Statistique et Informatique — UFR STAPS
Master 1 — Université de Paris X - Nanterre

TD 1. Initiation à R

0 Gestion du TD

R est logiciel informatique à ligne de commande. Pour bien comprendre son fonctionnement, il est nécessaire d'expérimenter par vous-même. Pour chaque concept développé, nous vous donnons un exemple (un code à recopier à l'identique) puis un travail à faire par vous même (rubrique "application").

Comme **R** ne garde pas la trace de vos calculs, tous les codes que vous tapez (et qui sont corrects, pas la peine d'inclure vos essais-erreurs) doivent être recopiés dans un fichier Word (grace à un copier coller). Puis, à la suite de vos codes, copiez les réponses de **R**.

Pour commencer, lancer **R** en cliquant sur l'icône Rgui (sur le bureau ou dans le menu *Démarrer*).

1 Variables et fonctions

Pour donner des instructions à **R**, il faut les taper dans une fenêtre appelée "R console". Tous les codes qui suivent sont donc à écrire dans cette fenêtre.

1.1 Comment fait-on un calcul ?

Pour faire un calcul, on le tape, tout simplement (sans signe égal devant). Par exemple, pour calculer 1 + 1, on tape `1+1`, puis on appuie sur la touche Entrée. Le symbole de la multiplication est le `*`, la division est `/`, la puissance est `^`.

Application : Calculez $1 + 1$; 5×3 ; 5^3 (utilisez troix lignes différentes).

1.2 Utilisation des fonctions

R permet d'utiliser toutes les fonctions mathématiques de base. Pour cela, il faut taper le nom de la fonction puis la valeur à calculer entre parenthèses. Par exemple, pour calculer log(10), on tape `log(10)`. Les fonctions de base sont : `exp()` pour l'exponentielle, `log()` pour le logarithme ; `sqrt()` pour la racine carrée (abréviation de SQuare RooT) ; `cos()` et `sin()` pour les fonctions trigonométriques.

Application : Calculez $exp(5)$; $\sqrt{60}$; $log(\sqrt{10})$

1.3 Utilisation des variables informatiques

R permet à l'utilisateur de *stocker* ses calculs dans la mémoire de l'ordinateur pour pouvoir l'utiliser plus tard. Chaque case mémoire est appelée *variable informatique* (par opposition à *variable statistique*) ou *variable* quand il n'y a pas d'ambiguïté.
Chaque variable informatique est désignée par un nom, nom choisi par l'utilisateur. Les noms de variables ne doivent jamais contenir d'accent ou d'espace. De plus, n'utilisez pas les fonctions **R** comme nom de variable (par exemple, ne tapez pas `sqrt <- 1+2`.
Pour mettre un calcul dans une variable (c'est à dire dans la mémoire), il faut utiliser une flèche : `age <- 5` signifie *"je place la valeur 5 dans la variable age"*. Le symbole `<-` est composé du signe inférieur puis du signe moins.

Exemple : je décide d'utiliser une variable appelée "cigarettes". Tapez `cigarettes <- 18`. La case mémoire `cigarettes` contient maintenant la valeur 18. Pour vérifier, il faut demander à **R** d'afficher le contenu de la variable `cigarettes`. Pour cela, tapez simplement `cigarettes`.

On peut naturement stocker en mémoire plusieurs variables. Tapez ensuite `dureeSejour <- 5`. Vérifier que la case mémoire `cigarettes` contient toujours 18. Puis vérifier que `dureeSejour` contient 5.
On peut utiliser des variables dans des calculs : tapez `cigarettes/dureeSejour`. Qu'obtenez-vous ?

Application stockez votre taille dans une variable dont vous choisirez le nom. Stockez ensuite la taille de votre voisin(e) dans une autre variable. Additionnez les deux variables et divisez le tout par 2 et stockez le résultat dans une troisième variable.

2 Lecture de données sous R

R permet de lire des données Excel. Pour cela, il faut préparer les données, puis préparer **R** et enfin lire les données :

2.1 Préparation des données

Il est interdit à un logiciel de lire les fichiers de format .xls (Excel, format propriétaire de microsoft). Par contre, **R** sait lire les fichiers au format .csv . Donc, nous allons préparer un fichier .csv. Pour cela :
– Dans le dossier *Mes Documents*, créez un dossier *M1StatInfo* (en respectant les majuscules et minuscules).
– Ouvrez Excel
– Saisissez le tableau suivant :

identifiant	contact	reaction
1	non	non
2	oui	non
3	non	non
4	oui	oui
5	non	non
6	oui	non

– Sauvegardez votre document au format .csv. Pour cela, allez dans *Enregistrez sous*. Dans type de fichier, choisissez *CSV (séparateur point-virgule)(*.csv)*. Puis sauvegardez votre fichier dans le répertoire *M1StatInfo*, sous le nom *essai*. Excel essaie de vous faire peur (forcément, quand on utilise un format non contrôlé par microsoft, Excel râle). Lisez la mise en garde puis cliquez sur OK.

Vous venez de créer un fichier .csv utilisable par **R**. Avant de passer à la suite, vérifiez qu'un fichier nommé *essai* de type csv vient bien d'être créé dans le répertoire *M1StatInfo*. Si ça n'est pas le cas, recommencez...

2.2 Lecture sous R

Pour lire un fichier sous **R**, il faut faire deux choses :
– Préciser à **R** le répertoire dans lequel le fichier se trouve.
– Procéder à la lecture du fichier.

