Peer reviewed exercise : Simpson's paradox

Introduction

We will be using the libraries ggplot2 and dplyr in this document.

```
library(dplyr)
library(ggplot2)
```

We are studying a dataset containing information gathered by two surveys conducted in 1977 and 1995 respectively. One sixth of the electorate was surveyed but the dataset we use in this analysis is restricted to women and, more specifically, to the 1314 that were categorized as "smoking currently" or "never smoked" (for the sake of simplicity). There were very few women in the initial dataset that were categorized differently (162 as "smoked but quit" and 18 for which data was not available). Each line in the dataset contains if the person smokes or not, if the person was alive at the time of the second survey and their age at the time of the first one.

Studying the mortality rate in both groups

We start by reading the data from the file.

```
data = read.csv(file = "Subject6_smoking.csv", sep = ",", header = T)
head(data)
```

```
##   Smoker Status  Age
## 1    Yes  Alive 21.0
## 2    Yes  Alive 19.3
## 3     No   Dead 57.5
## 4     No  Alive 47.1
## 5    Yes  Alive 81.4
## 6     No  Alive 36.8
```

First, we are going to compute the mortality rate among the two groups (smoking / non-smoking).

To do this, we add the variable *Death* which is equal to 1 if the person is dead at the time of the second survey, 0 otherwise. This makes computing the mortality rate easier as it is just the mean of this variable.

```
data %>% mutate(Death = case_when(
  data$Status == "Alive" ~ 0,
  data$Status == "Dead" ~ 1
)) -> data
```

We change the value in the column *Smoker* to improve readability in the plots later on

```
data[data == "Yes"] <- "Smoker"
data[data == "No"] <- "Non-Smoker"
```

As mentioned previously, we can compute the mortality rate by computing the mean of *Death* for both groups (smokers / non-smokers).

```
mortality_rate_smoking <- mean(data[data$Smoker == "Smoker",]$Death)
mortality_rate_smoking
```

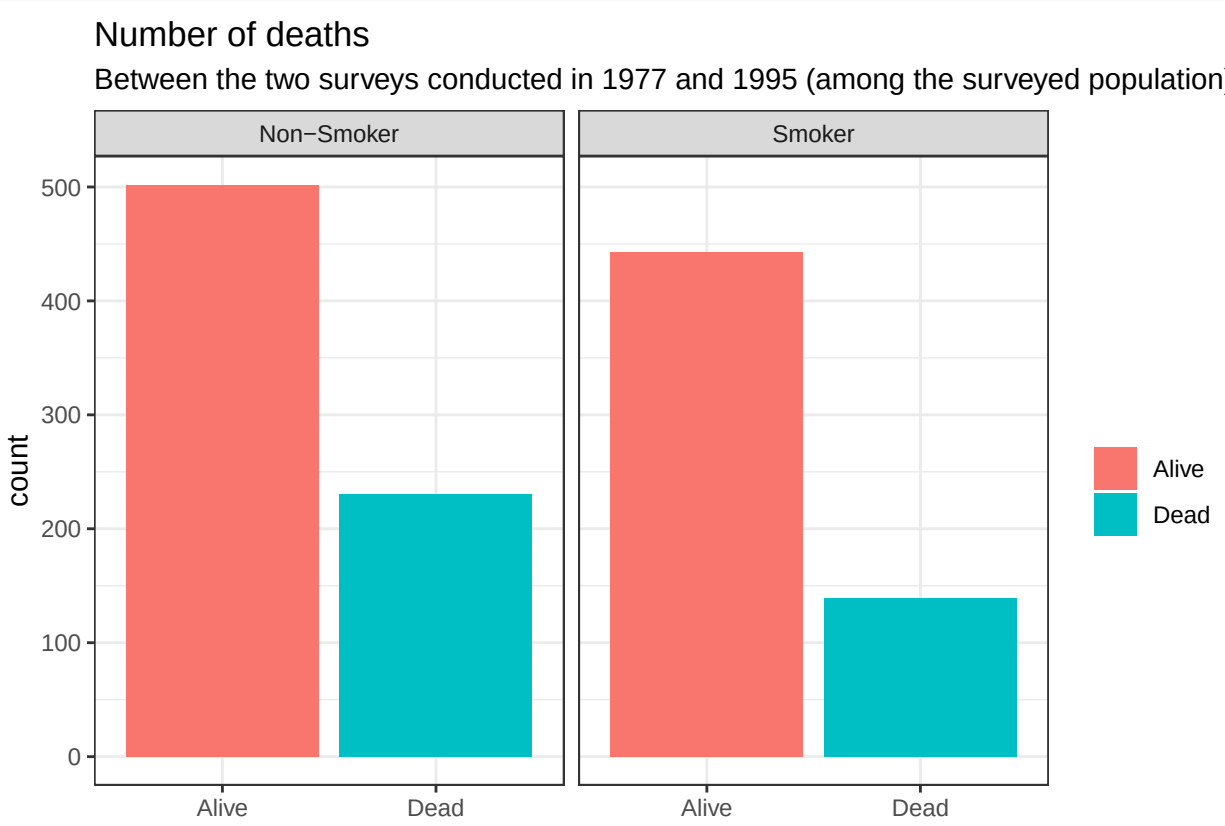
```
## [1] 0.2388316
```

```r
mortality_rate_nonsmoking <- mean(data[data$Smoker == "Non-Smoker",]$Death)
mortality_rate_nonsmoking
```

