Exercise day 2

Introduction to R for Basic Statistics

Alessandra Meddis

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.

Question 0 Load data into R :

```
dbf<-read.csv("https://raw.githubusercontent.com/AMeddis/
              IntrotoR-for-Basic-Statistics/main/data_exercise/follicle.csv")

dbp<-read.csv("https://raw.githubusercontent.com/AMeddis/
              IntrotoR-for-Basic-Statistics/main/data_exercise/patient.csv")
```

Remark: Remember to set your working directory with *setwd()* to name the script and save it in the correct folder

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

Question 1 : missing observations

1. Calculate mean and standard deviation of the diameter at Day0 (Be careful, there are some missing!)
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 at least one missing values)

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

```
2b. Check the dimension of the new data.frame
2c. Calculate mean and standard deviation of the diameter at Day0 from db.CC
2d. Compare results with the ones in point 1. Did something change? If yes, Why?
```

Question 2 : wide/long format Consider the data set without missing (db.CC) that you obtained in question 2a.

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

1. Respect to the follicles, are data in a wide or long format?

2. Convert data from wide to long (or viceversa) using the *reshape* function. Use *diameter* as column s name for the measurement of the follicles.

Question 3: Merge Follicles and Patients information in one data set 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
2. Create a categorical variable for age considering the intervals: 19-30, 30-35, 35-40. (Use the function *cut()*)

Question 4: descriptive statistics at baseline (Day 0)

1. Subset from the merged data only observation at baseline (Day 0)
2. Plot the histogram for the density of diameter at Day0
3. Create a Boxplot of diameter at Day 0 by Age category. Would you say that the follicle diameter is dependent on patient age?
4. Create a table with the counts of follicles per Disease and Treatment. Which is the most common disease?
5. Print the mean of diameter at Day0 by Disease (use *tapply()* or *aggregate()*).

   5a. Would you say that the follicle diameter is dependent on the disease?

   5b. Show the boxplot of diameter at time 0 by Disease.

Question 5: calculate the change in diameter from Day 0 We are interested in the follicle growth over time respect to time 0. We want to create a new variable *diam.change* that is calculated as the difference of follicles diameter at each time from time 0:

1. Take the subset at Day 0 (from 4.1). Create a data.frame with only the columns *Number* and *diameter*, where the variable of diameter is called *diameter0*.
2. Merge this data.frame and the long format of your data set (created in Question 3.1) by *Number*.
3. Create a new variable *diam.change* for the difference of diameter at each time point from Day 0.

Question 6: Descriptive on diameter change over time

1. Show the Boxplot of diameter by Day. Would you say that the diameter is growing over time?
2. Calculate the median diameter change by Day and Treatment (store the results in your environment, you need it for the next step). **use aggregate with formula: diam.change ~ Day + Treatment**

Question 7 (Optional)

1. Plot the diameter growth over time by treatment group:
   - 1a. rename the column "diam.change" with "median.change" of the data.frame obtained by the aggregate in 6.2
   - 1b. merge the data set with the one obtained in Question 5.4 (after creating diam.change)
   - 1c. create a plot of the difference from time 0 at varying of days:

```
- define the color by Treatment group

- precise as name of axis: x= "Days", y="median diameter's diff"

- add the legend
```

2. Calculate the relative change for the diameter((diameter - diameter0)/diameter0) and re-create the same plot.