

# Exercises day 2

Basic Statistics for health researchers 2022

November 2, 2022

## Warming up

Before starting the exercise below, learn from the R-demo of Lecture 2 (available from the course webpage):

1. Read and run the code.
2. Check that the output matches the results presented on the slides.
3. Do not hesitate to add your own comments into the script.

## Exercise A (hypothesis testing)

For each of the different data sets and corresponding task below do all the following steps:

- Formulate a research hypothesis and the corresponding null and alternative hypotheses, all in “plain English”.
- Describe the relevant statistical method(s) that you plan to use.
- Prepare the data set for analysis (if necessary)
- Use a suitable graph to get an idea of the data and the direction of the result (boxplot for “large” data sets, dotplot for “small”).
- Compute the relevant quantities according to the above choice of statistical method(s) (e.g. compute estimates of means and mean difference, confidence intervals, p-value).
- Visually check whether the main assumptions of the statistical methods seem fine to “safely” trust the results (if necessary).
- Conclude and report the results in one or a few clear sentences.

### Question 1

- **Data:** Milk data (available from the `nmle` package of R).
- **Task:** compare the average protein level of the milk at **6 weeks** after calving between cows fed with **Lupins only** and those fed with **Barley and Lupins**.

## Question 2

- **Data:** gene expression data (`alpha` available from the `coin` package of R).
- **Task:** compare the distribution of the level of expressed alpha synuclein mRNA of “intermediate” and “long” allele length, with a stringent control of the risk of false positive finding.

## Question 3

- **Data:** biometrics dose response data (`biom` available from the `DoseFinding` package of R).
- **Task:** compare the average response between dose=0 and:
  - dose=0.05
  - dose=0.2
  - dose=0.6
  - dose=1

with the aim to control the risk of making at least one false positive finding.

## Exercise B (power calculation)

A PhD student and her supervisor are planning a laboratory experiment. They aim to show a significant reduction in (average) tumor growth between treated and untreated mice.

They would like to plan the experiment to have (at least) 90% power to show a difference in tumor volume of (at least)  $3 \text{ mm}^3$ , while controlling the risk of false positive finding at the usual 5% level. They expect the standard deviation of the tumor volume to be (approximately)  $2 \text{ mm}^3$  in both groups. They plan to use a usual two-sided two-sample t-test and they think that it is fine to assume that the tumor volume distribution in each group is well approximated by a normal distribution.

### Question 1

How many treated and untreated mice should they plan to include in their study?

### Question 2

Using the sample size suggested by the result to the previous question, they wonder:

1. how much the power depends on their “best guess” of the standard deviation, i.e., what does the power of the planned study become if the standard deviation is actually  $2.5 \text{ mm}^3$  or  $3 \text{ mm}^3$  instead of  $2 \text{ mm}^3$ ?
2. what is the smallest difference they can hope to show with a “decent” power of 75%?