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

We plot the data in the following graph:

```r
ggplot(data) +
  facet_wrap(~ Smoker) +
  aes(x = Status, fill = factor(Status)) +
  geom_bar() +
  theme_bw() +
  labs(x = "", fill = "", title = "Number of deaths", subtitle = "Between the two surveys conducted in
```



We compute the standard deviation among the smokers ("manually" and with the built-in function `sd` to check that the result is correct).

```r
smokers <- data[data$Smoker == "Smoker",]$Death

sd_smokers <- 0
for (i in 1:length(smokers)) {
  sd_smokers <- sd_smokers + (smokers[i] - mortality_rate_smoking) ** 2
}
sd_smokers <- sqrt(sd_smokers / length(smokers))
sd_smokers
```

```
## [1] 0.4263696
```

```r
sd(smokers) * sqrt((length(smokers) - 1) / length(smokers))
```

```
## [1] 0.4263696
```

We do the same with the nonsmokers.

```
nonsmokers <- data[data$Smoker == "Non-Smoker",]$Death

sd_nonsmokers <- 0
for (i in 1:length(nonsmokers)) {
  sd_nonsmokers <- sd_nonsmokers + (nonsmokers[i] - mortality_rate_nonsmoking) ** 2
}
sd_nonsmokers <- sqrt(sd_nonsmokers / length(nonsmokers))
sd_nonsmokers
```

```
## [1] 0.4641995
```

```
sd(nonsmokers) * sqrt((length(nonsmokers) - 1) / length(nonsmokers))
```

```
## [1] 0.4641995
```

Finally, we can compute the confidence intervals for the two groups.

```
CI_smokers <- 2 * (sd_smokers / sqrt(length(smokers)))
c(mortality_rate_smoking - CI_smokers, mortality_rate_smoking + CI_smokers)
```

```
## [1] 0.2034844 0.2741788
```

```
CI_nonsmokers <- 2 * (sd_nonsmokers / sqrt(length(nonsmokers)))
c(mortality_rate_nonsmoking - CI_nonsmokers, mortality_rate_nonsmoking + CI_nonsmokers)
```

```
## [1] 0.2798930 0.3485223
```

The following table regroups all of the previously computed informations.

```
data %>%
  group_by(Smoker) %>%
  summarise_at(vars(Death), list(n = length, mortality_rate = mean, sd = sd)) %>%
  mutate(sd = sd * sqrt((n - 1) / n)) %>%
  mutate(CI_lower_bound = mortality_rate - 2 * sd / sqrt(n),
         CI_upper_bound = mortality_rate + 2 * sd / sqrt(n)) -> data_summary

data_summary
```

```
## # A tibble: 2 x 6
##   Smoker         n mortality_rate    sd CI_lower_bound CI_upper_bound
##   <chr>      <int>          <dbl> <dbl>          <dbl>          <dbl>
## 1 Non-Smoker   732          0.314 0.464          0.280          0.349
## 2 Smoker       582          0.239 0.426          0.203          0.274
```

We can see the two intervals do not overlap which means that we can say that the mortality rate among the smokers is lower than the one among nonsmokers with more than 90% of confidence.

