Exercise day 2 with solutions

Introduction to R for Basic Statistics

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

Exercise B: Reshaping data (Part II)

For this exercise we keep working with the data of Exercise of Day 1.

It is 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.

Question 0 Load data into R (use the *read.csv* function). **Remark: Remember to set your working directory with *setwd()*, or to define the correct path for the data**

1. Consider Data set *follicle*.
2. Visualize the first lines of the data.
3. Print the summary of the data. Is there any missing values? If yes, how many at Day0 and how many at Day8?

```
head(db_follicle)
```

```
##   Number Patient     Day0     Day2     Day4     Day6    Day8
## 1      1       1 101.4590 112.1605 135.1980 160.2315 161.500
## 2      2       1  89.8315 141.3770 165.4925       NA      NA
## 3      3       1  90.2835 116.9870 122.6500 127.7305 129.447
## 4      4       1 120.3145 148.8840 166.9970 170.0245 170.740
## 5      5       1  93.0085 112.3135 120.8550 120.9000 120.940
## 6      6       2  83.9085 100.3520 112.6875       NA      NA
```

```
summary(db_follicle)
```

```
##      Number         Patient          Day0             Day2
##  Min.   : 1.00   Min.   : 1.0   Min.   : 48.32   Min.   : 59.87
##  1st Qu.:18.25   1st Qu.: 4.0   1st Qu.: 62.74   1st Qu.: 81.50
##  Median :35.50   Median : 7.5   Median : 83.91   Median :105.03
##  Mean   :35.50   Mean   : 7.5   Mean   : 84.70   Mean   :106.85
##  3rd Qu.:52.75   3rd Qu.:11.0   3rd Qu.: 95.56   3rd Qu.:118.05
##  Max.   :70.00   Max.   :14.0   Max.   :194.47   Max.   :263.01
##                                 NA's   :3        NA's   :5
##       Day4             Day6             Day8
##  Min.   : 62.97   Min.   : 71.34   Min.   : 72.48
##  1st Qu.: 96.96   1st Qu.:102.64   1st Qu.:103.78
##  Median :120.51   Median :126.42   Median :129.45
##  Mean   :127.01   Mean   :139.46   Mean   :146.24
##  3rd Qu.:147.24   3rd Qu.:172.66   3rd Qu.:176.93
##  Max.   :299.65   Max.   :304.94   Max.   :318.88
```

```
##  NA's   :11       NA's   :24       NA's   :29
```

4. How many follicles have been collected by patient?

```
table(db_follicle$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
##  5  5  5  5  5  5  5  5  5  5  5  5  5  5
```

We have 5 follicles for each patient

Question 1

1. Calculate mean and standard deviation of the diameter at Day0

```
mean(db_follicle$Day0, na.rm=TRUE)
```

```
## [1] 84.70384
```

```
sd(db_follicle$Day0, na.rm=TRUE)
```

```
## [1] 27.6819
```

2. When we encounter into missing, we are often interested in the *complete case analysis* where we exclude patients with missing observations:

   2a. Use the *na.omit* function (excludes all rows that have one missing values) (Run the command *db.CC<-na.omit(NameofDataFrame)**)

   2b. check the dimension of the new data.frame

   2c. Calculate mean and standard deviation of the diameter at Day0 from db.CC . Compare results with the ones in point 1.

```
db.CC<-na.omit(db_follicle)
dim(db.CC)
```

```
## [1] 41  7
```

```
mean(db.CC$Day0)
```

```
## [1] 91.08129
```

```
sd(db.CC$Day0)
```

```
## [1] 31.47671
```

We can see that mean and standard deviation are different respect to the ones calculated in point 1. This is because the na.omit function is excluding all rows with at least one missing observation. However if one missing was at day 6, this is might not be missing at Day0. The mean calculation, with na.rm=TRUE is excluding only the missing at Day0, whereas the complete case consider only rows with observed measurement at each day

Question 2 For each follicle the diameter was measured at day 0,2,4,6,8.

1. Are the data in a wide or long format?

Data are in a wide format because we have one row for each follicle and several columns to indicate the measurement at different timepoints.

