Exercise day 1 with solutions

Introduction to R for Basic Statistics

Exercise A: Reshaping the data

For this exercise we will work with a subset of "follicle" data, collected from patients with cancer that had OTC (ovarian tissue cryopreservation). Follicles were cultured for 8 days and the diameter was collected every 2 days. The aim of the study was to compare the follicles growth among different treatment groups.

Information are collected in two data sets.

Data set *follicle* with:

- Patient: patient ID
- Number: follicle ID
- Day_x: follicle diameter at Day x

Data set *patient* with:

- Patient ID
- Treatment group
- Type of Disease
- Age at Day 0

Question 1 Download both data sets and load them into R.

```
db_follicle<-read.csv("~/Desktop/KVN2021/Course/IntrotoR/data_exercise/follicle.csv")
db_pat<-read.csv("~/Desktop/KVN2021/Course/IntrotoR/data_exercise/patient.csv")
```

Consider the data set *patient*.

1. How many Patients?

To check how many patients ar eincluded in the study we can check the dimension of the data frame where each line is referred to one patient.

```
dim(db_pat)
```

```
## [1] 14  4
```

We have 14 patients in total

2. Visualize the first lines of the data and print a summary of the data. What is the data type of each variable?

We can visualize the first 6 lines of the data using the function head()

```
head(db_pat)
```

```
##   Patient              Disease Treatment  Age
## 1       1        Breast_cancer       FBS 31.6
## 2       2        Breast_cancer       FBS 29.4
## 3       3  Neurological_cancer       FBS 19.3
## 4       4 Rheumatoid_arthritis       FBS 20.3
## 5       5           Mb_Hodgkin       hPL 29.8
```

```
## 6       6              Sarcoma       HSA 19.7
```

```
summary(db_pat)
```

```
##     Patient        Disease           Treatment             Age
##  Min.   : 1.00   Length:14          Length:14          Min.   :19.30
##  1st Qu.: 4.25   Class :character   Class :character   1st Qu.:29.50
##  Median : 7.50   Mode  :character   Mode  :character   Median :31.40
##  Mean   : 7.50                                         Mean   :29.74
##  3rd Qu.:10.75                                         3rd Qu.:32.25
##  Max.   :14.00                                         Max.   :37.10
```

Patient and Age are quantitative and are either numerical or integer, while Disease and Treatment are categorical variables and are listed as characters.

Question 2

Treatment and Disease are *characters*, can we understand from the summary how many different diseases are in the data?

1. Would it be better if Treatment and Disease were of a different type? If yes, which one?

For easier representation it is better to have factor for categorical variables, thus we can transform Disease and Treatment into factors.

2. Transform them into factor. You can use the function *factor()* in **R**.

```
db_pat$Treatment<-factor(db_pat$Treatment)
db_pat$Disease<-factor(db_pat$Disease)
```

3. Print the summary of the data, can you see any difference?

```
summary(db_pat)
```

```
##     Patient                          Disease  Treatment      Age
##  Min.   : 1.00   Brain_cancer            :1   FBS:11    Min.   :19.30
##  1st Qu.: 4.25   Breast_cancer           :8   hPL: 1    1st Qu.:29.50
##  Median : 7.50   Chronic_myeloid_leukemia:1   HSA: 2    Median :31.40
##  Mean   : 7.50   Mb_Hodgkin              :1             Mean   :29.74
##  3rd Qu.:10.75   Neurological_cancer     :1             3rd Qu.:32.25
##  Max.   :14.00   Rheumatoid_arthritis    :1             Max.   :37.10
##                  Sarcoma                 :1
```

The summary now shows the different levels of Treatment and Disease with the number of observations in that category.

Question 3

Age is a continuous covariate:

1. Show min,max and mean for age.

To check their min,max and mean value we could use the summary function, otherwise we use min,max and mean separetely

```
summary(db_pat$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.30   29.50   31.40   29.74   32.25   37.10
```

```
min(db_pat$Age)
```

```
## [1] 19.3
```

```
max(db_pat$Age)
```

```
## [1] 37.1
```

```
mean(db_pat$Age)
```

```
## [1] 29.74286
```

2. Calculate the mean of Age for each disease group (you can use **aggregate** or **tapply**)

When using *aggregate()*

```
aggregate(Age~Disease, db_pat, mean)
```

```
##                    Disease     Age
## 1             Brain_cancer 32.1000
## 2            Breast_cancer 32.9125
## 3 Chronic_myeloid_leukemia 31.9000
## 4               Mb_Hodgkin 29.8000
## 5      Neurological_cancer 19.3000
## 6     Rheumatoid_arthritis 20.3000
## 7                  Sarcoma 19.7000
```

