Around Simpson's Paradox

Benjamin Cathelineau

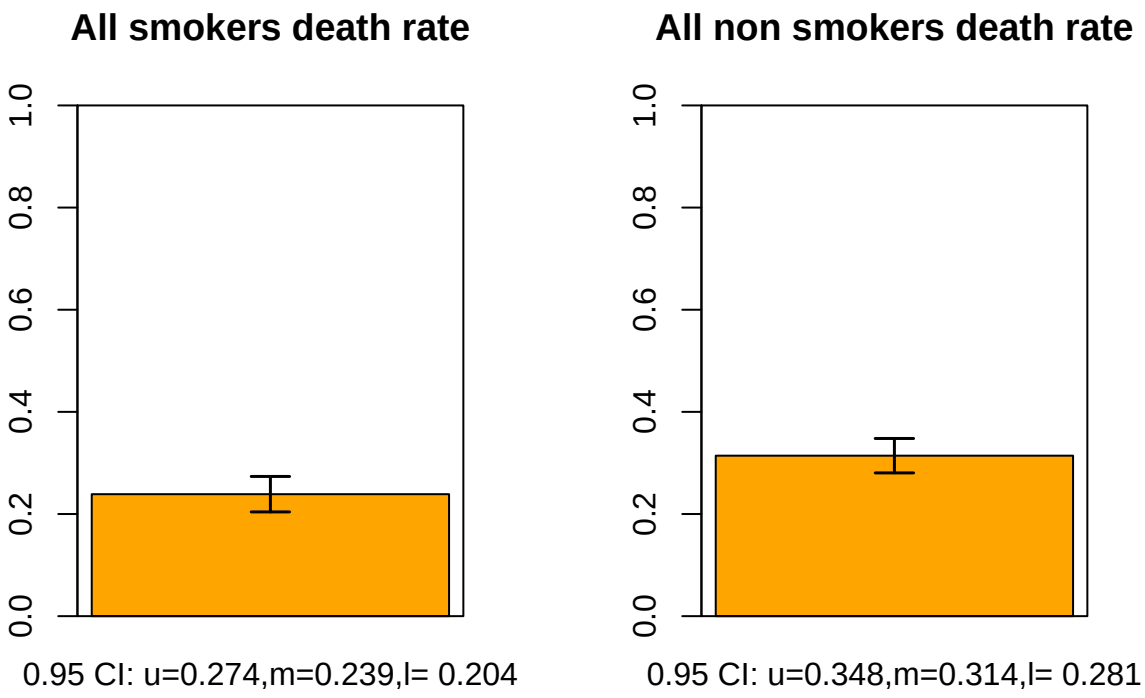
21/11/2021

Question 1

```
df = read.csv("Subject6_smoking.csv")


compute_confidence_interval_and_plot <- function(smoker_arg,df_arg, title_arg){
  alives= df_arg %>% filter(Status == "Alive"& Smoker== smoker_arg) %>% mutate(death_variable=0)
  dead= df_arg %>% filter(Status == "Dead"& Smoker== smoker_arg) %>% mutate(death_variable=1)
  all_member = rbind(alives,dead)

  my_ci= CI(x=all_member$death_variable,ci=0.95)
  # inspired by https://app-learninglab.inria.fr/moocrr/gitlab/moocrr-session3/moocrr-reproducibility-s
  death_rate<-c(my_ci[2])
  bp<-barplot(death_rate,col="orange",ylim=c(0,1),names.arg=sprintf("0.95 CI: u=%.3f,m=%.3f,l= %.3f",my
  arrows(bp,my_ci[1],bp,my_ci[3],lwd=1.5,angle=90,length=0.1,code=3)
}
```

We declare a function so that we can compute both rates (for smokers and non smokers), without repeating our code

```
par(mfcol=c(1,2))
compute_confidence_interval_and_plot(smoker_arg="Yes",df_arg = df, title_arg="All smokers death rate")
compute_confidence_interval_and_plot(smoker_arg="No",df_arg = df, title_arg="All non smokers death rate
```



The mortality rate is significantly higher for the group that is not smoking. In other words, in with this data, a woman who smoked in 1977 is less likely to have died in 1995 than a woman who did not smoke in 1977.

