Principles of Statistical Data analysis: relative bacteria abundance 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 14, 2019

Contents

Abbreviations

Abbreviation	
BMI	Body mass index
IQR	Interquartile range
KW test	Kruskal-Wallis test
WMW test	Wilcoxon-Mann-Withney test
RGA	Relative genus abundance
RSA	Relative species 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 the armpit microbiome.

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 are missing must be omitted from the data set. Observations with missing BMI or gender values do not necessarily need to be omitted.

Data inspection: Investigation of the distribution of patient age, gender and Body Mass Index (BMI). The objective of this step is to identify possible bias in the data. Distributions can be visualized with boxplots or histograms.

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

Investigate BMI and gender category: 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 RGA and between gender and RGA will be investigated with WMW-tests.

Make age categories: The age was categorized by making two groups with boundary at 40 years old: '$age \leq 40$' and '$age > 40$' (see 2.1.2).

Categorical age analysis: First, WMW-test is used to investigate a possible difference in Corynebacterium RGA distribution between the age groups[1].

Continuous age analysis: Study the association between RGA and age as a continuous variable was studied with the Kendall correlation coefficient. The Kendall correlation coefficient was calculated because the data contained many ties and the distribution showed a strong deviation from normality. TODO : ALL - Vraag: gingen we nu ook relabundance dichotomiseren? Is gebeurd in de code en er zijn enkele grafieken / test uitgevoerd, maar mss is het niet meer nodig? Eenvoud siert mss wel...

Investigation of RSA: Relative species abundance was studied both as a binary variable (detected or not) and a quantitative variable. First, the number of detected species per subject was studied by plotting a histogram. Next, Fisher exact tests were performed for each pair of species to investigate if certain species were more or less likely to occur together. Finally, Kendall correlation coefficients were calculated and tested as described higher to

[1] Since there are only two genera, the relative abundance of Staphylococcus genus is fully characterized with the abundance of Corynebacteria genus.

explore if relative abundance of one species was correlated with abundance of another species. No correction for multiple comparisons was done.

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.

Unless further specified, *Statistically significant* should be interpreted as $p \leq 0.05$ in this report.

1.2 Data cleaning and preparation

The provided 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 trailing spaces, e.g. *'F '* was re-encoded to *'F'*.

Missing age values: One observation did not include a value for age (nor gender). This observation was omitted, as was explained in §1.

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

After these operations, the final data-set used for the analysis in this document was obtained. The distribution of the observations over the categories BMI and Gender is shown in table 1.

Gender	BMI ≤ 25	BMI >25	Total
F	18	6	24
M	8	7	15
Total	26	13	39

Table 1: *Distribution over BMI and gender*

Figure 1, shows the heavy skewness of the patient's age distribution. There are few older patients in the data set.

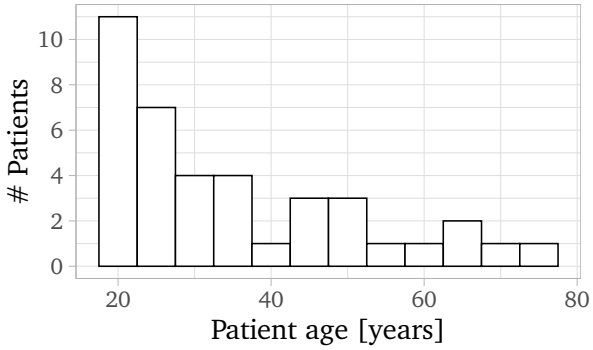
2 Results

2.1 Relative abundance of genera

Since the armpit microbiome only contains two genera of bacteria, the relative abundance of one defines the other. Malodor is caused mainly by Corynebacteria. The analysis focuses on the relative abundance of this genus. The abundance of the Staphylococcus genus is simply $100\% - RGA$ of Corynebacteria.

