Principles of Statistical Data analysis: Bacteria population in armpit

Group 9: Jan Alexander[1], Paul Morbée [1], Steven Wallaert [1] and Joren Vereken[1]

[1] *1/4 of effort for data analysis and reporting*

December 2, 2019

Abbreviations

Abbreviation	
BMI	Body mass index
KW test	Kruskal-Wallis test
WMW test	Wilcoxon-Mann-Withney test
RBA	Relative bacteria abundance

1 Methods

1.1 Analysis protocol

The objective of the analysis is to determine if there is an association between a persons age and the ratio of malodour causing bacteria in his or her armpit. An important note regarding the data: Only relative abundances of bacteria are known. In this data set, no information is available on absolute abundances of the different bacteria species. This analysis is performed based on the following protocol steps:

Data Cleaning: Correct badly encoded observations. Treat missing values (see 1.2). Since the objective of the analysis is to investigate an association between age and relative bacteria abundance, observations where these variables must be omitted from the data set. Observations with missing BMI or gender values do not necessarily need to be omitted.

Data inspection: Quantify the distribution of the observations over age, gender and Body Mass Index (BMI).

Synthesising bacteria genera: Sum the different species concentrations of both genera to obtain the concentration for both of the genera.

Make age categories: the initial strategy was categorize the patient's age in 4 categories with boundaries 30, 40 and 50 years old. This resulted in very unbalanced categories.

Categorical age analysis: First, Kruskal-Wallis is used to determine weather all three age categories belong to the same distribution.

Continuous age analysis: ...

investigate other associations: Apart from the main focus of this work, the patient's age, the data set contains two other variables: BMI and gender. The marginal association between BMI and RBA and between gender and RBA will be investigated with WMW-tests.

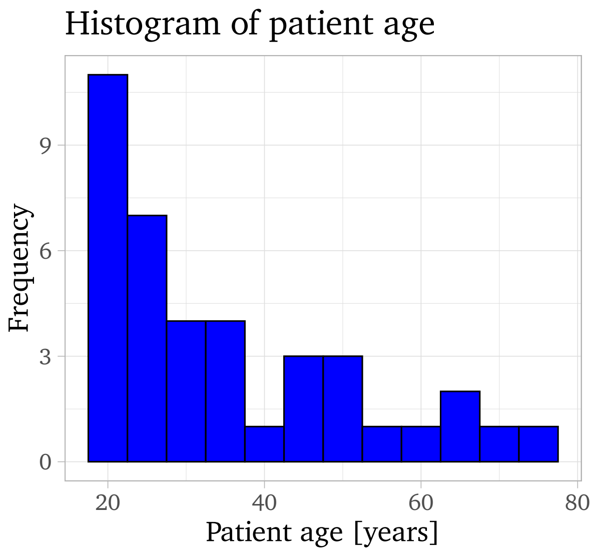

Figure 1: *Age Distribution*

1.2 Data cleaning and preparation

The data set contains 40 observations. The data was prepared for analysis in the following steps:

Badly encoded genders: The gender was badly encoded for some observations due to some trailing spaces.

Missing age values: 1 observation did not include a value for age or gender. This observation was scrapped.

Data anomalies: The relative quantities of both genera should end up to 100%. It was verified that this was the case.

Age categories:

