Exercise day 2

Introduction to R for Basic Statistics

Alessandra Meddis

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 Download both data sets from the course material webpage (folder data_exercise) and load them 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 and print a summary of the data.

Question 1

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) (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

   2d. Compare results with the ones in point 1.

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?
2. Convert data from wide to long (or viceversa). **Hint: You can use the function *reshape***
3. How many rows would we expect for each patient? Is it correct? (You can use the command *table(db$Patient)*)

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. Create a data.frame with Number (follicle ID) and the diameter at Day 0.
2. Rename the variable of diameter into *diameter0*
3. Merge this data.frame and the long format of your data set by *Number*

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

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 of *follicle* (Question 2) and *patient* to have baseline characteristics in one data.frame
2. Check if the number of observation for each Patient is correct (use *table()*)
3. Print min, max, median and standard deviation of diameter at Day0
4. Plot the histogram for the density of diameter at Day0

Question 5 Descriptive at Day 8

1. Calculate the average difference in diameter after 8 days.
2. Create the log-transformed variable for the difference in diameter at day 8.
3. Calculate mean and standard deviation for the log-transformed difference in diameter after 8 days in each treatment group.
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”
5. Create a boxplot for the difference in Diameter after 8 days by treatment group
    - specify three colors (one for each treatment group)