2.1.1 Basic statistics

Table 2 contains the basic statistics of the relative distributions of the different bacteria species. The RGA shows a strongly skewed distribution. The relative



min	Q1	median	mean	Q3	max
19	22	30	35	43	76

Figure 1: *Histogram of age distribution*

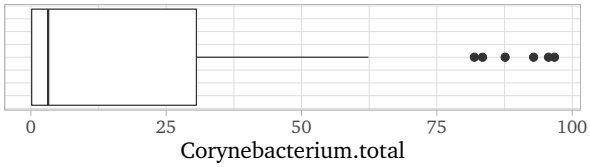


Figure 2: *Boxplot of the relative abundances of the Corynebacteria genus in all 39 observations*

Corynebacteria abundance is very low (only 3.2%) for at least half of the patients. Nevertheless, the armpit microbiome relative concentration of Corynebacteria can be as high as 97%. This is visualized in figure 2.

Within the same genus, the skewness of the data is clear from the second part of the table. The *Staphylococcus* species 2, 3 and 4 are completely absent in at least half the samples. Nevertheless, the RSA of *Staphylococcus.2* can reach up to 89.7%.

2.1.2 Age as a categorical variable

Due to the relatively low number of observations, it was chosen to deviate from the original plan to divide the patients in 4 categories divided at 30, 40 en 50 years old. This resulted in categories containing a very low number of observations. This would have yielded the power of the statistical tests undesirably low. TODO : All - Is this correct? Is this the best way to put it? Instead, the patients were split in 2 age categories: *40 and younger* and *over 40*. The age of 40 years is chosen because:

- The mean age of a Belgian is 40.8 years [2].
- 40 is close to half of the oldest patient age and the mean participant age.

The difference in RGA of Corynebacteria can be observed in figure 3 and from table 3. For most people of maximum 40 years old in this data set, a lower RGA

[2] Census België, 2011 - `https://www.census2011.be/data/fresult/age-avg_nl.html`

Genus	mean [%]	median [%]	sd [%]	min [%]	max [%]
Corynebacterium.total	21	3.2	33	0	96.7
Staphylococcus.total	79	96.8	33	3.3	100
Species	**mean [%]**	**median [%]**	**sd [%]**	**min [%]**	**max [%]**
Corynebacterium.1	6.6	0.2	20.1	0	96.2
Corynebacterium.2	11.7	0.2	25.4	0	87.2
Corynebacterium.3	1.4	0	5	0	30.2
Corynebacterium.4	1.2	0	3.9	0	18.4
Staphylococcus.1	71.2	93.7	36.2	1.2	100
Staphylococcus.2	6.3	0	20.1	0	89.7
Staphylococcus.3	0.5	0	1.4	0	6.7
Staphylococcus.4	1	0	3.6	0	21.6

Table 2: *Relative abundance statistics for both genera and all species TODO : JAL - Uniform formatting numbers*

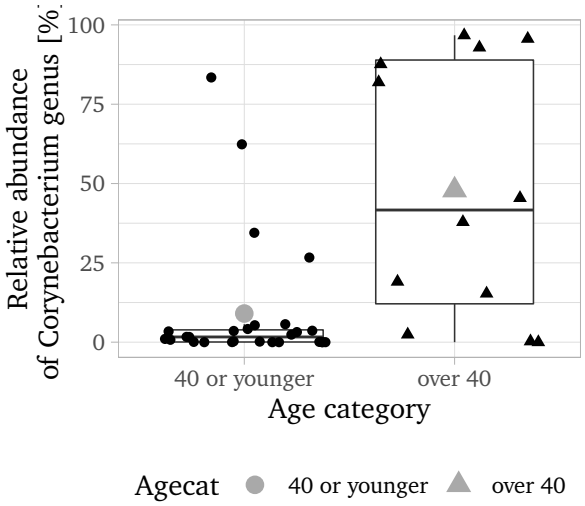


Figure 3: *Boxplot of Corybacteria RGA per age category. The points indicate the individual observations. Grey markers indicate the mean for both groups.*

