[Final] Portpolio

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

Yongku Kim, 2016150405, Department of Statistics, Korea University

BSMS222 Biostatistics

2021.12.17

1. Introduction

It is known that lung cancer has been mainly attributed to tobacco exposure. However, in East Asia, it's incidence is predominant 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 analysis confounding or not.

To analyze it in more details, I added age to the explanatory variable because age can affect to disease occurrence 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
```

The data shows us the information of the lung cancer patients. The information in consideration are age, gender, and tobacco exposure. 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
```

Now, we have the data which shows us the information about the age, gender, tobacco exposure of the lung cancer patients.

Let's briefly look at the details of the columns.

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

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

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

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

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

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

Because we need the information about the tobacco exposure, both ex-smokers and current smokers can be considered as those who are tobacco exposed. I will combine the two categories 'Current_smoker' and 'Ex-smoker' to 'Exposed' and 'Nonsmoke' to 'Not_Exposed'.

```
dat <- dat %>%
  mutate("Tobacco Exposure" =
           ifelse(`Smoking Status` == "Nonsmoke", "Not_Exposed", "Exposed")) %>%
  select("Gender","Age","Tobacco Exposure")
head(dat)
```

```
##   Gender      Age Tobacco Exposure
## 1   Male 73.77687      Not_Exposed
## 2 Female 52.97741      Not_Exposed
## 3   Male 72.75017          Exposed
## 4 Female 46.86105      Not_Exposed
## 5   Male 67.40589      Not_Exposed
## 6 Female 53.80424      Not_Exposed
```

2. Comparing Two Explanatory Variables

2-1. Age & Gender

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

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

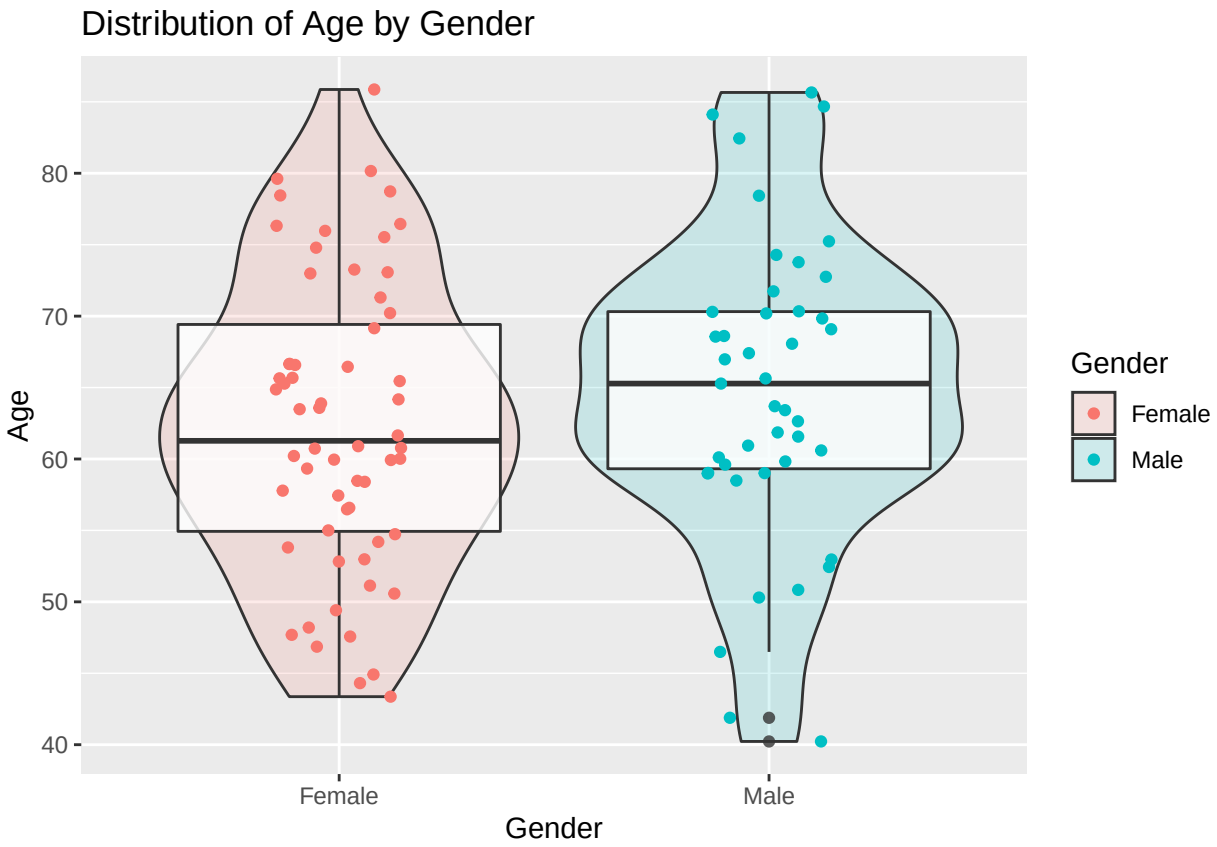
```
## `summarise()` has grouped output by 'Tobacco Exposure'. You can override using the `.groups` argumen
```

```
## # 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
```

