Biostatistics: Exercise 06

Beate Sick, Lisa Herzog

20.10.2020

Exercise 01: Cohort study

The table shows data from a cohort study representing the prevalence of hay fever and eczema in 11 year old children:

	Hay fever yes	Hay fever no	Total
Eczema yes	131	400	531
Eczema no	918	13 513	14 431
Total	1049	13 913	14 962

- Calculate the risk ratio (RR) and odds ratio (OR) for hay fever in people with Eczema compared to people without. Interpret the results. (Hint: Start by defining the variables a, b, c ,d as in the lecture slides. Then apply the formulas)

```
# define the variables (s. for example slide 8)

# y is the disease = hay fever, x is the exposure = eczema
a = 131 # P(y=hay fever|x=eczema)
b = 400 # P(y=no hay fever|x=eczema)
c = 918 # P(y=hay fever|x=no eczema)
d = 13513 # P(y=no hay fever|x=no eczema)

# Risk ratio:
# 1. Calculate the risk for hay fever in people with and without Eczema
(risk_with_eczema = a/(a+b))
```

```
## [1] 0.2467043
```

```
(risk_without_eczema = c/(c+d))
```

```
## [1] 0.06361306
```

```
# 2. risk ratio = risk_with_eczema/risk_without_eczema
(rr = risk_with_eczema/risk_without_eczema)
```

```
## [1] 3.878203
```

```
# The risk for hay fever is about 3.9 times higher in people
# with eczema compared to people without.


# Odds ratio:
(or = (a*d)/(b*c))
```

```
## [1] 4.820814
```

```
# The odds for hay fever is about 4.9 times higher for people
# with eczema compared to people without eczema.
```

- Calculate the risk ratio (RR) and odds ratio (OR) for Eczema in people with Hay fever compared to people without. Interpret the results. Do you notice anything interesting when you compare the results to the results of the previous subtask?

```
# define the variables (s. for example slide 8)
# y is the disease = eczema, x is the exposure = hay fever
a = 131 # P(y=eczema|x=hay fever)
b = 918 # P(y=no eczema|x=hay fever)
c = 400 # P(y=eczema|x=no hay fever)
d = 13513 # P(y=no eczema|x=no hay fever)


# Risk ratio:
# 1. Calculate the risk for hay fever in people with and without Eczema
(risk_with_eczema = a/(a+b))
```

```
## [1] 0.1248808
```

```
(risk_without_eczema = c/(c+d))
```

```
## [1] 0.02875009
```

```
# 2. risk ratio = risk_with_eczema/risk_without_eczema
(rr = risk_with_eczema/risk_without_eczema)
```

```
## [1] 4.343668
```

```
# The risk for eczema is about 4.4 times higher in people
# with hay fever compared to people without.


# Odds ratio:
(or = (a*d)/(b*c))
```

```
## [1] 4.820814
```

```
# The odds for eczema is about 4.9 times higher for people
# with hay fever compared to people without hay fever.


# We observe that the OR is the same (OR=4.9) no matter what
# variable is chosen as outcome. The risk ratio changes.
```

- Is there evidence for an association between Eczema and Hay fever? Choose an appropriate test. (R-Hint: The function `chisq.test()` expects a cross table)

```
# We choose a Chi2-Test to test for an association
# between eczema and hay fever
x = rbind(c(131,400), c(918,13513))
x
```

```
##      [,1]  [,2]
## [1,]  131   400
## [2,]  918 13513
```

```r
# Perform a Chi2-Test
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x
## X-squared = 260.54, df = 1, p-value < 2.2e-16
```

```r
# There is high evidence for an association between Eczema
# and hay fever (p-value<0.001).
```

Exercise 02: Case control study

We are using the data from a case control study investigating the association between coffee drinking and pancreatic cancer. The dataset `coffee.csv` can be downloaded from the website. You can read in the data using `read.csv(..., sep = ";", header = TRUE)`.

```r
# read in the data
dat = read.csv(paste0(dir, "data/coffee.csv"), sep=";", header=TRUE)
head(dat)
```

```
##   case sex coffee
## 1    0   0      1
## 2    0   0      1
## 3    0   0      1
## 4    0   0      1
## 5    0   0      1
## 6    0   0      1
```

- In general it is not possible to calculate the Risk Ratio from a Case control study. Why?

```r
# It is not possible to calcualte the RR because case control studies
# use disease based sampling. That is, we are able to estimate
# probabilities conditioned on disease status but not on exposure status.
```

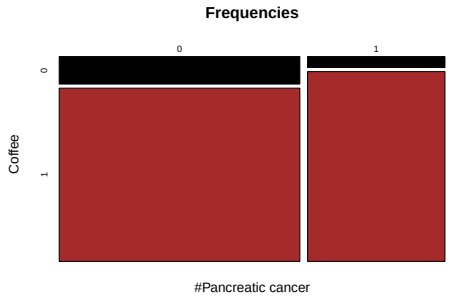
- To investigate your dataset, calculate a contingency table for coffee drinking vs. pancreatic cancer. Additionally, show the frequencies graphically. What do you observe? (R-Hint: `table()`, `mosaicplot()`)

```r
# We use a mosaic plot to investigate the association

# define a contingency table:
tab = table(cases = dat$case, coffee = dat$coffee)
tab
```

```
##      coffee
## cases   0   1
##     0  88 555
##     1  20 347
```

```r
# Then use a mosaic plot
mosaicplot(tab,
           col=c('black','brown'),
           xlab="#Pancreatic cancer",
           ylab="Coffee",
           main='Frequencies')
```



- Compute the OR and the corresponding 95% CI with the function `fisher.test()` measuring the association between coffee drinking and pancreatic cancer. Inetrpret the result. Is there evidence for an association between coffee drinking and pancreatic cancer?

```
# calculate a contingency table
tab = table(dat$case, dat$coffee)
tab
```

```
##
##       0   1
##   0  88 555
##   1  20 347
```

```
# OR and CI with fishers test
fisher.test(tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 3.028e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.641006 4.806538
## sample estimates:
## odds ratio
##   2.748557
```

```
# The odds to have pancreatic cancer is about 2.7 times higher
# for people who drink coffee compared to people who don't.
# The fisher test shows high evidence for an association between
# coffee drinking and pancreatic cancer.
```