Exercises day 1

Basic Statistics for health researchers 2023

October 23, 2023

Warming up

Before starting the exercise below, learn from the R-demo of Lecture 1 (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

For this exercise we will work with the Sickle Cell Disease (SCD) data (available from the course webpage).

Question 1

Read the description of the data and load the data into R. Visualize the first lines of the data and a summary of the data. Which variables have missing values?

Question 2

Using the `table()` function, count how many included subjects:

1. are men/women?
2. have Sickle Cell Disease (SCD)?

Question 3

1. Make two boxplots, one for those with SCD and one for those without, to compare the distribution of Systolic pressure in the two populations. Take the time to fine tune the plot (with appropriate title for x- and y-axis etc.).
2. What is the interpretation of each element of the plot (boxes, whiskers, dots...)?
3. Does the interpretation of the plot makes sense?

Question 4

1. Create a new variable `BMI` corresponding to the Body Mass Index (BMI) of each subject and add it to the data. **Hint:** you can use (and learn from) this code

```
d$BMI <- d$weight/(d$height/100)^2
```

2. Make a histogram to describe the distribution of BMI. Does it look normally distributed?

3. Compute the mean, standard deviation (sd), minimum, maximum, first and third quartile and median of BMI.

4. Which of the above descriptive statistics would you prefer to report in a publication? Explain why.

5. World Health Organization (WHO) regards a BMI below 18.5 as "underweight", "normal" if between 18.5 and 25, "overweight" if between 25 and 30 and "obese" if above 30 (for adults). How many subjects of each category do you observe in the data? **Hint:** you can use (and learn from) this code:

```
d$BMIgroup <- cut(d$BMI,
                  breaks=c(0,18.5,25,30,Inf),
                  include.lowest=TRUE,
                  right=FALSE)
```

6. Make a barplot to describe the frequencies of each group. **Hint:** you can first make a table, using the function `table()`. Then you can use `barplot()` to graphically show the frequencies.

7. Make similar barplots to describe the frequencies among subjects with and without SCD. What differences do you observe? **Hint:** to facilitate the comparison of the two barplots, you can use the `ylim` option in `barplot()`.

Question 5

1. Compute the mean and sd of the diastolic pressure among subjects with and without SCD.

2. Compute the 95% confidence interval of the mean diastolic pressure in the two populations of subjects with and without SCD. What can you conclude?

3. Compute the 95% prediction intervals of the mean diastolic pressure in the two populations of subjects with and without SCD. Write down your interpretation in "plain English".

Question 6

1. Compute the 95% confidence interval of the mean systolic pressure in the two populations of subjects with and without SCD. What can you conclude?
2. Same question, but use data from women only, i.e., compute the 95% CI of the mean systolic pressure in the two populations of women with and without SCD. What can you conclude?
3. Make a dotplot (stripchart) to visually compare the distribution of the observations of systolic pressure among female with and without SCD. Take the time to fine tune the plot (with appropriate title for x- and y-axis etc.).
4. Make two QQplots to visualize whether the distribution of the systolic pressure among females with and without SCD look appropriately normally distributed. etc.).
5. How many observations do you have from women with and without SCD?
6. Based on the results from the above two items, how comfortable are you with the interpretation of the 95% confidence? Why, i.e., are the main assumptions fulfilled?

Question 7

1. Create (and add to the data) the new variable `MAP` to define the Mean Arterial Pressure as the sum of the diastolic pressure (`Pdias`) plus one third of the difference between the systolic pressure (`Psys`) and the diastolic pressure (`Pdias`). In short,
$$\texttt{MAP} = \texttt{Pdias} + \frac{1}{3}\big(\texttt{Psys} - \texttt{Pdias}\big).$$
2. Make a histogram and a QQplot to visualize whether the distribution of the Mean Arterial Pressure looks normally distributed, among subjects without SCD.
3. To strengthen your interpretation, produce a "Wally plot" to compare the QQplot produced with the data with eight others produced from random samples drawn from a normal distribution. Conclude.
4. Estimate the "normal range" of MAP. Does it match with the following statement from wikipedia: "*MAP is normally between 65 and 110 mmHg*"[1].

Finalizing

Take the time to:

1. clean-up and indent your code.
2. add and clean-up comments. The aim is that you can remember what the code does (and why) next time you look at it, to help you perform similar analyses.
3. Fine tune all your plots to make then as close as possible to "publishable".

[1] `https://en.wikipedia.org/wiki/Mean_arterial_pressure`, retrieved 14 October 2020.

Additional challenge

If you have time and are up for more (and more "technical") challenges, try to produce a plot as close as possible to the plot below. Hint: you can use the option `add=TRUE` in the function `stripchart` to plot a stripchart on top of another.