Principles of Statistical Data analysis - Project 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*

November 30, 2019

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.

This analysis is performed based on the following protocol steps:

Data Cleaning: Correct badly encoded observations. Treat missing values (see 1.2).

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

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

Make age categories: Initially, we planned to make categories with boundaries at 30, 40, 50 and 60 years old. This classification resulted in very unbalanced numbers of observations in the classes. We chose to reduce the number of age categories and split at

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

Continuous analysis:

1.2 Data cleaning and preparation

The data set contains 40 observations. The data

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.

Include table with distributions

2 Results

2.1 Age as a continuous variable

2.2 Age as a categorical variable

2.2.1 Age category definition

3 Conclusions and discussions

3.1 R code