Practicals - Case control study

Epidemiological methods in medical research 2023

9 February 2023

Exercise 1: Warm-up about case-control design

1.1 Specificities of a case control study

1. What is the main difference between a cohort and a case control study?
2. What impact does the case-control design have on the statistical analysis?
3. What assumption is required for this analysis to provide unbiased estimates? How can this assumption be violated?

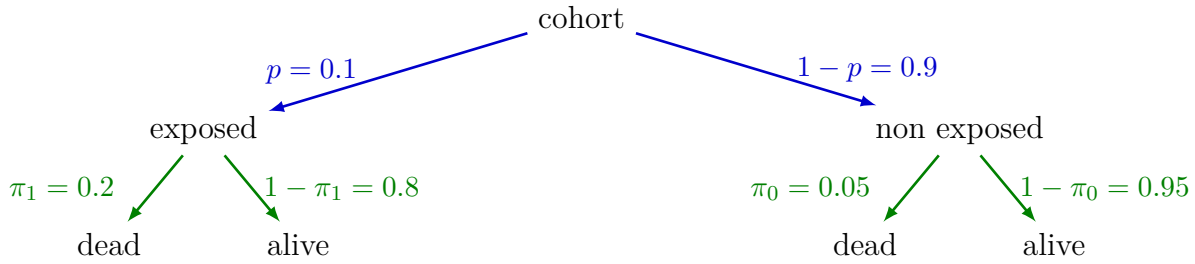
1.2 Case control as a cohort study

A case-control study can be described as a specific cohort study. This underlying cohort is sometimes refered to as 'study base'. To facilitate the description of the correspondence between cohort and case control designs we will consider:

- a binary exposure E leading to two groups: exposed and non-exposed
- a binary outcome Y: death or alive at time τ (e.g. 1 year)
- no covariate (in particular no confounder), no competing risk (no death) or loss to follow-up (emigration).

One can think about the situation where some people died of food poisoning. They all ate meat sold at a local butcher and we would like to investigate whether that could be the origin of the disease.

1. One (hypothetical) approach would be to go back in time the day before food poisoning, and include everybody in town in the study. We would get the following model of the study base:

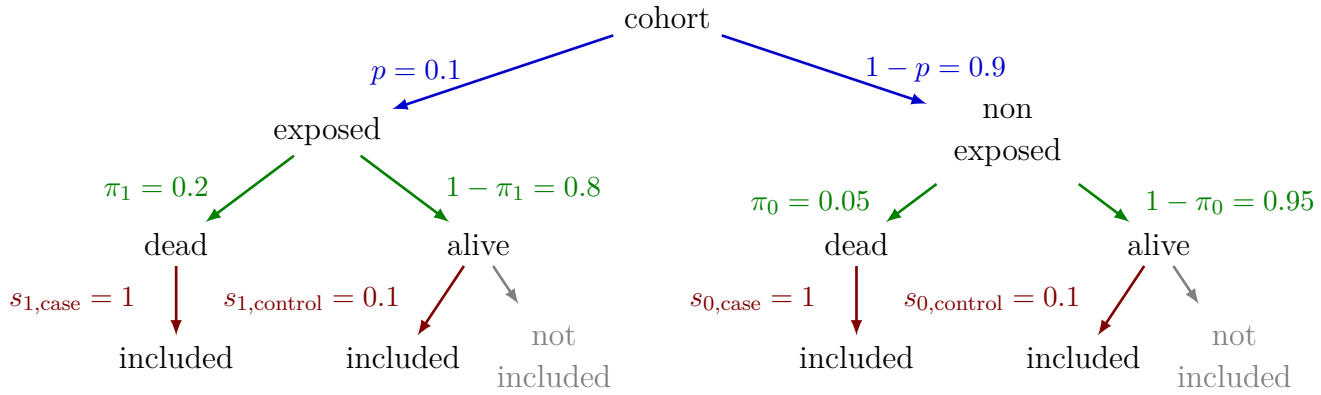


a) The town has 1000 inhabitants. How many would you expect[1]:

- die while having been exposed
- stay alive while having been exposed
- die while having not been exposed
- stay alive while having not been exposed

b) Deduce the risk difference, risk ratio, and odds ratio

2. In practice, we would instead do a case-control study, e.g. including only a subset of the controls. We can for instance take the list of inhabitants and randomly sample 10% of those who did not get sick:



a) How many would you now expect in the study[2]:

- die while having been exposed
- stay alive while having been exposed
- die while having not been exposed
- stay alive while having not been exposed

b) Deduce the risk difference, risk ratio, and odds ratio

[1] You can write it formally (using p, π_1, and π_0) or numerically.
[2] You can write it formally (using p, π_1, π_0, $s_{1,\text{case}}$, $s_{1,\text{control}}$, $s_{0,\text{case}}$, $s_{0,\text{control}}$) or numerically.

c) How do the fractions of cases and controls sampled among the exposed and non-exposed affect the estimation of the risk difference, risk ratio, and odds ratio? Would it be a good idea to include all the persons who died from food poisoning (regardless of whether they ate meat from the local butcher as cases) and as controls all (alive) customers of the local butcher plus 10% of the other (alive) inhabitants.

