Exercise day 1

Introduction to R for Basic Statistics

Aessandra Meddis

Exercise A:

For this exercise we will work with 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.

More information on the study and full data are available at https://doi.org/10.1016/j.rbmo.2023.06.011.

Information are collected in two data sets.

Data set *follicle* with:

- Patient: patient ID
- Number: follicle ID
- Day_x: follicle diameter at Day x

Data set *patient* with:

- Patient ID
- Treatment group
- Type of Disease
- Age at Day 0

Question 1 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("path")*, or to define the correct path for the data**

Focus on the data set *patient*:

1. Check the dimension of the data.frame. How many patients were included in the study?
2. Visualize the first lines of the data using the function *head()*.
3. Print a summary of the data. What is the data type of each variable?

Question 2 Treatment and Disease are *characters*, can we understand from the summary how many different diseases are in the data?

1. Would it be better if Treatment and Disease were of a different type? If yes, which one?
2. Transform them into factor. You can use the function *factor()* in *R*.
3. Print the summary of the data, can you see any difference?

Question 3 Age is a continuous covariate. We would like to have an idea of the age distribution of patients included in the study:

1. Show min,max and mean for age
2. Calculate the mean of age for each disease group (you can use *aggregate* or *tapply*)

3. Create a categorical variable for age with the median as cut-off for the two categories. We can code it using two different functions in **R**:

   3a. Use the function *cut* in **R**. Run the command *str(NameofDatabase)*, of which type is the new categorical variable?

   3b. Use the function *ifelse* in **R**. Run the command *str(NameofDatabase)*, of which type is the new categorical variable?

Question 4 We know focus on disease and treatment groups:

1. Show the proportion of patients in each treatment group (use *prop.table*)
2. Show the number of patients in each disease group
3. Show the number of patients by treatment group and disease (two-ways table)
4. We want to create a new variable *Cancer* that groups the Disease into *Breast cancer* and *Others*:

   4a. Run the command: $db\_patCancer < -ifelse(db_patDisease{=}{=}“Breast\_cancer”,“Breast\_cancer”,“Others”))$. What does it do?

   4b. Run the command:

```
db_pat$Cancer<-db_pat$Disease

db_pat$Cancer[db_pat$Disease!="Breast_cancer"]<-"Others"
```

We have a warning message.

Cancer is a factor, thus we cannot substitute a value with something different of its levels. This is coerced to NA:

```
table(db_pat$Cancer, useNA = "ifany")
```

If we run instead:

```
db_pat$Cancer<-as.character(db_pat$Disease)
db_pat$Cancer[db_pat$Disease!="Breast_cancer"]<-"Others"

table(db_pat$Cancer, useNA = "ifany")
```

5. Calculate the average age for each group respect to *Cancer*

Question 5 Consider only patients with Breast cancer:

1. Subset data for Breast cancer patients (use the function *subset*)
2. Show the number of patients for each age group
3. Calculate mean and standard deviation for age