Statistical Methods in Cancer Epidemiology using R

Janne Pitkäniemi

Faculty of Social Sciences, University of Tampere
Finnish Cancer Registry

Lecture 2

janne.pitkaniemi@cancer.fi>

Feb,10 2010

Contents

- Binary outcomes and proportions
- Comparative parameters of risks and their estimation
- Binomial regression models and comparative parameters
- Adjustment for confounding and evaluation of modification by binomial regression

Main R functions covered:

- `twoby2()` (Epi package)
- `glm()`
- `ci.lin()` (Epi package)

Outcomes in epidemiologic research

Epidemiologic studies address the occurrence of diseases and other health related phenomena:

- (a) cross-sectional: **prevalence** of diseases,
- (b) longitudinal: disease **incidence**, and mortality

Often we want to compare the prevalence or incidence of disease between two groups defined by a binary *risk factor* X

- $X = 1$: `exposed` $X = 0$: `unexposed`

Types of outcome variables

- *Binary* (0/1) variables at individual level
  - disease *status* at a *time point*
  - *change* of status, *event* or *transition* ({*e.g.*} *from healthy to diseased)*
- *Proportions* at group level
  - prevalence
  - incidence proportion or cumulative incidence,
- *Rates* of events
  - incidence or mortality rate (per 1000 y)
  - car accidents (per million km)
- *Time* to event
  - survival time (often censored)

Incidence and prevalence proportions}

- **Incidence proportion** (R) of a binary (0/1) outcome (disease, death etc.) over a fixed risk period is defined

$$R = \frac{D}{N} = \frac{\text{number of new cases during period}}{\text{size of population-at-risk at start}}$$

Also called {**cumulative incidence**} **(or even "risk").\ NB.**

This formula requires complete follow-up, i.e. no *{censorings}, and absence of {competing risks}.*

- ***Prevalence (proportion)*** *P of disease at time point t*

$$P = \frac{\textit{no. of existing cases at } t}{\textit{total population size at } t}.$$

Two-group comparison

- Binary risk factor X: exposed vs. unexposed.
- Summarizy results from cohort study with fixed risk period and no losses:

Exposure	Cases	Non-cases	Group size
yes	D_1	C_1	N_1
no	D_0	C_0	N_0
total	D_+	C_+	N_+

- Incidence proportions in the two exposure groups

$$R_1 = \frac{D_1}{N_1}, \qquad R_0 = \frac{D_0}{N_0}.$$

- These are crude *estimates* of the true *risks* π_1, and π_0 of outcome in the two exposure categories.

Example: Observational clinical study

Treatment failure in two types of operation for renal calculi (Charig *et al.* 1986. *BMJ* 292: 879-882)

- OS = open surgery (invasive)
- PN = percutaneous nephrolithotomy

Treatment group (j)	Failure (D_j)	Success (C_j)	Patients (N_j)	Failure-% (R_j)
OS ($j = 1$)	77	273	350	*22.0*
PN ($j = 0$)	60	290	350	*17.1*

Crude incidence proportions of treatment failure:

$$R_1 = 77/350 = 22.0\%, \qquad R_0 = 60/350 = 17.1\%$$

Risks and their comparative parameters

The **risk** or **probability** of binary outcome (*e.g.* new case of disease) in the `exposed` π_1 and in the `unexposed` π_0 as to binary risk factor X (values 1 and 0) are typically compared by

- risk difference $\theta = \pi_1 - \pi_0$
- risk ratio $\phi = \pi_1/\pi_0$
- odds ratio (risk odds ratio)

$$\psi = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

The odds ratio is close to the risk ratio when the risks are `small` (less than 0.1 – the `rare-disease assumption`).

