Principles of Statistical Analysis Project

Analysis of 8 microorganisms living in the armpits

written by:

Boudry Robin Tang Lin

Meron Mengistu Vandenbulcke Stijn

# 1. Methods

## 1.1 Data Cleaning & Transformation

Data of 40 subjects was provided, containing information about their BMI, age, gender and relative abundances of 8 bacterial species (4 Corynebacterium and 4 Staphylococcus) present in their armpit. 1 observation was removed due to missing data. The categorical variable ‘Gender’ had whitespaces before the letters, these were removed. For readability and easier coding the names ‘Corynebacterium 1’ were shortened to ‘Cb1’ as was ‘Staphylococcus 1’ shortened to ‘St1’.

Moreover, the continuous variable Age is made a discrete variable ‘Agecat’ as a categorical variable for the categorical analysis between Agecat and genus composition. The Agecat is a 3 level ordered factor variable made up by Junior (<=26), Adult (27<=age<40), Senior (>=40), these bins were chosen to have roughly equal sample sizes 14, 13 and 12 respectively.

## 1.2 Descriptive analysis

When making boxplots of the species (appendix) we see the relative abundance of Staphylococcus 1 being dominant with a mean of 71.21 and a median of 93.67. This skew is somewhat mediated by using a log transformation. There are 24 females and 15 males in the sample. When plotting the genus with age, we see Corynebacterium having a cluster of low values in the younger age group, and Staphylococcus having a cluster of high values in the younger age group.

## 1.2 Categorical variable analysis

## 1.2.1 Interaction among factors

We should confirm that there is no interaction among the other factors (Agecat, Gender, BMI) before we do the ANOVA for each of them. We used the two-way ANOVA and interaction graphs to check if the interactions are statistically significant. If the interactions between Agecat and other two factors are statistically significant, the Agecat variable cannot be analyzed. In this situation, we have to take consider of the interaction of both factors rather than the individual factor. If there are no interaction between each two of them, then we can do the one-way ANOVA to Agecat.

## 1.2.2 One-way ANOVA

The sample sizes of each Agecat are unequal, the normally one-way ANOVA and POST HOC test, such as Fisher’s LSD and Tukey’s HSD, are not available for this situation. So we choose the Kruskal–Wallis rank sum test with the null hypothesis that the location parameters of the distribution are the same in each group, with an alternative hypothesis that they differ in at least one. If the Kruskal-Wallis test returns a significant result, we perform a pairwise Wilcoxon rank sum test to find the groups that are significantly different from each other. Benjamini-Yekutieli is used for the multiple testing adjustment.

## 1.3 Relation of 8 relative abundances amongst each other

## 1.3.1 choice of test

We are confronted with a small sample size and since no information is given about the underlying distribution, we have to rely on non-parametric methods to make valid inferences. For correlation testing, we consulted Hauke & Kowoski (2011)’s comparison of correlation tests, and concluded that Spearman rank was suitable for our problem. To better understand the Spearman Rank correlation significance test, we simulated data with the same sample size (n1=n2=13) which resembles what we observed and with higher sample sizes. We noted that smaller sample sizes take higher correlation estimates for a significant p-value. We also checked the Type1 error in a simulation, and saw it being controlled at the specified 0.05 level for our Spearman Rank correlation test, similar to more established tests.

## 1.3.2 Hypothesis and testing procedure

The null hypothesis can be stated as follows: Spearman’s ρ = 0 (meaning that there is no monotonic association between the two variables) and alternative Spearman’s ρ≠0. This testing can be done with the help of a permutation test. For each combination of species a permutation distribution is made, by permuting the relative abundances and calculating the Spearman correlation. The package “Psych” allows us to use a faster approximation, which calculates a p-value and confidence intervals via the asymptotic t approximation using the formulas below. In the appendix we show that this test performs equivalent to a permutation test in our dataset.

## 1.3.3 Multiple testing adjustment

To adjust for the multiple testing problem, we use the Benjamini-Yekutieli (BY) method, this is a less conservative correction than Bonferroni or Holms-Bonferroni (HB), which gives us more statistical power. Correcting for multiple testing will lead to higher p-values. However in our case, the amount of significant pairs did not change whether BY or HB was used.

