Code

Confounded effects of drinking coffee on Deaths per thousands from Cancer

Author: Olle Gottfrid Holmberg

This data exploration on the confounded effects from drinking coffee on deaths per thousands from coronary heart disease (CHD) causing death is a part of a course on Data Analysis and Visualization in R at the Technical University of Munich (TUM) given by the Gagneur lab.

The following data exploration will distringuish the difference between correlation and causality between risk of drinking coffee on CDH.

We will see that consumption of coffee is strongly correlated with risk of dying from CDH. But, when taking the individuals consumption of sigarettes in consideration, it becomes clear that coffee is not the casual factor of daying in cancer, but smoking is.

set up

Hide

```
#genetic and health data Analysis - set up
library(ggplot2)
library(data.table)
library(magrittr) # Needed for %>% operator
library(tidyr)
library(plotly)
library(GGally)
library(gridExtra)
#insert your own data directory
DATA_DIR <- '/home/olle/Documents/University Courses/Data analysis and vizulisation
/Data visualization in R/exercises&lectures/data'
```

first look at the data

Hide

```
coffee_file <- file.path(DATA_DIR, 'coffee.csv')
coffee_dt <- fread(coffee_file, fill = T)
#disply dimension and first five rows
dim(coffee_dt)
```

```
[1] 9 5
```

Hide

```
head(coffee_dt,15)
```

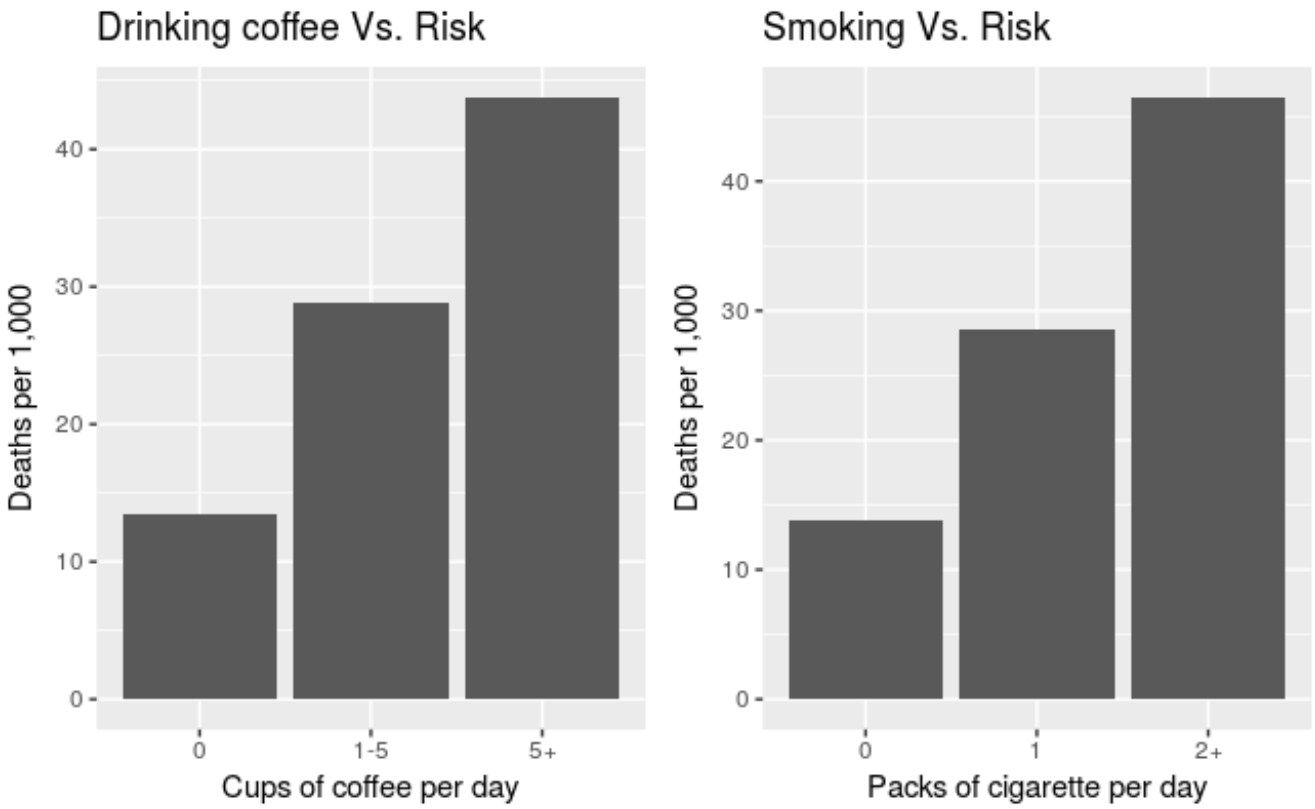
cups <chr>
0
1-5
5+
0
1-5
5+
0
1-5
5+

9 rows | 1-1 of 5 columns

correlation between drinking coffe, smoking sigarettes and risk of dying from cancer

