Homework 2

Lacey Gleason

2/21/2018

### Installing the car package

* install.packages(“quantreg”, dependencies=TRUE)
* install.packages(“car”, dependencies=TRUE)

### The Data - Davis dataset in the car package

The Davis dataset in the car package contains data on the measured and reported heights and weights of men and women engagedin regular exercise.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(car)

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

dataDavis <- car::Davis

Use tools within the dplyr package as much as possible to answer the following questions.

### **Question 1:** What kind of R object is the Davis dataset?

class(dataDavis)

## [1] "data.frame"

str(dataDavis)

## 'data.frame': 200 obs. of 5 variables:  
## $ sex : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...  
## $ weight: int 77 58 53 68 59 76 76 69 71 65 ...  
## $ height: int 182 161 161 177 157 170 167 186 178 171 ...  
## $ repwt : int 77 51 54 70 59 76 77 73 71 64 ...  
## $ repht : int 180 159 158 175 155 165 165 180 175 170 ...

The Davis dataset is a data.frame with 200 observations and 5 variables. The variables in the Davis dataset are sex, measured weight in kg, measured height in cm, reported weight in kg, and reported height in cm.

### **Question 2:** How many observations are in the Davis dataset?

dataDavis %>%  
 summarise(nObs = n())

## nObs  
## 1 200

There are two hundred observations in the Davis dataset.

### **Question 3:** For reported weight, how many observations have a missing value?

missingWeight <- is.na(dataDavis$repwt)  
sum(missingWeight)

## [1] 17

mean(missingWeight)

## [1] 0.085

For reported weight, 17 observations (8.5%) have a missing value.

### **Question 4:** How many observations have no missing values? *(HINT: find complete cases)*

completeObs <- complete.cases(dataDavis)  
table(completeObs)

## completeObs  
## FALSE TRUE   
## 19 181

One hundred and eighty-one observations have no missing values.

Create a subset containing only females.

dataFemale <- dataDavis %>%  
 filter(sex == "F")

### **Question 5:** How many females are in this subset?

dataFemale %>%  
 summarise(nObs = n())

## nObs  
## 1 112

There are 112 females in the subset containing only females.

*Now* return to the overall dataset with both males and females.

Body mass index is one way to quantify the amount of tissue mass (muscle, fat, and bone) in an individual, then categorize that person as *underweight*, *normal* *weight*, *overweight*, or *obese* according to that value.

We calculate the BMI as the **ratio of the weight in kilograms divided by the square of the height in meters**, and the categorization based on BMI is as follows:

#### BMI Categories

|  |  |
| --- | --- |
| Category | BMI range (kg/m2) |
| Underweight | <18.5 |
| Normal | 18.5 to <25 |
| Overweight | 25 to <30 |
| Obese | 30 or higher |

Create the BMI variable and then a variable to depict BMI category. Note that the height variable is in centimeters, and weight is in kg. You need to create the BMI variable using the correct formula.

dataDavis <- dataDavis %>%  
 mutate(BMI = (weight/((height/100)^2)))

dataDavis <- dataDavis %>%  
 mutate(BMIcat = if\_else(BMI<18.5, "Underweight",  
 if\_else(BMI<25, "Normal",   
 if\_else(BMI<30, "Overweight",   
 "Obese", "Missing"),"Missing"),"Missing"))

Now answer these questions:

### **Question 6:** What is the average BMI for these individuals?

dataDavis %>%  
 summarise(meanBMI = mean(BMI),  
 sdBMI = sd(BMI))

## meanBMI sdBMI  
## 1 24.70096 34.68423

The mean BMI is 24.7 kilograms per meter squared.

