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.

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 Import data into R.

1. Consider the data set on the follicles measurement.
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 - complete case analysis

1. Calculate mean and standard deviation of the diameter at Day2 (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 Day2 from db.CC
2d. Compare results with the ones in point 1. Is it different? 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 vice versa) using the *reshape* function. Use *diameter* as name for the variable of follicles measurement.

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. Create a table with the counts of follicles per Disease. Which is the most common disease?
3. Calculate the proportion of follicles by Treatment.
4. Print the mean of diameter at Day0 by Disease (use *tapply()* or *aggregate()*). Would you say that the follicle diameter is associated with the disease?

Question 5: 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.
4. Save the data.frame with the command *save(data, file="path/NameofData.rda")*, where *data* is the name of the data.frame in the environment, *path* is the path of where you would like to locate the data, *NameofData* is the name you would like to save the data with.

Question 6: Descriptive on diameter change over time

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

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

1. Calculate the relative change for the diameter ((diameter - diameter0)/diameter0)
2. Calculate the median relative change by Treatment at day 8. Would you conclude the same as in Question 6.2 ?

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