TP1 : Lecture et affichage de tableaux avec R Statistiques Multidimensionnelles – Oct.-Nov. 2013

Pour ces travaux pratiques, on utilisera le langage/logiciel R, lancé depuis un interprète de commande - ou avec une interface comme R Commander. Pour garder une trace des travaux effectués, on utilisera un fichier script, qui peut être exécuté dans R avec la fonction source. R est un logiciel libre, disponible (à l'adresse `www.r-project.org`) pour les systèmes d'exploitation les plus répandus. On peut accéder à une documentation complète de chaque fonction avec la fonction help. Le manuel d'introduction donne des explications plus simples sur les principales fonctionnalités du langage : voir `cran.r-project.org/doc/manuals/R-intro.pdf`.

Types de variables

Lors de ces travaux pratiques, on va manipuler des tables de données, décrivant souvent certains caractères pour plusieurs individus. On a en général un individu par ligne, et chaque colonne correspond à un caractère des individus, autrement dit à une *variable*. On peut distinguer au moins 3 types de variables :

variables numériques: les valeurs sont des nombres réels, qu'on peut comparer, et telles que l'écart entre deux valeurs a un sens. Par exemple : taille ou poids d'un individu.

variables ordinales: les valeurs sont des catégories, il y a un ordre entre ces catégories. Par exemple, une variable qui aurait pour valeurs possibles tres grand, grand, moyen, petit, tres petit.

variables nominales: les valeurs sont des catégories, et on n'a pas d'ordre entre ces catégories. Par exemple, une variable dont les valeurs seraient les nationalités d'individus.

Avec R, les variables nominales et ordinales sont représentées par un même type, le type factor.

Exercice 1 Télécharger les tables suivantes, repérer les individus et indiquer le type de chaque variable :
- `onu67_budget_temps.txt`
- `cesp92_budget_temps_multimedia.txt`
- `pbio.txt`
- `ronfle.txt`
- `presidentielles07_regions.txt`

Lecture de tableaux de données

read.table : c'est la fonction principale pour lire une table de données. Range les données dans un « data.frame ». Quelques paramètres (avec leur valeur par défaut) :

sep (" ") : le caractère qui sépare les colonnes ; le défaut " "représente à la fois les espaces et les tabulations

dec (".") : le caractère qui marque la séparation entre les parties entière et décimale des nombres

comment.char ("#") : le caractère qui indique une ligne qui ne doit pas être lue dans le fichier

colClasses: un vecteur qui permet de spécifier les types des variables / colonnes (en l'absence de ce paramètre, R fait « au mieux »).

header : un booléen qui indique si la première ligne contient les noms des colonnes

row.names : un vecteur donnant les noms des lignes

col.names : un vecteur donnant les noms des colonnes (si header=FALSE)

Remarque : si la première ligne contient un champs de moins que les autres, alors par défaut header=TRUE, la première ligne est utilisée pour les noms des colonnes, et la première colonne est utilisée pour les noms des lignes.

Exemple

$T \leftarrow$read.table("onu67_budget_temps.txt",colClasses=c("character",rep("numeric",10),rep("factor",4)))

(La fonction rep reproduit sont premier argument, par exemple rep("factor", 4) est équivalent à :
c("factor","factor","factor","factor").)

Exercice 2 Utiliser la fonction read.table pour lire les tables de l'exercice 1. On **vérifiera** que chaque table a été bien lue avec les commandes summary et head.

Manipulation des data.frame Si T est un tableau représenté par un data.frame(par exemple le résultat de read.table), on peut :

- avoir les noms des lignes et des colonnes avec les instructions row.names(T) et names(T) ; on peut changer ces noms, en affectant à row.names(T) ou names(T) un vecteur de chaînes de caractères (names(T) $\leftarrow \ldots$) ;
- extraire certaines lignes ou colonnes avec par exemple $T[, c(4, 8)]$ (pour la 4ème et la 8ème colonne) ou encore $T[3,]$ (la 3ème ligne) ;
- extraire certaines lignes qui vérifient certaines conditions, par exemple : $T[T\$\text{PAYS} == 1,]$ pour extraire les lignes pour lesquelles la variable PAYS vaut 1 ;
- rajouter des lignes et des colonnes avec les commandes rbind et cbind ;

Exercice 3 Quel est le résultat de l'instruction T$PAYS $== 1$, si T est la table du fichier onu67_budget_temps.txt ?

À l'aide des fonctions ifelse et cbind, rajouter à la table T une colonne qui contient les noms des pays de chaque individu ("USA", "Yougoslavie", "Autre Ouest" ou "Autre Est").

