Fiche TD avec le logiciel R : mab1

Familiarisation à R

S. Venner & E. Desouhant

Une introduction au logiciel R dans le cadre de l'UE Mathématiques Appliquées à la Biologie (MAB) de L3 BOP.

1 Première session de travail

Les lignes de code que vous devrez saisir dans la console R seront toujours écrites en rouge. Vous pouvez les copier/coller directement à partir du document PDF dont l'URL est donnée en pied-de-page.

1.1 Quelques commandes classiques dans R

Vous pourrez utiliser R comme une calculatrice. Les opérateurs * et / sont respectivement utilisés pour la multiplication et la division.

```
 3 + 5
[1] 8
 3^2 ; sqrt(9)
[1] 9
[1] 3
 log(1) ; exp(1)
[1] 0
[1] 2.718282
 factorial(3)
[1] 6
 cos(pi)
[1] -1
 # Test logique
 1 == 0 ; cos(pi) == 0
[1] FALSE
[1] FALSE
```

1.2 Importer de données

Un fichier de données peut être créé à partir de plusieurs logiciels mais doit être sauvegardé dans le format texte (`.txt`) pour l'utiliser avec R. Ici, vous allez utiliser des données sauvegardées sur un site internet.

1.2.1 Copie du site distant dans le dossier de travail

Enregistrez dans votre espace de travail le fichier de données « `balanin.txt` » que vous trouverez sur le site `ftp://pbil.univ-lyon1.fr/pub/cours/DESOUHANT/`.

1.2.2 Inspection du fichier avec un éditeur de texte

Ouvrez le fichier et explorez son contenu. Les espèces *Curculio glandium*, *C. elephas*, *C. pellitus* et *C. venosus* sont symbolisées par les lettres G, E, P et V. Les poids sont exprimés en mg et les mesures corporelles (LR et LC) en mm. « Semaine » représente le numéro de la semaine où une espèce a été observée sur l'arbre.

1.2.3 Lecture du fichier dans R

Lisez le fichier dans R en utilisant :

```
read.table("balanin.txt")
read.table("balanin.txt", header = TRUE)
```

Que signifie l'argument (`header = TRUE`) de la fonction `read.table()` ? Pour toute fonction que vous utiliserez, vous pouvez consulter la documentation. Consulter la fiche de documentation est une opération fondamentale que vous devrez systématiquement réaliser pour comprendre toute fonction. Exemple :

```
?read.table
```

Le nom de la fonction est suivi de ses arguments entre parenthèses. Le signe « = » est utilisé pour préciser les valeurs des paramètres. On peut passer les valeurs dans l'ordre (voir documentation de la fonction) sans donner leur nom, les passer dans le désordre en donnant leur nom (toute abréviation non ambiguë est acceptée). Les paramètres ont pour la plupart des valeurs par défaut.

1.2.4 Lecture directe d'un site distant

Vous pouvez directement charger les données sauvegardées dans le répertoire distant (*i.e.* sur internet).

```
read.table("ftp://pbil.univ-lyon1.fr/pub/cours/DESOUHANT/balanin.txt", header = TRUE)
```

1.3 Rappel des commandes précédentes

En utilisant la « flèche vers le haut » (↑) sur le clavier.

1.4 Sauvegarde du travail

1.4.1 Sauvegarde des commandes

Lorsque vous quittez R, une fenêtre vous demande « Sauvez une image de la session ? », cliquer sur oui vous permettra de rappeler vos commandes à la prochaine ouverture de Rlogo. en utilisant la « flèche vers le haut » (↑) sur le clavier.

1.4.2 Sauvegarde des résultats

Les sorties de R (résultats et figures) pourront (et devraient) être enregistrées dans un document de type traitement de texte. La mise en forme des sorties de R sera conservée en utilisant la police « Courier New » (police taille 10).

2 Objets

Voir la fiches tdr13 [1] pour compléments.