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

Let's visualize the data.

```
dat %>%
  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 two outliers which affect the minimum value of the age of male.

Now, let's analyze it in statistical methods. T-test will show us whether the mean of age in male and female statistically different or not. Because gender variable has two groups, F-test for the variance of two groups will be held first.

```
var.test(dat$Age ~ dat$Gender)
```

```
##
##  F test to compare two variances
##
## data:  dat$Age by dat$Gender
## F = 0.93255, num df = 59, denom df = 42, p-value = 0.795
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.522183 1.620654
## sample estimates:
## ratio of variances
##          0.9325541
```

F-test result show us that we can not reject the null hypothesis, which indicates that the two groups male and female have the same variance.

Now, let's progress to t-test for two samples.

```
t.test(dat$Age ~ dat$Gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$Age by dat$Gender
## t = -1.1519, df = 88.697, p-value = 0.2525
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -6.548494  1.742456
## sample estimates:
## mean in group Female   mean in group Male
##             62.44075             64.84377
```

T-test result shows us that we can not reject the null hypothesis, which indicates that there is no significance to conclude the true difference in means between male and female. In other words, means of male and female are the same at 95% significance level.

The result is quite different by only looking at the plot and analyzing it statistically. This means that the difference shown in the plot may be coincidence or not significant.

2-2. Age vs Tobacco Exposure

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

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

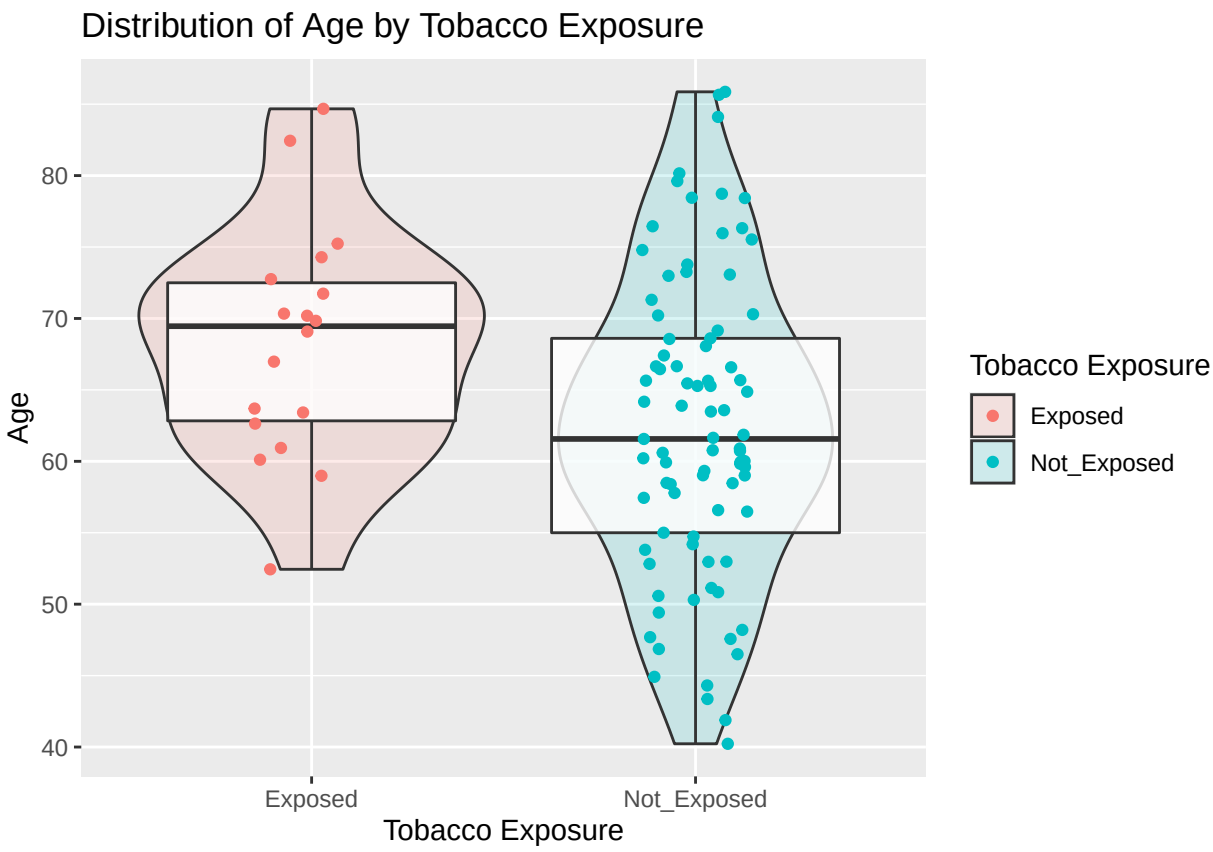
```
## `summarise()` has grouped output by 'Tobacco Exposure'. You can override using the `.groups` argumen
```

