MST - Logiciel R

TP n° 2

La librairie Rcmdr

La librairie ou *package* `Rcmdr` (appelée « R commander ») permet une utilisation conviviale du logiciel R grâce à une interface graphique où les traitements statistiques sont accessibles à travers un menu déroulant. `Rcmdr` ne propose pas toutes les fonctions disponibles sous R, mais ce sont en revanche les analyses statistiques les plus couramment utilisées en pratique.

Installation et chargement de la librairie Rcmdr

La librairie est en général installée par défaut lors de l'installation de R. Il est nécessaire néanmoins de charger `Rcmdr` à chaque nouvelle session R grâce la commande :

```
> library(Rcmdr)
```

Tant qu'aucun jeu de données n'est importé dans `Rcmdr`, aucune analyse statistique n'est possible ; seules les fonctionnalité du menu `Distribution` sont utilisables (générer des échantillons aléatoires selon une distribution, calculer des probabilités, représentation graphique de distributions usuelles).

Importer un jeu de données - Premières vérifications

Généralités sur les data.frame

- Dans R, un jeu de données est associé à un objet de type `data.frame`. C'est l'équivalent du tableau statistique *individus* x *variables*, format de base à partir duquel sont effectués la majorité des traitements statistiques.
- En important un tableau *individus* x *variables*, `Rcmdr` crée automatiquement un objet de type `data.frame`.
- On peut convertir une matrice en data.frame à l'aide de la fonction `as.data.frame`.

1. À partir du menu *Données*, importer le fichier texte `taillefum.txt` (séparateur de champ = tabulation ; séparateur décimal = virgule).

   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 R commander. On peut ainsi facilement réexécuter une commande ou modifier certains arguments de la fonction soumise.

2. **Inspection rapide des données**. Effectuer les opérations suivantes :
   - Visualiser le jeu de données (aucune modification des données possibles)
   - Éditer les données (modification possible des données ou des noms de variables)

— Avoir un aperçu des premières ou dernières lignes du tableau :
```
> head(taillefum) ; tail(taillefum)
```
— Connaître le nom des variables :
```
> names(taillefum)
```
— Afficher les dimensions du jeu de données :
```
> dim(taillefum)
```
— Afficher les valeurs d'une variable, par exemple la variable `Annee` :
`> taillefum$Annee` ou `> taillefum[,1]`
— Afficher les données d'un individu, par exemple le 3e :
```
> taillefum[3,]
```

Remarque : pour désigner une variable, on peut se dispenser de faire référence au jeu de données en utilisant au préalable la fonction `attach` :
```
> attach(taillefum)
> Annee
```

Gestion des données - Premiers traitements statistiques

1. Produire un **résumé statistique** des données (fonction `summary` dans R) :
   Statistiques → *Résumés* → *Jeu de données actif.*
   Repérer quelles sont les variables qui ont été importées comme quantitatives (type *numeric* dans R) et comme qualitatives (appelées *factor* dans R).
2. À l'aide du menu de gestion des variables, **convertir** la variable `Annee` en un facteur.
3. Identifier la variable comportant une **valeur manquante** (à laquelle R attribue le code interne `NA` (pour *Not Applicable*).
4. Pour **connaître l'indice** de l'individu présentant la valeur manquante, on utilise la fonction `which` :
   ```
   > which(is.na(Taille.Pere))
   ```
   Quels sont les indices des individus dont le père est plus petit que 1.65 m ?
5. À l'aide du menu *Données* → *Jeu de données actif*, supprimer l'individu présentant la valeur manquante. Appeler le nouveau jeu de données `TF`.
6. **Sélection d'un sous-ensemble de données** dans un dataframe.
   — Créer deux nouveaux jeux de données en sélectionnant le sous-ensemble des données relatives exclusivement aux filles et celui relatif aux garçons :
   ```
   > TF.fille <- subset(taillefum, subset = Sexe=="Fille")
   > TF.garcon <- subset(taillefum, subset = Sexe=="Garcon")
   ```
   Remarque : cette opération est possible également à travers le menu :
   Données → *Jeu de données actif* → *Sous-ensemble.*
   — Afficher les données relatives aux élèves mesurant plus de 1.80 m ; aux filles mesurant plus de 1.80 m.

