# Licence 3 – Méthodes statistiques 2020-2021

1. Ouvrir le fichier disponible à l’adresse suivante :

<http://www.fabricemonna.com/wp-content/uploads/2019/11/bresilL3.txt>

Il représente l’analyse de plusieurs lichens collectés au Brésil, aux alentours de Rio de Janeiro (les concentrations sont exprimées en µg g-1). RJ : centre de Rio, PB et PH (deux sites distincts) un parc naturel à 30 km au nord de Rio, et LUM : site de référence à plus de 100 km au Nord.

1. Combien le tableau crée contient-ils d’observations et de variables (utiliser une fonction R) ?
2. Donner, à l’aide d’une fonction R, la moyenne, MAX, MIN, Q1, la médiane, et Q3 des concentrations en Al, Ba, Ca, Cd, Cu, Fe, La, Mg, Mn, Na, Pb, S, Sr, Ti, Zn, Zr (et seulement ces variables).
3. Quel est l’écart type de ces mêmes variables (fonction apply, en combinaison avec sd)?
4. Représenter sous forme d’histogramme la distribution en Al et Ti (On prendra soin de nommer correctement les axes et de donner un titre aux deux graphes). Quelle est l’allure des deux distributions ?
5. Une transformée log peut être appliquée sur des données asymétriques pour récupérer un peu de normalité. Opérer cette transformation sur Al et Ti. Représenter à nouveau les distributions. Qu’observez-vous ?
6. Il existe en statistiques une fonction qui mesure l’asymétrie d’une distribution. Quelle est-elle ? Appliquer cette fonction aux données Al et Ti puis aux mêmes données transformées. Qu’observez-vous ?
7. Calculer la droite de régression (pente et ordonnée à l’origine de Al = f(Ti) (fonction lm). Représenter graphiquement cette droite de régression (en rouge) sur le diagramme XY. Les points représentant les données seront des triangles ouverts, pointant vers le haut.
8. La corrélation est-elle significative ? Justifiez votre réponse.
9. L’ordonnée à l’origine peut être considérée comme nulle ? Justifiez votre réponse.
10. Reprendre la question 8, mais cette fois avec Al et Ti, après transformation log.
11. Observez-vous une différence significative entre les deux sites du parc (PH et PB) pour le Pb ? On ne suppose pas que les données sont normalement distribuées, il faut donc mettre en place un test non paramétrique U de Mann-Whitney (fonction wilcox.test).
12. En fait, il s’avère qu’aucune différence significative ne peut être observée entre ces deux échantillons. Regrouper le Pb des échantillons PH et PB dans un même vecteur nommé *Pb\_parc*.
13. Observez-vous une différence significative entre les deux sites du parc regroupés (*Pb\_parc*) et RJ pour le Pb ? Qu’en est-il entre le parc (*Pb\_parc*) et LUM ? On ne suppose toujours pas que les données sont normalement distribuées. Avez-vous une explication ?
14. Réaliser une matrice de corrélation montrant les coefficients de corrélation pour les variables suivantes : Pb, Zn, Cd en utilisant tous les échantillons mesurés.
15. Même chose que 15, mais cette fois les cases de la matrice doivent contenir les diagrammes XY au lieu des valeurs du coefficient de corrélation (fonction pairs).
16. On s’intéresse maintenant plus particulièrement à RJ. Trouver l’intervalle de confiance de la teneur en Pb sur ce site en considérant que les données sont normalement distribuées, d’abord à 95%, puis à 99% de confiance.
17. Si la population sous-jacente n’est pas normalement distribuée, et que le nombre d’individus est petit, on peut toujours estimer l’intervalle de confiance par bootstrap. Faire cette estimation pour 95% de confiance.
18. Faire un ‘stripchat’ représentant la teneur en Pb en fonction des 4 sites étudiés (l’ajout d’un argument jitter sera apprécié). Est-il plus intéressant d’utiliser la teneur en Pb pour ce type de diagramme ou sa transformée log ? Pourquoi ?
19. Envoyer le script R à l’adresse suivante : Fabrice.Monna@u-bourgogne.fr. Le nom du fichier doit être NOM\_PRENOM.R, transmis dans la limite du temps imparti.