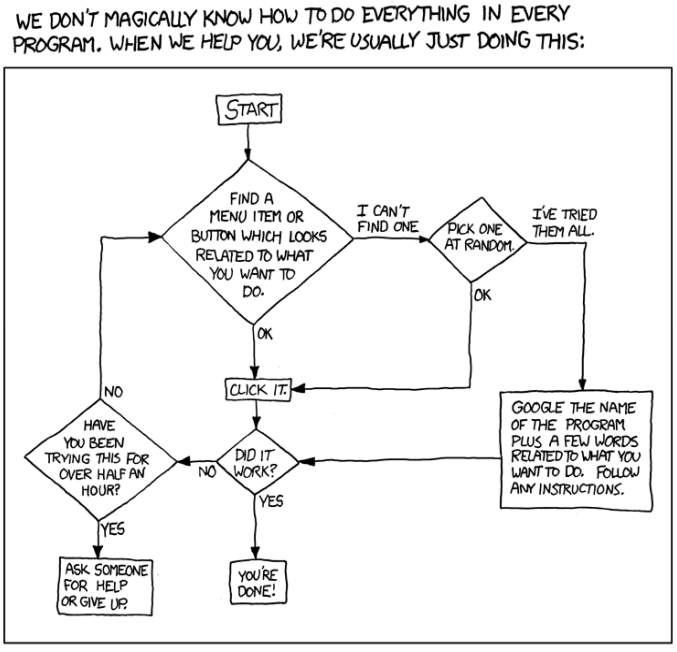
Introduction à 

Julien Barnier
Centre Max Weber
julien.barnier@ens-lyon.fr

Version 2.2
25 avril 2016



http://xkcd.com/627/

Table des matières

Partie 1

Introduction

1.1 À propos de ce document

Ce document a pour objet de fournir une introduction à l'utilisation du logiciel libre de traitement de données et d'analyse statistiques R. Il se veut le plus accessible possible, y compris pour ceux qui ne sont pas particulièrement familiers avec l'informatique.

Ce document est basé sur R version 3.2.5 (2016-04-14).

La page Web « officielle » sur laquelle on pourra trouver la dernière version de ce document se trouve à l'adresse :

http://alea.fr.eu.org/pages/intro-R

1.2 Licence

1.3 Remerciements

L'auteur tient à remercier Mayeul Kauffmann, Julien Biaudet, Frédérique Giraud, Joël Gombin pour leurs corrections et suggestions. Et un remerciement plus particulier à Milan Bouchet-Valat pour sa relecture très attentive et ses nombreuses et judicieuses remarques.

Joseph Larmarange[1] est l'auteur de l'ensemble des encadrés concernant RStudio et de nombreuses autres corrections et améliorations. Il maintient par ailleurs une autre version de ce document, enrichie de chapitres sur différentes méthodes statistiques :

https://github.com/larmarange/intro-r/tree/CoursM2

1. http://joseph.larmarange.net/

1.4 Conventions typographiques

Ce document suit un certain nombre de conventions typographiques visant à en faciliter la lecture. Ainsi les noms de logiciel et d'extensions sont indiqués en caractères sans empattement (R, SAS, Linux, questionr, ade4...). Les noms de fichiers sont imprimés avec une police à chasse fixe (`test.R`, `data.txt`...), tout comme les fonctions R (`summary`, `mean`, `<-`...).

