Are age, gender, and tobacco exposure confounding variables in the correlation analysis of lung cancer?

1. Introduction

It is known that lung cancer has been mainly attributed to tobacco exposure. However, in East Asia, it's incidence is predominent among women, especially to those who are non-smoking.

Does it mean that the occurrence of lung cancer in East Asia affected by both gender, and tobacco exposure? Or is it just the result of the fact that male tend to smoke more than female?

Confounding variable is the variable that can affect to both the explanatory variable and the response variable. In this case, lung cancer is the response variable and the others are the explanatory variables. The goal of this analysis is whether the explanatory variables in this anlaysis confounding or not.

To analyze it in more details, I added age to the explanatory variable because age can affect to disease occurence and tobacco exposure. So age, gender, and tobacco exposure are the variables in the analysis.

Now, let's download the data to use.

```
library(readxl)
```

```
## Warning:    'readxl'  R   4.1.1
```

```
dat <- read_excel("1-s2.0-S0092867420307431-mmc1.xlsx", sheet = 2)
dat <- as.data.frame(dat)
```

Let's take a look at the data.

```
head(dat)
```

```
##     ID Proteome_Batch Gender      Age Smoking Status Histology Type Stage
## 1 P002          B01-2   Male 73.77687       Nonsmoke            ADC    IB
## 2 P004          B01-4 Female 52.97741       Nonsmoke            SCC    IA
## 3 P005          B02-1   Male 72.75017 Current_Smoker            SCC    IA
## 4 P006          B02-2 Female 46.86105       Nonsmoke            ADC    IB
## 5 P007          B02-3   Male 67.40589       Nonsmoke            ADC   IIA
## 6 P009          B03-1 Female 53.80424       Nonsmoke            ADC   IIA
##   EGFR_Status Primary Tumor Location
## 1      others                   LUL
## 2   exon19del                   RLL
## 3          WT                   LUL
## 4          WT                   RLL
## 5          WT                   RLL
## 6       L858R                   LLL
```

Let's remove the unconsidered columns.

```
library(tidyverse)
```

```
## Warning:    'tidyverse'  R   4.1.1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning:    'ggplot2'  R   4.1.1
```

```
## Warning:    'tidyr'  R   4.1.1
```

```
## Warning:    'readr'  R   4.1.1
```

```
## Warning:    'purrr'  R   4.1.1
```

```
## Warning:    'dplyr'  R   4.1.1
```

```
## Warning:    'forcats'  R   4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dat %>% select("Gender", "Age", "Smoking Status") %>% head()
```

```
##   Gender      Age Smoking Status
## 1   Male 73.77687       Nonsmoke
## 2 Female 52.97741       Nonsmoke
## 3   Male 72.75017 Current_Smoker
## 4 Female 46.86105       Nonsmoke
## 5   Male 67.40589       Nonsmoke
## 6 Female 53.80424       Nonsmoke
```

Let's learn about more details of the columns.

```
dat1 <- dat %>% select("Gender", "Age", "Smoking Status")
dat1 %>% count(Gender)
```

```
##   Gender  n
## 1 Female 60
## 2   Male 43
```

```
dat1 %>% summarize(min_median_max = quantile(Age, c(0,0.5,1)))
```

```
##   min_median_max
## 1       40.22724
## 2       63.48528
## 3       85.86448
```

```
dat1 %>% count(`Smoking Status`)
```

```
##   Smoking Status  n
## 1 Current_Smoker  6
## 2      Ex-smoker 12
## 3       Nonsmoke 85
```

I will combine the current_smoker and ex-smoker to exposed and nonsmoke to not exposed.

```
dat2 <- dat1 %>%
  mutate(`Tobacco Exposure` =
           ifelse(`Smoking Status` == "Nonsmoke", "not exposed", "exposed"))
head(dat2)
```

```
##   Gender      Age Smoking Status Tobacco Exposure
## 1   Male 73.77687       Nonsmoke      not exposed
## 2 Female 52.97741       Nonsmoke      not exposed
## 3   Male 72.75017 Current_Smoker          exposed
## 4 Female 46.86105       Nonsmoke      not exposed
## 5   Male 67.40589       Nonsmoke      not exposed
## 6 Female 53.80424       Nonsmoke      not exposed
```

2. Comparing Two Explanatory Variables

2-1. Age & Gender

Let's look at the distribution of the age by gender.

