Exercise 5 - solution

Paul Blanche

Exercise A

Part I

Question 1

We first load the data and look at the "summary", as always.

```
load(url("http://paulblanche.com/files/smoking.rda"))
d <- smoking
summary(d)
```

```
##          age        smoker     death
##   below 45:628   no :732   no :945
##   45-54   :208   yes:582   yes:369
##   55-64   :236
##   above 65:242
```

We can see that 945 women among the 1314 were still alive 20 years after the initial survey. There were 582 smokers among the 1314.

Question 2

We first produce a 2 by 2 table.

```
table(smoker=d$smoker,death=d$death)
```

```
##        death
## smoker  no yes
##    no  502 230
##    yes 443 139
```

We observe:

- 502 non-smokers alive at 20 years after the initial survey

- 230 non-smokers dead within 20 years after the initial survey
- 443 smokers alive at 20 years after the initial survey
- 139 smokers dead within 20 years after the initial survey

This gives the following 20-year risk estimates for smokers and non-smokers:

```r
230/(502+230) # 20-year risk of death among non smokers
```

```
## [1] 0.3142077
```

```r
139/(443+139) # 20-year risk of death among smokers
```

```
## [1] 0.2388316
```

That is, 31.4% for non smokers and 23.9% for smokers. These results seem surprising: we know that smoking is unhealthy, hence we expected a higher risk of death for smokers.

Question 3

We now compute the corresponding "exact" binomial confidence intervals.

```r
binom.test(x=230,n=502+230) # non smokers
```

```
##
##  Exact binomial test
##
## data:  230 and 502 + 230
## number of successes = 230, number of trials = 732, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2807031 0.3492176
## sample estimates:
## probability of success
##              0.3142077
```

```r
binom.test(x=139,n=443+139) # smokers
```

```
##
##  Exact binomial test
##
## data:  139 and 443 + 139
## number of successes = 139, number of trials = 582, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2047323 0.2756061
## sample estimates:
## probability of success
##              0.2388316
```

We therefore have the following 20-year risk estimate and 95% CI:

- Non-smokers: 31.4% (28.1 ; 34.9)
- Smokers: 23.9% (20.5 ; 27.6)

The 95% confidence interval do not overlap, hence the condidence interval suggest that the direction of the result, that is, a higher risk of death for non-smokers, is not due to small sample random variation.

Question 3

We now perform a statistical hypothesis test to compare the risk among smokers and non-smokers. We choose to use the Fisher's exact test because we prefer to obtain "exact" results to approximate results when we can. Howerver, the large sample size can justify to use a large sample method such as the Pearson Chi-square test.

```
fisher.test(table(smoker=d$smoker,death=d$death))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(smoker = d$smoker, death = d$death)
## p-value = 0.002989
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5307485 0.8822128
## sample estimates:
## odds ratio
##  0.6850392
```

We obtain a p-value=0.003, smaller than 5%, hence we conclude to a significant association between smoking and 20-year risk of death.

Just out of curiosity, we also compute the p-value of the Pearson Chi-square test.

```
chisq.test(table(smoker=d$smoker,death=d$death))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(smoker = d$smoker, death = d$death)
## X-squared = 8.7515, df = 1, p-value = 0.003093
```

We see that the p-value is almost the same, which is due to the large sample size.

We now use the fonction **table2x2()** from the **Publish** package to estimate several association measures, with confidence intervals. We start with the survival probability difference:

```r
Tab1 <- table(smoker=d$smoker,death=d$death)
library(Publish)
```

```
## Loading required package: prodlim
```

```r
table2x2(Tab1,stats = c("table","rd"))
```

```
## _____________________________
##
## 2x2 contingency table
## _____________________________
##
##                  deathno     deathyes      Sum
## smokerno             502          230      732
## smokeryes            443          139      582
## --                    --           --       --
## Sum                  945          369     1314
##
## _____________________________
##
## Statistics
## _____________________________
##
##
## a= 502
## b= 230
## c= 443
## d= 139
##
## p1=a/(a+b)= 0.6858
## p2=c/(c+d)= 0.7612
##
## _____________________________
##
## Risk difference
## _____________________________
##
## Risk difference = RD = p1-p2 = -0.07538
## Standard error = SE.RD = sqrt(p1*(1-p1)/(a+b)+p2*(1-p2)/(c+d)) = 0.02463
## Lower 95%-confidence limit: = RD - 1.96 * SE.RD = -0.1237
## Upper 95%-confidence limit: = RD + 1.96 * SE.RD = -0.0271
##
## The estimated risk difference is -7.5%  (CI_95%: [-12.4;-2.7]).
```

We can first check that the survival probability estimates match the previous results. We have:

- Non-smokers: 68.6%, which is indeed equal to 100 - 31.4
- Non-smokers: 76.1%, which is indeed equal to 100 - 23.9

We finally conclude that the estimated survival chance difference is 7.5% (95% CI= [12.4;2.7]).

