

TP Bioinformatique: Statistiques avec

1 Objectifs de ce TP

Nous vous proposons au cours de ce TP d'étudier sous  (<http://cran.r-project.org/>) des données concernant différentes espèces d'oiseaux. Ce TP doit vous permettre de continuer votre initiation au langage , ainsi que de réaliser des analyses statistiques sur un jeu de données réel, et donc de réviser des notions importantes de statistiques.

2 Importation des données et quelques analyses simples

Dans cette partie, vous allez utiliser les données du fichier *espece_avec_poids.txt*, disponible à l'adresse http://pbil.univ-lyon1.fr/R/donnees/espece_avec_poids.txt. Les colonnes de ce fichier contiennent dans l'ordre un numéro d'identifiant de l'espèce, son nom commun, son nom latin, son régime alimentaire, le poids moyen et la taille moyenne observés. Pour l'ensemble du TP, si une question vous est posée et que vous ne savez pas quelle commande utiliser, employez l'aide de  et les différentes façons de chercher à l'intérieur pour trouver les commandes requises.

1. Importez le fichier sous , dans un objet appelé `tesp` en utilisant la fonction `read.table()` (regardez bien les paramètres de chaque fonction avant de l'utiliser, ainsi que les exemples d'utilisation grâce à l'aide de .
2. Utilisez la commande `colnames()` pour afficher les noms des colonnes.

Vous pouvez désormais analyser l'objet `tesp` :

3. Affichez les 5 premières lignes de l'objet `tesp`.
4. Calculez la moyenne et la variance du poids des oiseaux.
5. Listez les espèces d'oiseaux dont le poids est supérieur à 1 kg.
6. Affichez un sommaire des statistiques (moyenne, valeurs minimales, maximales, ...) des poids des oiseaux et interprétez.
7. Représentez graphiquement la taille des espèces en fonction de leur poids, grâce à la fonction `plot()`. Calculez la covariance et la corrélation entre la taille et le poids. Testez si le coefficient de corrélation est significativement différent de 0. Calculez les coefficients de la droite de régression à l'aide de la fonction `lm()`. Ajoutez une droite de régression à la figure à l'aide des commandes `abline()` et `lm()`.

8. Faites les mêmes opérations que précédemment mais cette fois-ci pour les logarithmes de chacune des deux colonnes. Quelles différences observez vous ?
9. Visualisez graphiquement l'impact de chacun des régimes alimentaires sur le poids des oiseaux grâce à la fonction `boxplot()`. Représentez le même type de graphe pour les données de poids en échelle logarithmique. Interprétez les résultats.

Pour le reste du TP, on adoptera la transformation log, c'est à dire que l'on travaillera toujours sur le logarithme des tailles et des poids.

3 Quelques tests statistiques

Avec le graphique `boxplot` vous avez pu constater visuellement des différences de poids en fonction du régime alimentaire des oiseaux. On souhaite maintenant savoir si ces différences sont significatives. Réfléchissez à un test statistique que vous pourriez appliquer pour comparer les moyennes de poids pour les différents régimes alimentaires pris deux à deux.

3.1 Test de comparaison de variances

Avant de faire un test de comparaison de moyennes vous devez vérifier si les variances des deux échantillons ne sont pas significativement différentes (test de l'homoscédasticité). La fonction qui réalise ce test dans \mathbb{R} est `var.test()`. Regardez la documentation de cette fonction, puis effectuez le test d'homoscédasticité sur les régimes alimentaires pris deux à deux. Quelles conclusions tirez-vous ?

3.2 Test de comparaison de moyennes

Quel test devez-vous appliquer maintenant pour comparer les moyennes des poids par catégorie de régime alimentaire ?

1. Quelle est l'hypothèse nulle de ce test ?
2. Quelles sont les différentes conditions d'application ?
3. Trouvez la commande de ce test sous \mathbb{R} .
4. Utilisez cette commande pour comparer les poids des oiseaux ayant différents régimes alimentaires.
5. Quelles sont vos conclusions ?

3.3 Impact d'un facteur sur un autre

L'analyse précédente n'est pas la bonne méthode quand on veut vérifier l'influence d'un facteur (le régime) sur une autre (le poids). Analysez maintenant l'impact du facteur régime sur le facteur poids :

1. Décrivez ces deux facteurs (quantitative-qualitative, fixe-aléatoire...)
2. Quel type de test allez vous appliquer et pourquoi ?
3. Quelle est la commande \mathbb{R} correspondante ?
4. Quelles sont les conditions d'application ?
5. Commentez le résultat.