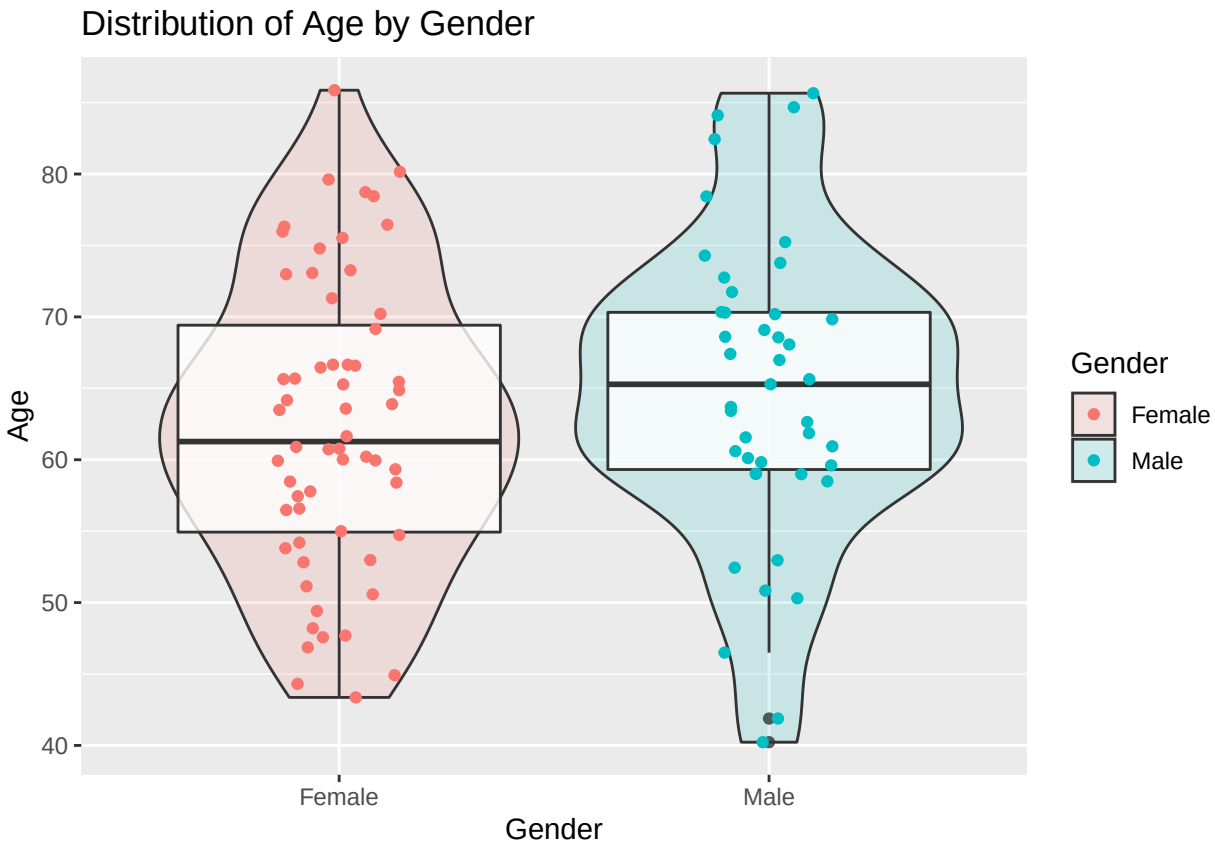
```
dat2 %>%
  group_by(Gender) %>%
  summarize(min_median_max = quantile(Age, c(0,0.5,1)))
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the `.groups` argument.
```

```
## # A tibble: 6 x 2
## # Groups:   Gender [2]
##   Gender min_median_max
##   <chr>           <dbl>
## 1 Female           43.4
## 2 Female           61.3
## 3 Female           85.9
## 4 Male             40.2
## 5 Male             65.3
## 6 Male             85.7
```

It's hard to know the detailed distribution of the age by gender by above summary statistics. Let's visualize the data.

```
dat2 %>%
  ggplot(aes(Gender, Age)) +
  geom_violin(aes(fill = Gender), alpha = 0.15) +
  geom_boxplot(fill = "White", alpha = 0.8) +
  geom_jitter(aes(col = Gender), width = 0.15) +
  ggtitle("Distribution of Age by Gender")
```



We can easily see that the age of male tend to be higher than that of female. There are outliers which affect the minimum value of the age of male.

2-2. Age vs Tobacco Exposure

Let's look at the distribution of the age by tobacco exposure.

```
dat2 %>%
  group_by(`Tobacco Exposure`) %>%
  summarize(min_median_max = quantile(Age, c(0,0.5,1)))
```

```
## `summarise()` has grouped output by 'Tobacco Exposure'. You can override using the `.groups` argumen
```

```
## # A tibble: 6 x 2
## # Groups:   Tobacco Exposure [2]
##   `Tobacco Exposure` min_median_max
##   <chr>                      <dbl>
```

```
## 1 exposed                      52.4
## 2 exposed                      69.5
## 3 exposed                      84.7
## 4 not exposed                  40.2
## 5 not exposed                  61.6
## 6 not exposed                  85.9
```

We can see much easier than age vs gender that those who are exposed to tobacco tend to be older than who are not.

2-3. Gender vs Tobacco Exposure

Let's look at the distribution of the tobacco exposure by gender.

```
dat2 %>%
  group_by(Gender) %>%
  count(`Tobacco Exposure`)
```

```
## # A tibble: 3 x 3
## # Groups:   Gender [2]
##   Gender `Tobacco Exposure`     n
##   <chr>  <chr>              <int>
## 1 Female not exposed           60
## 2 Male   exposed               18
## 3 Male   not exposed           25
```

We can find out that all of the females are not exposed to the tobacco, and about 42% of the males are exposed to tobacco and 58% are not.

