Home Exercises 1

25.9.2023

You can do the exercises using R markdown by directly writing your answers below each question (remember to use R chunk for code in order that Rstudio knows to interpret it as code and run it in R). Return your solutions as a PDF or HTML file by saving from R markdown (option arrow on the right hand side of **Knit** button). If you Knit to Word document, then save as PDF in Word before returning. Note that your solutions must contain both the R code and the output from R when the code was run including plots asked.

If you are not able to use R markdown, you can write the code in a new R script file (Choose File -> New File -> R script) and run the commands on the console and by copying the relevant output to your solution file, which may be a text file of any type (such as MS Word). This is not recommended, however, since it is complicated and error prone.

Write your name at the beginning of the file as "author:".

1. Return to Moodle by **09.00am, Mon 25.9.** (to section "BEFORE").
2. Watch the exercise session in Moodle available by 10am on Mon 25.9.
3. If you observe during the exercise session that your answers need corrections, return a corrected version to Moodle (to section "AFTER") by 09.00am Mon 2.10.

Problem 1. Patients A and B have been measured for a biomarker at six time points and the values are

A: 2.5, 8.7, 1.1, 4.4, 2.5, 6.1

B: 5.4, 3.3, 4.6, 5.2, 3.7, 4.3

(i) Make vectors A and B of the six values for each patient.

(ii) Compute the difference A - B at the six time points. (Result has six values, one for each time point.)

(iii) Which patient has a larger difference between his/her maximum and minimum measurement?

(iv) Which patient has a larger standard deviation among his/her six measurements?

Problem 2. Download file "systbp_ldlc.txt" to the same directory where you have this .Rmd file. It contains blood pressure and LDL-cholesterol values. Read it in using command

```
x = read.table("systbp_ldlc.txt", as.is = TRUE, header = TRUE)
```

NOTE: If Rstudio can't find the file, it is likely not in the same directory as the .Rmd file. Then either move `systbp_ldlc.txt` to the correct directory, or specify the full path to the file as `read.table("full_path/systbp_ldlc.txt", as.is = TRUE, header = TRUE)`.

Run the following commands: `head(x)` to see first 6 lines; `dim(x)` to see dimensions, that is, no. of rows and no. of columns, and `str(x)` to see the structure of `x`.

The data object `x` is of R data type `data.frame` (as `str()` showed above) which combines a set of vectors into columns of a table. Here we have 3 columns and 965 rows. Each row is one individual and each column

is a variable measured on the individuals. To extract the second column, `systbp`, as its own vector you can do any of the following: `bp = x[,2]` (picks the 2nd column), or `bp = x[,"systbp"]` (picks the column named "systbp") or `bp = x$systbp` (picks the column named "systbp"). Choose one of them and make a histogram of `bp` values.

Find mean, median, sd and range of `bp` values.

What is the blood pressure value below which 5% of the sampled population are (and above which are the remaining 95% of the sampled individuals)?

Problem 3. Continue with the data frame `x` that you read in in Problem 2.

(i) Make a `table()` of column "sex" to see how many males (1) and females (2) there are.

(ii) We would like to look at the distribution in males. We can make a new `data.frame` called `x.males` that picks only the rows that have `sex` equal to 1 by command `x.males = x[x$sex == 1,]`. NOTE there are two `=` symbols to denote the comparison operation: Interpretation is that we assign `x.males` to be such part of original `x` for which it holds that the `sex` column equals to 1. Pick the males into their own `data.frame` using notation above and plot histogram of `ldlc` in males.

(iii) Another way to subset a `data.frame` is by `subset()` function that takes three arguments: 1st = original data.frame, 2nd = logical expression defining which rows to keep, 3rd = which columns to return (or if not defined, then returns all columns). Thus, to do the same as above in part (ii), you could do `x.males = subset(x, sex == 1)`. Use `subset()` to pick the **females** from `x` and plot their `ldlc` as a histogram.

Compute mean `ldlc` in males and in females.

Problem 4. Suppose that a treatment helps about 30% of patients. You are treating 50 patients. You want to evaluate how many of them will benefit.

(a) Plot the theoretical distribution of how many will benefit by using `dbinom()` to get the probabilities and follow lectures (Example 1.2) to make a barplot.

(b) Compute the probability that at most 10 patients will benefit.

(c) Compute the probability that at least 20 patients will benefit