Fondements statistiques pour la science des données

TP n°1. Logiciel R – Introduction à R et Rcmdr

1 Installation du logiciel R, environnement de travail

1. Les fichiers nécessaires à l'installation du logiciel R sont distribués à partir du site Internet CRAN (Comprehensive R Archive Network) à l'adresse suivante : `http://www.r-project.org/`
2. Sélectionner un miroir CRAN puis **télécharger** et **installer** le logiciel en fonction de votre système d'exploitation.
3. R démarre chaque session au sein d'un **répertoire de travail** créé par défaut. Par exemple, sous Windows : `C:\Users\Utilisateur\Documents`.
   On peut connaître celui-ci grâce à la commande `getwd()` puis éventuellement le modifier par l'intermédiaire du menu Fichier ou à l'aide de la commande `setwd`; par exemple : `setwd("D:/Cours/AD/TP R")`
4. Lorsqu'on **quitte** une session R, le logiciel propose de sauvegarder l'ensemble du travail réalisé au sein de deux fichiers :
   `.RData` : environnement de travail (objets R, variables, fonctions, données),
   `.Rhistory` : historique des commandes exécutées.

2 Premiers pas sous R

1. Ouvrir le logiciel R. Dans une première étape, les commandes seront saisies dans la fenêtre `R console`.
2. R permet de réaliser les opérations habituelles d'une calculatrice à l'aide des opérateurs usuels : `+, -, *, /, exp, log, ^`, etc. Exécutez par exemple les commandes suivantes :
   ```
   > exp(log(5))^2
   > 1/sqrt(2*pi)*exp(-1/2)
   ```
3. Pour affecter une valeur à un objet, on utilise indifféremment le symbole égal (`=`) ou une "flèche" à gauche (`<-`) :
   ```
   > x = 45
   > y <- 5
   > Y <- 9
   ```
   Remarques : R différencie minuscules et majuscules. Le séparateur décimal est le point.

4. Saisir le nom de la variable permet d'en lire le contenu :
   `> x`
   `> x/y/Y`
5. Il est possible de soumettre plusieurs commandes sur une même ligne :
   `> factorial(5) ; choose(32,4) ; round(pi,4) ; floor(sqrt(2))`
6. Affecter une valeur à un objet et afficher simultanément le contenu :
   `> (s = sum(1:10))`
7. Il est possible de naviguer dans l'historique des commandes précédemment exécutées à l'aide des flèches ↑ ou ↓ du clavier.

3 La fenêtre de script

En début de session R, il est vivement conseillé d'ouvrir une fenêtre de script afin de saisir les différentes commandes R à soumettre et de les gérer plus facilement : conserver les commandes importantes, effacer les commandes erronées ou inutiles, insérer des commentaires, exécuter plusieurs commandes simultanément, programmer ou écrire ses propres fonctions.

1. Ouvrir un **nouveau script**. Dans la fenêtre de script, écrire les deux expressions ci-dessous en langage R puis exécuter les.
   $\left(9^2 + \frac{19^2}{22}\right)^{\frac{1}{4}}$; $\frac{99^2}{2206\sqrt{2}}$
   Remarque. L'exécution d'une commande dans la fenêtre de script s'effectue à l'aide du raccourci clavier CTRL R (en situant le curseur n'importe où sur la commande). Pour soumettre plusieurs commandes simultanément, sélectionner l'intégralité des commandes puis CTRL R.
2. À l'aide du symbole `#`, ajouter un **commentaire** après chaque expression. Tout ce qui suit ce symbole est ignoré par R.
3. **Sauvegarder** le script dans le répertoire de travail courant puis fermer le script.
4. **Exécuter** l'intégralité du script à l'aide de la commande
   `source("script.R", echo=T)`