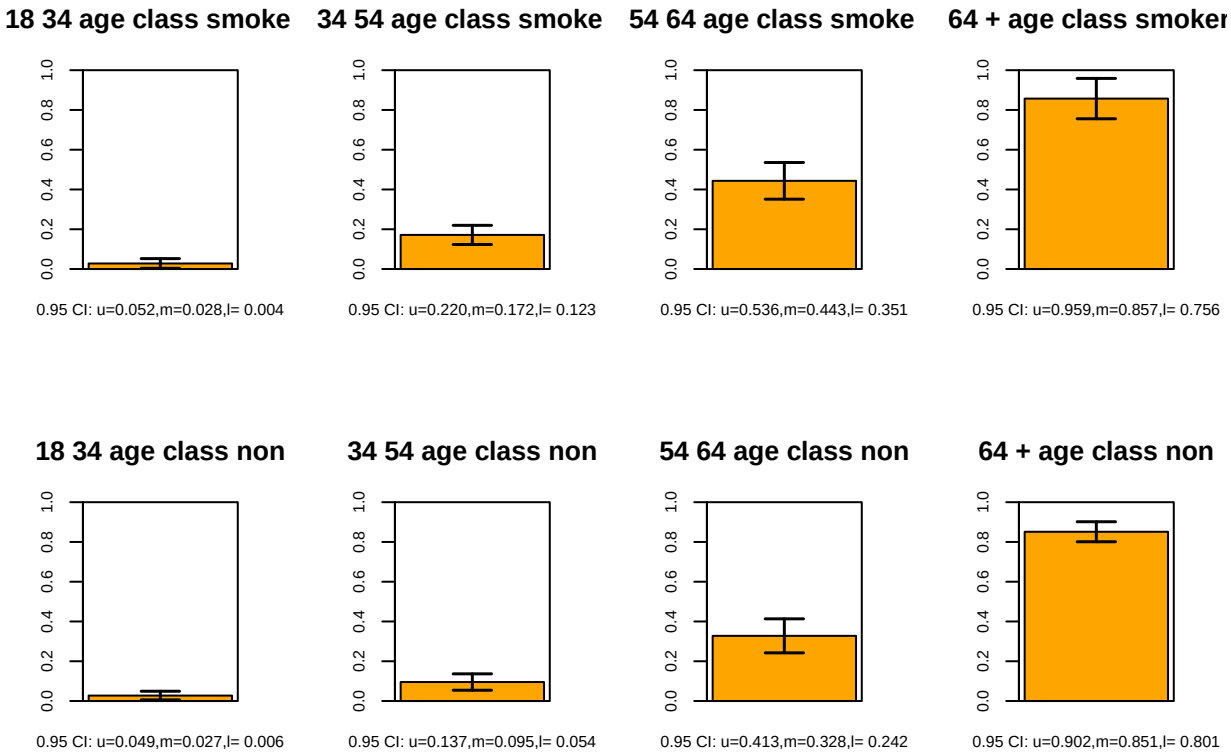
Of course, this is very surprising because it is now known that smoking cigarette increases the risk of death, trough various mechanisms, such as increased risk of cancer and cardiovascular disease. For more details, consult the relevant wikipedia article.

Question 2

We will use the recommended age grouping.

```
class1834 = df %>% filter(Age >= 18 & Age < 34)
class3454 = df %>% filter(Age >= 34 & Age < 54)
class5464 = df %>% filter(Age >= 54 & Age < 64)
class64 = df %>% filter(Age >= 64)

mortality_ratio_and_ci <- function(df_arg, title_arg){
compute_confidence_interval_and_plot(smoker_arg="Yes",df_arg = df_arg, title_arg = paste(title_arg,"smo
compute_confidence_interval_and_plot(smoker_arg="No",df_arg = df_arg, title_arg = paste(title_arg,"non"

}

par(mfcol=c(2,4),cex.axis=0.8)


mortality_ratio_and_ci(class1834,"18 34 age class")
```

```
mortality_ratio_and_ci(class3454,"34 54 age class")
mortality_ratio_and_ci(class5464,"54 64 age class")
mortality_ratio_and_ci(class64,"64 + age class")
```