Lorsqu'on présente des commandes saisies sous R et leur résultat, la commande saisie est indiquée avec une police à chasse fixe et précédée de l'invite de commande `R>` :

```
R> summary(rnorm(100))
```

Le résultat de la commande tel qu'affiché par R est également indiqué dans une police à chasse fixe :

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.59600 -0.61620  0.04094  0.03073  0.74090  2.09300
```

Lorsque la commande R est trop longue et répartie sur plusieurs lignes, les lignes suivantes sont précédées du symbole `+` :

```
R> coo <- scatterutil.base(dfxy = dfxy, xax = xax, yax = yax, xlim = xlim, ylim = ylim,
+     grid = grid, addaxes = addaxes, cgrid = cgrid, include.origin = include.origin)
```

1.5 Présentation de R

R est un langage orienté vers le traitement de données et l'analyse statistique dérivé du langage S. Il est développé depuis une vingtaine d'années par un groupe de volontaires de différents pays. C'est un logiciel libre [2], publié sous licence GNU GPL.

L'utilisation de R présente plusieurs avantages :

- c'est un logiciel *multiplateforme*, qui fonctionne aussi bien sur des sytèmes Linux, Mac OS X ou Windows ;
- c'est un logiciel *libre*, développé par ses utilisateurs et modifiable par tout un chacun ;
- c'est un logiciel *gratuit* ;
- c'est un logiciel très puissant, dont les fonctionnalités de base peuvent être étendues à l'aide d'extensions [3] ;
- c'est un logiciel dont le développement est très actif et dont la communauté d'utilisateurs ne cesse de s'élargir ;
- c'est un logiciel avec d'excellentes capacités graphiques.

Comme rien n'est parfait, on peut également trouver quelques inconvénients :

- le logiciel, la documentation de référence et les principales ressources sont en anglais. Il est toutefois parfaitement possible d'utiliser R sans spécialement maîtriser cette langue ;
- il n'existe pas encore d'interface graphique pour R équivalente à celle d'autres logiciels comme SPSS ou Modalisa [4]. R fonctionne à l'aide de scripts (des petits programmes) édités et exécutés au fur et à mesure de l'analyse, et se rapprocherait davantage de SAS dans son utilisation (mais avec une syntaxe et une philosophie très différentes). Ce point, qui peut apparaître comme un gros handicap, s'avère après un temps d'apprentissage être un mode d'utilisation d'une grande souplesse.

2. Pour plus d'informations sur ce qu'est un logiciel libre, voir : http://www.gnu.org/philosophy/free-sw.fr.html
3. Il en existe actuellement plus de 8000, disponibles sur le *Comprehensive R Archive Network* (CRAN) : http://cran.r-project.org/

— comme R s'apparente davantage à un langage de programmation qu'à un logiciel proprement dite, la courbe d'apprentissage peut être un peu « raide », notamment pour ceux n'ayant jamais programmé auparavant.

1.6 Philosophie de R

Deux points particuliers dans le fonctionnement de R peuvent parfois dérouter les utilisateurs habitués à d'autres logiciels :

— sous R, en général, on ne voit pas les données sur lesquelles on travaille ; on ne dispose pas en permanence d'une vue des données sous forme de tableau, comme sous Modalisa ou SPSS. Ceci peut être déroutant au début, mais on se rend vite compte qu'on n'a pas besoin de voir en permanence les données pour les analyser ;
— avec les autres logiciels, en général la production d'une analyse génère un grand nombre de résultats de toutes sortes dans lesquels l'utilisateur est censé retrouver et isoler ceux qui l'intéressent. Avec R, c'est l'inverse : par défaut l'affichage est réduit au minimum, et c'est l'utilisateur qui demande à voir des résultats supplémentaires ou plus détaillés.

Inhabituel au début, ce fonctionnement permet en fait assez rapidement de gagner du temps dans la conduite des analyses.

4. Certaines extensions ou logiciels proposent cependant des interfaces graphiques plus ou moins généralistes. Voir la section A.4, page 122

Partie 2

Prise en main

L'installation du logiciel proprement dite n'est pas décrite ici mais indiquée dans l'annexe A, page 121. On part donc du principe que vous avez sous la main un ordinateur avec une installation récente de R, quel que soit le système d'exploitation que vous utilisez (Linux, Mac OS X ou Windows).

Le projet RStudio tend à s'imposer comme l'environnement de développement de référence pour R, d'autant qu'il a l'avantage d'être libre, gratuit et multiplateforme. Son installation est décrite section A.5 page 122.
Les astuces et informations spécifiques à RStudio seront présentées tout au long de ce document dans des encadrés similaires à celui-là.
RStudio peut tout à fait être utilisé pour découvrir et démarrer avec R.

2.1 L'invite de commandes

Une fois R lancé, vous obtenez une fenêtre appelée *console*. Celle-ci contient un petit texte de bienvenue ressemblant à peu près à ce qui suit [1] :