2.1 Les vecteurs

L'objet de base est le vecteur. Affecter une valeur à un objet de nom « `w` » à l'aide de la flèche `<-` (combinaison de `<` et de `-`) :

```
w <- 6
```

Afficher le contenu de l'objet :

```
w
[1] 6
```

Déterminer la nature de l'objet :

```
is.factor(w)
[1] FALSE
is.data.frame(w)
[1] FALSE
is.vector(w)
[1] TRUE
# w est un vecteur
```

Générer des séquences :

```
w <- 1:7
w
[1] 1 2 3 4 5 6 7
w <- seq(-1, 10, 0.5)
w
 [1] -1.0 -0.5  0.0  0.5  1.0  1.5  2.0  2.5  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5
[17]  7.0  7.5  8.0  8.5  9.0  9.5 10.0
seq(from = -1, to = 10, by = 0.5)
 [1] -1.0 -0.5  0.0  0.5  1.0  1.5  2.0  2.5  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5
[17]  7.0  7.5  8.0  8.5  9.0  9.5 10.0
```

Créer des combinaisons et déterminer le mode d'un vecteur :

```
w <- c(2, 5, 8, 3) # "c" pour "concaténation"
w
[1] 2 5 8 3
mode(w)
[1] "numeric"
# w est un vecteur numérique
w1 <- c(2, 5, 8, "a")
w1
```

1. http://pbil.univ-lyon1.fr/R/pdf/tdr13.pdf

```
[1] "2" "5" "8" "a"
 mode(w1)
[1] "character"
 # w1 est un vecteur de caractère et c'est aussi un nouvel objet !
 w2 <- c(TRUE, TRUE, FALSE, TRUE)
 w2
[1]  TRUE  TRUE FALSE  TRUE
 mode(w2)
[1] "logical"
 # w2 est un vecteur logique
```

Dresser la liste des objets créés et supprimer des objets :

```
 ls()
[1] "w"  "w1" "w2"
```

Détruire l'objet w et vérifier qu'il a bien été éliminé :

```
 rm(w)
 ls()
[1] "w1" "w2"
```

2.2 Les tableaux

```
 balanin <- read.table("ftp://pbil.univ-lyon1.fr/pub/cours/DESOUHANT/balanin.txt", header = TRUE)
```

Recherchez la nature de l'objet « balanin ».

```
 is.vector(balanin)
 is.matrix(balanin)
 is.data.frame(balanin)
```

L'objet « balanin » correspond-il à un vecteur (`vector`), une matrice (`matrix`) ou à un tableau (`data.frame`) ? Un « data.frame » est un tableau dont les colonnes peuvent avoir des « natures » variées (variables quantitatives, qualitatives, logiques). Une matrice ne contient qu'un type de variables. L'objet « balanin » appartient à la classe des « data.frame ».

Faire apparaître les titres des colonnes uniquement :

```
 names(balanin)
```

Faire apparaître le début de votre jeu de données :

```
 head(balanin)
```

Extraire des éléments d'un tableau

```
 balanin[1:4, ]
 balanin[2:6, c(2, 6)]
```

Quel est le sens des valeurs entre crochets ?

```
 dim(balanin)
 nrow(balanin)
 ncol(balanin)
```

À quoi correspondent les fonctions `dim()`, `nrow()` et `ncol()` ? Faire apparaître les poids, longueur de rostre (`LR`) et longueur de corps (`LC`) des 5 derniers individus du tableau :

```
    poids  LR  LC
352  27.4 6.3 6.5
353  24.7 5.7 6.2
354  21.1 5.5 5.8
355  20.2 3.1 5.5
356  24.3 3.3 6.2
```

Les variables (ou noms des colonnes) d'un `data.frame` sont directement accessibles par leur nom, ou bien, par la syntaxe suivante : `nomObjet$nomVariable` (exemple : `balanin$poids`). Créez un objet appelé `titi` dans lequel vous placerez les poids mesurés.

```
titi <- balanin$poids
titi
```