Include table with distributions

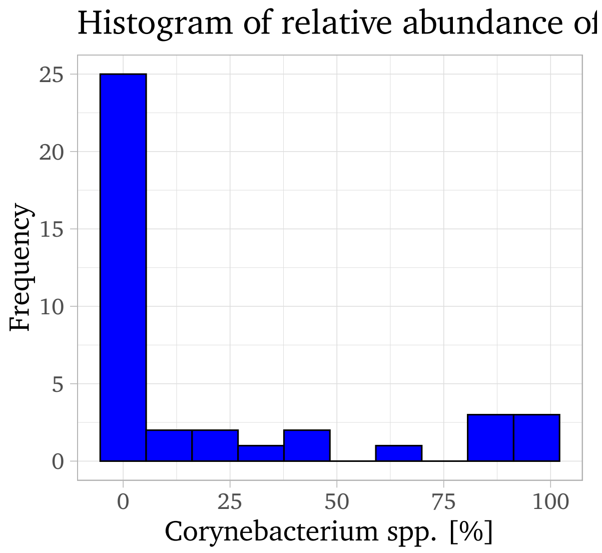

Figure 2: *Corynebacterium total distribution*

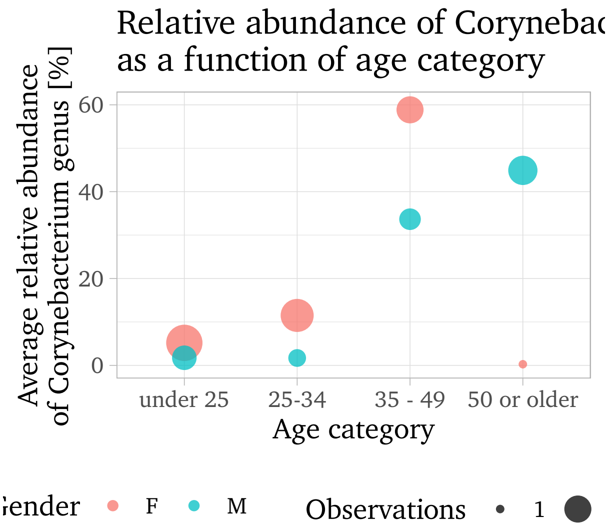

Figure 4: *Gender*

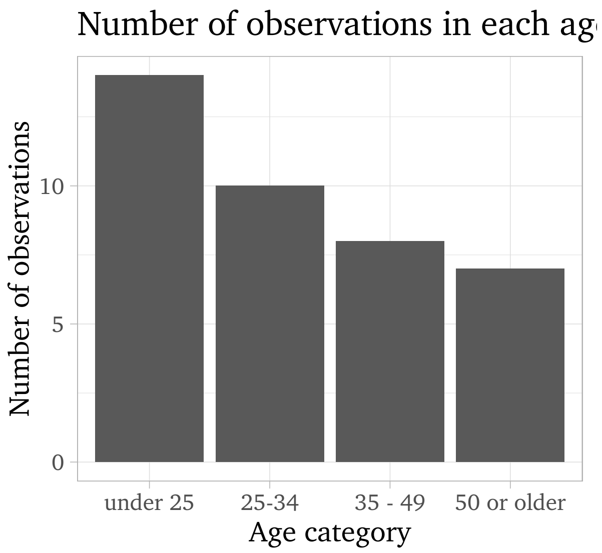

Figure 3: *Age cat distribution*

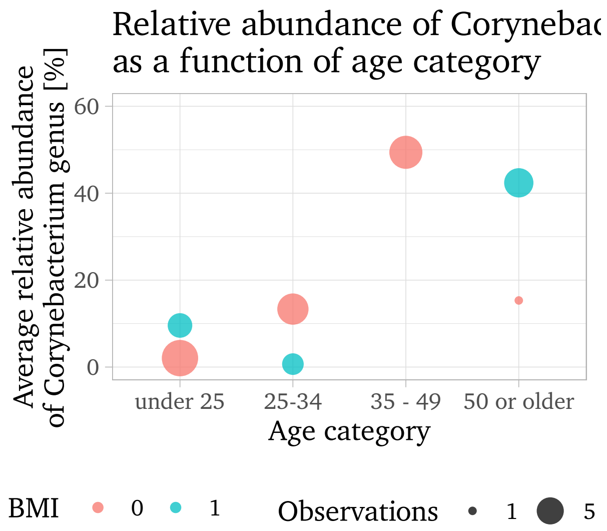

Figure 5: *BMI*

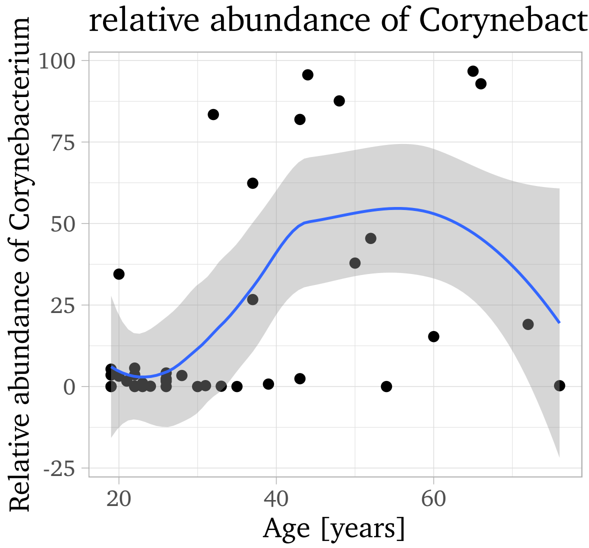

Figure 6: *plot scatterplot Age corby*

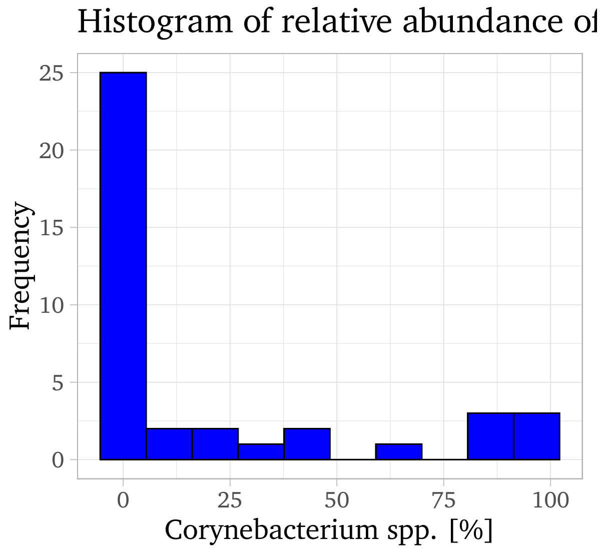

Figure 7: *plot Coryne totalDist*