# 2. Results

## 2.1 Age as a categorical variable

## 2.1.1 Interaction analysis

The interaction graphs show that there might be a possible interaction between Agecat and Gender/BMI for both Staphylococcus and Corynebacterium. We fitted an analysis of variance model (aov) for CB/ST~Agecat\*Gender and CB/ST~Agecat\*BMI. The values for Staphylococcus and Corynebacterium are the same, this makes sense as relative abundances are complementary.

The Age category is significant for both gender and BMI with p-values of 0.00159 and 0.00133 respectively, while none of the other terms are significant. From this we conclude there is a significant relation between the relative abundance of Staphylococcus/Corynebacterium and age category that is not influenced by either gender or BMI.

Hence we can conclude the effect of the categorical age classes stays significant on the presence of Staphylococcus, even when adding Gender and BMI with their respective interactions.

## 2.1.2 Kruskal-Wallis and Wilcoxon rank sum tests

The Kruskal-Wallis rank sum test returned a p-value of 0.008403 for Staphylococcus in relation to the age categories and a value of 0.01557 for Corynebacterium. There is a significant difference between the age categories for both genus. As there is a significant difference, we performed a pairwise Wilcoxon rank sum test with results Junior:Adult p-value of 1, Junior:Senior 0.014 and Adult:Senior 0.052. There is a significant difference in distribution of the relative abundance between the junior and senior age category, however the p-value for the adult and senior category is close to 0.05.

## 2.2 Age as a continuous variable

To verify if genus composition changes with age we use smooth line plot for each species in the two genus groups. However it is noticed that availability of relative abundances of all four species are not equal. For instance, from Corynebacterium , Corynebacterium.2 is most dominant species, followed by Corynebacterium.1 ( figure9, appendix). In the Corynebacterium.1, the relative abundancy keeps constant value of 0% at young ages and suddenly increases at around 30 year. It reaches maximum (approximately 41%)in age between 40 and 60 years, where after 60 year it gradually drops to 0%. In Staphylococcus genus the dominant species is staphylococcus.1 (figure9, appendix). In opposite to Corynebacterium at young age it attains higher relative abundance (greater than 75%), which gradually decreases till age of 40 year. The relative abundance remains constant at its minimum (below 50 ) between age of 40 and 60 years, and increases for ages older than 60 year.

For ease of interpretation and comparison between two Genus groups, we aggregate each of the four species in to one respective group, Corynebacterium and Staphylococcus. The result shows that the distribution of aggregated genus groups follows the dominant species distribution. As a result for Corynebacterium ( figure 1, B) the relative abundancy gets minimum at the early age of around 20, and suddenly increases at around 24 year. It gets maximum approximately at age 43 and 60 years. It drops gradually for ages older than 60 year. In case of Staphylococcus ( figure 1, A) the relative abundancy is high at early age of about 20 and gradually decreases with increasing age. Lower relative abundancy for this genus is attained when age is between 40 and 60. For older ages than 60 the relative abundancy starts to increase again. It is clearly seen that the change in relative abundancy with age is not linear. Consequently the correlations between the aggregated relative abundancy and age is calculated using the spearman correlation test in R. correlation coefficient obtained for Corynebacterium is 0.40 with p-value of 0.012 and confidence interval of (0.09 to 0.63) and Staphylococcus has a coefficient of -0.41 with p-value of 0.009 and a confidence interval of (-0.64 to -0.11). There is a positive association between the relative abundances and age for Corynebacterium and a negative one for Staphylococcus.