We now compute the ratio of the survival probabilities.

```
table2x2(Tab1,stats = c("rr"))
```

```
##
## ------------------------------
##
## Risk ratio
## ------------------------------
##
## Risk ratio = RR = p1/p2 = 0.9010
## Standard error = SE.RR = sqrt((1-p1)/a+(1-p2)/c)= 0.9010
## Lower 95%-confidence limit: = RR * exp(- 1.96 * SE.RR) = 0.8427
## Upper 95%-confidence limit: = RR * exp(1.96 * SE.RR) = 0.9633
##
## The estimated risk ratio is 0.901 (CI_95%: [0.843;0.963]).
```

The estimated survival ratio is 0.901 (95% CI=[0.843;0.963]). Equivalently, we can also say that the chance of survival is 1-0.901= 9.9% lower for non smokers than for smokers (95% CI=[3.7;15.7]).

Finally, we now compute the survival odds ratio.

```
table2x2(Tab1,stats = c("or"))
```

```
##
## ------------------------------
##
## Odds ratio
## ------------------------------
##
## Odds ratio = OR = (p1/(1-p1))/(p2/(1-p2)) = 0.6848
## Standard error = SE.OR = sqrt((1/a+1/b+1/c+1/d)) = 0.1257
## Lower 95%-confidence limit: = OR * exp(- 1.96 * SE.OR) = 0.5353
## Upper 95%-confidence limit: = OR * exp(1.96 * SE.OR) = 0.8761
##
## The estimated odds ratio is 0.685 (CI_95%: [0.535;0.876]).
```

The estimated survival odds ratio is 0.685 (95% CI=[0.535;0.876]).

Question 5

We now produce a barplot to compare the number of included women in each age group between smokers and non smokers.

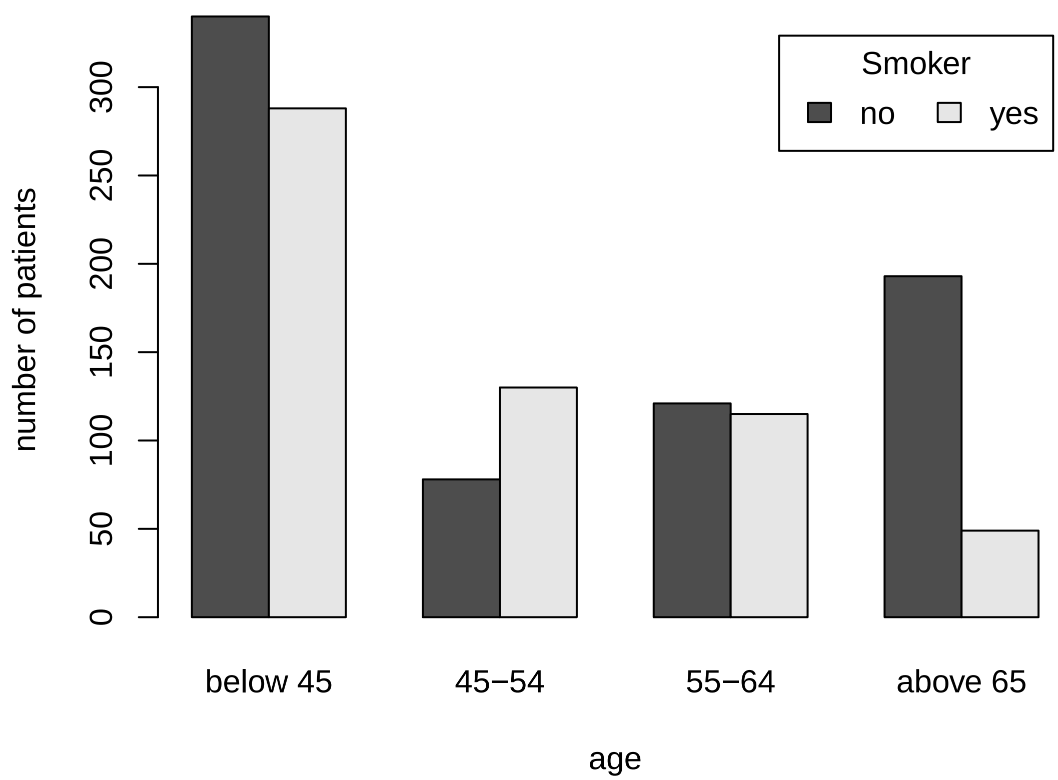
We first compute the frequency table.

```
Tab2 <- table(d$smoker,d$age)
Tab2
```

```
##
##       below 45 45-54 55-64 above 65
##   no       340    78   121      193
##   yes      288   130   115       49
```