```
R version 3.2.4 (2016-03-10) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

(...)
```

suivi d'une ligne commençant par le caractère > et sur laquelle devrait se trouver votre curseur. Cette ligne est appelée l'*invite de commande* (ou *prompt* en anglais). Elle signifie que R est disponible et en attente de votre prochaine commande.

1. La figure 2.1 page ci-contre montre l'interface par défaut sous Windows.

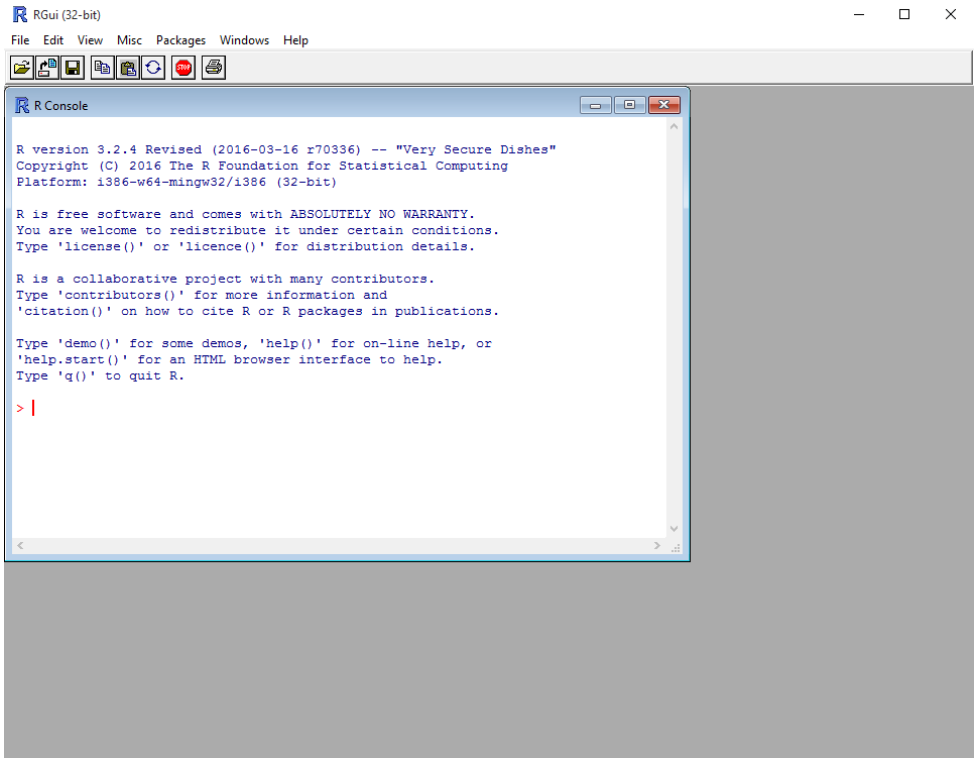


FIGURE 2.1 – L'interface de R sous Windows au démarrage

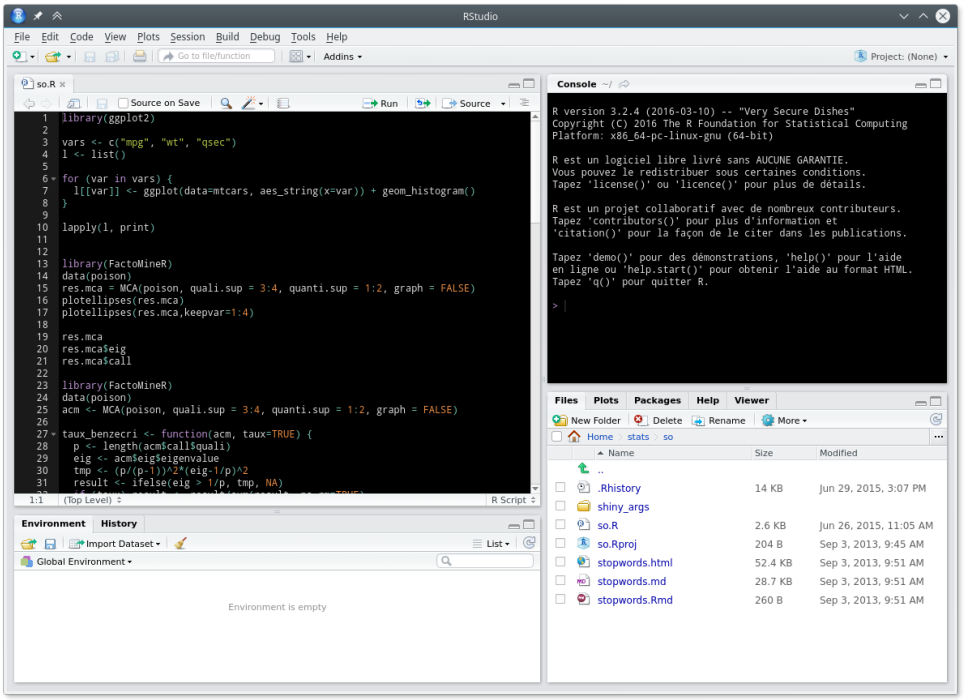


FIGURE 2.2 – L'interface de RStudio au démarrage

L'interface de RStudio se présente différement (voir figure 2.2). Elle est divisée en quatre parties. Le quadrant haut-gauche est dédié aux fichiers sources (scripts). Le quadrant haut-droite fournit des informations sur vos données en mémoire et votre historique. Le quadrant bas-droite vous permet naviguer dans votre répertoire de travail, affiche l'aide, vos graphiques et les extensions disponibles. Enfin, la *console* est affichée en bas à gauche. C'est elle qui nous intéresse pour le moment. Nous aborderons les autres quadrants plus loin dans ce document.

Nous allons tout de suite lui fournir une première commande :