Pour préciser le répertoire :
– Allez dans le menu déroulant *Fichier* puis *Changer le répertoire courant...*
– Sélectionnez votre répertoire (le répertoire *M1StatInfo* dans *Mes Documents*, puis cliquez sur OK.
R sait maintenant a quel endroit il doit allez lire les fichiers de donner. Pour la lecture proprement dite, on utilise l'instruction `read.csv2("nom_de_fichier.csv")`

Dans notre cas : `read.csv2("essai.csv")`. **R** lit le fichier et l'affiche à l'écran.
Pour pouvoir manipuler ce fichier (et faire des statistiques dessus), il faut le stocker dans une variable de type un peu spécial qu'on appelle *data.frame*. Pour cela, on fait comme lorsqu'on voulait stocker des nombres, on utilise `<-`. Tapez `donnees <- read.csv2("essai.csv")`. Il ne se passe rien a l'écran, mais *donnee* contient maintenant les colonnes que vous aviez tapez sous Excel. Pour vérifiez que c'est bien le cas, tapez simplement `donnees`, **R** affichera le contenu de *donnee*.
donnee est une variable un peu particulière puisqu'elle contient plusieurs colonnes et plusieurs lignes. C'est pour cela que ce type de variable n'est pas appelé variable mais est appelé *data.frame*.

Application : sous Excel, créez un fichier contenant trois colonnes :

identifiant	age	taille
1	18	1,86
2	19	1,92
3	18	1,85
4	20	1,89
5	19	1,75
6	19	1,90

Note : sous excel, les nombres décimaux sont notés avec une virgule (notation française) ; sous R, ils sont notés avec un point (notation anglo-saxone). Heureusement, la conversion de 1,86 en 1.86 est faite automatiquement par la fonction `read.csv2()`

En reprenant les étapes précédentes, importez ce tableau sous **R** (le fichier devra s'appeler *essai2* ; au final, les données devront être dans le data.frame *donnees2*).

2.3 Travailler sur un tableau de donnée

Pour accéder aux données, on dispose de plusieurs méthodes :
- Pour afficher toutes les données, il suffit de taper le nom du data.frame qui les contient : dans notre cas : `donnees`
- Pour afficher une seule case, on tape *donnees* suivi du numéro de la ligne, une virgule, le numéro de la colonne, le tout entre crochet : `donnees[4,2]` donnera la valeur de la quatrième ligne et deuxième colonne.
- Pour afficher une seule ligne, on tape *donnees* suivi du numéro de la ligne et une virgule le tout entre crochet : `donnees[4,]` donne la quatrième ligne.
- Pour afficher une seule colonne, on tape *donnees* suivi d'une virgule et du numéro de la colonne, le tout entre crochet derrière : `donnees[,2]` donne la deuxième colonne. On peut aussi taper *donnees* suivi du caractère $ puis le nom de la colonne : `donnees$contact`

Application : affichez la colonne des tailles du deuxième fichier, puis la sixième ligne du deuxième fichier.

3 R, premiers indices

3.1 Sur nos exemples

Pour nos études, nous commencerons toujours par charger un fichier de données dans **R**. Nous stockerons ce fichier dans un data.frame. Chaque ligne de ce fichier correspondra à un individu de notre étude. Chaque colonne sera une variable statistique. Dans *donnees*, les variables statistiques sont *identifiant*, *contact* et *reaction*. Rappel : on peut accéder à ces variables en tapant `donnees$identifiant` , `donnees$contact` et `donnees$reaction` .

R va ensuite permettre de travailler sur les variables. La première opération est de compter les effectifs de chaque variable. Pour cela, on doit utiliser l'instruction `table()` . Pour obtenir les effectifs de la variable contact, tapez `table(donnees$contact)` .

Application : pour chacune des quatre variables *contact*, *reaction*, *age* et *taille* :
- Dressez, grâce à une fonction **R**, le tableau des effectifs.
- Y a-t-il des cas où dresser le tableau des effectifs n'apporte aucune information ?

On peut ensuite représenter graphiquement le tableau des effectifs. Pour cela, on utilise l'instruction `barplot()` : tapez `barplot(table(donnees$contact))` .

Application : pour chaque variable (y compris celle pour laquelle le tableau des effectifs n'apporte pas d'information), donnez une représentation graphique.

Enfin, on peut dresser le tableau des effectifs d'une variable *relativement* à une autre. On l'appelle le *tableau croisé* (ou tableau croisé dynamique sous Excel. Vous souvenez-vous du temps qu'il fallait pour le faire sous Excel ? Attention, sous **R**, ça décoiffe...) Pour obtenir un tableau croisé, on utilise l'instruction `table()` en lui donnant le nom des deux variables à croiser : `table(donnees$contact,donnees$reaction)`

3.2 Sur un cas reel

Nous allons maintenant travailler sur un cas réel. Il s'agit d'un extrait de l'enquête EPO2008 "Connaissance du SIDA".

- Sur le site, téléchargez le fichier *M1StatTD1exemple.csv* et enregistrez-le dans le répertoire *M1StatInfo*.
- Lisez ce fichier à partir de **R**.
- Dressez le tableau des effectifs de chacune des variables.
- Représentez graphiquement les effectifs de chacune des variables. Y a-t-il des cas ou une telle représentation ne présente pas d'intérêt ?
- Dressez le tableau des effectifs croisés pour chaque couple de variable (Sexe et NiveauDEtude ; Sexe et SituationConjugale ; ...) Y a-t-il des cas ou il ne présente pas d'intérêt ?

Rédigez une rapide conclusion.

The end