## Antibiotics associated diarrhea I

Antibiotics associated Diarrhea is a worldwide economically important disease in human. The disease is Caused by disturbance in the commensal bacteria that is followed by either colonization of pathogenic bacteria named Clostridioides difficile, referenced to as CDI (C. difficile infection) or diarrhea caused by the disturbance itself, referenced to as AAD (Antibiotics associated Diarrhea), or experienced no Diarrhea (ND).

To ad dress this a prospective clinical trial across 6 European countries (France, Netherland, Spain, Romania, Germany, Greece) was conducted and ~ 1000 patients were enrolled. The inclusion criteria were the fact they are CDI negative, haven’t taken Antibiotics prior to the enrolment. Patients received on of a three classes of Antibiotics and were sampled on Day1 (just before the antibiotics) and again on Day6, then they were monitored for 90 days and their stool samples (in case of diarrhea) were collected.

* A: PBL: penicillin + beta-lactamase inhibitor.
* OBL: other beta-lactamase antibiotics.
* FQN: fluoroquinolones.

The samples were sequenced using 16S microbial profiling and their microbial richness (Chao) and evenness (Shannon) and beta-diversity distance (between D1 and D6, Jclass) was assessed.

We are interested in the following outcomes:

* Changes in between the baseline samples across various countries, gender and age-stratas
* Changes over time (D1, D6 and Stool) for each antibiotics
* Changes across antibiotics over time

. The data can be read as shown in the next chunck of R code: load("AAD.RData")

## Data reading

Read in the dataset and analyze the skim through the data variables.

## Descriptive statistics

* Summarize your data and calculate the following: mean, median, minimum, maximum, first and third quartile (for each variable).
* For the categorical variable existing, calculate a frequency table.
* Calculate the correlation coefficient (D1 Shannon and D6 Shannon*)* and (D1 Chao D6 Chao)

## Graphics

* Generate a bar chart of a categorical variable for the Outcome (AAD, CDI ,ND)
* Generate a bar chart graph with mean Outcome in BPL, FQ, OBL
* Make a histogram of a continuous variable: “D1 Chao” as well as “D6 Chao”.
* Make a scatterplot of 2 continuous variables D1 Chao and D6 Chao, and add the regression lines for each antibiotics
* Make a boxplot of Jacard distance and a separate boxplots per *Antbiotics* (as.factors).

## Outlier detection

* Explore the data for any existing outliers, identify them (do NOT remove them if found).
* What do you think?

## Testing for normality/ homoscedasticity (for all features mentioned above)

* Check the normality using two methods
* Check the homoscedasticity using two methods.
* What do you think?

1. Statistical Inference

* Calculate the 90%, 95%, 99% confidence interval for the means of Chao and Shannon each *Antbiotics*.
* How would you describe those inferences and what do you observe in terms of the interval width when request higher confidence (i.e. 99% C.I.)?

1. Hypothesis testing

* We hypothesis that Chao/Shannon at day 6 different from day1 for CDI. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework
* Assess whether the previous test assumptions have been met for the test.
* We hypothesis that Chao “different” in the group receiving BPL *Antibiotics* compared to the FQ antibiotics B. Can you test this hypothesis assuming heteroscedasiticy
* Assess the previous test assumption.
* We hypothesis that Chao is different between the different Antibiotics **overtime**. Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity).

1. Linear model

* Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above)
* Calculate and interpret a 95% confidence interval of the regression slope , (bonus)
* Estimating the average CHAO change for with changing the Antibiotics
* Fit a linear regression to the data and interpret the regression coefficient (for the one of the hypotheses mentioned above), taking into account repeated measures!