Exercise 2: Case study: BCG study

We will now re-visit the BCG study where we look at the survival of children depending on their vaccination status. The dataset contains the following variables:

- `age`: age group
- `scar`: vaccination status
- `status`: death (`case`) or alive (`con1000`, `conall`, `conmatch`). `conall` refer to the whole control population, `con1000` to 1000 randomly selected controls, and `conmatch` to age matched controls.
- `n`: number of children of a given age, vaccination status, and survival status or control group

```
library(Epi)
bcg <- read.table("http://publicifsv.sund.ku.dk/~pka/epidata/bcgalldata.
    txt", header=TRUE)
bcg$status <- as.factor(bcg$status)
bcg$scar <- factor(bcg$scar, labels = c("No","Yes"))
bcg$age <- as.factor(bcg$age)
str(bcg)
```

```
'data.frame':        56 obs. of  4 variables:
 $ age   : Factor w/ 7 levels "1","2","3","4",..: 1 2 3 4 5 6 7 1 2 3 ...
 $ scar  : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 1 1 ...
 $ status: Factor w/ 4 levels "case","con1000",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ n     : int  1 14 22 28 19 11 6 1 11 28 ...
```

We start by computing the age adjusted log-odds ratio for the vaccination effect when considering all controls:

```
bcg.all <- bcg[bcg$status %in% c("case","conall"),]
e.glmAll <- glm(status=="case" ~ age + scar,
                family = binomial(link="logit"),
                weight = n, data = bcg.all)
summary(e.glmAll)$coef
```

```
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -8.8800380  0.7102641 -12.502445 7.238979e-36
age2         2.6235363  0.7349321   3.569767 3.572990e-04
age3         3.5831114  0.7212206   4.968121 6.760471e-07
age4         3.8241284  0.7237070   5.284084 1.263355e-07
age5         3.9001565  0.7252749   5.377487 7.553260e-08
age6         4.1556320  0.7233366   5.745087 9.187360e-09
age7         4.1576390  0.7221755   5.757103 8.556965e-09
scarYes     -0.5470646  0.1409100  -3.882370 1.034434e-04
```

We can extract the estimated probabilities (in %) by age group doing:

```
grid <- unique(bcg.all[,c("age","scar")])
grid$fit.all <- 100*predict(e.glmAll, newdata = grid, type = "response")
pfit.all <- reshape(grid[,c("age","scar","fit.all")],
                    direction = "wide", v.names = "fit.all",
                    idvar = "age", timevar = "scar")
names(pfit.all) <- c("age","bcg.yes","bcg.no")
pfit.all
```

```
  age     bcg.yes     bcg.no
1   1 0.008050566 0.01391195
2   2 0.110857986 0.19142721
3   3 0.288888618 0.49820138
4   4 0.367333904 0.63312291
5   5 0.396235665 0.68279358
6   6 0.510980794 0.87978949
7   7 0.512002127 0.88154145
```

1. Investigate whether the vaccine effect could be age dependent

   a) fit a logistic model with interaction between age and vaccine status (see `e.glmAllI` below). Output the fitted probabilities and compare them to the one above (`pfit.all`).

```
e.glmAllI <- glm(status=="case" ~ age + age:scar,
                 family = binomial(link="logit"),
                 weight = n, data = bcg.all)
```

b) Use a likelihood ratio test to formally quantify the evidence for an interaction effect.

c) Output the age-specific odds ratios for the vaccine effect with their confidence intervals. Can we conclude about the absence of an age dependent vaccine effect?

2. Redo the analysis when using the 1000 randomly sampled controls. What is the impact of the change of control group on the validity and precision of the estimates?

3. [Extra] Same question with the matched controls. What happen if you omit age from the model?

4. [Extra] Consider this alternative parametrisation for the logistic model with interaction:

```
e.glmAllI.bis <- glm(status=="case" ~ age + scar + scar:age,
                     family = binomial(link="logit"),
                     weight = n, data = bcg.all)
summary(e.glmAllI.bis)$coef
```

```
               Estimate Std. Error     z value     Pr(>|z|)
(Intercept)  -8.9349820   1.000066 -8.93439374 4.094169e-19
age2          2.4589892   1.044596  2.35400992 1.857212e-02
age3          3.6347023   1.017853  3.57095105 3.556874e-04
age4          4.0077284   1.031060  3.88699843 1.014914e-04
age5          4.1317810   1.024959  4.03116535 5.550098e-05
age6          4.1391915   1.013972  4.08215596 4.461984e-05
age7          4.2200991   1.010742  4.17524950 2.976596e-05
scarYes      -0.4339847   1.414290 -0.30685687 7.589523e-01
age2:scarYes  0.3204617   1.470644  0.21790579 8.275025e-01
age3:scarYes -0.1054519   1.442808 -0.07308792 9.417362e-01
age4:scarYes -0.3082730   1.448795 -0.21277889 8.314994e-01
age5:scarYes -0.4467520   1.450323 -0.30803615 7.580548e-01
age6:scarYes  0.2344074   1.455936  0.16100121 8.720925e-01
age7:scarYes -0.1773891   1.479581 -0.11989142 9.045692e-01
```

The p-value for `scarYes` is rather large, 0.76. So why when using the `anova` function do we get a rather low p-value for `scar`?

```
anova(e.glmAllI.bis, test = "Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: status == "case"

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        27     3504.0
age       6  200.659        21     3303.3 < 2.2e-16 ***
scar      1   15.297        20     3288.0 9.187e-05 ***
age:scar  6    3.600        14     3284.4    0.7306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```