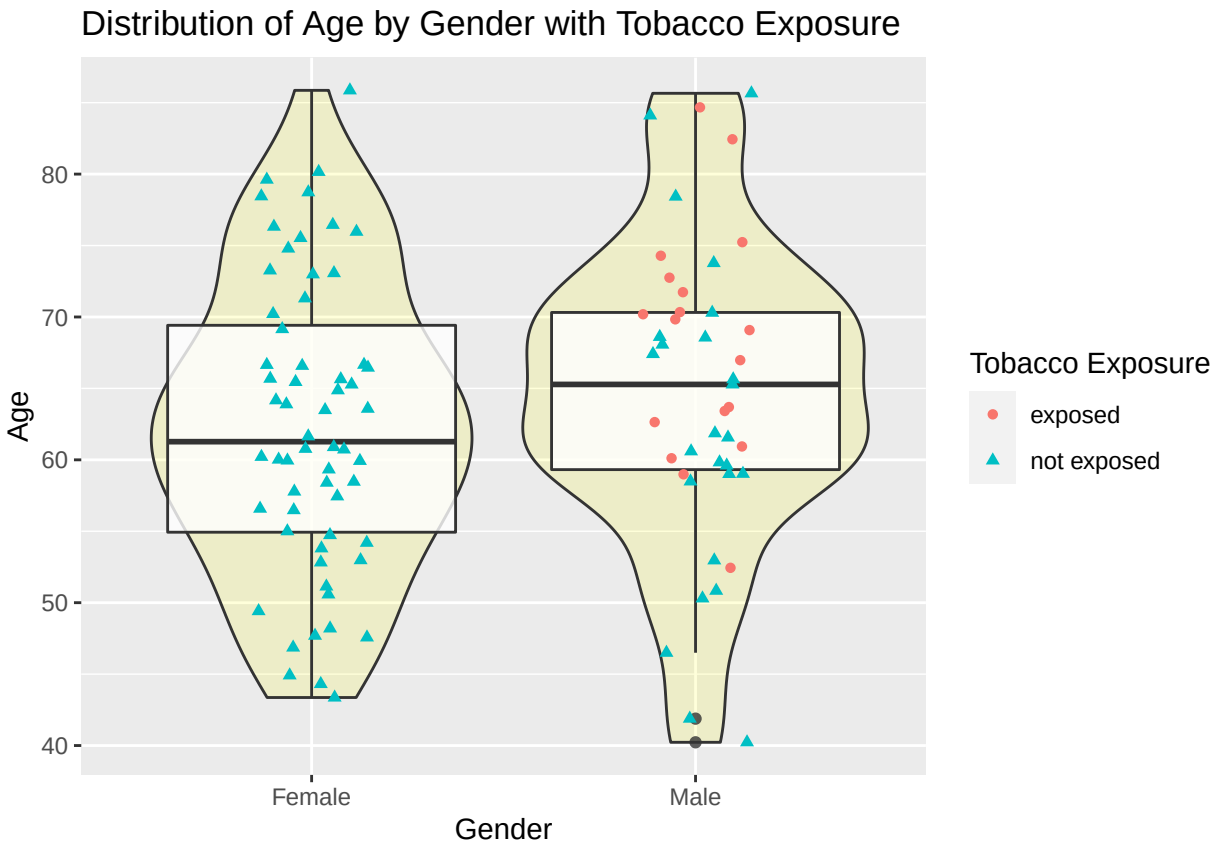
3. Conclusion

As we can look at the above results, we can say that

1. age of the male are older than female
2. age of smokers are older than non-smokers
3. females are much less exposed to tobacco than males

Let's try to visualize all three variables.

```
dat2 %>%
  ggplot(aes(Gender, Age)) +
  geom_violin(fill = "yellow", alpha = 0.15) +
  geom_boxplot(alpha = 0.8) +
  geom_jitter(aes(col = `Tobacco Exposure`, shape = `Tobacco Exposure`), width = 0.15) +
  ggtitle("Distribution of Age by Gender with Tobacco Exposure")
```



Lung cancer in East Asia is known to occur more frequently to female and non-smokers. And the probability of lung cancer gets higher if we live longer just like other diseases. However, we could find out from the result of the analysis that the age of male are higher than that of female, and higher for smokers than that of non-smokers. By this data, we can say that lung cancer in East Asia occurs more to females than males even if they are younger. Also, we can say that lung cancer in East Asia occurs more to non-smokers than smokers even if they are younger. However, females are much likely to be not exposed to tobacco, so we can not easily conclude that gender and tobacco exposure are completely separated variables.

To conclude, we can say that age is not confounding variable to gender and tobacco exposure for the analysis of lung cancer in East Asia. In the other hand, we can say that gender and tobacco exposure are confounding variables to each other.