

L2 Biostatistiques-Bioinformatique

Statistiques avec

Marc Bailly-Bechet






Nous vous proposons d'étudier sous  des données concernant le sommeil chez les mammifères, et ses liens avec leur environnement et leur mode de vie [1]. 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.

Table des matières

| | | |
|----------|---|----------|
| 1 | Importation des données et quelques analyses simples | 1 |
| 2 | Quelques tests statistiques | 4 |
| 2.1 | Test de comparaison de variances | 4 |
| 2.2 | Test de comparaison de moyennes | 5 |
| 2.3 | Impact d'un facteur sur une variable | 5 |
| 2.4 | Impact de deux facteurs sur une variable | 6 |
| 2.5 | Indépendance de deux facteurs | 6 |
| | Références | 6 |

1 Importation des données et quelques analyses simples

Dans cette partie, vous allez utiliser les données du fichier `TP_bioinfo_L2_sleep.txt`, disponible à l'adresse https://pbil.univ-lyon1.fr/R/donnees/TP_bioinfo_L2_sleep.txt. 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 `sleep` 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  et de votre cours).

2. Utilisez la commande `colnames()` pour afficher les noms des colonnes de ce jeu de données.

```
[1] "Species"      "Body.weight" "Brain.weight" "Total.sleep"  "Danger.index"
[6] "Order"
```

Les différentes colonnes du fichier contiennent, pour 40 espèces de mammifères, dans cet ordre :

Species Le nom de l'espèce.

Body.weight Le poids moyen, en kilos.

Brain.weight Le poids moyen du cerveau, en grammes.

Total.sleep La durée moyenne du sommeil, en heures.

Danger.index Le niveau relatif de danger auquel est exposé l'espèce, de par son mode de vie (exposé ou non) et ses prédateurs naturels.

Order L'ordre cladistique dans lequel est rangée l'espèce : *A* pour les Artiodactyles, *C* pour les Carnivores, *D* pour les Périssodactyles, *P* pour les Primates, *R* pour les Rongeurs et *S* pour les Soricomorphes (anciennement classés comme Insectivores).

Vous pouvez désormais analyser l'objet `sleep` :

3. Affichez les 5 premières lignes de l'objet `sleep`.

```
      Species Body.weight Brain.weight Total.sleep Danger.index Order
1 African.giant.pouched.rat      1.000         6.6         8.3      Average      R
2 Arctic.Fox      3.385         44.5        12.5     Very.Low      C
3 Arctic.ground.squirrel      0.920         5.7        16.5     Average      R
4 Baboon      10.550        179.5         9.8         High      P
5 Brazilian.tapir    160.000        169.0         6.2         High      D
```

4. Calculez et enregistrez dans un vecteur `rapport` le rapport du poids du cerveau sur le poids du corps (attention aux unités) : quelle est l'espèce qui a le rapport le plus élevé ? Qu'en concluez-vous ?

```
[1] Ground.squirrel
40 Levels: African.giant.pouched.rat Arctic.Fox Arctic.ground.squirrel ... Vervet
```

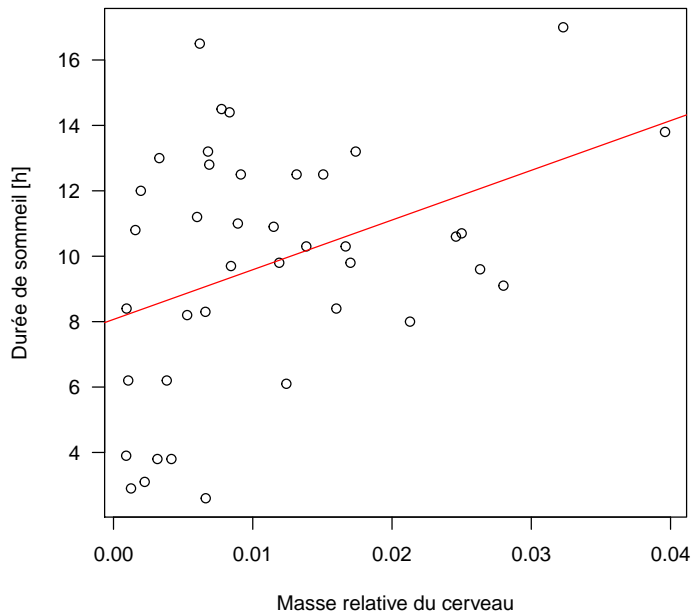
5. Listez les espèces qui dorment moins de 6 heures par jour. Ont-elles un point commun ?

```
[1] Cow      Donkey  Goat    Horse   Roe.deer Sheep
40 Levels: African.giant.pouched.rat Arctic.Fox Arctic.ground.squirrel ... Vervet
```

