SUJET DE STAGE DE MASTER 2 EN STATISTIQUE

Contexte. Lors de la planification d'une expérience biologique, la détermination du nombre minimal d'individus nécessaires est une question difficile, un compromis entre puissance statistique et contraintes pratiques. De plus, il est connu que certains facteurs extérieurs peuvent affecter la reproductibilité des résultats. Dans le cadre de ses missions de support aux chercheurs de l'Institut Pasteur, le Hub de Bioinformatique et Biostatistique travaille à optimiser la qualité des données générées lors d'expérimentations biologiques et augmenter la reproductibilité des résultats sur les aspects statistiques. Ce projet se divise en 2 axes:

(1) cartographie des effets confondants : par exemple, la distribution des unités expérimentales, peut contribuer jusqu'à 25% à la variabilité observée sur certaines données. En négligeant cette source de variabilité, le risque de fausse découverte peut être multiplié par deux. On s'éloigne alors des fameux 5% que l'on croit contrôler. Heureusement, une planification expérimentale adaptée permet de contrôler ces effets indésirables, augmentant ainsi la reproductibilité des résultats. Ainsi, nous cherchons à établir un catalogue des effets indésirables impliqués lors des expérimentations, aussi appelés effets "batchs". Il est connu que certains facteurs peuvent affecter la reproductibilité des résultats obtenus lors des expérimentations, mais il est aussi établi que l'effet de ces facteurs peut être controlés, en optimisant les plans d'expériences. En pratique, nous cherchons à établir une liste la plus complète possible de ces facteurs, spécifiques aux domaines de recherche (infectueux, microbiote, comportement ...) et quantifier leur impact sur la variabilité de différents types de mesures. Certains effets batchs sont déjà reconnus, comme les dates auxquelles sont réalisées les prélèvements ou l'expérimentateur qui réalise ce prélèvement. Cependant, la quantification de l'impact de ces facteurs sur la reproductibilité des mesures n'est pas bien établie.

(2) développement d'une application Shiny de calcul de puissance adaptée aux besoins des chercheurs en biologie : la version actuelle gère les cas simples de calcul de puissance (comparaison de moyennes, ANOVA, distribution des individus par unité expérimentale...). L'objectif est d'enrichir cette application de cas particuliers comme le calcul de puissance pour les cas déséquilibrés et le pooling et de proposer des extensions aux utilisateurs plus avancés, comme le calcul de puissance pour les analyses multivariées de type analyse en composantes principales ou la planification de données de séquençage

Objectifs du stage. Le stage proposé s'inscrit dans le contexte général de l'analyse de puissance, appliquée à l'expérimentation en biologie. Le/La stagiaire participera donc à :

- la réalisation d'un travail bibliographique sur les questions d'analyse de puissance en expérimentation biologique,
- l'étude de cartographie des effets confondants actuellement en cours dans l'équipe
- l'enrichissement de l'application Shiny de calcul de puissance développée par l'équipe
- en fonction de l'avancement du projet, rédaction d'un article scientifique

Compétences requises. Bonnes connaissances en statistiques (tests, modélisation, analyses multivariées), développement en langage `R`, la connaissance du package Shiny serait un plus. Intérêt pour la recherche et la biologie.

Entité d'accueil. Le stage se déroulera à l'Institut Pasteur (Paris 15ème), au Hub de Bioinformatique et Biostatistique, à partir de février/mars 2020, pour une durée de 4 à 6 mois. Une gratification est prévue (environ 550 euros par mois).

Encadrement et contact.

- Emeline Perthame (ingénieure de recherche, Institut Pasteur)
emeline.perthame@pasteur.fr
- Pascal Campagne (ingénieur de recherche, Institut Pasteur)
pascal.campagne@pasteur.fr