And we now plot these counts.

```
barplot(Tab2,beside=TRUE,
        xlab="age",
        ylab="number of patients",
        legend=TRUE,
        args.legend=list(title="Smoker",ncol=2))
```



Interestingly, we observe that very few women above 65 are smokers. Although the number of smokers and non-smokers is not the same also in the other age groups, the differences are not as large.

Question 6

We now produce a barplot to compare the 20-year survival probability in each age group.

We first compute the frequency table, and then compute the survival probabilities.

```
Tab3 <- table(d$death,d$age) # counts
Tab3
```

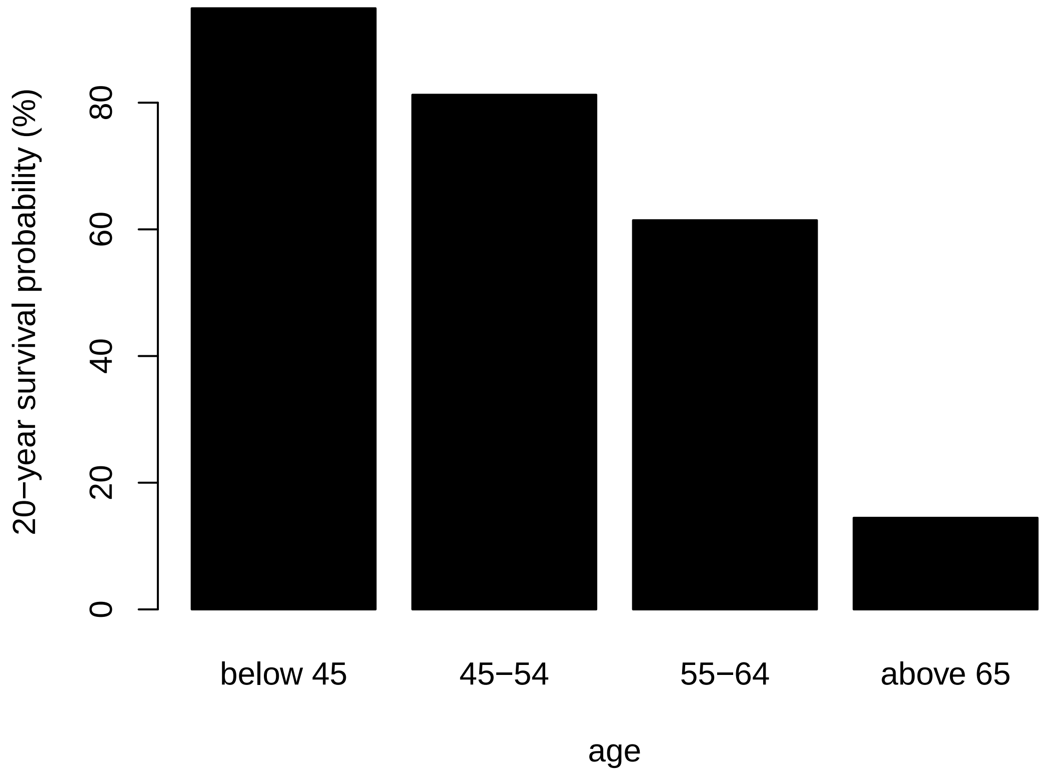
```
##
##       below 45 45-54 55-64 above 65
##   no       596   169   145       35
##   yes       32    39    91      207
```

```
Tab4 <- prop.table(Tab3,margin=2) # proportions per age
Tab4
```

```
##
##         below 45      45-54      55-64   above 65
##   no  0.94904459 0.81250000 0.61440678 0.14462810
##   yes 0.05095541 0.18750000 0.38559322 0.85537190
```

We are now ready to plot the 20-year survival probabilities in each age group.

```
barplot(100*Tab4[1,],
        xlab="age",
        ylab="20-year survival probability (%)",
        col="black")
```



We can observe that the older the women, the less likely they survive 20 years. This seems to make perfect sense.

