of the vagina. Here $A = 10\%$ and $H = 20\%$. The credible interval for the log-odds ratio (beta in the program) is $(2.57, 7.86)$ with the median being 4.43, based upon 10000 iterations with a burn-in of 5000.
Example: Follow-up data for males from the Scottish Heart Health Study (SHHS). This considers smoking status as a confounder in the relationship between the outcome (CHD) (Coronary Heart Disease) and the risk factor (household owner or renter). Renting, as against ownership, is considered the exposure.

Of 1770 non-smoking owner-occupiers, 48 had CHD as compared with 33 out of 956 non-smoking renters. Among 707 smoker owner-occupiers, 29 had CHD while among the 950 smokers who also rented their housing 52 had CHD.

Consider Program 3.6 in the Congdon suite of programs. There are two models that compute Relative Risk (RR) and Odds-Ratio (OR) using the Mantel-Haenszel assumption of the log RR and log OR having underlying normal distributions with a common mean for each strata level.

**Homework:** Explain exactly what Program 3.6 is doing. Run Program 3.6 to obtain posterior distributions of RR, OR and the stratum level means.
Attributable risk measures health consequences of an association between a risk factor and a disease. It is defined as:

\[
\text{ARE} = \frac{\text{Risk for exposed} - \text{Risk for Unexposed}}{\text{Risk for exposed}}
\]

ARE stands for attributable risk in the exposed subpopulation. If \( I_0 \) and \( I_1 \) are the incidence rates among those not exposed and exposed respectively, then \( RR = \frac{I_1}{I_0} \) and \( ARE = \frac{(I_1 - I_0)}{I_1} \) or, equivalently, \( (RR - 1)/RR \).

ARP stands for attributable risk in the total population and is given by:

\[
\text{ARP} = \frac{(RR - 1)p}{1 + (RR - 1)p}
\]

where \( p \) is the population exposure rate.

**Homework:** Examine Program 3.7 in the Congdon suite. A prospective study compared death rates in 701768 person-years observed for current cigarette smokers among US veterans and 1015999 person-years for non-smokers. Observed deaths in the two groups - over the follow-up period - are 1116 and 426 respectively. Write down clearly how this data is being modelled, what the different variables are and the related distributions. Run this program to obtain posterior analysis of ARE and ARP.