Exercise day 2 with solutions

Introduction to R for Basic Statistics

Exercise B: Reshaping data (Part II)

Consider the data we used for 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.

You can find the data about follicles measurement here: "https://raw.githubusercontent.com/AMeddis/IntrotoR-for-Basic-Statistics/refs/heads/main/data_exercise/db_follicle.csv"

You can find the data about patients information here:

"https://raw.githubusercontent.com/AMeddis/ IntrotoR-for-Basic-Statistics/refs/heads/main/data_exercise/db_patient.csv"

In the following we will refer with *dbf* the data about the follicle measurements and with *dbp* the data with patients information.

Question 0 Load data into R :

1. Consider Data set *dbf*.
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(dbf)
```

```
##   Number Patient Treatment     Day0     Day2     Day4     Day6    Day8 AMH
## 1      1       1       FBS 125.3315 155.2745 173.0715 176.4855 177.167 <35
## 2      2       1       hPL  95.8205 105.9625 129.2270 148.7755 158.258 <35
## 3      3       1       hPL  68.9195  84.2325 114.0155 130.9890 146.407 <35
## 4      4       1       hPL  92.8125 120.5930 166.5025 177.1645 180.073 <35
## 5      5       1       hPL  75.2505  95.0710 110.6600       NA      NA <35
## 6      6       2       FBS  83.9085 100.3520 112.6875       NA      NA <35
```

```
summary(dbf)
```

```
##      Number         Patient      Treatment              Day0       
##  Min.   : 1.00   Min.   : 1.0   Length:70          Min.   : 36.05  
##  1st Qu.:18.25   1st Qu.: 4.0   Class :character   1st Qu.: 57.21  
##  Median :35.50   Median : 7.5   Mode  :character   Median : 67.28  
##  Mean   :35.50   Mean   : 7.5                      Mean   : 72.66  
##  3rd Qu.:52.75   3rd Qu.:11.0                      3rd Qu.: 83.25  
##  Max.   :70.00   Max.   :14.0                      Max.   :236.53  
##                                                                    
##       Day2             Day4             Day6             Day8       
##  Min.   : 49.67   Min.   : 61.09   Min.   : 79.19   Min.   : 87.53  
##  1st Qu.: 69.39   1st Qu.: 87.85   1st Qu.:102.86   1st Qu.:113.22  
##  Median : 80.66   Median :108.74   Median :122.83   Median :131.27
##  Mean   : 89.76   Mean   :112.54   Mean   :130.68   Mean   :139.01
##  3rd Qu.:100.95   3rd Qu.:127.48   3rd Qu.:142.62   3rd Qu.:147.59
##  Max.   :309.04   Max.   :339.76   Max.   :353.19   Max.   :373.36
##  NA's   :8        NA's   :16       NA's   :26       NA's   :31
##      AMH
##  Length:70
##  Class :character
##  Mode  :character
##
##
##
##
```

4. Use the command *table(dbf$patient)*, interpret the numbers.

```
table(dbf$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 Day2

```
mean(dbf$Day2, na.rm=TRUE)
```

```
## [1] 89.75802
```

```
sd(dbf$Day2, na.rm=TRUE)
```

```
## [1] 35.8144
```

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 Day2 from db.CC .

   2d. Compare results with the ones in point 1. Did something change? If yes, Why?

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

```
## [1] 39  9
```

```
mean(db.CC$Day2)
```

```
## [1] 94.83172
```

```
sd(db.CC$Day2)
```

```
## [1] 42.03498
```

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 might not be missing at Day0. The mean calculation, with na.rm=TRUE is excluding only the missing at Day2, whereas the complete case consider only follicles that have all observed measurements (for all days)*

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 time points.

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

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

Question 3 We would like to create one data set with all characteristics of patients at baseline.

1. Merge the data set obtained in Question 2.2 and *dbp* to add baseline characteristics in the data.frame

```
db_all<-merge(db_long, dbp, by="Patient")
```

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

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

Question 4: descriptive statistics at baseline (Day 0)

1. Subset from the merged data only observation at baseline (Day 0)

```
db0<-subset(db_all, Day=="0")
```

2. Create a table with the counts of follicles per Disease. Which is the most common disease?

```
table(db0$Disease)
```

```
##
##              Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                         4                       20                        5
##                Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##                         3                        3                        1
##                   Sarcoma
##                         3
```

Most patients have breast cancer

3. Calculate the proportion of follicles by Treatment.

```
prop.table(table(db0$Treatment))
```

```
##
##       FBS       hPL       HSA       UCP
## 0.1794872 0.5384615 0.1794872 0.1025641
```

4. Print the mean of diameter at Day0 by Disease (use *tapply()* or *aggregate()*). Would you say that the follicle diameter is dependent on the disease?

```
tapply(db0$diameter,db0$Disease,mean,na.rm=TRUE)
```

```
##            Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                63.81750                 75.60873                 70.14100
##              Mb_Hodgkin      Neurological_cancer     Rheumatoid_arthritis
##                77.09333                 65.45833                236.53500
##                 Sarcoma
##                73.61667
```