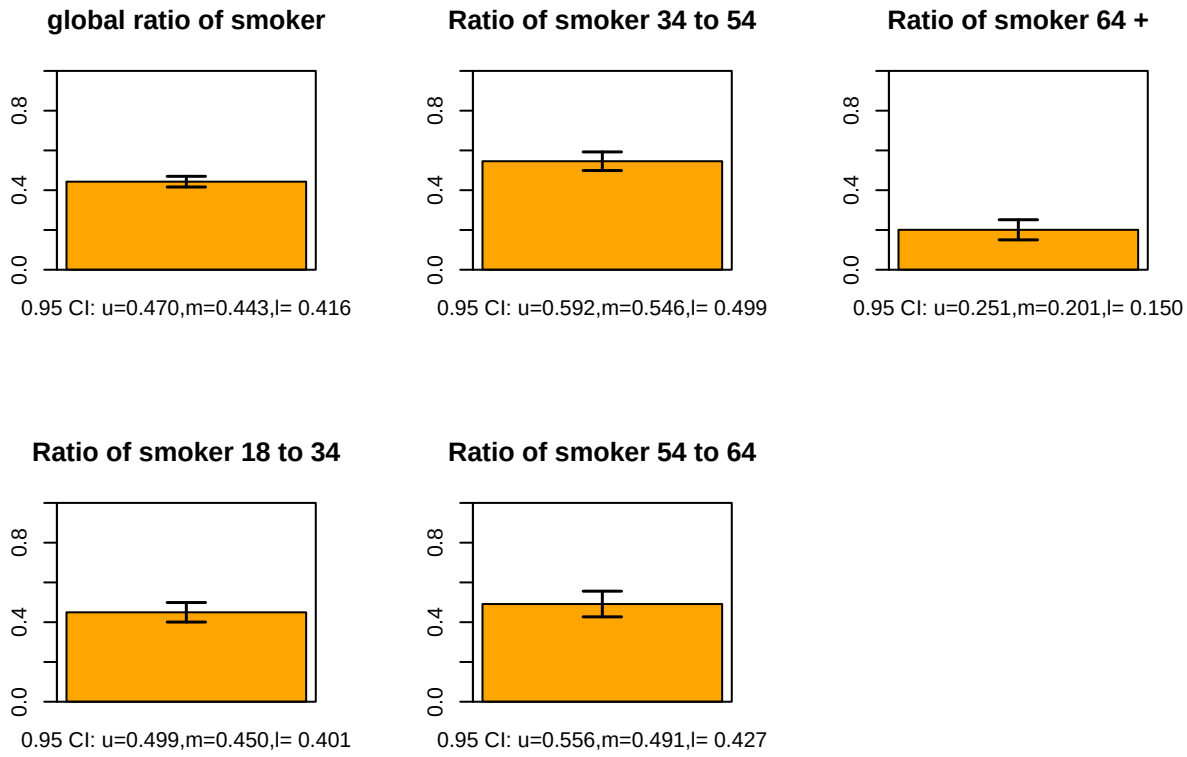
This is very surprising, because, as we saw in question 1, the mortality rate was higher for *non smoker*. But now, after organizing the data in age classes, for every single class, the mortality is higher for the *smoker* group. So there is seemingly direct contradiction.

What is happening

To figure out what is happening, we need more information. Currently, we have the global death ratio for smoker and non smoker. We also have the death ratio for separate age class, and that's where we find a contradiction. An interesting information to have would be the ratio of smoker, especially in separate age class because we might find that people of different age have different smoking habit. I developed a function to compute to ratio of smoker.

```
compute_smoker_ratio <- function(df_arg, title_arg){
  smokers= df_arg %>% filter( Smoker== "Yes") %>% mutate(smoker_variable=1)
  non_smokers= df_arg %>% filter(Smoker== "No") %>% mutate(smoker_variable=0)
  all_member = rbind(smokers,non_smokers)

  my_ci= CI(x=all_member$smoker_variable,ci=0.95)
  # inspired by https://app-learninglab.inria.fr/moocrr/gitlab/moocrr-session3/moocrr-reproducibility-s
  smoker_rate<-c(my_ci[2])
  bp<-barplot(smoker_rate,col="orange",ylim=c(0,1),names.arg=sprintf("0.95 CI: u=%.3f,m=%.3f,l= %.3f",my
  arrows(bp,my_ci[1],bp,my_ci[3],lwd=1.5,angle=90,length=0.1,code=3)
```

```
}
par(mfcol=c(2,3))

compute_smoker_ratio(df_arg = df, title_arg = "global ratio of smoker")
compute_smoker_ratio(df_arg = class1834, title_arg = "Ratio of smoker 18 to 34")
compute_smoker_ratio(df_arg = class3454, title_arg = "Ratio of smoker 34 to 54")
compute_smoker_ratio(df_arg = class5464, title_arg = "Ratio of smoker 54 to 64")
compute_smoker_ratio(df_arg = class64,title_arg = "Ratio of smoker 64 +" )
```



