Principles of Statistical Data Analysis

Project 2019-2020

1 Introduction

The aim of this project assignment is to demonstrate that you can

1. learn new statistical methods from the literature
2. perform an informative descriptive analysis of a dataset using informative graphs and tables
3. perform a focused statistical data-analysis using R
4. report accurately about your findings
5. work in a team.

The data that have been analysed for this project come from a study performed by LabMET of Ghent University, but because of confidentially issues parts of the data have been altered. The study focuses on malodor generating bacteria in the armpits. For more details on the project, see `http://www.drarmpit.com`. In short, the bad smell that may come from armpits is caused by bacteria living in the armpits. The community of all microorganisms living together at a particular locus (here: armpit) is known as the microbiome. Recent studies have showed that two bacterial genera always occurred in the armpit: staphylococci and corynebacteria, as 77% of all sequences were assigned to those two bacterial groups. The corynebacteria are known to cause malodor in the armpit, whereas staphylococci are known to cause no significant malodor. In this project we will continue to work with those two bacteria obtained from a sample of 40 subjects (which were randomly selected from a larger sample of 200 volunteers).

The analyses gave relative abundances of 8 microorganisms: 4 Staphylococcus species and 4 Corynebacteria species. For the non-biologists: Staphylococcus and Corynebacteria are "genera" and each "genus" may have many "species". Both genus and species refer to the hierarchical taxonomic rank of organisms in the tree of life. For each of the 8 species the dataset contains the relative abundances (rescaled to make them sum to 100%).

Here the research questions are formulated as follows.

1. Does the genus composition change with age? To answer this question, we ask you to implement two strategies:

   - strategy 1: make age discrete (i.e. make age groups).
   - strategy 2: keep age as a continuous variable.

   After implementing and discussing the results of the two strategies, you also have to discuss the differences (advantages/disadvantages) between both strategies.

2. How are the relative abundances of the 8 species related to one another?

You are asked to analyse the data so as to answer these research questions. We are aware of the fact that the best analysis here might come in the form of a regression model, which we are not currently allowing. The challenge here is to understand what can and cannot be usefully learned from simpler techniques, inference and measures (means, variances, correlations,...) as developed in the Prinstat course. To promote good statistical practice, please write out the protocol of your data-analysis approach before actually analyzing the data. Adapt the protocol, if needed, after the descriptive statistics have been examined. For answering the second research question you will need to know more about hypothesis tests concerning correlations. Part of the assignment is to consult the literature to find out more about this topic so that you can apply it for answering the research questions. You may limit your search to one of the correlation coefficients that is calculated by the `cor` function of R. For this second question, you should also critically discuss the added value of using correlation measures and related hypothesis tests to study associations.

The dataset `armpit.txt` contains the following variables:

- Corynebacterium.1, Corynebacterium.2, Corynebacterium.3, Corynebacterium.4. The relative abundances (%) of the 4 Corynebacterium species.
- Staphylococcus.1, Staphylococcus.2, Staphylococcus.3, Staphylococcus.4. The relative abundances (%) of the 4 Staphylococcus species. (note that the numbering of Staphylococcus is not related to the numbering of Corynebacterium - they just indicate different species).
- The age (years) and the gender (F: female; M: male) of the subject.
- A body mass index indicator (0: BMI ≤ 25; 1: BMI > 25).

2 Practicalities

Perform all statistical analyses that you find necessary to support your conclusions and to make them fully informative. You should write your solution to the project as a scientific paper of at most 5 pages (excluding appendices). There is no need to introduce the problem setting again, so it is sufficient to include a methods section, a results section and a conclusion. Thus even the materials you will consult for understanding hypothesis tests for correlations should not be summarized in the report (except for a brief description and motivation in the Methods section, but no longer than for the other methods). Your report should be well structured, concise and to the point, but also accurate and provide complete information. In particular, conclusions should be reported with quantitative results and the reader should be able to understand how you analyzed the data without checking the software code (if this is not the case for one of your analyses, you may be assigned a zero score for that analysis, even if it was correct). Moreover, you should report whether the required assumptions for the analyses are met; when they are not, then attempt to give guidance how this might have affected the results (e.g., do you expect p-values to be under/overestimated). You should not repeat any statistical theory in your report (not even for the hypothesis tests related to the correlation). You should think carefully what results you will report. Any analysis results that you decide not to include in the final report but that were relevant during the data analysis process can be reported in a separate Appendix. Be careful in selecting graphs and tables to present in the body of your report; only informative graphs and tables that are essential for the argumentation towards your conclusions should be included. Other graphs and tables may be moved to an appendix. Also the cleaned R code of your analyses must be provided in an appendix and submitted as a separate file.

Except for the hypothesis tests about the correlations, for which you need to consult the literature, you are not expected to use methods that were not covered in the course (although you may have seen methods in other courses that are relevant for this project assignment). You are asked to conduct the project in groups of 4 students (groups less than 4 students are not allowed) - you can form the groups yourself. Feel free to divide some of the work, but make sure all students in the group understand all answers and all agree on the final results presented. After submission of the project you will be asked to grade your team members on their participation to the group work. Your own grade will take this part into account.

The project will be evaluated in two ways:

1. **Oral defense.** As a group you are asked to present your main results during a 'storm presentation' of 4 minutes (you can decide if one or multiple people present). If you want you can use slides or a poster, but know

that your time cannot exceed 4 minutes (after 4 minutes we will interrupt the presentation). Following this presentation, we will ask questions to all group members related to the project. In addition to the project content, we can also ask more general questions related to the content of the course (e.g. explain in your own words a fundamental concept such as a type I error, p-value, confidence interval, sampling distribution, etc). The presentations take place in **Week 13 (December 16-20, 2019)**.

2. **Written report.** Via Ufora you have to submit your final report as a PDF file with the name of the file should be the family names of the people in the group, ending with Project, e.g. 'DeNevePlevoets-Project.pdf'. Your R code should be commented and submitted as a separate file with the same name as the pdf report, but with the .R extension (e.g. 'DeNevePlevoets-Project.R'). We encourage you to modify your analyses and report in the case you received feedback during the oral defense. The final report should be submitted no later than **January 6, 2020 23h59**.

Good luck!