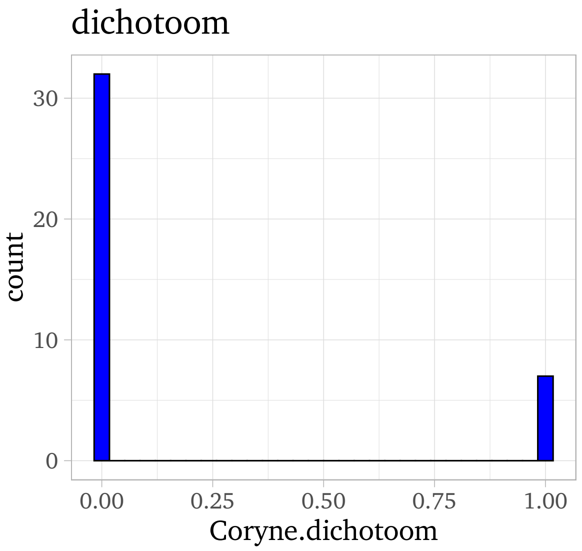

Figure 8: *Dichotoom*

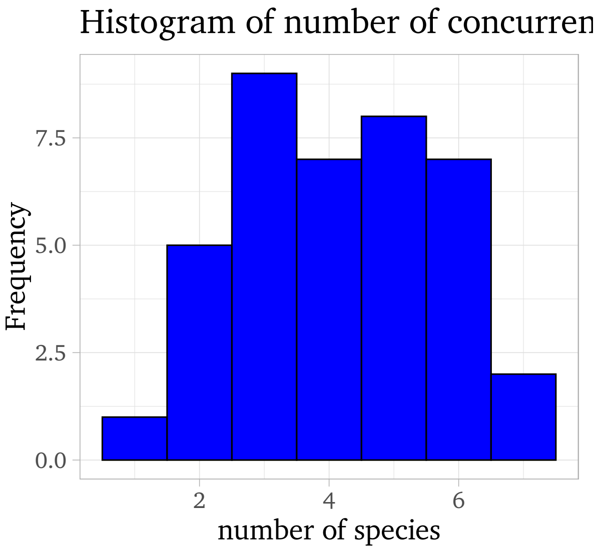

Figure 9: *JointOccurance*

2 Results

2.1 Age as a continuous variable

2.2 Age as a categorical variable

3 Conclusions and discussions

3.1 R code

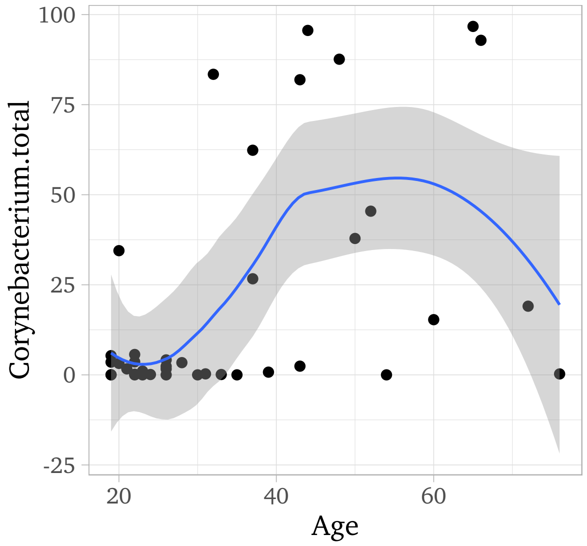

Figure 10: *plot1*

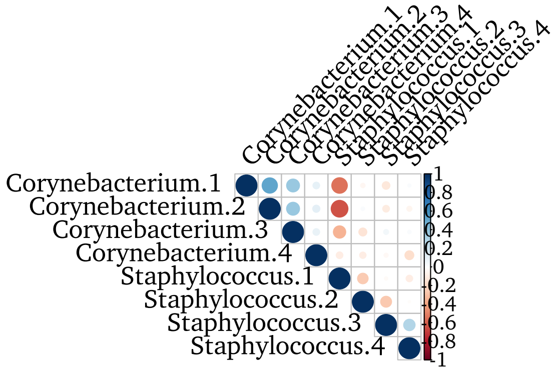

Figure 11: *correlation*

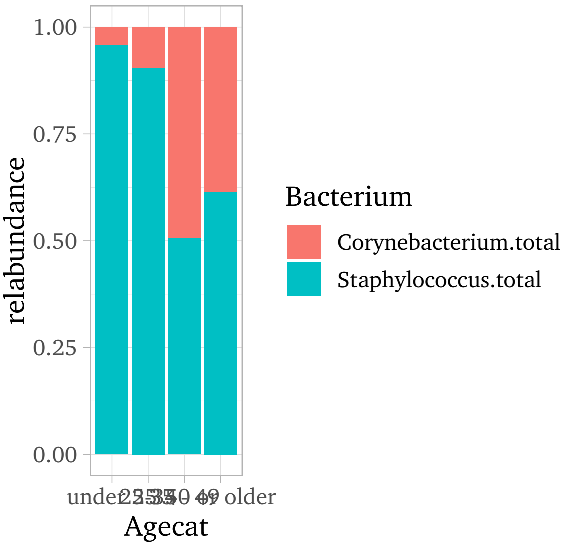

Figure 12: *plot3*