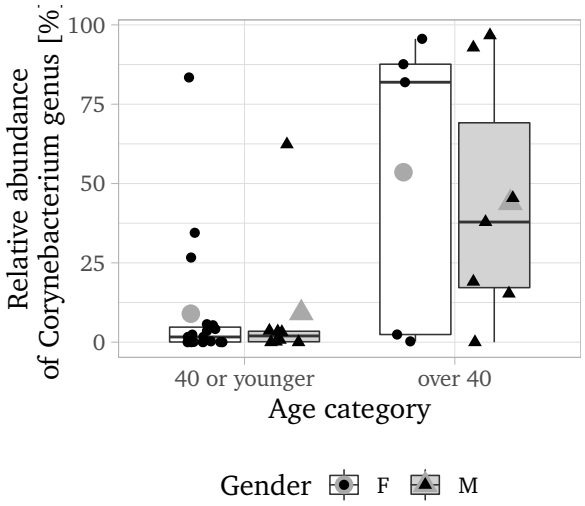


Figure 4: *Boxplot of Corybacteria RGA for both age category and gender. The points indicate the individual observations. Grey markers indicate the mean for both groups.*

(<10%) of Corynebacteria is observed (with 4 exceptions). The distribution of RGA of Corynebacteria for people over 40 years old shows more variability, the IQR is higher. The center of this distribution (both the median and the mean) is higher. The (two-sided) WMW-test indicates a p-value < 0.01, indicating that H_0 (no location shift between category *'40 or younger'* and category *'over 40'*) should be rejected.

Apart from the patient's age, two other patient characteristics are known:

1. Whether the patient's BMI is over 25.
2. The patient's gender.

Before concluding we observed a location shift between the relative bacteria genus abundances in the armpit microbiome in different age categories, we wish to investigate the possible influence the two other mentioned parameters may have. Secondly, we want to investigate if there is a significant difference between both age categories in BMI or gender composition.

In table 4, the distribution of the patients in the categories is shown.

TODO : JAL - Distribution observations (BMI and gender) within age categories –> KW

2.1.3 Age as a continuous variable

TODO : Steven, Joren - Tekst en tabel

TODO : Paul - Tekst, code en figuren voor bootstrap methode

2.2 Relative abundance of species

Look again at table 2. See figure 6.

		Corynebacteria genus, relative abundance					
Age category	n	mean	min	median	max	n present	% present
≤ 40	27	9	0	1.6	83.4	21	77.8
> 40	12	47.9	0	41.7	96.7	11	91.7

Table 3: *Difference in relative abundance of the Corynebacteria genus between the two age categories*

BMI	Gender	≤ 40	>40
BMI ≤ 25	F	14	4
BMI ≤ 25	M	6	2
BMI > 25	F	5	1
BMI > 25	M	2	5
Total		27	12

Table 4: *Number of patients in each combination of categories*

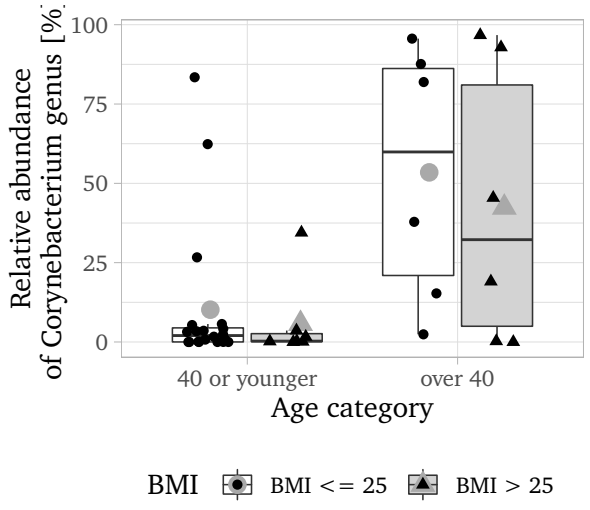


Figure 5: *Boxplot of Corybacteria RGA for both age category and BMI. The points indicate the individual observations*