Afbeelding met tekst, kaart

Automatisch gegenereerde beschrijving

Figure 1: Smooth line plots for Staphylococcus (A) and Corynebacterium (B) plotting their relative abundances (%) in relation to age (year)

## 2.3 Relation of relative abundances of the 8 species to one another

The relation of the relative abundances of the species was investigated by using the spearman rank correlation (see methods). The heatmaps (figure 2) show that overall the correlation coefficients are small, the six significant pairs in the non-adjusted heatmap are [Cor 1 with Cor 2, Cor 3], [Cor 2 with Cor 3], [Staph 1 with Cor 1, Cor 2, Cor 3]. Adjusting for multiple testing leaves three significant pairs [Cor 1 with Cor 2] and [Staph 1 with Cor 1, Cor 2] (Figure 3). The confidence intervals however are very wide, therefore we cannot conclude anything about the magnitude of the association, only about the direction (Figure 3). From the significant pairs, Staph 1 is negatively correlated with Cor 1 and 2, while Cor 1 and 2 are positively associated with each other.

Afbeelding met schermafbeelding

Automatisch gegenereerde beschrijving

Figure 2: Heatmaps showing the Spearman’s ρ in colour with its p-value as a number. A: p-values not adjusted B:p-values adjusted with the Benjamini-Yekutieli method

Afbeelding met schermafbeelding

Automatisch gegenereerde beschrijving

Figure 3: Confidence intervals Spearman’s ρ for each of the species pairs, adjusted with the BY method. Significant confidence intervals in red

# 3. Conclusion

We received a dataset consisting of 8 species, four from the genus Corynebacterium and four from Staphylococcus. In this report we investigated if there is an association between the relative abundances of the genus and age. This analysis was performed with age as a categorical variable consisting of junior, adult and senior categories and using age as a continuous variable.

The categorical analysis showed that there is indeed a significant difference between these three groups in terms of relative abundances per genus and age category. Further analysis showed a significant difference in relative abundance distributions between junior and senior age category. The p-value for the comparison between adult and senior returned a p-value of 0.052, although it is close to the 0.05 cut-off we could not reject the null hypothesis. It might however be interesting to further investigate this difference.

The continuous analysis showed that there is a significant monotonic relationship between age and relative abundance per genus, with a positive association for Corynebacterium and a negative association for Staphylococcus. The confidence intervals however are fairly wide (0.09 to 0.63) and (-0.64 to -0.11), therefore we can only conclude that there might be a weak to a mild association.

There was also a relation between the relative abundances of the 8 species, with three pairs having significant association [Cor 1 positive association with Cor 2] and [Staph 1 negative association with Cor 1, Cor 2]. A possible cause for this is if Staph 1 grows in the same microbial environment as Cor 1 and 2 and are therefore competitors. Cor 1 and 2 are positively associated, this could be because of either a symbiotic relation or because the environment in which they thrive is slightly different from one another, which allows them to grow together and not compete. However the confidence intervals here are even wider than the continuous analysis we can only conclude the direction but not the magnitude of the association.

As the observations were anonymized it is not possible to identify their origin, in case all the participants live in Belgium then these results can only be concluded for humans that live in Belgium. In addition, in this analysis most of the people are younger than 40, and we recommend repeating the analysis for a homogenous data sample.

# 4. Appendix

## 4.1 Descriptive analysis

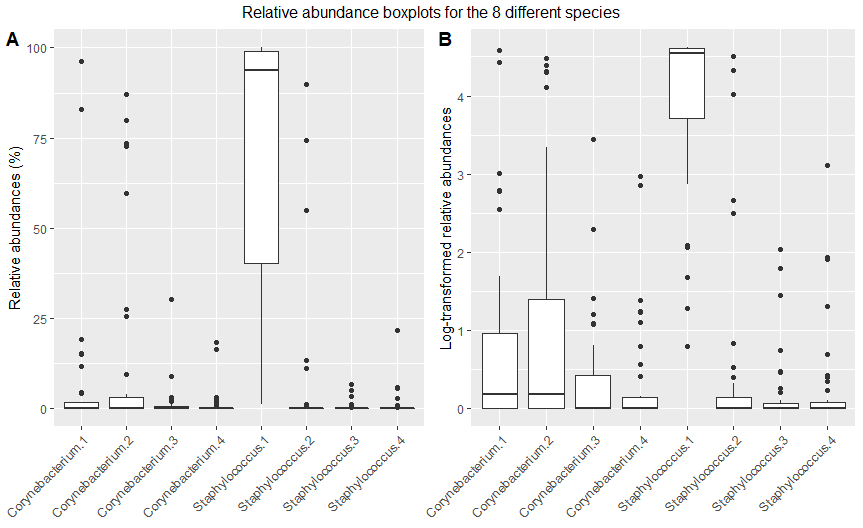


Figure 4: boxplots of the relative abundances (in % for A, Log-transformed for B) for the 8 different species.