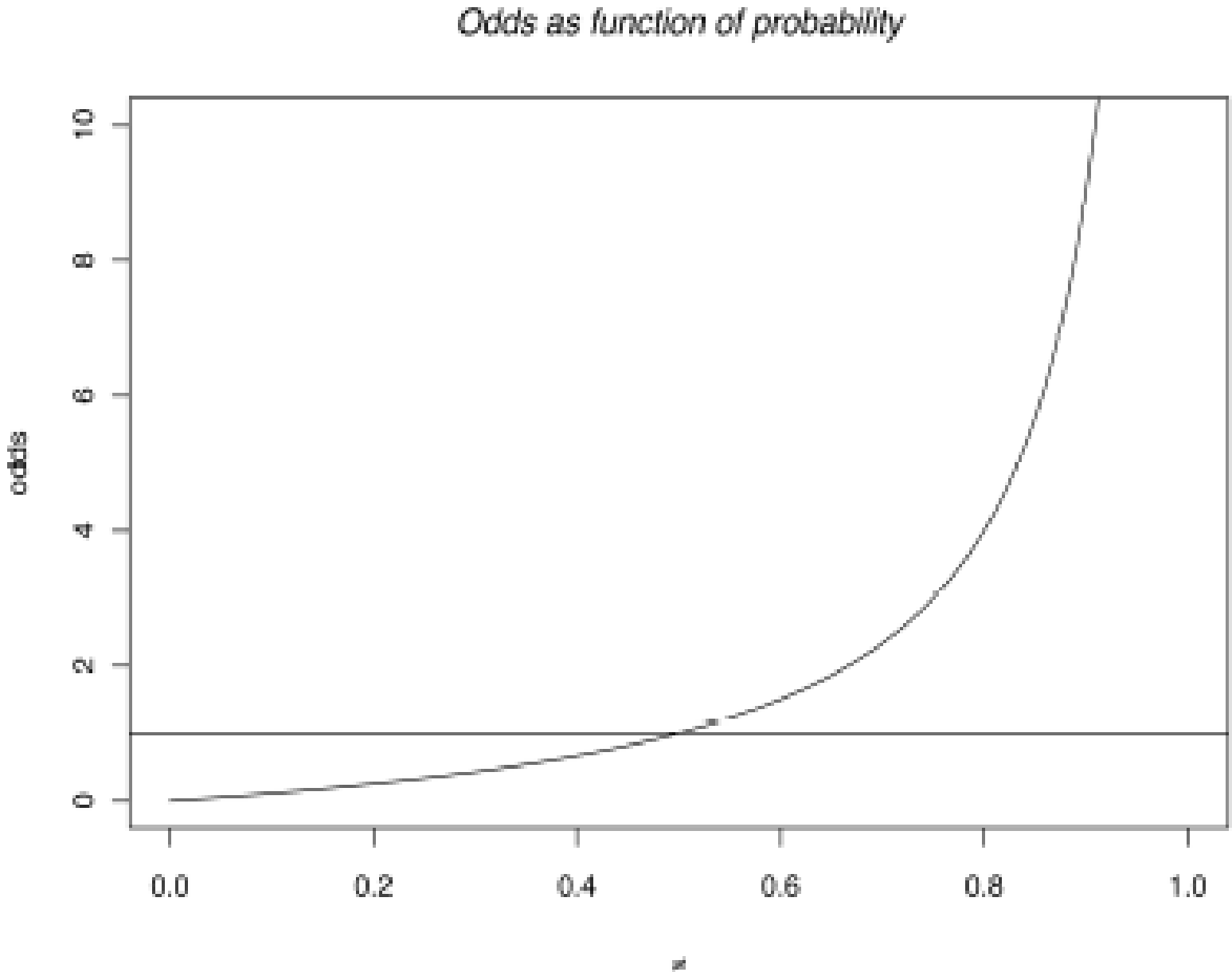
Odds and Odds Ratio (OR)

The **odds (Ω)** is the probability of binary outcome $P(Y = 1) = \pi$ divided by the the probability of binary outcome $P(Y = 0) = 1 - \pi$.

$$\Omega = \frac{\pi}{1 - \pi}$$

- Odds of 2.5 means that the probability of Y=1 (success) is two and half times higher than the probability of Y=0 (failure)
- Odds 0.5 means that success probability of success is 50% of the probability of failure
- Odds of 1 implies that probability of both outcomes 0.5 (equal)

Probability and odds



Risks and comparative parameters estimated

The risks π_1 and π_0 are estimated by empirical incidence proportions $R_1 = D_1/N_1$, and $R_0 = D_0/N_0$.