Question 7

We now produce a barplot to compare the 20-year survival probability in each age group between smokers and non smokers. First, we compute the survival probabilities in each age group.

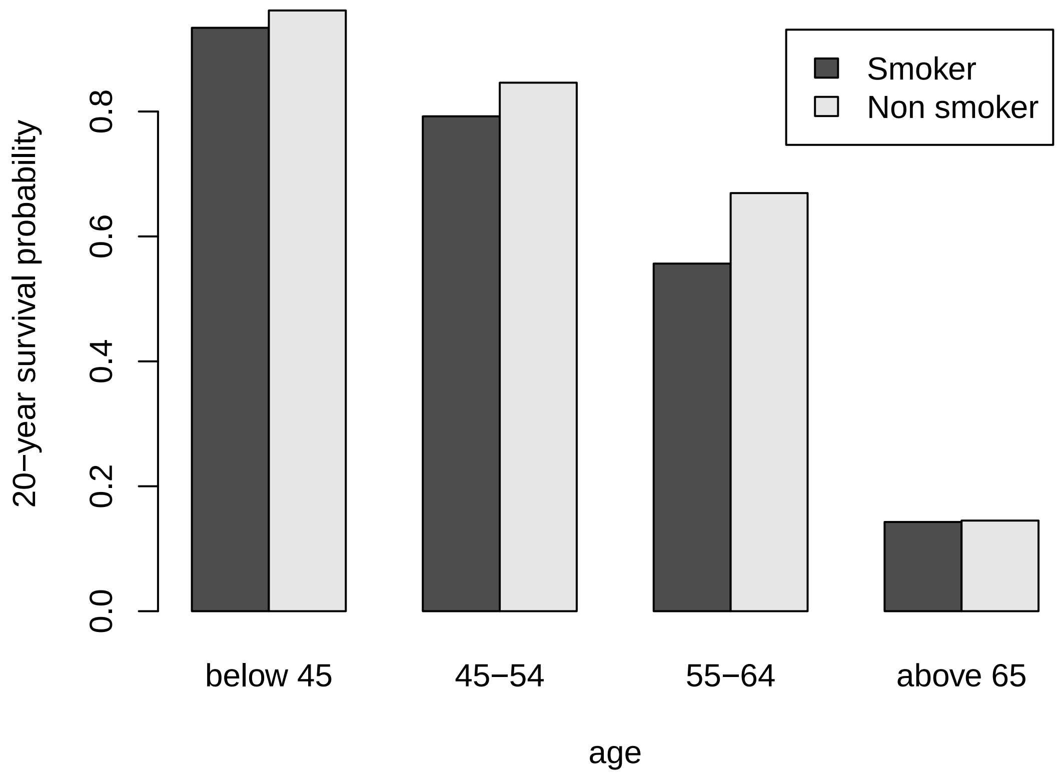
```
Tab3smokers <- table(d$death[d$smoker=="yes"],
                     d$age[d$smoker=="yes"]) # counts for smokers
Tab4smokers <- prop.table(Tab3smokers,
                          margin=2) # proportions per age for smokers
Tab3NonSmokers <- table(d$death[d$smoker=="no"],
                        d$age[d$smoker=="no"]) # counts for non smokers
Tab4NonSmokers <- prop.table(Tab3NonSmokers,
                             margin=2) # proportions per age for non smokers
# merge the results into one unique matrix
Tab5 <- rbind(Tab4smokers[1,],Tab4NonSmokers[1,])
rownames(Tab5) <- c("Smoker","Non smoker")
Tab5
```

```
##              below 45     45-54     55-64  above 65
## Smoker      0.9340278 0.7923077 0.5565217 0.1428571
```

```
## Non smoker 0.9617647 0.8461538 0.6694215 0.1450777
```

We are now ready to produce the plot.

```
barplot(Tab5,
        beside=TRUE,
        xlab="age",
        ylab="20-year survival probability",
        legend=TRUE)
```



Question 8

We have seen that smokers are younger than non-smokers in this data set. Especially, most of the women older than 65 are non-smokers. Because we have also seen that the 20-year risk of death is much higher for women older than 65 than the others (as expected), we might suspect that the survival difference between smokers and non-smokers is mostly driven by the fact that those who do not smoke are older.

Smoking is known to be unhealthy, but it is probably not as life-threatening as becoming "old", which is what suggests the plot obtained at question 7. Hence the results of question 3: when comparing young smokers to old non-smokers, the data suggest that old non-smokers die more.