Hide

```
# this is the confounded association of coffee
plot_1 <-ggplot(coffee_dt, aes(cups, coffee_risk_margin)) +  geom_bar(stat='identit
y') +
  labs(x = "Cups of coffee per day",y = "Deaths per 1,000", title= "Drinking coffe
e Vs. Risk")
# this is the true one from smoking
plot_2 <- ggplot(dt, aes(packs, cig_risk_margin)) +
  geom_bar(stat='identity') +
  labs(x = "Packs of cigarette per day",
  y = "Deaths per 1,000", title= "Smoking Vs. Risk")
grid.arrange(plot_1, plot_2, ncol=2)
```

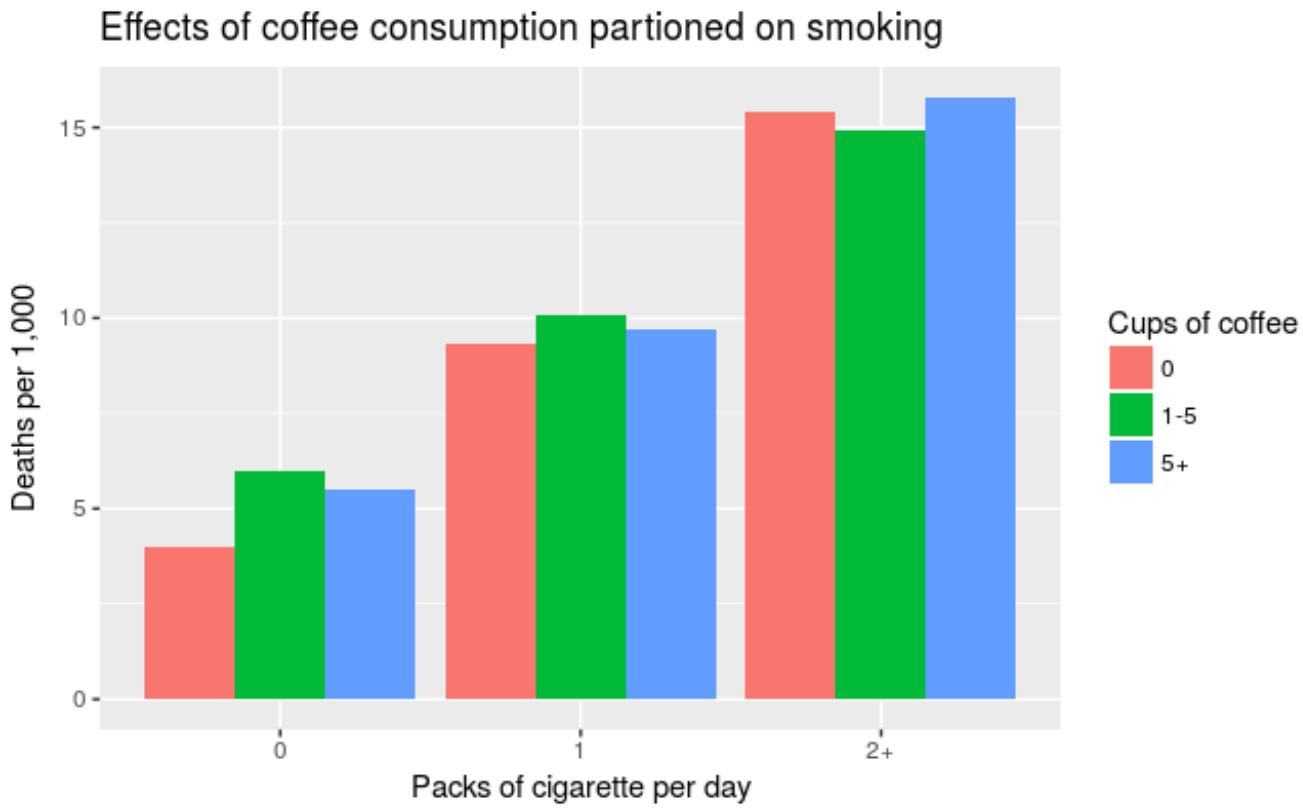


We clearly see that the number of cups of coffee and consumption of sigarettes per day is higly corrrelated with the risk of dyin from CDH. From this graph only, one could inferr that drinking coffee and consumption of sigarettes are equally “bad” for CDH. As we will see in the below graph, this is not the case.

Partitioning on Packs of sigarettes per day

Hide

```
# coffee effects are always the same within each smoking group
ggplot(dt, aes(packs, risk, fill = cups)) +
  geom_bar(stat='identity', position='dodge') +
  labs(x = "Packs of cigarette per day",
  y = "Deaths per 1,000", title = "Effects of coffee consumption partioned on smoki
ng") +
  guides(fill = guide_legend(title = "Cups of coffee"))
```



Here one sees that consumption of coffee does not yield an increased risk for CDH. Individuals that consume more than 5 cups of coffee per day still have very low Death per year from CDH as long as they dont smoke. The difference on Deaths per year does not seem to be affected at all by the individuals coffee consumption, but solely by Packs of sigarettes smoked per day.

This dta exploration of the coffee data set using ggplot2 in R shows the importance of clearly understanding confounded effects in data. Differing between correlation and causality in statistics is, as this short analysis shows, crucial to understand a given problem.