```
R> 2 + 3

[1] 5
```

Bien, nous savons désormais que R sait faire les additions à un chiffre[2]. Nous pouvons désormais continuer avec d'autres opérations arithmétiques de base :

```
R> 8 - 12

[1] -4

R> 14 * 25

[1] 350

R> -3/10

[1] -0.3
```

Une petite astuce très utile lorsque vous tapez des commandes directement dans la console : en utilisant les flèches *Haut* et *Bas* du clavier, vous pouvez naviguer dans l'historique des commandes tapées précédemment, que vous pouvez alors facilement réexécuter ou modifier.

Sous RStudio, l'onglet *History* du quadrant haut-droite vous permet de consulter l'historique des commandes que vous avez transmises à R. Un double-clic sur une commande la recopiera automatiquement dans la console. Vous pouvez également sélectionner une ou plusieurs commandes puis cliquer sur *To Console*. Voir également (en anglais) : `http://www.rstudio.com/ide/docs/using/history`

Lorsqu'on fournit à R une commande incomplète, celui-ci nous propose de la compléter en nous présentant une invite de commande spéciale utilisant les signe `+`. Imaginons par exemple que nous avons malencontreusement tapé sur `Entrée` alors que nous souhaitions calculer `4*3` :

2. La présence du `[1]` en début de ligne sera expliquée par la suite, page 13.