It is obviously quite surprising as we would normally expect the mortality rate among the smokers to be higher. However, this mortality rate does not take into account the cause of death.

```
# Number of old (65 years old or more) people in the smoker group and the percentage of this group it r
c(nrow(data[data$Smoker == "Smoker" & data$Age >= 65.0,]),
nrow(data[data$Smoker == "Smoker" & data$Age >= 65.0,]) / data_summary[data_summary$Smoker == "Smoker",]
```

```
## [1] 49.00000000  0.08419244
```

```
# Number of old (65 years old or more) people in the non-smoker group and the percentage of this group
c(nrow(data[data$Smoker == "Non-Smoker" & data$Age >= 65.0,]),
nrow(data[data$Smoker == "Non-Smoker" & data$Age >= 65.0,]) / data_summary[data_summary$Smoker == "Non-
```

```
## [1] 193.0000000   0.2636612
```

We can see in the result above that there are much more old people in the non-smoker group than in the other one. The mortality rate in this group may therefore be increased by natural deaths.

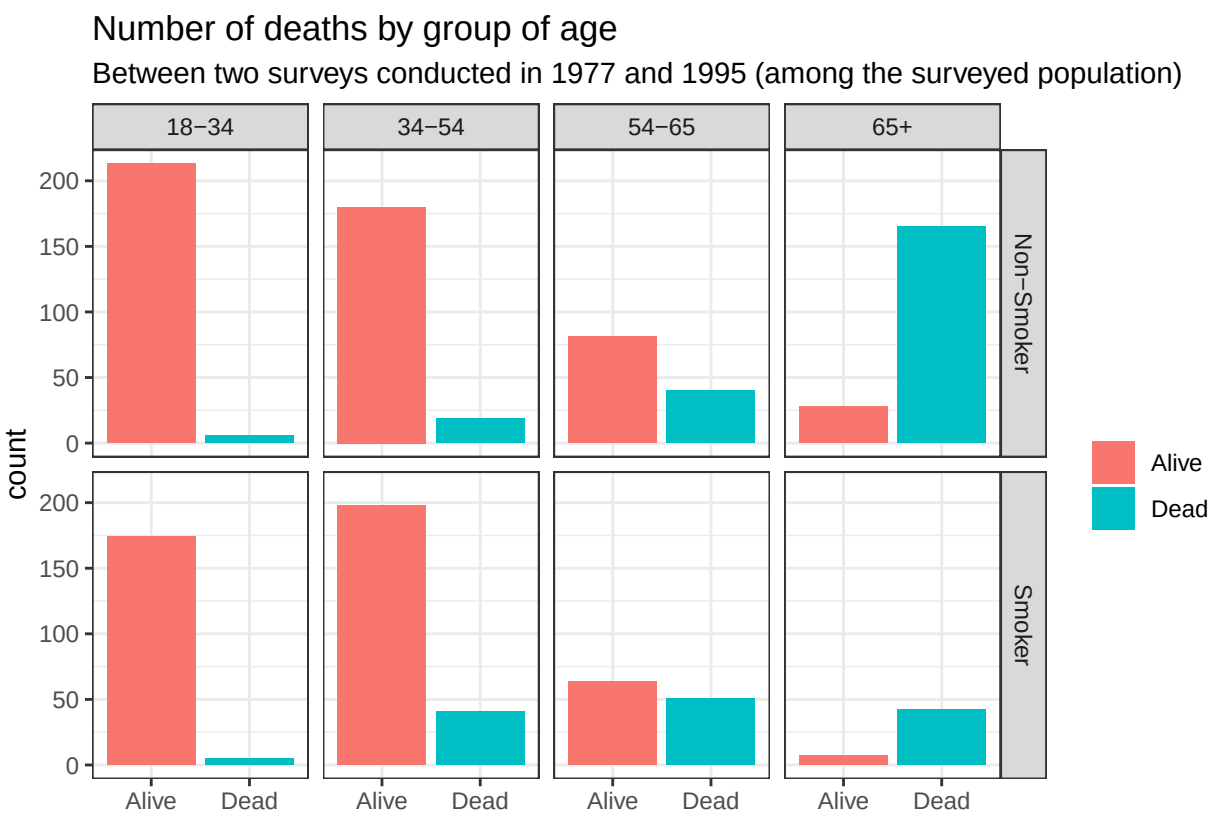
Splitting the groups into age classes

We split both groups into 4 age classes : 18 to 34 years old, 34 to 54, 54 to 65 and above 65.

```
data %>% mutate(age_class = case_when(
  data$Age >= 18.0 & data$Age < 34.0 ~ "18-34",
  data$Age >= 34.0 & data$Age < 54.0 ~ "34-54",
  data$Age >= 54.0 & data$Age < 65.0 ~ "54-65",
  data$Age >= 65.0 ~ "65+",
)) -> data2
```

We plot the data in the following graph:

```
ggplot(data2) +
  facet_grid(Smoker ~ age_class) +
  aes(x = Status, fill = factor(Status)) +
  geom_bar() +
  theme_bw() +
  labs(x = "", fill = "", title = "Number of deaths by group of age", subtitle = "Between two surveys c
```



The following table shows, for each class, the number of people in this class, the mortality rate, the standard deviation and the confidence interval.

```
data2 %>%
  group_by(Smoker, age_class) %>%
  summarise_at(vars(Death), list(n = length, mortality_rate = mean, sd = sd)) %>%
```

```
  mutate(sd = sd * sqrt((n - 1) / n)) %>%
  mutate(CI_lower_bound = mortality_rate - 2 * sd / sqrt(n),
         CI_upper_bound = mortality_rate + 2 * sd / sqrt(n)) -> data2_summary

data2_summary
```