Part II

Below we examplify how to do it for the age group 45-54.

```
Tab1a <- table(smoker=d$smoker[d$age=="45-54"],death=d$death[d$age=="45-54"])
table2x2(Tab1a,stats = c("table","rr"))
```

```
## ------------------------------
##
## 2x2 contingency table
## ------------------------------
##
##              deathno    deathyes    Sum
## smokerno          66          12     78
## smokeryes        103          27    130
## --                --          --     --
## Sum              169          39    208
##
## ------------------------------
##
## Statistics
## ------------------------------
##
##
## a= 66
## b= 12
## c= 103
## d= 27
##
## p1=a/(a+b)= 0.8462
## p2=c/(c+d)= 0.7923
##
## ------------------------------
##
## Risk ratio
## ------------------------------
##
## Risk ratio = RR = p1/p2 = 1.0680
## Standard error = SE.RR = sqrt((1-p1)/a+(1-p2)/c)= 1.0680
## Lower 95%-confidence limit: = RR * exp(- 1.96 * SE.RR) = 0.9385
## Upper 95%-confidence limit: = RR * exp(1.96 * SE.RR) = 1.2153
##
## The estimated risk ratio is 1.068 (CI_95%: [0.938;1.215]).
```

We can proceed similarly for the other age groups and obtain:

- Age group below 45: estimated risk ratio 1.030 (95% CI= [0.992;1.069]).
- Age group 45-54: estimated risk ratio 1.068 (95% CI= [0.938;1.215]).
- Age group 55-64: estimated risk ratio 1.203 (95% CI= [0.979;1.478]).
- Age group above 65: estimated risk ratio 1.016 (95% CI= [0.472;2.186]).

There is no significant difference in any age group, although there seems to be a systematic trend towards a higher 20-year risk of death for smokers. The fact that the results are not significant is probably due to a lack of power. The sample size of each age group is not very large.

Question 10

The results of Part II can be thought as more interesting because the two groups that we compare (smokers versus non smokers) are expected to be more similar with respect to everything but smoking, as the women of the two groups have a similar age. Hence we can expect that the association between smoking and survival that we estimated are closer to causal associations than those of Part I, although we cannot rigouroulsy claim that they are indeed causal.

Exercise B

Question 1

```r
power.prop.test(p1=0.40,p2=0.60,power=0.85)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 110.6668
##              p1 = 0.4
##              p2 = 0.6
##       sig.level = 0.05
##           power = 0.85
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

We need to include 111 women in each treatment group, hence we need to include 222 women in total.

Question 2

```
power.prop.test(n=111,p1=0.40,p2=0.5)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 111
##              p1 = 0.4
##              p2 = 0.5
##       sig.level = 0.05
##           power = 0.3210212
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The power of the study drops to only 32%, if we include 222 women (111 per group) and if the treatment results only in 50% chance of pregnancy.

```
power.prop.test(n=111,p1=0.40,p2=0.55)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 111
##              p1 = 0.4
##              p2 = 0.55
##       sig.level = 0.05
##           power = 0.6106339
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The power of the study drops to 61%, if we include 222 women (111 per group) and if the treatment results only in 55% chance of pregnancy.

```
power.prop.test(n=111,p1=0.40,power=0.75)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 111
##              p1 = 0.4
##              p2 = 0.5760567
##       sig.level = 0.05
##           power = 0.75
##     alternative = two.sided
```

```
##
## NOTE: n is number in *each* group
```

The smallest improvement in chance of pregnancy that we can hope to show with this sample size and a decent power of 75% is 18%, i.e. 58% chance of pregnancy with treatment verus 40% without treatment (if the chance of pregnancy without treatment is indeed 40%) .

Exercise C

Question 1

```r
TabCCS1 <- data.frame(Case=c(47,687),Control=c(36,2364))
rownames(TabCCS1) <- c("Multiple birth","Singleton")
table2x2(TabCCS1,stats=c("table","or"))
```

```
## -----------------------------
##
## 2x2 contingency table
## -----------------------------
##
##                     Case     Control      Sum
## Multiple birth        47          36       83
## Singleton            687        2364     3051
## --                    --          --       --
## Sum                  734        2400     3134
##
## -----------------------------
##
## Statistics
## -----------------------------
##
##
## a= 47
## b= 36
## c= 687
## d= 2364
##
## p1=a/(a+b)= 0.5663
## p2=c/(c+d)= 0.2252
##
## -----------------------------
##
## Odds ratio
```