Premiers traitements statistiques

1. À partir du jeu de données `TF`, produire une **représentation graphique** pour chacune des variables quantitatives :
   — à l'aide d'un graphe faisant apparaître l'ensemble des valeurs individuelles : graphe indexé ou de type `stripchart` (non proposé dans le menu) ;
   — à l'aide d'une représentation graphique par boîte à moustaches. Peut-on considérer qu'il existe des valeurs « extrêmes » ? Si oui, pour quelles variables ? Identifier les individus présentant ces valeurs.
2. Sur une même feuille graphique, représenter les **histogrammes** des trois variables quantitatives l'un au dessus de l'autre. Indication : utiliser une échelle identique pour les trois variables (en ajoutant l'argument `xlim` à la fonction `Hist`).
3. Sur une nouvelle feuille graphique, faire de même avec les **fonctions de répartition** empiriques.
4. Produire une représentation graphique pour chaque variable qualitative (sur une même feuille graphique).
5. À l'aide du menu *Statistique* → *Résumé*, calculer la moyenne, l'écart type, le coefficient de variation et les quartiles pour toutes les variables continues. Quelle est la variable présentant la plus forte dispersion ?
6. Fournir les distributions de fréquences de chacune des variables qualitatives.

Gérer et recoder les variables

1. **Découpage d'une variable** continue en classe.
   — Découper la variable `Taille` en quatre classes de même amplitude (= de *taille* égale dans Rcmdr). Appeler la nouvelle variable `Taille.4Ca` et choisir les amplitudes comme noms de niveaux.
   — Découper en quatre classes de même effectif la variable `Taille`. Appeler la nouvelle variable `Taille.4Ce` et choisir les amplitudes comme noms de niveaux.
   — Proposer une représentation graphique pour les nouvelles variables obtenues.
2. **Calcul d'une nouvelle variable**. On estime qu'il est possible de prédire approximativement la taille d'un enfant à partir de celle de ses parents en calculant la moyenne des tailles des parents puis en ajoutant (respectivement en retranchant) 6.5 cm si c'est un garçon (respectivement si c'est une fille). Calculer cette nouvelle variable (qu'on appellera `Taille.Pred`) de la façon suivante :
   — À partir de la variable `Sexe`, créer une nouvelle variable prenant les valeurs $+1$ ou -1 selon que l'élève est un garçon ou une fille. (Indication : menu *Recoder des variable* ; entrer les deux directives de recodage : `"Fille" = -1` et `"Garcon" = 1`). Appeler `Sexe.Num` la variable créée.
   — Calculer la nouvelle variable `Taille.Pred` selon la formule donnée plus haut (menu : *calculer une nouvelle variable*).
   — À l'aide d'une représentation graphique par nuage de points, évaluer dans quelle mesure la variable calculée est une bonne estimation de la taille réelle d'un élève. Ajouter sur ce graphique la première bissectrice (fonction `abline()`).
3. Calculer les **données centrées réduites** pour toutes les variables quantitatives. Produire ensuite la moyenne et l'écart type des nouvelles variables (dites standardisées.)
4. **Enregistrer** le jeu de données `TF` au format interne de R (fichier `.RData`).

Comparaison de sous-populations, liaison entre deux variables

1. **Comparer des sous-populations**.
   — Fournir un résumé statistique de la `Taille` (des élèves), par sexe. Construire ensuite une représentation graphique par boîte à moustaches permettant de comparer la variable `Taille` selon le `Sexe`. Ajouter un titre au graphique ainsi que des couleurs aux boîtes de dispersion obtenues.
   — Faire de même pour la taille de la mère.
   — Comparer la taille moyenne d'un élève selon les années, puis selon la combinaison des deux facteurs (Sexe, Annee) (menu : *tableau de statistique*). Associer une représentation graphique à chacune de ces comparaisons.
2. Introduction à la notion de **corrélation linéaire**. On souhaite savoir si la taille d'un élève est plus ou moins liée à la taille de sa mère ou à la taille de son père.
   — Sélectionner le jeu de données `TF.garcon`.
   — À l'aide d'une matrice de nuage de points, représenter les liaisons entre les trois variables `Taille`, `Taille.Mere` et `Taille.Pere`.
   — Selon vous, la taille d'un élève est-elle davantage liée à celle de son père ou de sa mère ? Que peut-on dire de la liaison entre la taille du père et celle de la mère ?
   — Calculer les coefficients de corrélation linéaire entre tous les couples de variables (*Résumé* → *Matrice de corrélation*).
3. Étudier la **liaison entre deux facteurs**. On souhaite étudier si le fait qu'un élève fume dépend ou non du statut de fumeur de ses parents.
   — Sélectionner le jeu de données global `TF`.
   — Créer un tableau croisé (appelé aussi tableau de contingence) entre les variables `Parents` (en ligne) et `Etat` (en colonne). Demander l'affichage des profils lignes. Commenter.
   — Comparer le profil Fumeur / Non fumeur d'un élève selon le statut de fumeur des parents en soumettant la commande graphique :

```
> barplot(t(rowPercents(.Table)[,1:2]), beside=TRUE,
xlab="Parents", cex.lab=1.3, col=c("darkorange3","darkslateblue"),
legend.text = paste("Etat",attr(t(.Table), "dimnames")[[1]], sep=":") )
```