assignment2

2022-09-14

Question1

You will need the file Human_data.txt. Place this file on your working environment.

Read the file Human_data.txt and store this in an object called human. Before reading the data, check whether the file has a header. If it does, use the appropriate argument in the read function to include the header. Look at the structure of human using the function str.

```
library(psych)
path = "Human_data.txt"
human <- read.table(path,header = TRUE)
str(human)
```

```
## 'data.frame':    500 obs. of  10 variables:
##  $ Index         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender        : chr  "M" "F" "M" "F" ...
##  $ age           : int  22 33 46 24 37 31 38 38 21 31 ...
##  $ Ocupation     : chr  "Nothing" "Nothing" "Work" "student" ...
##  $ Head_size     : num  34.4 28 27 24.8 30.1 26.6 25.6 25.6 27.6 23.6 ...
##  $ Height_cm     : num  206 163 162 156 173 ...
##  $ Weight_kg     : num  105.3 71.3 94.7 56 103.3 ...
##  $ Salary        : int  0 0 19268 2034 14829 10586 11272 13048 2068 12326 ...
##  $ blood_type    : int  4 4 4 3 2 3 4 2 1 3 ...
##  $ Sugar_in_blood: num  95.2 83.5 92.7 95.8 114.1 ...
```

(b)The body mass index (BMI) is defined as a person's weight in kilograms divided by the square of height in meters. Add a column named bmi to the data frame with the value of this index for each subject. Count how many subjects have BMI above 30.

```
human <- within(human, bmi <- Weight_kg/ (Height_cm/100)^2 )
sum(human$bmi > 30,na.rm = TRUE)
```

```
## [1] 108
```

c. Calculate mean and standard deviation for bmi according to Gender. Compare these results and comment. Plot bmi against age, color the dots by Gender, and comment.

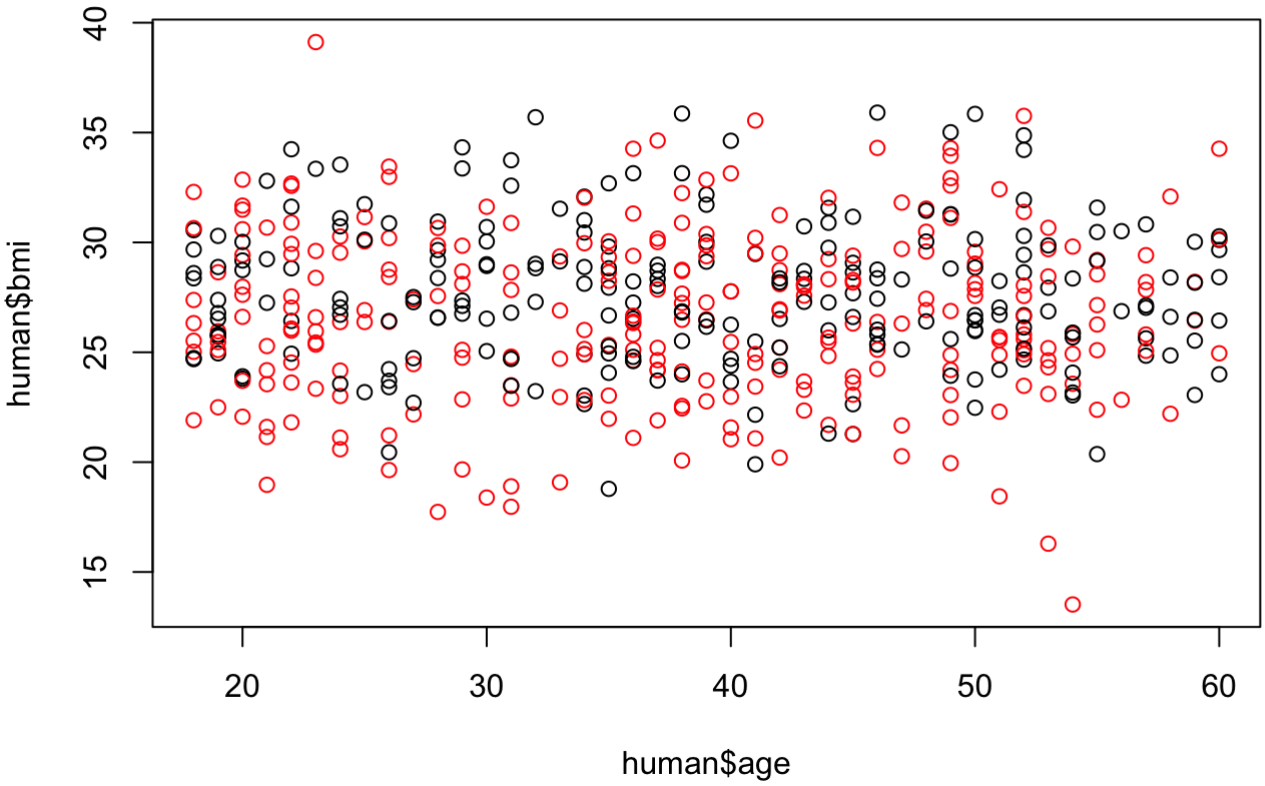
```
aggregate(human$bmi,list(human$Gender),FUN=mean)
```

```
##   Group.1        x
## 1       F 26.60612
## 2       M 27.68282
```

```
aggregate(human$bmi,list(human$Gender),FUN=sd)
```

```
##   Group.1        x
## 1       F 3.888583
## 2       M 3.253710
```

```
plot(human$bmi ~ human$age,col= ifelse(human$Gender=="F", "red", "black"))
```