4 Résumé numérique d'une petite série statistique

1. Saisir la série statistique ci-dessous (à noter la fonction `c` qui permet sous R de saisir une série de valeurs sous la forme d'un vecteur) :
   `> age = c(26, 25, 23, 19, 22, 21, 26, 27, 22, 18, 31, 32, 38, 27, 24, 25)`
2. Lister le contenu du vecteur `age` :
   `> age`
3. Il est très simple de transformer les valeurs d'un vecteur :
   `> age^2 ; sqrt(age) ; log(age)`
4. Sélectionner des valeurs de la série par leurs indices, par des conditions logiques :
   `> age[3] ; age[10:12] ; age[c(2,4,6)] ; age[-(1:5)] ; jeune = age[age<20]`
5. Les commandes suivantes permettent successivement de donner le nombre de valeurs de la série, de les ordonner, d'en obtenir le min, le max, la somme, la moyenne, la médiane, l'écart-type et la variance :
   `> length(age) ; sort(age) ; min(age) ; max(age) ; sum(age)`

```
> mean(age) ; median(age) ; sd(age) ; var(age)
```
Remarque : la variance calculée par R correspond à l'estimation non biaisé de la variance (avec $\texttt{sd} = \sqrt{\texttt{var}}$).

6. Soumettre la fonction générique `summary(age)` qui produit un résumé statistique succinct des valeurs de la série.
7. Que représente la valeur suivante ?
```
> mean(age^2) - mean(age)^2
```

5 Représentation graphique d'une série statistique

1. La commande générique `plot` produit un graphe dit indexé. Les valeurs sont tracées selon leur rang (ou indice) dans la série :
```
> plot(age)
```
Reproduire le même graphe, mais avec les valeurs triées.
Modifier le type de tracé des valeurs en ajoutant à la fonction l'argument `type` :
```
> plot(sort(age), type="h")
```
2. Une représentation plus classique des valeurs individuelles : représentation axiale (ou nuage de points) :
```
> stripchart(age)
```
3. Comparer le graphe précédent avec celui fourni par la commande suivante :
```
> stripchart(age, method="stack", pch=20, col="blue")
```
4. Construire un histogramme :
```
> hist(age)
```
5. Un histogramme plus joli :
```
> hist(age, col="steelblue4", border="white", main="Histogramme de la variable
Age", xlab="Age de l'enquêté", ylab = "effectif")
```
Remarque : la commande `colors()` affiche toutes les couleurs disponibles sous R.
6. Un histogramme avec découpage en classes d'amplitudes inégales :
```
> hist(age, breaks=c(18,20,24,27,32,38), main="Histogramme de la variable Age",
col="steelblue4", border="white", xlab="Age de l'enquêté", ylab = "densité")
```
7. Tracer une boîte à moustaches :
```
> boxplot(age)
```
8. Une belle boîte à moustaches rose et cintrée :
```
> boxplot(age, col="pink", horizontal=TRUE, notch=TRUE, xlab="Age de l'enquêté")
```
9. Le paramètre graphique `mfrow` permet de partitionner une feuille graphique ; par exemple :
```
> par(mfrow=c(3,3))
```
Plus précisément, on a ici la possibilité d'avoir 9 graphiques disposés selon 3 lignes et 3 colonnes. Après avoir soumis la commande précédente, représenter tous les graphiques précédents dans une même fenêtre.

6 L'installation de nouveaux packages (ou librairies)

Pour des réaliser des traitements statistiques spécifiques (analyse de données, tests d'hypothèses, etc.) ou pour appliquer des méthodes spécifiques à un domaine d'application (économétrie, données de survie, etc.), il est parfois nécessaire d'installer de nouvelles libraries. Cela se fait en deux temps :

1. **Installation** du package (en le téléchargeant depuis un miroir CRAN). Par exemple :
   `> install.packages("FactoMineR", dependencies=TRUE)`
2. **Chargement** du package pour qu'il puisse être utilisé dans la session courante :
   `> library(FactoMineR)`

L'étape d'installation (téléchargement) est faite une fois pour toutes, mais le chargement de la librairie doit être effectuée à chaque nouvelle session de R. Ces deux opérations peuvent être réalisées à partir du menu `Packages`.