```
## # 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
```

It seems that the age tends to be higher for 'Exposed', but it is hard to know the detailed distribution of the age by tobacco exposure by above summary statistics.

Let's visualize the data.

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



It seems quite clear that age of 'Exposed' is higher than 'Not_Exposed'.

Let's analyze it statistically for detailed analysis.

```
var.test(dat$Age ~ dat$`Tobacco Exposure`)
```

```
##
##  F test to compare two variances
##
## data:  dat$Age by dat$`Tobacco Exposure`
## F = 0.59683, num df = 17, denom df = 84, p-value = 0.229
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3076565 1.3986355
## sample estimates:
## ratio of variances
##          0.5968304
```

F-test result show us that we can not reject the null hypothesis, which indicates that the two groups 'Exposed' and 'Not_Exposed' have the same variance.

Now, let's progress to t-test for two samples.

```
t.test(dat$Age ~ dat$`Tobacco Exposure`)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$Age by dat$`Tobacco Exposure`
## t = 2.6387, df = 30.429, p-value = 0.013
## alternative hypothesis: true difference in means between group Exposed and group Not_Exposed is not
## 95 percent confidence interval:
##   1.338376 10.479670
## sample estimates:
##     mean in group Exposed mean in group Not_Exposed
##                  68.32033                  62.41131
```

T-test result shows us that we can reject null hypothesis, which indicates that the means between two groups are significantly different at 95% significance level.

We can conclude that those who are exposed to tobacco tend to have higher age.

2-3. Gender vs Tobacco Exposure

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

```
dat %>%
  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.

Let's analyze it statistically.

```
x <- c(60,25)
n <- c(60,43)
prop.test(n=n, x=x, alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 27.602, df = 1, p-value = 7.453e-08
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2748975 1.0000000
## sample estimates:
##    prop 1    prop 2
## 1.0000000 0.5813953
```

Proportion test result shows us that we can reject the null hypothesis, which indicates that the proportion of female non-smokers is higher than male at 95% significance level.

We can conclude that male is more likely to smoke than female.