Graphiques bi-variés

plot : c'est la fonction de base pour afficher des points avec deux coordonnées. Elle re-créée une graphique à chaque appel, et place des points selon deux axes.

legend : cette fonction permet d'ajouter une légende.

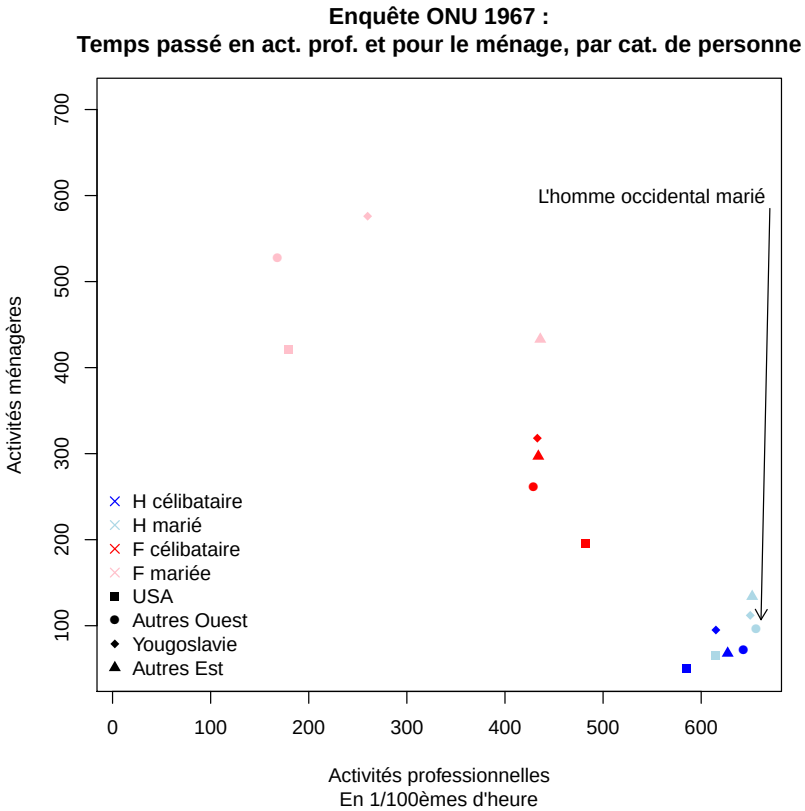
Exemple

plot($T[, c(1,3)]$, pch = ifelse(T\$SEX == 1, 25, 26), col = ifelse(T\$SEX == 1, "blue", "red"))
legend("bottomleft", c("H", "F"), col = c("blue", "red"), pch = c(25, 26))

Le premier paramètre de la fonction plot – si c'est une table ayant au moins deux colonnes – ou les deux premiers paramètres – si ce sont des vecteurs – donnent les coordonnées des points à dessiner. Autres paramètres :

pch : un (vecteur d')entier(s) indiquant le(s) symbole(s) à utiliser ;
col : un (vecteur de) couleur(s) indiquant le(s) couleur(s) à utiliser ("black" par défaut, on peut utiliser "blue", "red",..., la commande colors() retourne la liste des couleurs possibles) ;
xlim, ylim : les limites des axes horizontaux et verticaux (par défaut, les limites sont calculées de manières à faire rentrer exactement tous les points) ;
xlab, ylab : les labels des axes horizontaux et verticaux (par défaut, les noms des variables) ;
main : le titre du graphique (en gras, centré au-dessus du graphique) ;
sub : le sous-titre du graphique (centré au-dessous du graphique).

Exercice 4 Créer le graphique ci-contre, avec des couleurs distinctes pour distinguer les hommes des femmes, ainsi que leur statut marital. (Il faut utiliser les fonctions text, pour d'ajouter du texte à un graphique, et arrows, pour ajouter des flèches (la fonction lines permet d'ajouter des lignes).



Exercice 5 Tester l'instruction suivante :
pairs($T[, 1:10]$).
Quelles variables semblent corrélées (positivement ou négativement) ?

Créer des graphiques dans des fichiers C'est possible à l'aide des fonctions pdf, tiff, png, jpeg entre autres. Par exemple :

pdf(mon_fichier.pdf) ; ... *instructions graphiques* ... ; dev.off()

Pour aller plus loin Le manuel d'introduction à R (cran.r-project.org/doc/manuals/R-intro.pdf) contient une liste de fonctions graphiques, ainsi qu'une liste de paramètres que peuvent accepter ces fonctions (page 68).