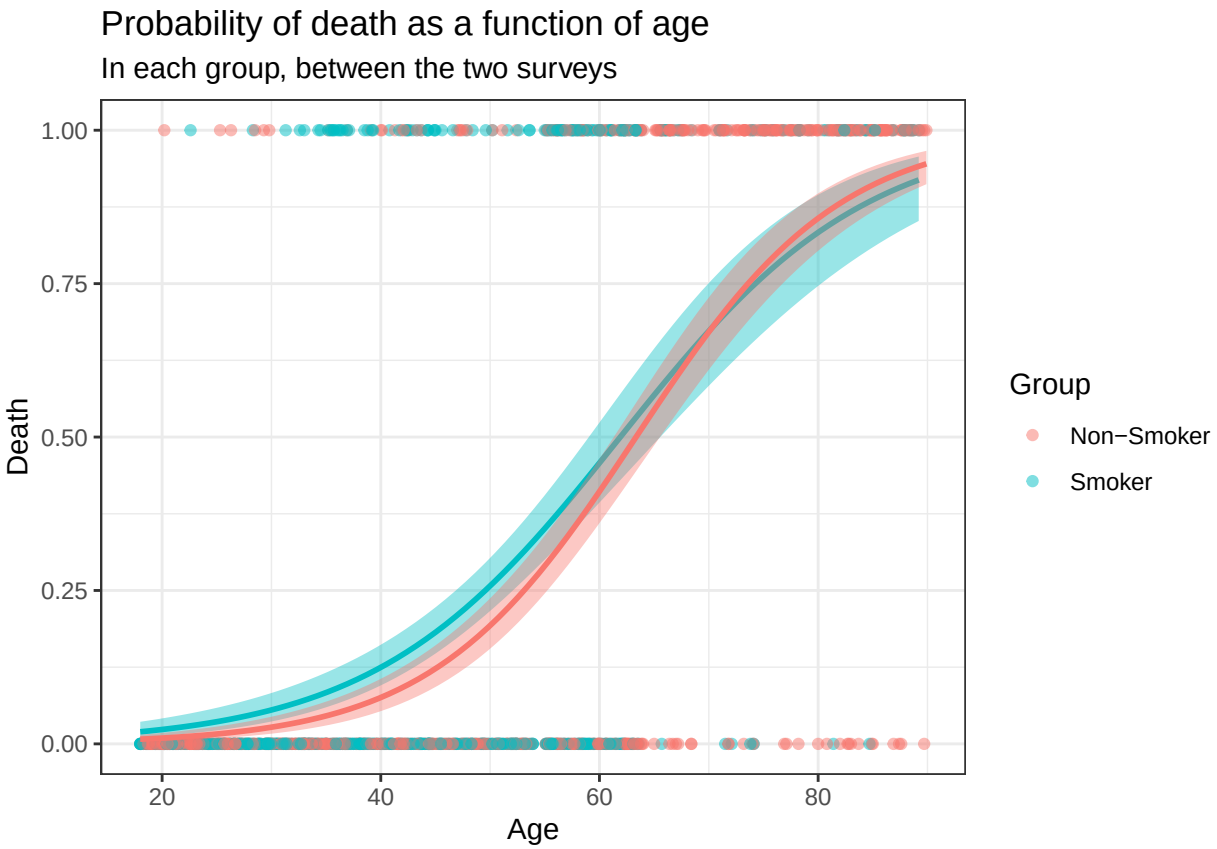
```
## # A tibble: 8 x 7
## # Groups:   Smoker [2]
##   Smoker     age_class     n mortality_rate    sd CI_lower_bound CI_upper_bound
##   <chr>      <chr>     <int>          <dbl> <dbl>          <dbl>          <dbl>
## 1 Non-Smoker 18-34       219         0.0274 0.163        0.00534         0.0495
## 2 Non-Smoker 34-54       199         0.0955 0.294        0.0538          0.137
## 3 Non-Smoker 54-65       121         0.331  0.470        0.245           0.416
## 4 Non-Smoker 65+         193         0.855  0.352        0.804           0.906
## 5 Smoker     18-34       179         0.0279 0.165        0.00330         0.0526
## 6 Smoker     34-54       239         0.172  0.377        0.123           0.220
## 7 Smoker     54-65       115         0.443  0.497        0.351           0.536
## 8 Smoker     65+          49         0.857  0.350        0.757           0.957
```

The mortality rate for the classes 34-54 and 54-65 are much higher for smokers than non-smokers. The other two classes have a similar mortality rate in both groups (although it is still greater by a small amount). However, the confidence intervals do not allow us to conclude here as they all overlap.

We can see the Simpson's Paradox appear here: the conclusion we could make from this plot/table is the opposite of the one we made without the age classes.

Logistic regression

```
ggplot(data, aes(x = Age, y = Death, col=Smoker)) +
  geom_point(alpha=0.5) +
  geom_smooth(data = data[data$Smoker == "Smoker",], method="glm", col="#00BFC4", method.args = list(fa
  geom_smooth(data = data[data$Smoker == "Non-Smoker",], method="glm", col="#F8766D", method.args = lis
  theme_bw() +
  labs(colour = "Group", title = "Probability of death as a function of age", subtitle = "In each group
```



Looking at the curves, we see that the mortality rate in this sample is higher for smokers up to approximately 70 years old and then it starts being the opposite (which is close to what we got in the previous section). However, we still cannot conclude on the harmfulness of smoking as the confidence intervals overlap everywhere.

With more measurements to reduce the confidence interval we could maybe say that the mortality rate is lower for non-smokers under the age of 50 as the confidence interval barely overlaps.

It is likely that the dataset we are using does not contain the measurements we would need to conclude on the kind of things (for example, we could compute the life expectancy of a smoker vs non-smoker if we had the date of death).