TODO : Paul - herhalen van theorie in de tekst was eigenlijk specifiek verboden in de opgave. Mag dit wat worden herschreven om de expliciete theorie eruit te laten? In the absence of regression the consideration of correlation seems promising to compare the possible associations between the different genus and species. The study of the literature revealed that there are basically 3 major correlation methods, namely Pearson (r coefficient), Spearman (ρ coefficient) and Kendall (τ coefficient). Pearson's r is essentially a measure of linearity and requires the data sets to be continuous. Moreover, for r to be significant, normality of the data sets is assumed. Which clearly is not the case for the different abundances TODO : all - (QQ-plots tonen ?). And if we consider nominal data (the categories of Age for example) Pearson's r is no longer of use. Spearman ρ converts the data sets to ranks which are afterwards subjected to a Pearson correlation test to calculate the ρ value. Spearman's ρ appears to be sensitive to ties in the data, which is unfortunately the case in the armpit relative abundances (many zeroes). In this case literature points out that Kendall's τ is more appropriate as a measure of association. Especially when the number of observations is limited. Kendall's τ basically calculates a measure of concordance between couples of data elements from the different sets. So for these reasons we opt for the use of Kendall's τ measure. To assert correlation between data sets we need to calculate a p-value for the hypothesis test H0: $\tau = 0$ versus Ha: $\tau \neq 0$. If the number of observations is > 10 this can be done by doing a z-test, which we have implemented. We have also calculated the CI. As the standard R-functions to calculate correlations (such as cor() or cor.test()) do not provide confidence intervals we obtained these through a bootstrapping procedure.

In table 7 TODO : JAL - insert table we depict the significant correlations ($\alpha \leq 0.05$) in the armpit data frame:

Generally we observe that, if there is a significant correlation, the Staphilococcus species and Corynebacterium species internally correlate positively. Between the genus and their species, where significant, they associate in a negative way. This is explained by the aging process discussed earlier but we have also to consider that the different species add up to 100% which in itself reinforces the negative correlation between the two genus. This is by definition shown by the almost perfect negative correlation between the totals of the two genus. Surprisingly, as a secondary effect, we see that Age and Gender (ordered F, M) positively correlate