It is immediately apparent that the oldest age class smokes much less than all the other age classes.

Given this, I can give a tentative explanation of what is going on:

1. Age is, as far as I know, the most important factor in increasing the probability of death, in the general population.
2. People of older age are therefore, everything else being equal, much more likely to die.
3. In this study it is revealed that older people (in the 64 + age class), are much less likely to be smokers than the younger age classes.
4. However, due to their age, they are much more likely to die than the younger age classes.

The age classes are not exactly the same size but they are all in the same order of magnitude.

And that's why we see this effect. Old people are much more likely to die in the first place.

The conclusion is that Age is more important to your health than smoking, even if smoking is still very important and should be avoided for a good health.

According to this youtube video that I watched recently, the same phenomena will most likely occur at some point with the covid 19 vaccine. The vaccine lowers the risk of death from covid 19 (lower risk of infection and lower risk of death when infected). But since older people are more vaccinated and much more likely to die in the first place, we might see a negative association between the vaccine and covid 19 protection. In other words, unvaccinated people will have less probability of dying of covid 19 than vaccinated people, even if the vaccine is effective in protecting people. That is, again,because older people are much more likely to die in the first place, and they are more vaccinated than younger people.

Question 3

```
# inspired by https://gitlab.ensimag.fr/vaudeyj/mosig-smpe/-/blob/master/Peer%20evaluated%20exercise/Pe
dead= df %>% filter( Status== "Dead") %>% mutate(death_variabe=1)
alives= df %>% filter(Status== "Alive") %>% mutate(death_variabe=0)
all_member = rbind(dead,alives)
all_smokers = all_member %>% filter(Smoker == "Yes")

regression_smoker = glm(all_smokers$death_variabe~all_smokers$Age,family = binomial(link = logit))


summary(regression_smoker)
```

```
##
## Call:
## glm(formula = all_smokers$death_variabe ~ all_smokers$Age, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0745  -0.6464  -0.3756  -0.2013   2.6560
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.508106   0.466221  -11.81   <2e-16 ***
## all_smokers$Age  0.088977   0.008721   10.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 639.89  on 581  degrees of freedom
## Residual deviance: 480.41  on 580  degrees of freedom
## AIC: 484.41
##
## Number of Fisher Scoring iterations: 5
```

```
all_non_smokers = all_member %>% filter(Smoker == "No")

regression_non_smoker = glm(all_non_smokers$death_variabe~all_non_smokers$Age,family = binomial(link = l

summary(regression_non_smoker)
```