```
## ------------------------------
##
## Odds ratio = OR = (p1/(1-p1))/(p2/(1-p2)) = 4.4925
## Standard error = SE.OR = sqrt((1/a+1/b+1/c+1/d)) = 0.2257
## Lower 95%-confidence limit: = OR * exp(- 1.96 * SE.OR) = 2.8866
## Upper 95%-confidence limit: = OR * exp(1.96 * SE.OR) = 6.9918
##
## The estimated odds ratio is 4.492 (CI_95%: [2.887;6.992]).
```

The odds ratio (95% confidence limits) is 4.49 (2.89;6.99) for multiple birth compared to singleton.

Question 2

```
TabCCS2 <- data.frame(Case=c(47,687),Control=c(10999,722267))
rownames(TabCCS2) <- c("Multiple birth","Singleton")
table2x2(TabCCS2,stats=c("table","or"))
```

```
## ------------------------------
##
## 2x2 contingency table
## ------------------------------
##
##                    Case     Control        Sum
## Multiple birth       47       10999      11046
## Singleton           687      722267     722954
## --                   --          --         --
## Sum                 734      733266     734000
##
## ------------------------------
##
## Statistics
## ------------------------------
##
##
## a= 47
## b= 10999
## c= 687
## d= 722267
##
## p1=a/(a+b)= 0.0043
## p2=c/(c+d)= 0.001
##
## ------------------------------
```

```
##
## Odds ratio
## -----------------------------
##
## Odds ratio = OR = (p1/(1-p1))/(p2/(1-p2)) = 4.4925
## Standard error = SE.OR = sqrt((1/a+1/b+1/c+1/d)) = 0.1511
## Lower 95%-confidence limit: = OR * exp(- 1.96 * SE.OR) = 3.3411
## Upper 95%-confidence limit: = OR * exp(1.96 * SE.OR) = 6.0406
##
## The estimated odds ratio is 4.492 (CI_95%: [3.341;6.041]).
```

The odds ratio (95% confidence limits) is 4.49 (3.34;6.04) for multiple birth compared to singleton. The result is almost the same here. The estimated odds ratio is the same and the conficence interval is almost the same, although slightly narrower, [3.34;6.04] versus [2.89;6.99]. The same estimated odds ratio could be expected, because the odds ratio does not depend on the prevalence and because odds ratios are known to be possible to estimate without bias with case-control studies. The fact that the length of the confidence interval is almost the same also makes sense. Indeed, the standard error for the logarithm of the odds ratio, which we use to compute the length of the confidence intervals, is $1/a + 1/b + 1/c + 1/d$, where a, b, c and d are the counts in the 2 by 2 table. Hence, it is mostly the "small" values "a" and "c" which matter, i.e. the numbers of cases for exposed and non exposed. Wether the values of "c" and "d" are "large" or "very large" does not matter much.

Question 3

We **WRONGLY** estimate the risk ratio with the data from the case-control study.

```
table2x2(TabCCS1,stats=c("rr"))
```

```
##
## -----------------------------
##
## Risk ratio
## -----------------------------
##
## Risk ratio = RR = p1/p2 = 2.5148
## Standard error = SE.RR = sqrt((1-p1)/a+(1-p2)/c)= 2.5148
## Lower 95%-confidence limit: = RR * exp(- 1.96 * SE.RR) = 2.0601
## Upper 95%-confidence limit: = RR * exp(1.96 * SE.RR) = 3.0699
##
## The estimated risk ratio is 2.515 (CI_95%: [2.060;3.070]).
```

We **CORRECTLY** estimate the risk ratio with the data from the case-control study.

```
table2x2(TabCCS2,stats=c("rr"))
```

```
##
## ------------------------------
##
## Risk ratio
## ------------------------------
##
## Risk ratio = RR = p1/p2 = 4.4776
## Standard error = SE.RR = sqrt((1-p1)/a+(1-p2)/c)= 4.4776
## Lower 95%-confidence limit: = RR * exp(- 1.96 * SE.RR) = 3.3340
## Upper 95%-confidence limit: = RR * exp(1.96 * SE.RR) = 6.0135
##
## The estimated risk ratio is 4.478 (CI_95%: [3.334;6.013]).
```

The results are different when we estimate the risk ratio using either the cohort data or the case-control study data. This is because, as we discussed during the lecture, we **cannot** correctly estimate the risk ratio using case-control data (we can only estimate odds ratio). But of course we can with cohort data.