Projet de biostatistiques

1. Introduction

GMRepo[1] (Data Repository for Human Gut Microbiota) est un jeu de données de métagénomes de microbiote intestinal humain issu de la compilation de plusieurs bases de données publiques contenant les données brutes d'études ayant associé des communautés microbiennes mesurées chez des sujets humains et associés à des traits phénotypiques (données de santé). Le but de cette compilation est de permettre de faire des études sur la dynamique du microbiote humain et son éventuel lien avec des pathologies ou des traits phénotypiques. Les données de microbiotes sont issues obtenues par métagénomique ou métabarcoding (par exemple sur le gène 16S) sur des échantillons prélevés chez des patients.

Le jeu de données dont vous disposez est un sous-ensemble simplifié de ce jeu de données.

2. Contexte

Le microbiote intestinal est l'ensemble des microorganismes qui peuplent le système digestif chez l'humain. Les fonctions des microorganismes dans le système digestif sont nombreuses (métabolisme, production d'énergie, immunité) mais certains microorganismes peuvent également causer des maladies. La diversité de la communauté microbienne est par exemple associée à une limitation du risque de colonisation du tube digestif par des espèces pathogènes (dysbiose). Néanmoins l'écologie des communautés microbienne est complexe car elle fait intervenir de nombreuses bactéries et leurs propres prédateurs (bactériophages), et elle est encore mal comprise.

Il existe des études nombreuses sur le lien entre santé humaine et microbiotes mais ici également le sujet est complexe. La variabilité dans la communauté d'un sujet à l'autre ouvre la voie à de nombreuses études tentant d'associer telle ou telle espèce à des maladies. Par exemple, la bactérie *Escherichia coli* est commune chez l'humain mais la nature de son interaction (mutualiste, commensale ou parasite) dépend de la souche considérée. Les pathologies telles que la maladie de Crohn, le diabète ou les maladies cardiovasculaires ont pu être associées à des caractéristiques ou des changements de microbiotes.

3. Description des variables

Dans un premier temps, il s'agit de se familiariser avec le jeu de données. Téléchargez le jeu de données au format `.csv` sur Moodle ainsi que les métadonnées et importez-le dans `R`. Vérifiez que le fichier est correctement lu (les colonnes sont bien identifiées, il n'y a pas de décalage de colonne, les nombres sont bien identifiés comme des nombres etc.).

Il n'est pas attendu que les réponses à ces questions soient consignées dans votre rapport, elles sont seulement là pour vous aider à commencer à manipuler les données

a. Combien y a-t-il de variables et d'observations ? Parmi les variables, lesquelles sont numériques ou catégorielles ? Parmi les variables numériques, y en a-t-il que vous pourriez plutôt classer comme variables catégorielles ?
b. Pour chacune des variables, donner
   - le nombre d'observations manquantes
   - (variables numériques) la moyenne, l'écart-type, la médiane et l'étendue des valeurs

[1] https://gmrepo.humangut.info/home

- une représentation graphique de la distribution.

c. Représenter graphiquement l'indice de masse corporelle (IMC, *BMI* en anglais) en fonction du sexe, puis du pays d'origine, puis du régime alimentaire et enfin de l'état de santé.
d. En tenant compte de la distribution de l'IMC, quel type de test statistique pourriez-vous utiliser pour établir une éventuelle association entre une des variables précédentes et l'IMC. Réalisez ce test statistique pour une des variables de votre choix.
e. Représenter graphiquement l'abondance mesurée de Bifidobacterium en fonction du sexe, puis du pays d'origine, puis du régime alimentaire et enfin de l'état de santé.
f. Le même test que celui pour l'IMC pourrait-il être utilisé ici ? Pourquoi ?

4. Projet

5. Évaluation

L'évaluation de votre travail se fera sur une seule analyse, vous devrez donc :

- choisir une question d'intérêt à laquelle il est possible de répondre via le jeu de données disponible (vous pouvez vous inspirer de celles posées dans ce document, ou en proposer une nouvelle),
- poser les hypothèses,
- réaliser les tests statistiques nécessaires,
- produire les graphiques pertinents illustrant votre question d'intérêt,
- analyser les résultats et proposer une réponse,
- effectuer une synthèse.