**Protocol Prinstat** Robin Boudry

**0.Data cleaning & prepping:**  
-remove observations with missing outcome values (1 obs)  
-clean data (fe Gender seems to have whitespace next to at least 1 obs, making R read it as separate class)  
-Create reasonable discrete age group  
-Create aggregated groups for organisms  
-Check classes of variables  
-Check if data assumptions hold (do all organism rows truly sum to 100?)  
  
-Create function based on cor test literature (maybe in methods section)

**1.Descriptive analysis  
-**Rapport on univariate frequencies (age, BMI, Gender, organisms)  
- Rapport on bivariate frequencies, in line with research questions  
-Rapport on possible multivariate frequencies, supporting research question  
-Rapport concise, clear, visual. Develop insight for hypothesis testing  
-Rapport normality of variables and subgroups (in appendix, will need for testing)

**2. Hypothesis testing***1.Does the genus composition change with age? (strat 1 & strat 2)  
2. How are the relative abundances of the 8 species related to one another?*

-Small sample, need of assumption verification and Nonparametric robust testing

**-**Before using created cor test function, test Type1, Power for simulated data as to understand strength/weaknesses of this test. Compare to known tests *(analogues HW3)*

-Rapport results

**3. Conclusions**-Support, refute Hypotheses based on testing  
-Do descriptives offer insight in (failed/signif) hypotheses?  
-What could have been different, would allow for more firm conclusions? What should new research aim to obtain based on what we saw?

**0. Amendments**Rapport any amendments we made to agreed protocol and reasons for it.

**Timeline (suggestion)**

Presentation 16-20 dec

Rapport 06 jan

**Tasks (suggestion)**

-Reasearch and develop cor test

-Compare cor.test with known tests

-Good extensive descriptive analysis

-Perform meaningfull tests, rapport, compare, conclude

-Review

-Prepare presentation

-Prepare rapport

Workflow (suggestion):

-10/12 analyses done and reviewed

-1/12, test developed and tested, prepping & descriptives done (2,2)  
 -8/12, hyp testing done, rapported and concluded (2)  
 -9&10/12 review the result, agree on, amend where needed (2)

-13/12 presentation done (2)

-27/12 rapport done (4)

#Summary of assignment  
0.Objectives (relevant for analysis)

1. learn **new statistical methods** from the literature

2. perform an **informative descriptive analysis** of a dataset **using informative**

**graphs and table**s

3. perform a **focused statistical data-analysis** using R

4. **report** **accurately** about your \_ndings

0.1 Data info

2 bacteria obtained from a sample of 40 subjects

(which were randomly selected from a larger sample of 200 volunteers).

analyses gave relative abundances of 8 microorganisms: 4 Staphylococcus

species and 4 Corynebacteria species. Age, gender, BMI (0: BMI \_ 25; 1: BMI > 25). Also included.

For each of the 8 species the dataset contains the relative

abundances (rescaled to make them sum to 100%). -> each row sums to 1 over the 8 organisms

1.Research Question

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 di\_erences (advantages/disadvantages) between

both strategies.

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

1.1 How to answer the question

Do this **without** using Linear **regression**, but

For answering the second research question you will need to know more about hy-

pothesis **tests** **concerning** **correlations**. Part of the assignment is to **consult** the

**literature** to \_nd out more about this topic so that you can apply it for answer-

ing 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.

*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).*