Crude estimates of comparative parameters

- **incidence proportion difference** $\quad \mathrm{RD} = R_1 - R_0$
- **incidence proportion ratio** $\quad \mathrm{RR} = R_1/R_0$
- **incidence odds ratio**

$$\mathrm{OR} = \frac{R_1/(1-R_1)}{R_0/(1-R_0)}$$

NB. To remove *confounding*, the estimated must be adjusted for relevant *confounders*.

Example: OS vs. PN (cont'd)

Crude estimates of true risk difference θ, risk ratio ϕ, and \ odds ratio ψ between OS and PN:

$$\begin{aligned}
\mathrm{RD} &= \frac{77}{350} - \frac{60}{350} = 0.22 - 0.171 = +\mathbf{0.049}\ (+4.9\%) \\
\mathrm{RR} &= \frac{77/350}{60/350} = \frac{77/60}{350/350} = \frac{0.22}{0.171} = \mathbf{1.283} \\
\mathrm{OR} &= \frac{77/273}{60/290} = \frac{0.22/(1-0.22)}{0.171/(1-0.171)} = \mathbf{1.363}
\end{aligned}$$

PN appears more successful than OS.

Is this (a) true, (b) due to bias, or (c) due to chance?

Example: OS vs. PN (cont'd)

Standard error of log(RR), 95% error factor (EF) of RR, and 95% CI for true risk ratio ϕ:

$$\begin{aligned} \mathrm{SEL} &= \sqrt{\frac{1}{73} + \frac{1}{60} - \frac{1}{350} - \frac{1}{350}} \\ &= \mathbf{0.1547} \end{aligned}$$

$$\begin{aligned} \mathrm{EF} &= \exp\{1.96 \times 0.1547\} \\ &= \mathbf{1.3543} \end{aligned}$$

$$\begin{aligned} \mathrm{CI} &= [1.2833/1.3543, 1.2833 \times 1.3543] \\ &= [\mathbf{0.9476}, \mathbf{1.7380}]. \end{aligned}$$

Estimating comparative parameters in R

- A multitude of R functions in several packages are readily available for point estimation and CI calculation using either "exact" or/and various approximative methods.

- We shall here demonstrate the use of function `twoby2()` in the `Epi package`. It applies the simple Wald approximations as described above, but for

  - risk difference: the Newcombe method is used, and

  - odds ratio: the `exact` conditional method is also available.

- Hence, similar results are expected as obtained above.

Use of function `twoby2()`

- Loading the `Epi` package:

```
library(Epi)
```

- Reading the counts of the 2 x 2-table into a matrix:

```
counts <- c(77, 273, 60, 290)
tab <- matrix( counts, nrow=2, byrow=T)
```

- Viewing the contents of the matrix/table:

```
tab
```

```
     [,1] [,2]
[1,]   77  273
[2,]   60  290
```

Make 2 by 2 table

- Calling the function with `tab` as its argument:

```
twoby2(tab)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Col 1
Comparing : Row 1 vs. Row 2

      Col 1 Col 2    P(Col 1) 95% conf. interval
Row 1    77   273      0.2200    0.1797   0.2664
Row 2    60   290      0.1714    0.1355   0.2146

                                  95% conf. interval
            Relative Risk: 1.2833    0.9476   1.7380
         Sample Odds Ratio: 1.3632    0.9362   1.9851
Conditional MLE Odds Ratio: 1.3626    0.9206   2.0237
    Probability difference: 0.0486   -0.0103   0.1071

            Exact P-value: 0.1272
        Asymptotic P-value: 0.1061
------------------------------------------------------
```

Analyses based on binary regression model

Crude estimates and CIs for the comparative parameters can also be obtained by fitting appropriate **binary regression models** for the numbers D_j or proportions R_j.