2. Convert data from wide to long. **Hint: You can use the function *reshape***

```
db_long<-reshape(db_follicle, direction="long",
                 idvar="Number",
                 varying=c("Day0","Day2","Day4","Day6","Day8"),
                 timevar="Day",v.names=c("diameter"),
                 times=c("0","2","4","6","8"))
```

3. How many rows would we expect for each patient? Is it correct? (You can use the command *table(db$Patient)*)

```
table(db_long$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

We would expect 25 observations for patient, 5 follicles times 5 repetitions (0,2,4,6,8)

Question 3 We are interested in the follicle growth over time. We can calculate the diameter difference from time 0 at each time point:

1. Subset observations at Day 0. Create a data.frame with only columns *Number* and *diameter*

```
day0<-subset(db_long, Day==0)
day0<-day0[, c("Number","diameter")]
```

2. Rename the variable of diameter into *diameter0*

```
colnames(day0)<-c("Number","diameter0")
```

3. Merge this data.frame and the long format of your data set by *Number*.

```
db_join=merge(db_long,day0, by="Number")
```

4. Create a new variable for the difference of diameter at each time point.

```
db_join$diam.change<-db_join$diameter-db_join$diameter0

head(db_join)
```

```
##   Number Patient Day diameter diameter0 diam.change
## 1      1       1   0 101.4590  101.4590      0.0000
## 2      1       1   8 161.5000  101.4590     60.0410
## 3      1       1   6 160.2315  101.4590     58.7725
## 4      1       1   4 135.1980  101.4590     33.7390
## 5      1       1   2 112.1605  101.4590     10.7015
## 6      2       1   2 141.3770   89.8315     51.5455
```

Question 4 Descriptive at baseline (Day0). We want to create one data set with all characteristics of patients at baseline.

1. Merge the two data sets: long version from Question 3 and *patient* to have baseline characteristics in one data.frame

```
db_all<-merge(db_join, db_pat, by="Patient")
```

2. Check if the number of observation for each Patient is correct (use *table()*)

```
table(db_all$Patient)
```

```
##
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

3. Print mean and standard deviation of diameter at Day0 by Disease (use *tapply()* or *aggregate()*)

```
tapply(db_all$diameter0,db_all$Disease,mean,na.rm=TRUE)
```

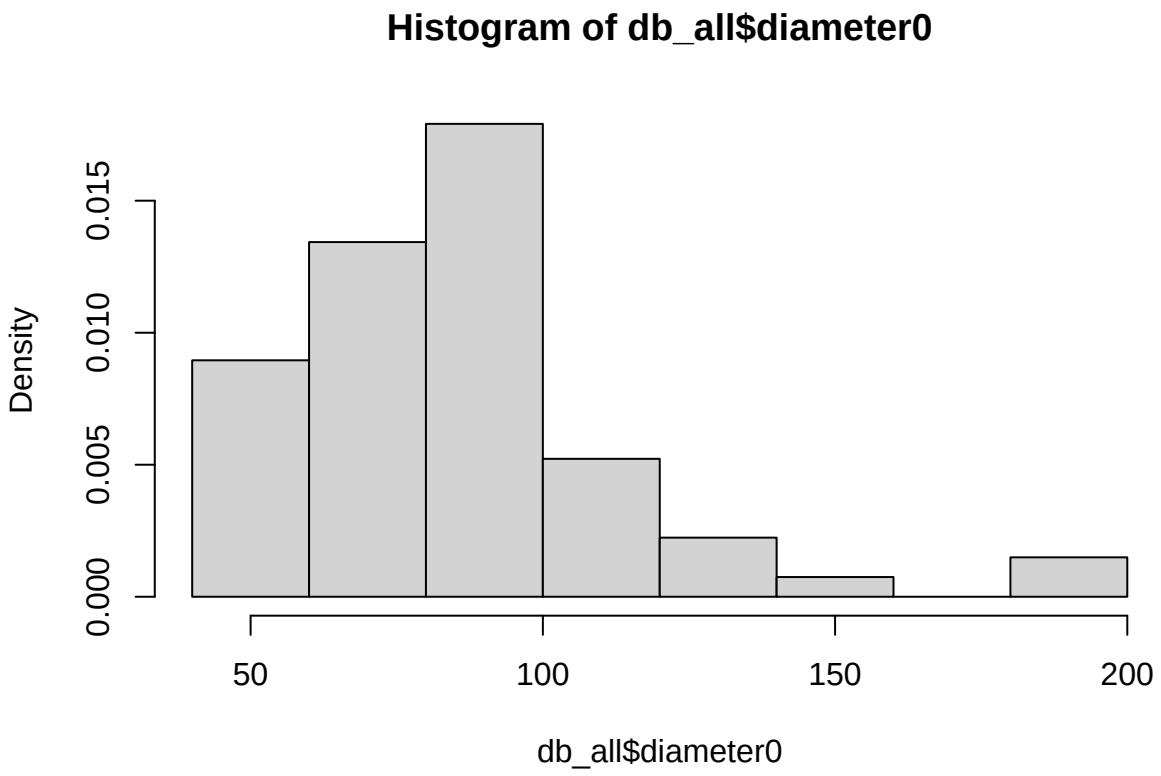
```
##             Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                 72.26000                 83.02593                112.77750
##               Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##                110.39300                 77.91750                 57.64700
##                  Sarcoma
##                 92.42000
```

```
tapply(db_all$diameter0,db_all$Disease,sd,na.rm=TRUE)
```

```
##             Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                20.504452                25.603763                10.002122
##               Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##                43.165471                22.097280                 2.776617
##                  Sarcoma
##                 8.460070
```