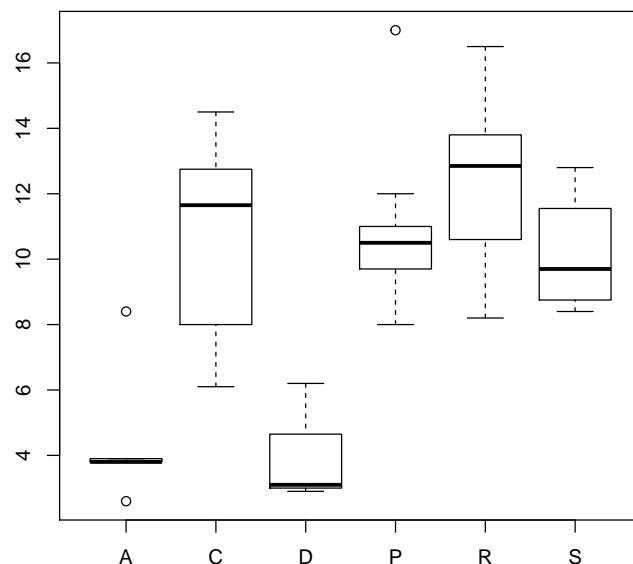
6. Affichez un sommaire des différentes colonnes du jeu de données.

```
      Species      Body.weight      Brain.weight
African.giant.pouched.rat: 1  Min. : 0.0050  Min. : 0.140
Arctic.Fox                 : 1  1st Qu.: 0.3888  1st Qu.: 5.375
Arctic.ground.squirrel    : 1  Median : 3.7875  Median : 41.850
Baboon                    : 1  Mean   : 55.5347  Mean   : 144.717
Brazilian.tapir           : 1  3rd Qu.: 52.9950  3rd Qu.: 176.000
Cat                       : 1  Max.   :521.0000  Max.   :1320.000
(Other)                   :34
Total.sleep              Danger.index Order
Min. : 2.60      Average : 7      A: 5
1st Qu.: 8.15    High    : 8      C: 8
Median :10.30    Low     : 7      D: 3
Mean   : 9.79    Very.High: 6    P:10
3rd Qu.:12.50    Very.Low :12    R:10
Max.   :17.00                    S: 4
```

7. Représentez graphiquement la durée de sommeil en fonction de la variable `rapport` calculée précédemment, grâce à la fonction `plot()`. Calculez la covariance et la corrélation entre ces variables. Vérifiez que vous retrouvez bien le coefficient de corrélation si vous le calculez à partir de la covariance. Testez si le coefficient de corrélation est significativement différent de 0 (vous pouvez vérifier manuellement votre réponse à l'aide de vos tables statistiques, en sachant que $t_{38,5\%} = 2.024$ et $t_{38,1\%} = 2.711$). 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. Visualisez graphiquement l'importance de la phylogénie des espèces sur la durée du sommeil grâce à la fonction `boxplot()`, dont la syntaxe est décrite dans le TP précédent. Interprétez les résultats.



2 Quelques tests statistiques

Avec le graphique `boxplot` vous avez pu constater visuellement des différences de durée du sommeil en fonction de l'ordre cladistique. On souhaite maintenant savoir si ces différences sont significatives. Réfléchissez à un test statistique que vous pourriez appliquer pour comparer les moyennes des durées de sommeil des différents ordres pris deux à deux (par exemple pour comparer les Primates et les Carnivores).