Application : installer puis charger la librairie `e1071` de R.

Utilisation de la librairie Rcommander

7 Importation d'un jeu de données

1. Consulter le fichier de données `temperature.txt`
2. Enregistrer ce fichier sur votre ordinateur puis retourner sous le logiciel `R`.
3. Charger la librairie (ou *package*) Rcmdr : `> library(Rcmdr)`
4. Importer le jeu de données `temperature.txt`
   Données → *Importer des données* → *depuis un fichier texte, le presse papier ou une URL*. Donner un nom au jeu de données (par exemple *temp*), renseigner le séparateur de champs (tabulation), le séparateur décimal (virgule) puis OK. Sélectionner le fichier `temperature.txt`.
   Essayer également d'importer les données depuis le presse papier (sélectionner et copier les données dans le navigateur) ou en indiquant une adresse URL (clic droit sur le nom de fichier dans le navigateur, copier l'adresse du lien)

Remarque : à chaque opération effectuée à partir du menu déroulant, la ligne de commande correspondante en langage R est affichée automatiquement dans la fenêtre de script de Rcommander. Pour exécuter à nouveau une commande, il suffit cliquer avec la souris sur la commande (n'importe où) puis de valider avec le bouton `Soumettre` (ou de faire `CTRL R` au clavier).

8 Premiers traitements statistiques

1. **Visualiser le jeu de données** (bouton *Visualiser*). Noter la colonne *Ville* qui apparaît comme une variable quelconque et non comme l'identificateur des individus. Fermer la fenêtre.
2. Déclarer la première colonne du tableau comme **identificateur** des individus : *Données* → *Jeu de données actif* → *nom des cas*. Sélectionner la variable *Ville*.
3. Visualiser à nouveau les données : le nom des villes doit maintenant apparaître dans une colonne grisée. La définition d'un identificateur permet entre autre d'afficher les noms des individus, et non des numéros, pour certains graphiques ou traitements statistiques.
4. **Renommer la variable** *aout* en l'appelant désormais *août* (soit par le menu *Données* → *Gérer les variables*, soit en éditant les données et en cliquant sur le nom de variable.)

5. **Ajouter une nouvelle variable** donnant pour chaque ville sa température maximale de l'année : *Données → Gérer les variables dans le jeu de données actif → Calculer une nouvelle variable.*
6. **Découper en classes** la variable *latitude* : on choisira un découpage en trois classes d'effectifs égaux avec les étendues comme nom de modalités.
7. Faire de même avec la variable *Longitude*.
8. Ajouter au jeu de données les 12 variables correspondant aux températures mensuelles **centrées - réduites** : *Données → Gérer les variables dans le jeu de données actif → standardiser les variables.*
9. Produire un **résumé statistique** de chacune des variables du jeu de données : *Statistiques → Résumé → Jeu de données actif.*
10. Produire les **statistiques descriptives** pour toutes les variables quantitatives. Durant quel mois les températures des 15 villes sont-elles les plus proches les unes des autres ?
11. **Comparaison graphique de moyennes** : comparer la température moyenne en février en fonction des 3 catégories de latitudes : *Graphes → Graphe des moyennes → Facteur=latitude, Variable réponse = février, barre d'erreur=écarts-types.*
12. Construire un graphique similaire au précédent à l'aide de **boîtes à moustaches en parallèle** : *Graphes → Boîtes de dispersion → variable=février, Graphe par groupe=latitude.*

8.1 Quitter une session R

À l'issue d'une session de travail sous R (ou Rcmdr), il est conseillé :

1. Lorsque des modifications ont été apportées au jeu de données importé (transformation, recodage, calcul de nouvelles variables, etc.) de sauvegarder celui-ci sous le format interne de R (.RData). Lors d'une prochaine session sous R, ce fichier pourra alors être récupéré au travers du menu *Données → Charger un jeu de données.*
2. De sauver intégralement ou partiellement les commandes importantes soumises pendant la session (menu : *Sauver le script*).