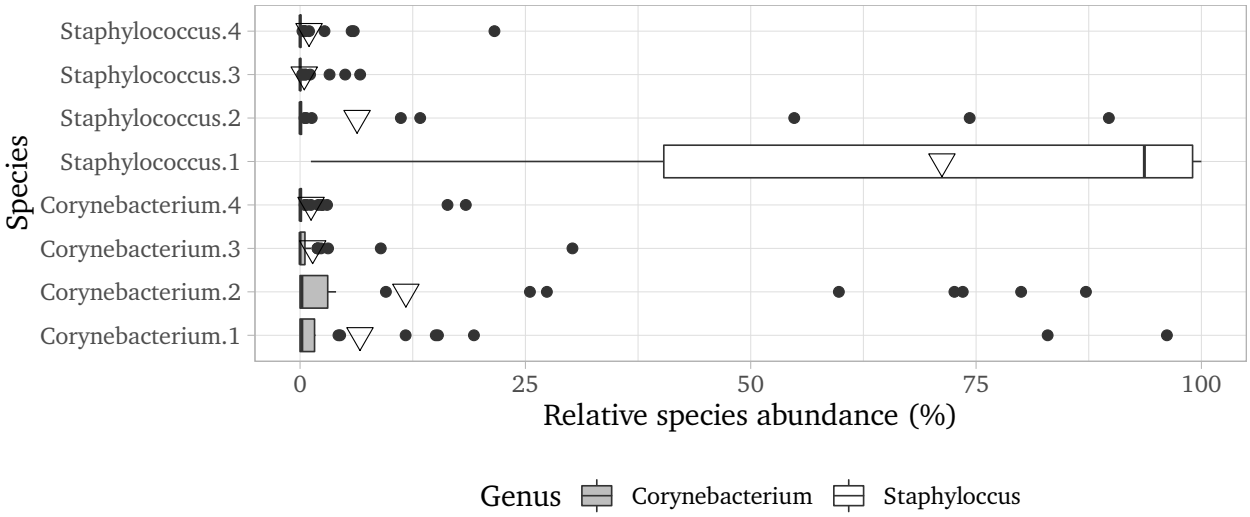


Figure 6: *Boxplots indicating the variation of relative abuncances of different species in the data*

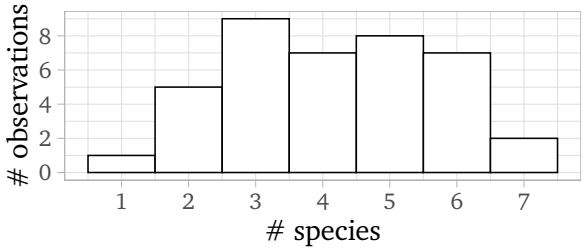


Figure 8: *Number of species in individual observations*

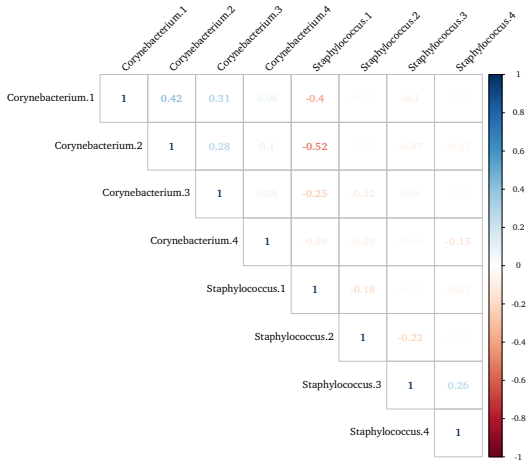


Figure 7: *Correlation plot TODO : all - mss beter letters in zwart en achtergrond van vakken in kleur?*

which confirms that woman are poorly represented in the older Age categories.

3 Conclusions and discussions

TODO : All - Discussion In this study, the association between relative bacteria abundance in the armpit microbiome and age was studied. This study provides an indication that the relative abundance of the Corynebacterium genus tends to increase with age. No evidence for influence of BMI and gender on the relative genus abundance in the armpit microbiome could be found in the data set.

It is important to remark that the study was realized on a data set which was highly biased towards younger individuals. Although biological insight can be gained, the results should not be generalized to the total population.