3. Conclusion

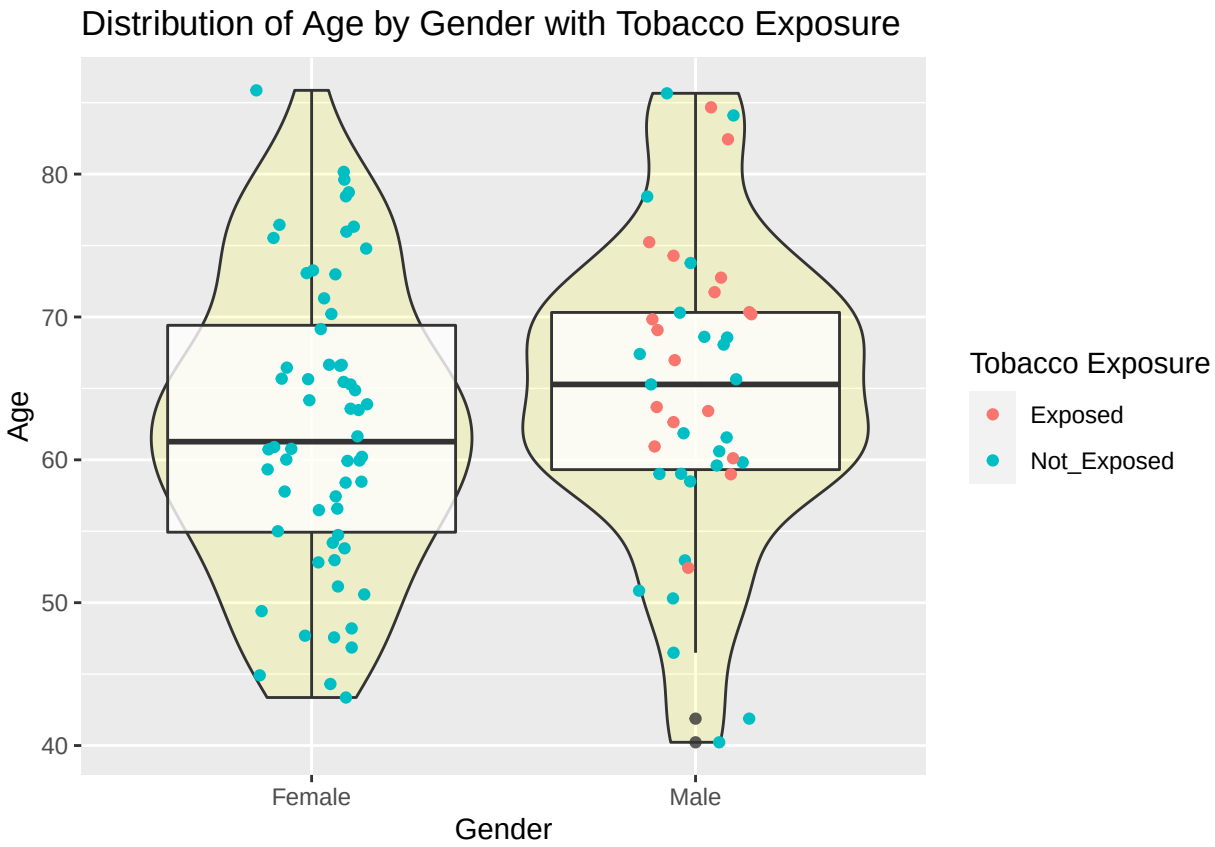
As we can look at the above results, we can say that

in the data analyzed,

1. age of the male and female are same
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.

```
dat %>%
  ggplot(aes(Gender, Age)) +
  geom_violin(fill = "yellow", alpha = 0.15) +
  geom_boxplot(alpha = 0.8) +
  geom_jitter(aes(col = `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.

The result shows us that age and gender have no significant correlation and as a result they are not confounding variables to each other.

Also, higher age may get the disease easier than lower age, but non-smokers are likely to have lower age but tend to have lung cancer more frequently. We can say that lung cancer in East Asia occurs more to non-smokers than smokers even if they are younger. As a result, they are confounding variables to each other, but the result of the interaction is not different from the previous result and it is a significant point of the analysis.

However, we could find out that female are more likely to smoke less than male, so two variables can not be easily treated apart. We can not conclude that gender and tobacco exposure both affect probability of lung cancer at the same strength. As a result, gender and tobacco exposure are confounding variables to each other and it has to be considered carefully to say that both variables have significant meaning for lung cancer.