2.1 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 `R` est `var.test()`. Regardez la documentation de cette fonction, puis effectuez le test d'homoscédasticité de la durée du sommeil entre les différents ordres pris deux à deux. Quelles conclusions tirez-vous ? Avec le couple (C, T) vous devez obtenir le résultat suivant :

```
F test to compare two variances
data:  sleep$Total.sleep[sleep$Order == "C"] and sleep$Total.sleep[sleep$Order == "P"]
F = 1.7059, num df = 7, denom df = 9, p-value = 0.4471
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4064411 8.2276955
sample estimates:
ratio of variances
 1.705852
```

2.2 Test de comparaison de moyennes

Quel test devez-vous appliquer maintenant pour comparer les moyennes des durées de sommeil entre chaque couple d'ordres ?

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 .
4. À l'aide de cette commande, comparez le temps de sommeil entre les Artiodactyles et les Carnivores, puis entre les Primates et les Rongeurs.


```
Welch Two Sample t-test
data:  sleep$Total.sleep[sleep$Order == "C"] and sleep$Total.sleep[sleep$Order == "A"]
t = 4.135, df = 10.641, p-value = 0.001778
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.874605 9.475395
sample estimates:
mean of x mean of y
 10.675      4.500
```

```
Welch Two Sample t-test
data:  sleep$Total.sleep[sleep$Order == "P"] and sleep$Total.sleep[sleep$Order == "R"]
t = -1.1445, df = 17.83, p-value = 0.2675
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.65954  1.07954
sample estimates:
mean of x mean of y
 10.90      12.19
```

5. Quelles sont vos conclusions ?

2.3 Impact d'un facteur sur une variable

L'analyse précédente n'est pas la bonne méthode quand on veut vérifier l'influence *globale* d'un facteur (l'ordre) sur une variable quantitative (la durée du sommeil). Analysez maintenant l'impact du facteur "ordre" sur la variable "Durée du sommeil" :

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

```
          Df Sum Sq Mean Sq F value Pr(>F)
sleep$Order  5  314.9   62.98    9.69 8.07e-06 ***
Residuals  34  221.0    6.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Faites la même analyse en regardant non plus si la position phylogénétique est significativement reliée à la durée du sommeil, mais si la durée du sommeil dépend du niveau relatif de danger de l'espèce. Que concluez-vous ? Ce résultat est-il compatibles avec le résultat précédent ?

2.4 Impact de deux facteurs sur une variable

On veut maintenant prendre en compte à la fois l'effet de la position phylogénétique des espèces et le niveau relatif de danger de leur environnement, simultanément.

1. Quel type de test allez vous appliquer et pourquoi ?
2. Quelles sont les conditions d'application ?
3. La commande `aov` pour réaliser une ANOVA 2 de la variable `y` en fonction des facteurs `x` et `z` est `aov(y~x*z)`. Adaptez cette commande à votre situation et réalisez le test.

```

              Df Sum Sq Mean Sq F value    Pr(>F)
sleep$Danger.index      4  352.3   88.08  13.079 1.07e-05 ***
sleep$Order              5   21.1    4.22   0.626   0.682
sleep$Danger.index:sleep$Order  7    7.6    1.08   0.161   0.991
Residuals              23  154.9    6.73
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

4. Le résultat est-il en accord avec ce que vous avez vu précédemment ?

2.5 Indépendance de deux facteurs

Pour expliquer les décalages entre les résultats des ANOVA1 et de l'ANOVA2, on va regarder si les deux facteurs de l'ANOVA 2 sont indépendants.

1. Quel type de test allez vous appliquer et pourquoi ?
2. La commande pour obtenir une table de contingence croisée des deux variables `x` et `y` est `table(x,y)`. À Partir de cela, trouvez et effectuez le test nécessaire.
3. Les deux facteurs sont ils indépendants ? Cela vous permet-il d'expliquer vos résultats précédents ?

Références

- [1] T. Allison and D. V. Cicchetti. Sleep in mammals : ecological and constitutional correlates. *Science*, 194 :732–734, 1976.