midterm

jason grahn

10/26/2018

# Problem 1:

## N. 1

Download smoking.csv (or smoking.txt) and read corresponding data into R. Example command in R: MyData <- read.csv(file="path/TheDataIWantToReadIn.csv",header=TRUE, sep=",")Make sure to include the code/command.

library(readr)  
smoking <- read\_csv("smoking.csv")

## Parsed with column specification:  
## cols(  
## gender = col\_character(),  
## age = col\_integer(),  
## maritalStatus = col\_character(),  
## highestQualification = col\_character(),  
## nationality = col\_character(),  
## ethnicity = col\_character(),  
## grossIncome = col\_character(),  
## region = col\_character(),  
## smoke = col\_character(),  
## amtWeekends = col\_integer(),  
## amtWeekdays = col\_integer(),  
## type = col\_character()  
## )

#View(smoking)

## N. 2

How many observations are there in this data set? How many variables, and what are they?

#how many observations and variables are shown with glimpse()  
glimpse(smoking)

## Observations: 1,691  
## Variables: 12  
## $ gender <chr> "Male", "Female", "Male", "Female", "Fema...  
## $ age <int> 38, 42, 40, 40, 39, 37, 53, 44, 40, 41, 7...  
## $ maritalStatus <chr> "Divorced", "Single", "Married", "Married...  
## $ highestQualification <chr> "No Qualification", "No Qualification", "...  
## $ nationality <chr> "British", "British", "English", "English...  
## $ ethnicity <chr> "White", "White", "White", "White", "Whit...  
## $ grossIncome <chr> "2,600 to 5,200", "Under 2,600", "28,600 ...  
## $ region <chr> "The North", "The North", "The North", "T...  
## $ smoke <chr> "No", "Yes", "No", "No", "No", "No", "Yes...  
## $ amtWeekends <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 15, NA,...  
## $ amtWeekdays <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 12, NA,...  
## $ type <chr> NA, "Packets", NA, NA, NA, NA, "Packets",...

The glimpse() function shows us both the number of variables and rows. The number of rows in the data is 1691, and the number of variables is 12. They are:

\* gender  
\* age  
\* maritalStatus  
\* highestQualification  
\* nationality  
\* ethnicity  
\* grossIncome  
\* region  
\* smoke  
\* amtWeekends  
\* amtWeekdays  
\* type

#What is the 300th observation of nationality?  
smoking$nationality[[300]]

## [1] "British"

## N. 3

Create a numerical summary for age and compute the interquartile range.

summary(smoking$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 34.00 48.00 49.84 65.50 97.00

age.iqr <- describe(smoking$age, IQR = TRUE) %>%   
 select(IQR)

The numerical summary for age is shown above. The IQR is 31.5.

Compute the relative frequency distribution for gender.

table(smoking$gender)/nrow(smoking)

##   
## Female Male   
## 0.5706682 0.4293318

How many males are in the sample? Include both the code/command and the output/graph.

male.count <- nrow(  
 smoking %>%   
 filter(gender == "Male")  
)

There are 726 males in the sample.

## N. 4

Using numerical summaries and a side-by-side box plot, determine if male smokers are as old as female smokers. Include both the code/command and the output/graph.

s.g <- smoking %>%   
 filter(smoke == "Yes") %>%   
 select(age, gender)  
  
#group the male data together  
s.g.m <- s.g %>% filter(gender == "Male")  
  
#group the female data together  
s.g.f <- s.g %>% filter(gender == "Female")

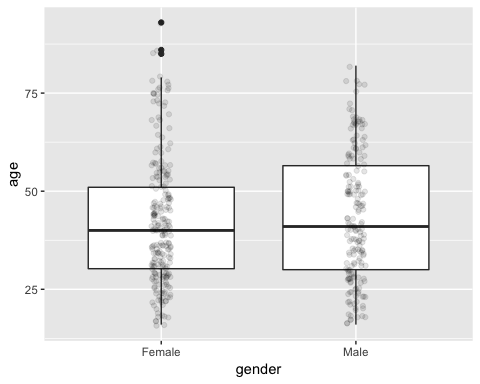
#age summary for male smokers  
summary(s.g.m$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 30.00 41.00 43.35 56.50 82.00

#age summary for female smokers  
summary(s.g.f$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 30.25 40.00 42.21 51.00 93.00

s.g %>%   
 ggplot(aes(x = gender, y = age)) +   
 geom\_boxplot() +  
 geom\_jitter(alpha = 0.1, width = 0.05)



## N. 5

Create a bar chart or frequency table for maritalStatus, what is the proportion for Divorced, Single, Married, and Widowed, respectively? What can you interpret from these numbers? Include both the code/command and the output/graph.

table(smoking$maritalStatus)/nrow(smoking)

##   
## Divorced Married Separated Single Widowed   
## 0.09520993 0.48018924 0.04021289 0.25251331 0.13187463

We can interpret from these numbers that most people in the smokers data frame are married. Single is next, then widowed.

# Problem 2

# N. 1

Suppose we’re flipping an unfair coin that we know only lands heads 30% of the time. Please simulate this flip 10 times, what is the proportion of heads?

If you simulate this flip 100 times, what is the proportion of heads now? Include both the code/command and the output/graph.