Relative bacteria abundance in armpit microbiome

Principles of statistical data analysis
Group 9: Jan Alexander, Paul Morbée , Steven Wallaert and Joren Verbeke

UNIVERSITEIT GENT

Problem statement

Dataset `armpit.txt`: 40 subjects with 11 variables:

- 2 bacteria genera with each 4 species;
- subject **age** in years;
- subject **gender**: *'M'* or *'F'*;
- subject **BMI** indication: $BMI \leq 25$ or $BMI > 25$.

Study objective:

1. Investigate association between malodor causing bacteria and age.
2. Investigate association between different bacteria species.

Remarks:
Unknown subject: data intake step required to write protocol.
Only relative abundances of bacteria.

Protocol

Data intake & cleaning: Correct badly encoded observations and treat missing values.

Data inspection: Investigation of the distribution of patient age, gender and Body Mass Index (BMI).

Synthesising bacteria genera: Relative abundance of a genus is the sum of the relative abundances of the species of this genus.

Make age categories: Two groups with boundary at 40 years old: '$age \leq 40$' and '$age > 40$'.

Categorical age analysis: Wilcoxon MW-test to investigate a possible difference in Corynebacterium rel. abundance between the age groups.

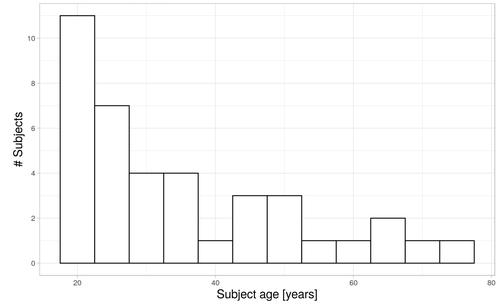
Investigate BMI and gender category: Investigate possible confounding with Kruskal-Wallis test.

Continuous age analysis: calculation of Kendall correlation coefficient between age and rel. abundance Corynebacterium.

Investigation of relative species abundance: Studied both as a binary variable (detected or not) with Fisher exact tests and a quantitative variable with Kendall-τ correlation.

Dataset statistics

Age distribution of the subjects is not normal. It does not correspond to the age distribution of the Belgian population.

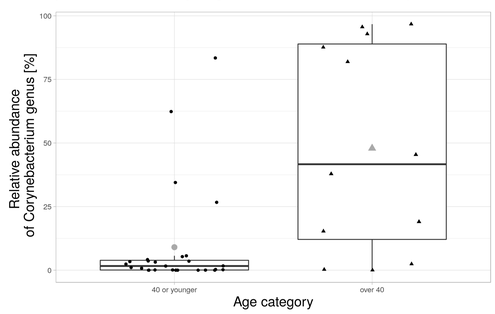


Genera abundance vs age as categorical variable

Subjects are categorized in *40 or younger* and *over 40*.

BMI	Gender	$\leq$ 40 y.o.	>40 y.o.	Total
BMI ≤ 25	F	**14**	**4**	18
BMI ≤ 25	M	**6**	**2**	8
BMI > 25	F	**5**	**1**	6
BMI > 25	M	**2**	**5**	7
Total		27	12	39

H_0 : *no location shift between age groups* could be rejected with $p < 0.01$.



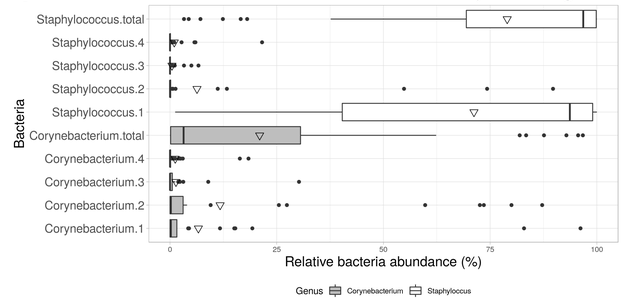
Kruskal-Wallis test: No ground in data to reject H_0 for interaction groups between BMI and gender.

Genera abundance vs age as continuous variable

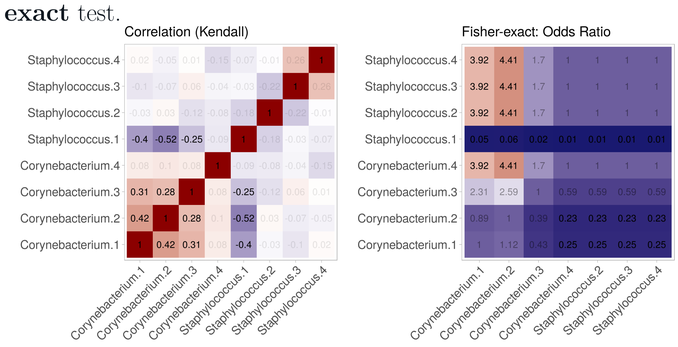
Association investigated with **Kendall-τ**: Non-normal distribution & high number of ties (0 values):
$CI_{95\%}(\tau) = [0.05; 0.48]$ with $p < 0.025$

Association between species

Species show extreme distributions: Practically 0 or high.



Association between species is investigated with **Kendall-τ** correlation coefficient and by treating the species detection as a binary variable. In the second case, the relation is investigated with **Fisher-exact** test.



Remark: For relative variables, the correlation coefficient is influenced by making the values relative.

References

- Chen, P. Y., & Popovich, P. M. (2002). Correlation, Parametric and Nonparametric Measures pag 84.
- Introduction to categorical data analysis; - `in Ysebaert M., Introductory Statistics, ICES Ghent University 2010-2011`
- Census België, 2011 - `https://www.census2011.be/data/fresult/age-avg_nl.html`