4. Plot the histogram for the density of diameter at Day0

```
hist(db_all$diameter0, prob=TRUE)
```

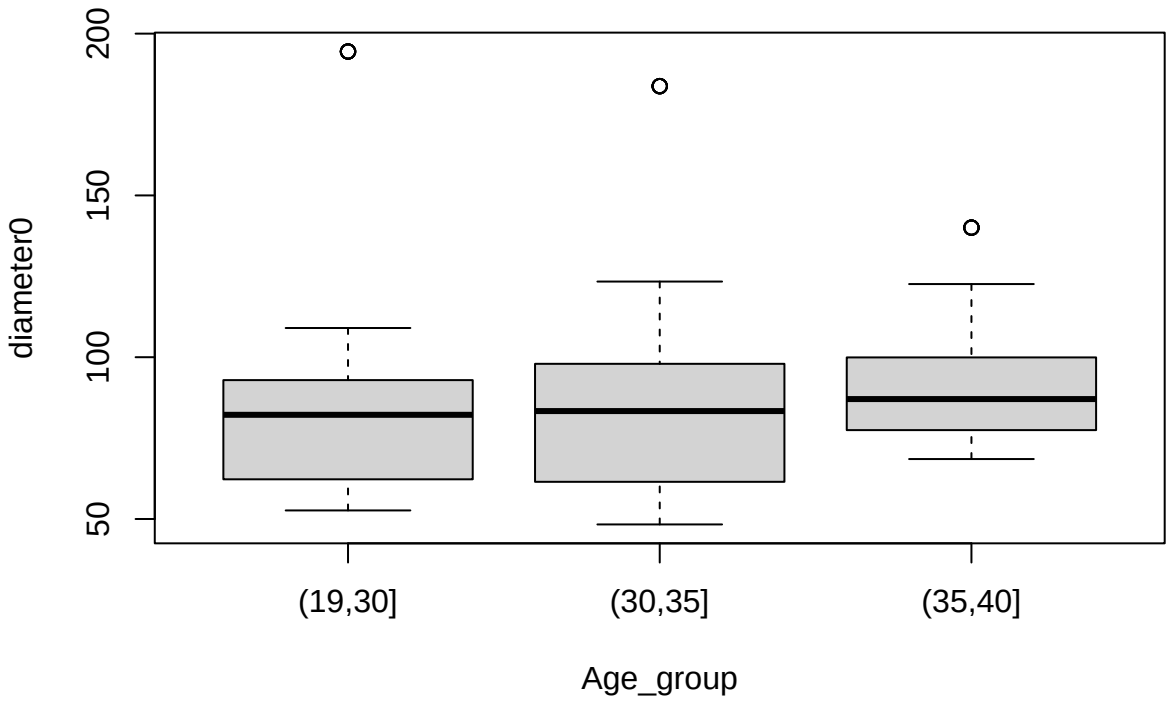


5. Create a categorical variable for age considering the intervals: (19,30], (30,35], (35,40] (use the function *cut()*)

```
db_all$Age_group<-cut(db_all$Age, breaks=c(19,30,35,40))
```

6. Create a Boxplot of diameter at Day 0 by Age category.

```
boxplot(diameter0~Age_group, db_all)
```



Question 5 Descriptive at Day 8

1. Calculate the average difference in diameter after 8 days.

```
#Option 1
tapply(db_all$diam.change,db_all$Day, mean, na.rm=TRUE)
```

```
##        0        2        4        6        8
##  0.00000 21.47157 40.04552 49.75852 55.16035
```

```
#Option2