## Afbeelding met tekst, kaart Automatisch gegenereerde beschrijving

Figure 5: Scatter plot relating the relative abundances (%) of the geni to age

## 4.3 Age as categorical variable

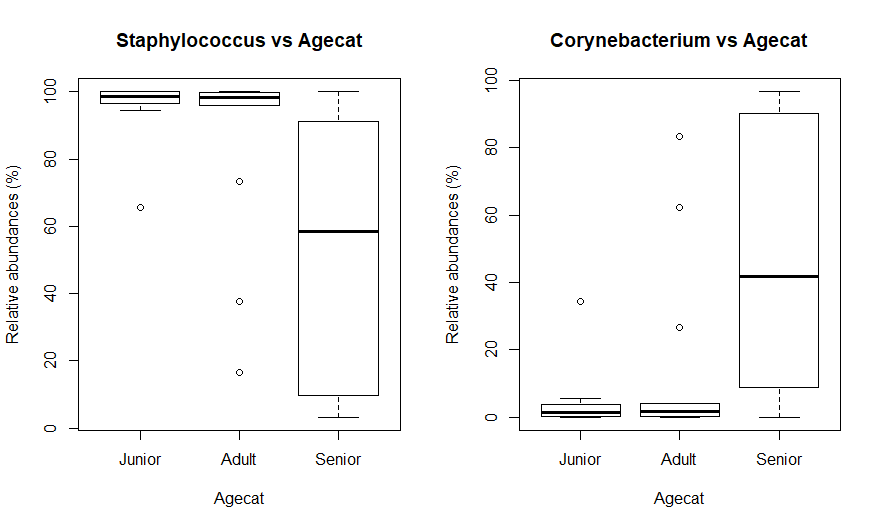


Figure 6: Boxplots of relative abundances (%) of Staphylococcus and Corynebacterium vs age categories

Afbeelding met kaart, tekst

Automatisch gegenereerde beschrijvingAfbeelding met kaart

Automatisch gegenereerde beschrijving

Figure 7: Interaction plots of the relative abundances (%) of Staphylococcus and Corynebacterium vs Gender/BMI

Tabel 1: Analysis of variance for Staphylococcus relative abundance ~ Age category \* gender

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ST~Agecat\*Gender | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Agecat | 2 | 13244 | 6622 | 7.885 | 0.00159 |
| Gender | 1 | 95 | 95 | 0.113 | 0.73865 |
| AgeCat:Gender | 2 | 244 | 122 | 0.145 | 0.86531 |
| Residuals | 33 | 27715 | 840 |  |  |

Tabel 2: Analysis of variance for Staphylococcus relative abundance ~ Age category \* BMI

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ST~Agecat\*BMI | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Agecat | 2 | 13244 | 6622 | 8.152 | 0.00133 |
| Gender | 1 | 337 | 337 | 0.415 | 0.52404 |
| AgeCat:Gender | 2 | 910 | 455 | 0.560 | 0.57636 |
| Residuals | 33 | 26806 | 821 |  |  |

Tabel 3: Analysis of variance for Corynebacterium relative abundance ~ Age category \* gender

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CB~Agecat\*Gender | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Agecat | 2 | 13244 | 6622 | 7.885 | 0.00159 |
| Gender | 1 | 95 | 95 | 0.113 | 0.73865 |
| AgeCat:Gender | 2 | 244 | 122 | 0.145 | 0.86531 |
| Residuals | 33 | 27715 | 840 |  |  |

Tabel 4: Analysis of variance for Corynebacterium relative abundance ~ Age category \* BMI