Seems like follicles of patients with Rheumatoid_arthritis have bigger diameter from start

Question 5 : We are interested in the follicle growth over time. We calculate the diameter difference from time 0 at each time point:

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

```
day0<-db0[, c("Number","diameter")]
```

1. Rename the variable of diameter into *diameter0*

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

2. Merge this data.frame and the long format of your data set (created in Question 3.1) by *Number*.

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

3. Create a new variable "diam.change" for the difference of diameter at each time point.

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

head(db_join)
```

```
##   Number Patient Treatment AMH Day diameter       Disease  Age Age_group
## 1      1       1       FBS <35   0 125.3315 Breast_cancer 31.6   (30,35]
## 2      1       1       FBS <35   2 155.2745 Breast_cancer 31.6   (30,35]
## 3      1       1       FBS <35   8 177.1670 Breast_cancer 31.6   (30,35]
## 4      1       1       FBS <35   6 176.4855 Breast_cancer 31.6   (30,35]
## 5      1       1       FBS <35   4 173.0715 Breast_cancer 31.6   (30,35]
## 6      2       1       hPL <35   0  95.8205 Breast_cancer 31.6   (30,35]
##   diameter0 diam.change
## 1  125.3315      0.0000
## 2  125.3315     29.9430
## 3  125.3315     51.8355
## 4  125.3315     51.1540
## 5  125.3315     47.7400
## 6   95.8205      0.0000
```

4. Save the data

```
save(db_join, file="~/Desktop/KVN2021/Course/IntrotoR/data_exercise/db_join.rda")
```

Question 6: Descriptive of diameter change over time

1. Print the median diameter change over time. Would you say that the diameter is growing over time?

```
aggregate(diam.change~ Day, db_join, median)
```

```
##   Day diam.change
## 1   0      0.0000
## 2   2      15.5850
## 3   4      39.0500
## 4   6      53.4650
## 5   8      57.8395
```

- yes, it seems like the diameter is growing over time, even if the growth does not seem to be linear*.

2. Calculate the median diameter change by Day and Treatment. Would you say that the treatment has an effect on the diameter change? Which treatment has the largest effect? and which one the smallest? **use aggregate with formula: diam.change~Day + Treatment**

```
aggregate(diam.change~ Day + Treatment, db_join, median)
```

```
##    Day Treatment diam.change
## 1    0       FBS     0.00000
## 2    2       FBS    21.11000
## 3    4       FBS    29.47000
## 4    6       FBS    45.94500
## 5    8       FBS    49.55500
## 6    0       hPL     0.00000
## 7    2       hPL    15.31300
## 8    4       hPL    39.65000
## 9    6       hPL    62.06950
## 10   8       hPL    71.46500
## 11   0       HSA     0.00000
## 12   2       HSA    15.58500
## 13   4       HSA    34.38000
## 14   6       HSA    43.05000
## 15   8       HSA    54.66000
## 16   0       UCP     0.00000
## 17   2       UCP    12.53850
## 18   4       UCP    30.81750
## 19   6       UCP    42.16500
## 20   8       UCP    49.87925
```

Question 7 (Optional): Descriptive on the relative change of the diameter

1. Calculate the relative change for the diameter ((diameter - diameter0)/diameter0)

```
db_join$diam.Rchange<-(db_join$diameter-db_join$diameter0)/(db_join$diameter0)
```

2. Calculate the median relative change by Treatment at day 8. Would you conclude the same as in Question 6.2 ?

```
dbday8<-subset(db_join, Day==8)
aggregate(diam.Rchange~ Treatment, dbday8, median)
```

```
##   Treatment diam.Rchange
## 1       FBS    0.7222550
## 2       hPL    0.9180270
## 3       HSA    0.6492109
## 4       UCP    0.8123566
```

Question 8 (Optional): Survival of follicles When information on the follicle are missing, it is because the follicle did not survive until day 8. We want to calculate the survival probability at day 8 to see which treatment is more aggressive or which disease is less resistant to the treatment.

1. Consider the full data *dbf*. We want to create a variable *status* equal to 1 if the follicle survived the 8 days and 0 otherwise. We can use the *is.na* function that provides TRUE if value is missing and FALSE otherwise.

```
dbf$status<-ifelse(is.na(dbf$Day8),0,1)
```

2. Calculate the probability for the follicle to survive the 8 days (status=1)

```
sum(dbf$status)/nrow(dbf)
```

```
## [1] 0.5571429
```

3. Calculate the proportions of follicles surviving when treated by UCP and by HSA. Check the result in 7.2, how would you discuss the findings?

```
db_UCP<-subset(dbf, Treatment=="UCP")
sum(db_UCP$status)/nrow(db_UCP)
```

```
## [1] 0.2666667
```

```
db_HSA<-subset(dbf, Treatment=="HSA")
sum(db_HSA$status)/nrow(db_HSA)
```

```
## [1] 0.5833333
```

It seems like there is a survival bias. Among the survivors (complete case) UCP seems to lead to a larger growth of diameter, however only 27% of the follicles actually survive.