3.1 R code

TODO : all - code cleanup afterwards

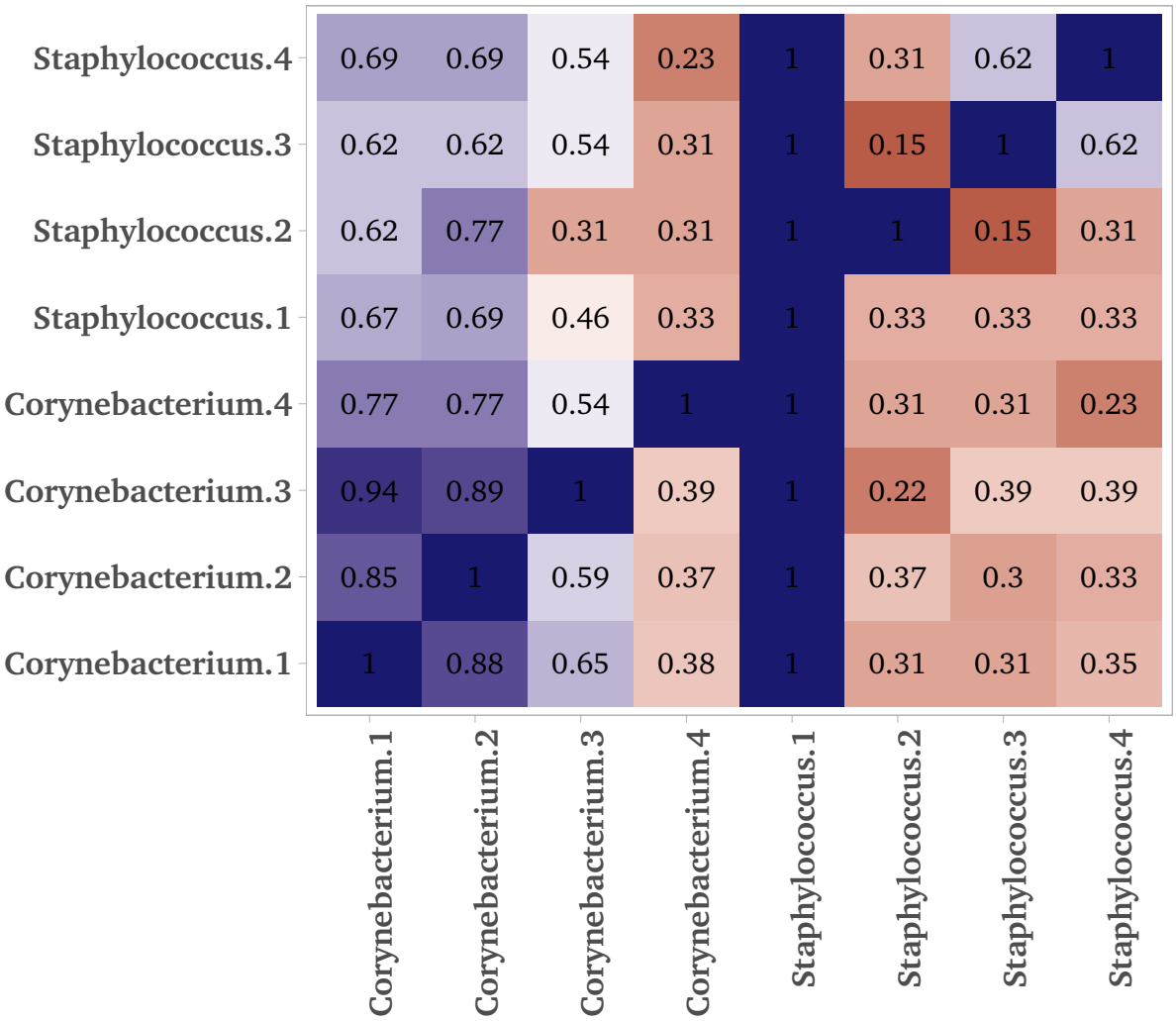


Figure 9: *conditional Probability of joint occurance*

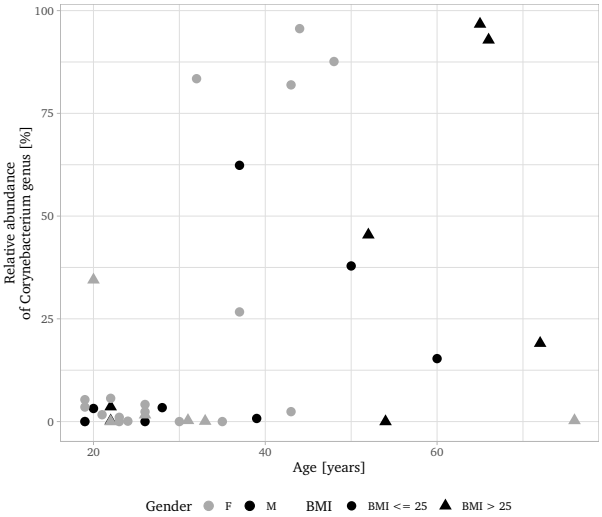


Figure 10: *Scatterplot of the Corynebacteria abundance*

4 Unused graphs