2.3 Les fonctions

2.3.1 Fonctions statistiques élémentaires

Testez les différentes commandes ci-après :

```
length(balanin$poids)
min(balanin$poids)
max(balanin$poids)
range(balanin$poids)
mean(balanin$poids)
median(balanin$poids)
var(balanin$poids)
sum(((balanin$poids - mean(balanin$poids))^2)/length(balanin$poids))
```

À quoi correspond la fonction `var()` ? Déterminer si la variance donnée par la fonction `var()` correspond à celle de l'échantillon ou à la variance estimée de la population.

```
hist(balanin$poids)
hist(balanin$poids, nclass = 4)
```

2.3.2 Fonctions génériques

Documentez-vous sur les fonctions `summary()` et `plot()`. Les fonctions génériques peuvent s'appliquer à plusieurs types d'objets très différents entre eux. On peut donc appliquer ces fonctions aux `data.frame` ou à un seul vecteur.

```
summary(balanin$LR)
summary(balanin$espece)
summary(balanin)
```

À quoi correspondent les résultats obtenus ?

```
plot(balanin$LR)
plot(balanin$LC, balanin$LR)
plot(balanin$sexe, balanin$poids)
plot(balanin$poids, balanin$sexe)
plot(balanin)
plot(balanin$espece, balanin$semaine)
plot(balanin$espece, balanin$sexe)
```

Que représentent les graphiques obtenus ? Vous pouvez enregistrer les graphiques obtenus au format pdf en sélectionnnant « sauver sous » → « PDF » dans l'onglet « fichier ».

2.3.3 Les facteurs

Les valeurs de la colonne `sexe` sont-elles reconnues comme des valeurs numériques, des chaînes de caractères ou des modalités d'un facteur ?

```
balanin$sexe
is.numeric(balanin$sexe)
is.character(balanin$sexe)
is.factor(balanin$sexe)
```

Les facteurs correspondent à des variables qualitatives. À quoi sert la fonction `levels()` ?

```
levels(balanin$sexe)
table(balanin$sexe)
tapply(balanin$semaine, balanin$espece,mean)
tapply(balanin$semaine, balanin$espece,var)
tapply(balanin$semaine, balanin$espece: balanin$sexe,var)
```

On peut extraire une partie des données correspondant à une modalité d'un facteur et les exploiter indépendamment du reste du jeu de données :

```
g <- balanin[balanin$espece == "G",]
```

Examinez ce que contient ce nouvel objet `g`.

```
gf <- balanin[balanin$espece == "G" & balanin$sexe == "F", ]
gm <- balanin[balanin$espece == "G" & balanin$sexe == "M", ]
```

Décrivez les résultats qui ne concernent que les mâles de *C. glandium* :

```
par(mfrow = c(1, 2))
hist(gm$LC)
```

Représentez le nuage de points correspondant à la taille des rostres en fonction de la taille du corps chez mâles de *C. glandium*. La commande `par(mfrow = c(1, 2))` sert à préparer la fenêtre de graphique. Ici la fenêtre est préparée pour recevoir 2 graphiques disposés sur 1 ligne 2 colonnes. Pour revenir à une fenêtre graphique « normale », utilisez la commande : `par(mfrow = c(1, 1))`.

3 Exploration des données

3.1 Introduction

Vous allez à présent explorer les jeux de données en réalisant des statistiques descriptives, notamment avec des représentations graphiques. À l'issue de cette phase exploratoire, vous proposerez une démarche d'analyse des données permettant de répondre aux questions qui vous sont posées dans l'énoncé (contexte biologique). Vous pourrez enregistrer votre travail dans un fichier Word (rappel : copier et coller les sorties de R qui vous paraissent pertinentes avec la police Courier New).