Special cases of **generalized linear models** (GLM) with

- (i) **random part**: D_j is assumed to obey the binomial distribution or {**family**} **with index** N_j **and probability** π_j**,**
- (ii) **systematic part**: **linear predictor** $\eta_j = \alpha + \beta X_j$, in which $X_j = 0$ for unexposed and $X_j = 1$ for exposed,
- (iii) **link function**: $g(.)$ that connects the probability π_j and the systematic part η_j by:

$$g(\pi_j) = \eta_j = \alpha + \beta X_j$$

Link functions and comparative parameters

General model: $g(\pi_j) = \alpha + \beta X_j$ for the risks by binary X

- **identity** link: $g(\pi_j) = \pi_j = \alpha + \beta X_j$:\ $\Rightarrow \beta = \pi_1 - \pi_0 = \theta$, \ $=$ risk difference btw $X_j = 1$ and $X_j = 0$

- **logarithmic** link: $g(\pi_j) = log(\pi_j)$\
  $\Leftrightarrow \pi_j = \exp(\alpha + \beta X_j) = e^{\alpha} e^{\beta X_j}$\
  $\Rightarrow \beta = \log(\pi_1) - \log(\pi_0) = \log(\phi)$, \ $\Rightarrow e^{\beta} = \phi =$ risk ratio btw exposed and unexposed,

- **logit** link: $g(\pi_j) = \log[\pi_j/(1-\pi_j)]$

$$\Leftrightarrow \quad \pi_j = \frac{1}{1 + \exp\{-(\alpha + \beta X_j)\}} = \text{expit}(\alpha + \beta X_j)$$

Logit model for odds ratio

Substituting logit function for $g(.)$ and values of X_j we get

$$\begin{aligned} \log\left(\frac{\pi_0}{1-\pi_0}\right) &= \alpha = \text{baseline logit} \\ \log\left(\frac{\pi_1}{1-\pi_1}\right) &= \alpha + \beta. \end{aligned}$$

This implies

$$\begin{aligned} \pi_0 &= \frac{1}{1+\exp(-\alpha)}, \quad \pi_1 = \frac{1}{1+\exp\{-(\alpha+\beta)\}}, \\ \beta &= \log\left\{\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right\} \\ e^{\beta} &= \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \psi = \underline{\text{odds ratio}} \text{ btw exp'd and unexp'd.} \end{aligned}$$

Fitting binary regression models in R

Function `glm()`

- estimation method: **maximum likelihood** (ML),computation algorithm: IWLS.

Key arguments of `glm()`:

- model formula: *response* $\sim$ expression of regressors
- `weights` = group sizes N_j when proportions R_j are given as the response (outcome) variable,
- `family = binomial(link = 'log')`, if risk ratio,
  `family = binomial(link = 'logit')`, if odds ratio,
  `family = binomial(link = 'identity')`, if risk diff'ce
  is the parameter of interest.

Example - Colorectal screenning RHS study

Table 3 Descriptive and comparative statistics for colorectal cancer incidence and mortality, by study arm and sex

	All		Males		Females	
	Screening	Control	Screening	Control	Screening	Control
Number of persons	180 210	180 282	89 712	89 807	90 498	90 475
Person-years	805 480	805 693	395 614	395 851	409 866	409 843
Deaths						
All causes	8000	7963	5486	5453	2514	2510
Colorectal cancer	170	164	93	106	77	58
Patients with colorectal cancer						
Number of patients	903	811	525	477	378	334
Incidence rate per 100 000 person-years	112.4	100.9	133.1	120.8	92.4	81.7
Person-years	2285	1805	1318	1016	967	790
Deaths	202	190	123	127	79	63
Expected number of deaths*	25.0	21.1	18.6	15.7	6.4	5.5
Excess number of deaths†	177.0	168.9	104.4	111.3	72.6	57.5
CRC incidence rate ratio (95% CI)	1.11 (1.01 to 1.23)	1	1.10 (0.97 to 1.25)	1	1.13 (0.98 to 1.31)	1
Mortality rates per 100 000						
All causes	993	988	1387	1378	613	612
Non-colorectal cancer causes	972	968	1363	1351	595	598
Colorectal cancer	21.1	20.4	23.5	26.8	18.8	14.2
Excess mortality due to CRC	21.7	21.0	26.4	28.1	17.7	14.0
Mortality rate ratios						
All causes	1.00 (0.97 to 1.04)	1	1.01 (0.97 to 1.05)	1	1.00 (0.95 to 1.06)	1
Non-colorectal cancer causes	1.00 (0.97 to 1.04)	1	1.01 (0.97 to 1.05)	1	0.99 (0.94 to 1.05)	1
Colorectal cancer	1.04 (0.84 to 1.28)	1	0.88 (0.66 to 1.16)	1	1.33 (0.94 to 1.87)	1
Excess mortality due to CRC	1.05 (0.84 to 1.30)	1	0.94 (0.71 to 1.24)	1	1.26 (0.88 to 1.80)	1

*Calculated according to the mortality of colorectal cancer-free persons stratified by sex, age, calendar year, arm and participation status.
†The difference between the observed and the expected number of deaths.
CRC, colorectal cancer.