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CB~Agecat\*BMI | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Agecat | 2 | 13244 | 6622 | 8.152 | 0.00133 |
| Gender | 1 | 337 | 337 | 0.415 | 0.52404 |
| AgeCat:Gender | 2 | 910 | 455 | 0.560 | 0.57636 |
| Residuals | 33 | 26806 | 821 |  |  |

## 4.2 Age as continuous variable

Afbeelding met kaart, tekst

Automatisch gegenereerde beschrijving

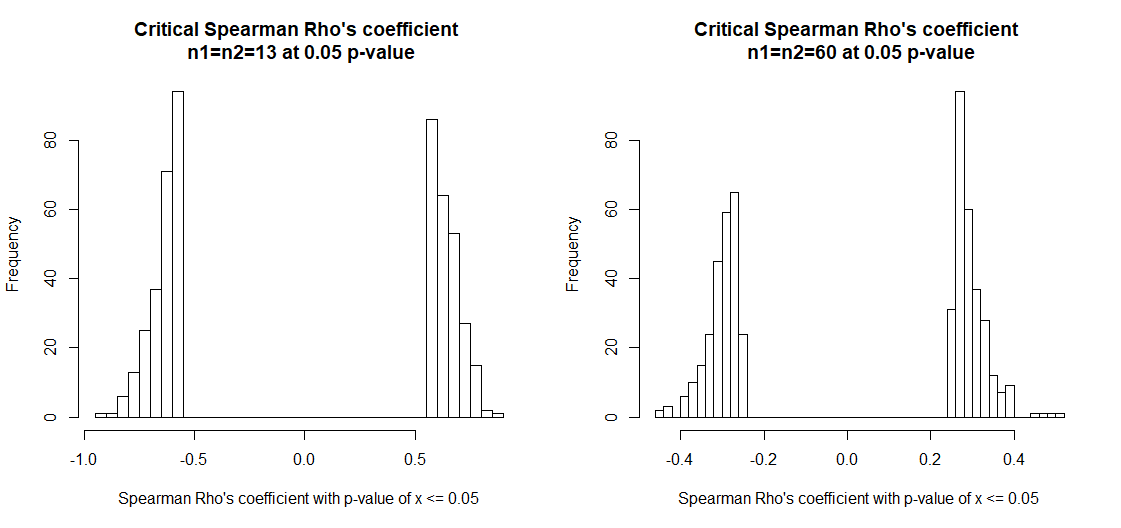
Figure 8: Smooth line plots for each Corynebacterium species in relation to age

Afbeelding met tekst, kaart

Automatisch gegenereerde beschrijving

Figure 9: Smooth line plots for each Staphylococcus species in relation to age

## 4.2 Spearman Rank test analysis



Afbeelding met schermafbeelding

Automatisch gegenereerde beschrijving

## 4.2 Comparison permutation test and asymptotic T approximation

To show that this test performs equivalent to a permutation test with n = 10000, we performed a permutation test for the correlations between [Corynebacterium 1, Staphylococcus 1] and [Staphylococcus 3, Staphylococcus 4], these were chosen as in the former their distributions are not alike, while in the latter case they are more similar. From the results we can see that both permutation distributions are normally distributed, aside from the permutation distribution between Staph 3 and 4 having a small skew (Figures 1 and 2). The p-values are in the same range for both tests, for Cor 1 vs Staph 1, the t-test returned a p-value of 0.000334 and 0.00048 for the permutation. In the second comparison the results are 0.06973 and 0.0701 respectively. The asymptotic t approximation was used as this is a quicker method and the package allows for easy creation of confidence intervals and adjusting for multiple testing.

Afbeelding met tekst, kaart

Automatisch gegenereerde beschrijving

Figure 10: Histogram (A) and QQ-plot (B) showing the permutation distribution with n = 10 000, between Corynebacterium 1 and Staphylococcus 1

Afbeelding met tekst, kaart

Automatisch gegenereerde beschrijving

Figure 11: Histogram (A) and QQ-plot (B) showing the permutation distribution with n = 10 000, between Staphylococcus 3 and Staphylococcus 4