When using *tapply()*

```
tapply(db_pat$Age, db_pat$Disease, mean)
```

```
##             Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                  32.1000                  32.9125                  31.9000
##               Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##                  29.8000                  19.3000                  20.3000
##                  Sarcoma
##                  19.7000
```

3. Create a new variable in the data.frame for a categorical variable for age with the median as cut-off for the two categories

3a. Use the function *cut* in **R**. Run the command *str(NameofDatabase)*, of which type is the new categorical variable?

```
db_pat$age.cat<-cut(db_pat$Age, breaks=c(15,32,40), labels=c("<32",">32"))
str(db_pat)
```

```
## 'data.frame':    14 obs. of  5 variables:
##  $ Patient  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Disease  : Factor w/ 7 levels "Brain_cancer",..: 2 2 5 6 4 7 1 2 2 3 ...
##  $ Treatment: Factor w/ 3 levels "FBS","hPL","HSA": 1 1 1 1 2 3 3 1 1 1 ...
##  $ Age      : num  31.6 29.4 19.3 20.3 29.8 19.7 32.1 34.8 37.1 31.9 ...
##  $ age.cat  : Factor w/ 2 levels "<32",">32": 1 1 1 1 1 1 2 2 2 1 ...
```

From the internal structure of the data.frame we can notice that age.cat created by the cut function is a factor

3b. Use the function *ifelse* in **R**. Run the command *str(NameofDatabase)*, of which type is the new categorical variable?

```
db_pat$age.cat2<-ifelse(db_pat$Age<32, "<32",">32")
str(db_pat)
```

```
## 'data.frame':    14 obs. of  6 variables:
##  $ Patient  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Disease  : Factor w/ 7 levels "Brain_cancer",..: 2 2 5 6 4 7 1 2 2 3 ...
##  $ Treatment: Factor w/ 3 levels "FBS","hPL","HSA": 1 1 1 1 2 3 3 1 1 1 ...
##  $ Age      : num  31.6 29.4 19.3 20.3 29.8 19.7 32.1 34.8 37.1 31.9 ...
##  $ age.cat  : Factor w/ 2 levels "<32",">32": 1 1 1 1 1 1 2 2 2 1 ...
##  $ age.cat2 : chr  "<32" "<32" "<32" "<32" ...
```

From the internal structure of the data.frame we can notice that age.cat2 created by the ifelse function is a character

Question 4

We know focus on disease and treatment groups:

1. Show the proportion of patients in each treatment group

```
prop.table(table(db_pat$Treatment))
```

```
##
##        FBS        hPL        HSA
## 0.78571429 0.07142857 0.14285714
```

2. Show the number of patients in each disease group

```
table(db_pat$Disease)
```

```
##
##             Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                        1                        8                        1
##               Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##                        1                        1                        1
##                  Sarcoma
##                        1
```

3. Show the number of patients by treatment group and disease (two-ways table)

```
table(db_pat$Disease,db_pat$Treatment)
```

```
##
##                            FBS hPL HSA
##   Brain_cancer               0   0   1
##   Breast_cancer              8   0   0
##   Chronic_myeloid_leukemia   1   0   0
##   Mb_Hodgkin                 0   1   0
##   Neurological_cancer        1   0   0
##   Rheumatoid_arthritis       1   0   0
##   Sarcoma                    0   0   1
```

4. We want to create a new variable *Cancer* that groups the Disease into *Breast cancer* and *Others*:

```
db_pat$Cancer<-ifelse(db_pat$Disease=="Breast_cancer","Breast_cancer","Others")
```

we use the ifelse function to create the variable Cancer which write "Breast cancer" when the condition is veryfied and "Others" otherwise. This creates a character variable.

5. Calculate the mean of Age for each group respect to *Cancer*

```
tapply(db_pat$Age, db_pat$Cancer, mean)
```

```
## Breast_cancer        Others
```

```
##      32.91250      25.51667
```

```
aggregate(Age~Cancer, db_pat, mean)
```

```
##          Cancer      Age
## 1 Breast_cancer 32.91250
## 2        Others 25.51667
```

Question 5

Consider only patients with Breast cancer.

1. Subset data for Breast cancer patients

```
db_BC<-subset(db_pat, Cancer=="Breast_cancer")
```

2. Show the number of patients for each Age group

```
table(db_BC$age.cat)
```

```
##
## <32 >32
##   4   4
```

3. Calculate mean and standard deviation for Age

```
mean(db_BC$Age)
```

```
## [1] 32.9125
```

```
sd(db_BC$Age)
```

```
## [1] 2.700496
```