### **Question 7:** How do these individuals fall into the BMI categories (what are the frequencies and relative %’s)?

dataDavis %>%   
 count(BMIcat)

## # A tibble: 4 x 2  
## BMIcat n  
## <chr> <int>  
## 1 Normal 143  
## 2 Obese 4  
## 3 Overweight 35  
## 4 Underweight 18

library(janitor)  
library(knitr)  
dataDavis %>%  
 janitor::tabyl(BMIcat) %>%  
 knitr::kable()

|  |  |  |
| --- | --- | --- |
| BMIcat | n | percent |
| Normal | 143 | 0.715 |
| Obese | 4 | 0.020 |
| Overweight | 35 | 0.175 |
| Underweight | 18 | 0.090 |

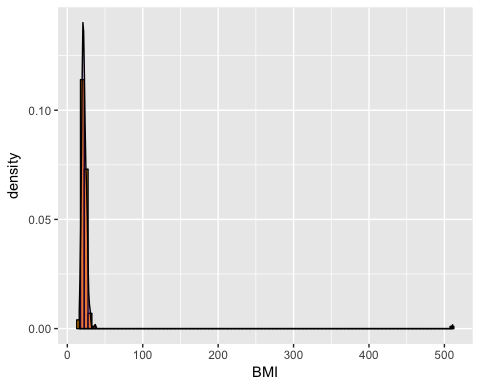
There are 18 (9.0%) individuals in the underweight category. There are 143 (71.5%) individuals in the normal category. There are 35 (17.5%) individuals in the overweight category. There are four (2.0%) individuals in the obese category.

## Test your graphing skills using ggplot2

Using the Davis dataset from the car package, create the following graphics/figures using ggplot() and associated geom\_xxx() functions.

### **Question 8:** Create a histogram of BMI.

library(ggplot2)  
  
 ggplot(dataDavis, aes(x=BMI)) +   
 geom\_histogram(aes(y=..density..),  
 binwidth=5,  
 colour="black", fill="orange") +  
 geom\_density(alpha=.2, fill="purple")

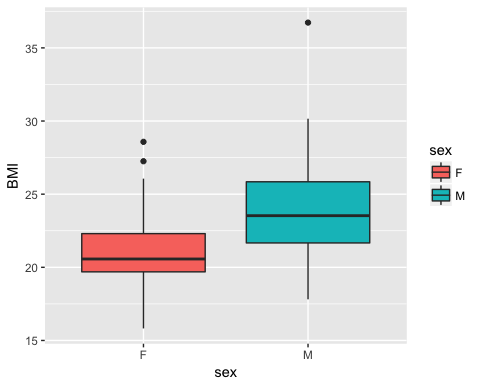


*What do you notice about the distribution (any outliers or skewness)?* Yes, there is a very high outlier that is above 500.

# Remove high outlier in BMI - it would not be possible to have this value  
dataNoOut<-dataDavis[!(dataDavis$BMI>204),]

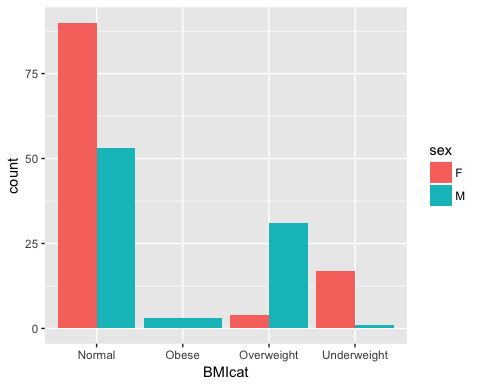
### **Question 9:** Create side-by-side boxplots of the BMI distributions by gender

# A basic box with the conditions colored  
ggplot(dataNoOut,   
 aes(x=sex, y=BMI, fill=sex)) +   
 geom\_boxplot()



### **Question 10:** Create a clustered bar chart of the BMI categories by gender

# cluster barchart of BMI categories by gender  
dataNoOut %>% ggplot(aes(x=BMIcat, fill=sex)) +  
 geom\_bar(position = "dodge")

 *(note: the y-axis should be counts)*

## Link for assignment on Github

<https://github.com/lpgleason/N741Homework2.git>