3.2 L'étalement des émergences sur plusieurs années chez *C. elephas*

Une étude antérieure a montré que chez *C. glandium*, la probabilité qu'une larve produite une année donnée émerge du sol sous la forme d'un adulte l'année suivante est de 0.02 dans les communautés suivies, les émergences étant essentiellement réalisées la 2e année qui suit le développement larvaire. Un suivi de terrain a été réalisé pour étudier l'étalement des émergences des adultes chez *C. elephas*. La fréquence d'émergence la première année qui suit le développement des larves dans le fruit a été déterminée sur 80 arbres échantillonnés. Par arbre, 50 larves tirées aléatoirement ont été suivies jusqu'à leur émergence. Le fichier `balanin_emergence.txt` [2] présente les effectifs en individus qui sont sortis au bout de 1 an pour chaque arbre échantillonné.

Importez les données du fichier `balanin_emergence.txt` dans un objet appelé `émergence`.

```
summary(émergence)
```

```
     Arbre            NbAd
Min.   : 1.00   Min.   :10.00
1st Qu.:20.75   1st Qu.:17.00
Median :40.50   Median :18.00
Mean   :40.50   Mean   :18.27
3rd Qu.:60.25   3rd Qu.:20.25
Max.   :80.00   Max.   :25.00
```

Quelle est la nature de la variable aléatoire étudiée ? Quelle représentation graphique pourriez-vous envisager pour une telle variable ?

Essayez les commandes suivantes :

```
barplot(table(émergence))
nrow(émergence)
barplot(table(émergence)/nrow(émergence))
```

Quel est l'intérêt de la seconde représentation graphique relativement à la première ? Comment procéderiez-vous pour tester l'hypothèse selon laquelle la probabilité de sortir au bout d'un an est la même chez *C. glandium* et chez *C. elephas* ?

3.3 Morphométrie des femelles adultes chez *C. glandium* et *C. elephas*

Utilisez les objets `femelle`, `gf` et `ef` correspondant respectivement à des tableaux ne comprenant que les données concernant les femelles des 2 espèces (*C. glandium* et *C. elephas*), les femelles des *C. glandium* et les femelles des *C. elephas* (*cf.* : extraction d'une partie des données via les facteurs).

```
femelle <- balanin[(balanin$espece == "E" | balanin$espece == "G") & balanin$sexe == "F", ]
```

Précisez l'effectif des femelles de chacune des deux espèces. Calculez la moyenne, la variance et la médiane de la longueur du corps et du rostre chez les 2 espèces et réalisez des boîtes à moustaches (fonction `boxplot()`) représentant ces données. Construisez des histogrammes concernant la longueur du rostre de l'ensemble des individus puis des deux espèces séparément. Quel est l'intérêt

2. URL : `ftp://pbil.univ-lyon1.fr/pub/cours/DESOUHANT/balanin_emergence.txt`

de ces histogrammes ? Comment comparer les longueurs des rostres entre les femelles des deux espèces ? À quels résultats vous attendez-vous ? Représentez sur un même graphique le nuage de points concernant la relation entre la longueur du rostre et la longueur du corps.

```
plot(femelle$LC, femelle$LR, type = "n")
points(gf$LC, gf$LR, col = "blue")
points(ef$LC, ef$LR, col = "red")
```

Interprétez ces résultats. Comment comparer les femelles de ces deux espèces en prenant en compte ces deux variables ?

3.4 Périodes de présence sur l'arbre des femelles de *C. glandium* et *C. elephas*

Après avoir précisé les effectifs de chacune de ces espèces, vous allez décrire et représenter graphiquement les périodes de présence de ces espèces (variable `semaine`). Quels sont les outils qui vous paraissent les plus pertinents pour décrire et représenter cette variable ? Représentation des périodes de présence des espèces *C. glandium* et *C. elephas*

```
table(balanin$espece,balanin$semaine)
periodeEG <- table(balanin$espece, balanin$semaine)[c(1,2), ]
periodeEG
barplot(periodeEG, beside = TRUE, las = 2, col = c("white","black"))
```

Décrivez la répartition temporelle des deux espèces. Comment envisagez-vous de comparer les périodes de présence sur l'arbre de ces deux espèces ?