aggregate(diam.change~Day, db_all, mean, na.rm=TRUE)
```

```
##   Day diam.change
## 1   0     0.00000
## 2   2    21.47157
## 3   4    40.04552
## 4   6    49.75852
## 5   8    55.16035
```

```
#Option 3
db8days<-subset(db_all,Day==8)
mean(db8days$diam.change, na.rm=TRUE)
```

```
## [1] 55.16035
```

2. Add to the data.frame the log-transformed variable for the difference in diameter at day 8.

```
db8days$diam.change.log<-log(db8days$diam.change)
```

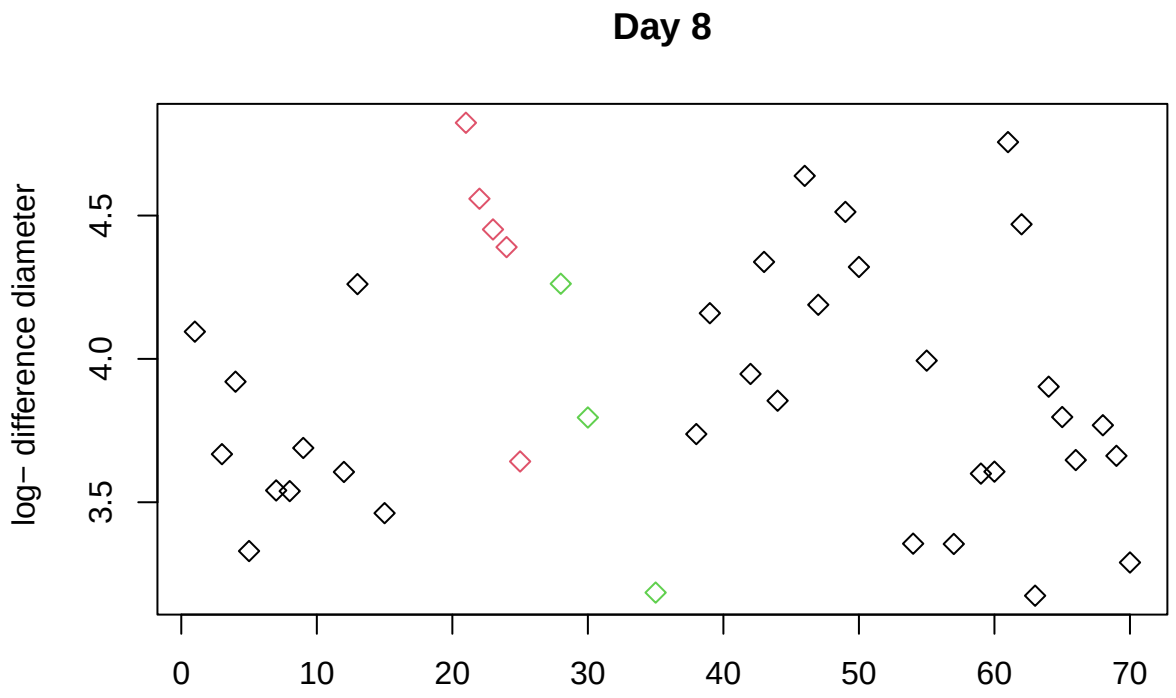
3. Calculate mean and standard deviation for the log-transformed difference in diameter after 8 days in each treatment group.

```
tapply(db8days$diam.change.log,db8days$Treatment, mean, na.rm=TRUE)
```

```
##      FBS      hPL      HSA
## 3.854136 4.373207 3.747541
```