Figure 2: Caption for the picture.

Example - data to R

CRC parient mortality between arms

- two observations (one for each treatment group),
- three variables:
  - `treat` = arm, with values 1 = Screen, 0 = control,
  - `d` = number of CRC (patient) deaths D_j,
  - `npat` = number of patients N_j,
- variable vectors defined:

```
treat <- c(0, 1)
fail <- c(190, 202)
npat <- c(811, 903)
prop <- fail/npat
c(prop*100,prop[2]/prop[1])
```

```
[1] 23.4278668 22.3698782  0.9548406
```

Estimation of risk ratio

- Defining the *model object*:

```
RRmodel <- glm( prop ~ treat, family=binomial(link='log'),
```

- Estimation results extracted by function `ci.lin()` in `Epi` (two columns of the whole output omitted for clarity)

```
library(Epi)
round( ci.lin(RRmodel, Exp=T)[, -(3:4)], 4)
```

```
             Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)   -1.4512 0.0635    0.2343 0.2069 0.2653
treat         -0.0462 0.0887    0.9548 0.8024 1.1362
```

- The estimate of β is $\widehat{\beta} = -0.0462 = \log(0.9548)$, and that of mortality (ratio) ratio ϕ is $\mathrm{RR} = \exp(-0.0462) = 0.9548$.
- Estimate of α is $\widehat{\alpha} = -1.4512 = \log(0.2342787) = \log(R_0)$.

Estimation of odds ratio

```
ORmodel <- glm( prop ~ treat,
       fam = binomial(link='logit'), weights=npat)
```

```
round( ci.lin(ORmodel, Exp=T)[, -(3:4)], 4)
```

```
             Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)   -1.1843 0.0829    0.3060 0.2601 0.3599
treat         -0.0599 0.1151    0.9418 0.7516 1.1802
```

- The estimate of β is $\widehat{\beta} = -0.0599 = \log(0.9418)$, and the estimated ψ is OR $= \exp(-0.0599) = 0.9418$.
- The estimate of α is $\widehat{\alpha} - 1.1843 = \log(0.3060)$, in which $0.3060 = 0.2342787/(1 - 0.2342787)$ is the estimated baseline odds $R_0/(1 - R_0)$.

Estimation of risk difference

```
RDmodel <- glm( prop ~ treat,
   fam=binomial(link='identity'), w=npat)
```

```
round( ci.lin(RDmodel), 3)
```

```
            Estimate StdErr      z     P   2.5% 97.5%
(Intercept)    0.234  0.015 15.752 0.000  0.205 0.263
treat         -0.011  0.020 -0.520 0.603 -0.050 0.029
```

- Again, same results obtained as with *e.g.* `twoby2()`, although CIs are here based on Wald statistic.
- **NB**. Fitting binomial model with this link easily fails with more complicated models, especially involving continuous variables.

Same with twoby2

```
counts<-c(fail[2],npat[2]-fail[2],fail[1],npat[1]-fail[1])
counts
```

```
[1] 202 701 190 621
```

```
twoby2(matrix( counts, nrow=2, byrow=T))
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Col 1
Comparing : Row 1 vs. Row 2

      Col 1 Col 2    P(Col 1) 95% conf. interval
Row 1   202   701      0.2237    0.1977   0.2520
Row 2   190   621      0.2343    0.2064   0.2647

                                    95% conf. interval
            Relative Risk:  0.9548    0.8024   1.1362
         Sample Odds Ratio:  0.9418    0.7516   1.1802
Conditional MLE Odds Ratio:  0.9419    0.7468   1.1881
    Probability difference: -0.0106   -0.0505   0.0291

             Exact P-value: 0.6048
         Asymptotic P-value: 0.6026
------------------------------------------------------
```