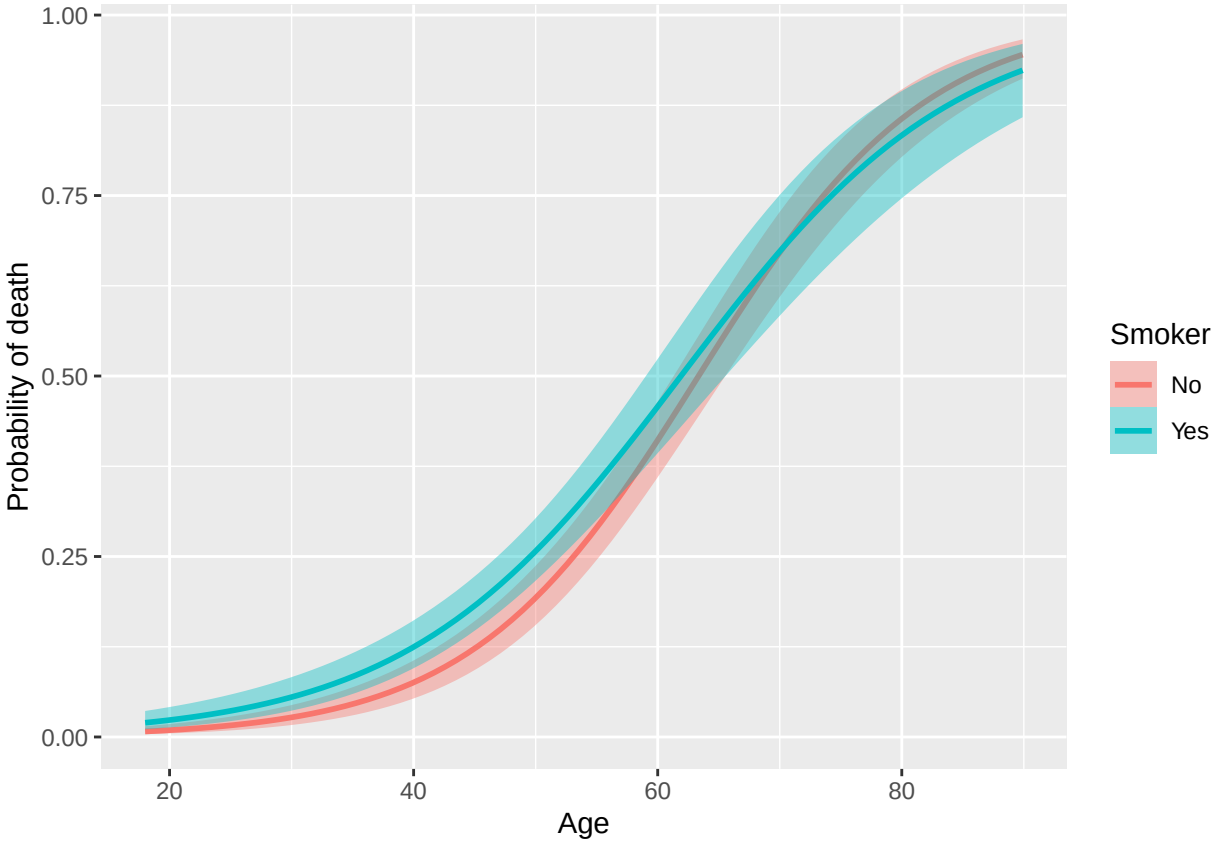
```
##
## Call:
## glm(formula = all_non_smokers$death_variabe ~ all_non_smokers$Age,
##     family = binomial(link = logit))
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4019  -0.5179  -0.2003   0.4728   3.0457
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -6.795507   0.479430  -14.17   <2e-16 ***
## all_non_smokers$Age  0.107275   0.007806   13.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 911.23  on 731  degrees of freedom
## Residual deviance: 519.08  on 730  degrees of freedom
## AIC: 523.08
##
## Number of Fisher Scoring iterations: 6
```

For both group the age is positively correlated with the probability of dying in the 20 years period of the study. In other words, the more the age of a person increases, the more likely it is that the person will die in the 20 year period.

```
ggplot(all_member)+
  aes(x = Age) +
  aes(y = death_variabe) +
  aes(group = Smoker) +
  stat_smooth( method="glm", se=TRUE, fullrange=TRUE,level=.95,
               method.args = list(family=binomial), aes(color=Smoker, fill = Smoker)) +
  ylab("Probability of death")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



As we have seen already, the age is positively correlated with the probability of dying, which is why we see this shape. Unfortunately, with this data we cannot give any kind of conclusion on the danger of smoking. Indeed it is standard in modern science to use a 5% $p - value$. In other words, to make the claim that smoking increases the probability of death, we would need the 95 % confidence interval to not overlap. However here the confidence interval overlap for every age. We could lower the confidence to 70 % and it would probably work but at this point we won't be very confident in what we say.

Fortunately, the terrible impact of smoking on people health is already well established anyways.