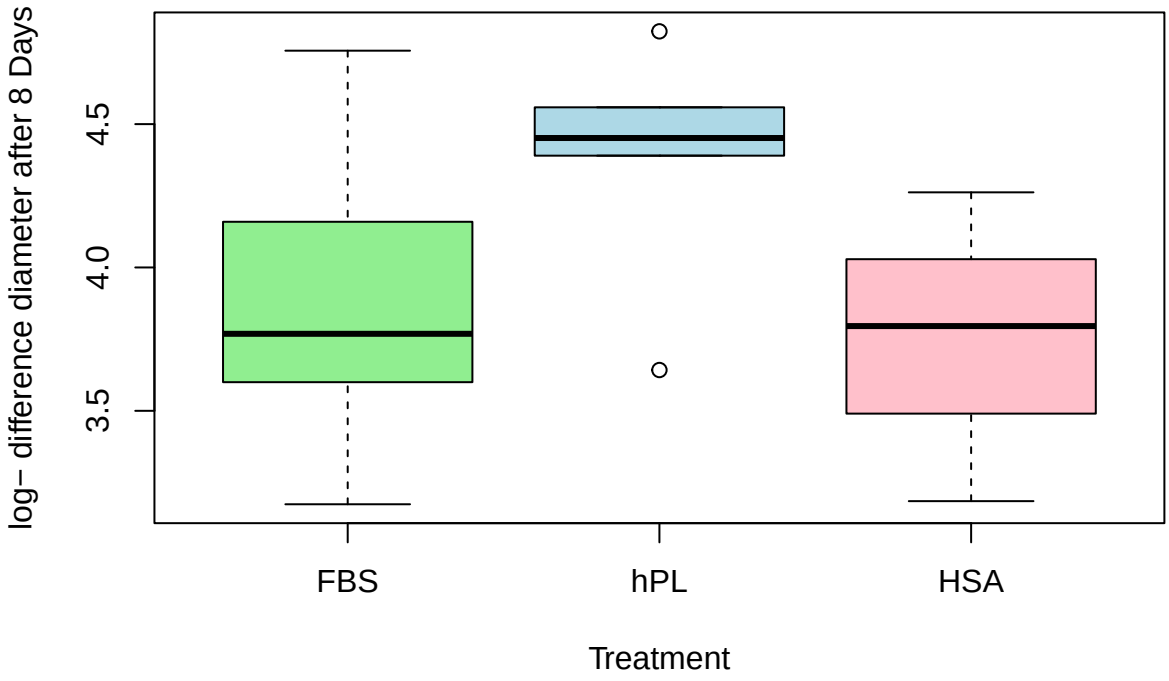
4. Create a scatterplot of the difference in diameter after 8 days :
    - defining the color by treatment group
    - specify one type of point with pch (you can choose)
    - precise name of axis: x= ” “, y=” log-difference diameter”
    - define the main title for the plot: “Day 8”

```
plot(db8days$diam.change.log, col=factor(db8days$Treatment), pch=5,
     xlab=" ", ylab=" log- difference diameter", main="Day 8")
```



5. Create a boxplot for the difference in Diameter after 8 days by treatment group
    - specify three colors (one for each treatment group)

```
boxplot(diam.change.log ~ Treatment, db8days, col=c("lightgreen","lightblue","pink"),
       ylab="log- difference diameter after 8 Days")
```


log− difference diameter after 8 Days
3.5
4.0
4.5
FBS
hPL
HSA
Treatment