The average BMI of the male is 1 greater than that of the female, and the standard deviation of males was smaller than that of females. It can be seen that many females have low BMI.

d. Using subset, create a new data frame from human with the variables Head_size, Height_cm, Weight_kg for subjects with age between 30 and 50 (both inclusive) and head size bigger than 26. Call this new data frame human1.

```
human1 = subset(human, human$age <= 50 &  human$age>= 30 &  human$Head_size > 26,sele
ct=Head_size:Weight_kg)
```

e. Use the function apply twice to calculate the mean and standard deviation for each of the three variables in human1. Call the vectors you obtain human.mean and human.sd.

```
human.mean = apply(human1,2,mean)
human.sd = apply(human1,2,sd)
```

f. Use the function sweep twice, first to subtract the mean for each variable to the values in human1 and then to divide by the standard deviation. Store the result in a data frame named human.std.

```
tmp = sweep(human1, 2, human.mean, FUN = "-")
human.std = sweep(tmp,2,human.sd,FUN = "/")
```

g. The previous procedure is known as standardization. The resulting columns in the human.std should now have mean zero and variance equal to one. Verify this using apply.

```
std.mean = apply(human.std,2,mean)
std.sd = apply(human.std,2,sd)
print(std.mean)
```

```
##    Head_size    Height_cm    Weight_kg
## 6.606647e-16 6.125368e-16 3.787337e-16
```

```
print(std.sd)
```

```
## Head_size Height_cm Weight_kg
##         1         1         1
```

Question 2

For this question you will use again the file human that you created in the first question. (a) Use the function split on the file human with second argument Gender and store the result in an object called human2. Describe this object.

```
human2 = split(human, human$Gender)
```

human2 is a list which has two list. First list stores Male information, second list stores female's information.

b. Using the data in human2 obtain a numerical summary (summary) for the variable Salary for males and females and compare.

```
summary(human2$F$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    2063   10660    8849   13364   44232
```

```
summary(human2$M$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    2152   11178    9836   14404   27354
```

Median, Mean and 3rd quantile of male salary are higher than female.

c. Use again the function split on human but now you want to use two variables for splitting the data, Gender and Work. Look at the help for this function to find out how to do this. Call the resulting object human3. Describe the file human3.

```
human3 = split(human,list(human$Gender,human$Ocupation))
```

human3 is a list have 6 list, first is female without work , second is male without work, third is female student, fourth is male student, fifth is female worker, sixth is male worker.

d. Using the data in human3 obtain numerical summaries for the variable Salary for males and females that work and compare.

```
summary(human3$F.Work$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10011   11130   12772   13960   15696   44232
```

```
summary(human3$M.Work$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10012   11245   12904   14386   16163   27354
```

Male workers have more Mean/Median salaries than female workers. However, the difference is smaller for all males compared with females.

e. The function cut divides the range of values of a continuous variable into intervals and creates a factor according to which interval the values fall. You have to use this function to divide the range of salaries in the file human into three intervals, according to the following scheme: below 8000 is low, between 8000 and 18000 is medium, and more than 18000 is high. Call the resulting factor sal. Use the function table to count how many subjects fall in each category.

```
sal = cut(human$Salary,breaks = c(-Inf,8000,18000,Inf),labels = c("low","medium","hig
h"))
table(sal)
```

```
## sal
##    low medium   high
##    185    272     43
```

f. Using the factor sal and the variable Gender, split the file human and call the resulting file human4. Using this file, obtain numerical summaries for the variable Salary for males and females that have a high salary and compare.

```
human4 = split(human,list(sal,human$Gender))
summary(human4$high.F$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18025   18678   20351   22980   21671   44232
```

```
summary(human4$high.M$Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18905   19930   21776   22272   24252   27354
```

we can find that in high salary people, females have higher mean salaries. This conclusion is different from the previous comparison.