```
4 *
```

On peut alors compléter la commande en saisissant simplement `3` :

```
R> 4 *
+ 3

[1] 12
```

Pour des commandes plus complexes, il arrive parfois qu'on se retrouve coincé avec un invite `+` sans plus savoir comment compléter la saisie correctement. On peut alors annuler la commande en utilisant la touche `Echap` ou `Esc` sous Windows. Sous Linux on utilise le traditionnel `Control + C`.

À noter que les espaces autour des opérateurs n'ont pas d'importance lorsque l'on saisit les commandes dans R. Les trois commandes suivantes sont donc équivalentes, mais on privilégie en général la deuxième pour des raisons de lisibilité du code.

```
R> 10+2
R> 10 + 2
R> 10       +       2
```

2.2 Des objets

2.2.1 Objets simples

Faire des opérations arithmétiques, c'est bien, mais sans doute pas totalement suffisant. Notamment, on aimerait pouvoir réutiliser le résultat d'une opération sans avoir à le resaisir ou à le copier/coller.

Comme tout langage de programmation, R permet de faire cela en utilisant des *objets*. Prenons tout de suite un exemple :

```
R> x <- 2
```

Que signifie cette commande ? L'opérateur `<-` est appelé *opérateur d'assignation*. Il prend une valeur quelconque à droite et la place dans l'objet indiqué à gauche. La commande pourrait donc se lire *mettre la valeur 2 dans l'objet nommé* `x`.

On va ensuite pouvoir réutiliser cet objet dans d'autres calculs ou simplement afficher son contenu :

```
R> x + 3

[1] 5

R> x

[1] 2
```

Par défaut, si on donne à R seulement le nom d'un objet, il va se débrouiller pour nous présenter son contenu d'une manière plus ou moins lisible.

On peut utiliser autant d'objets qu'on veut. Ceux-ci peuvent contenir des nombres, des chaînes de caractères (indiquées par des guillemets droits ") et bien d'autres choses encore :

```
R> x <- 27
R> y <- 10
R> foo <- x + y
R> foo

[1] 37

R> x <- "Hello"
R> foo <- x
R> foo

[1] "Hello"
```

Les noms d'objets peuvent contenir des lettres, des chiffres (mais ils ne peuvent pas commencer par un chiffre), les symboles `.` et `_`, et doivent commencer par une lettre. R fait la différence entre les majuscules et les minuscules, ce qui signifie que `x` et `X` sont deux objets différents. On évitera également d'utiliser des caractères accentués dans les noms d'objets, et comme les espaces ne sont pas autorisés on pourra les remplacer par un point ou un tiret bas.
Enfin, signalons que certains noms courts sont réservés par R pour son usage interne et doivent être évités. On citera notamment `c, q, t, C, D, F, I, T, max, min`...

2.2.2 Vecteurs

Imaginons maintenant que nous avons interrogé dix personnes au hasard dans la rue et que nous avons relevé pour chacune d'elle sa taille en centimètres. Nous avons donc une série de dix nombres que nous souhaiterions pouvoir réunir de manière à pouvoir travailler sur l'ensemble de nos mesures.

Un ensemble de données de même nature constituent pour R un *vecteur* (en anglais *vector*) et se construit à l'aide d'un opérateur nommé `c`[3]. On l'utilise en lui donnant la liste de nos données, entre parenthèses, séparées par des virgules :

```
R> tailles <- c(167, 192, 173, 174, 172, 167, 171, 185, 163, 170)
```

Ce faisant, nous avons créé un objet nommé `tailles` et comprenant l'ensemble de nos données, que nous pouvons afficher :

```
R> tailles

 [1] 167 192 173 174 172 167 171 185 163 170
```

3. `c` est l'abbréviation de *combine*. Le nom de cette fonction est très court car on l'utilise très souvent.

Dans le cas où notre vecteur serait beaucoup plus grand, et comporterait par exemple 40 tailles, on aurait le résultat suivant :

```
R> tailles

 [1] 144 168 179 175 182 188 167 152 163 145 176 155 156 164 167 155 157
[18] 185 155 169 124 178 182 195 151 185 159 156 184 172 156 160 183 148
[35] 182 126 177 159 143 161 180 169 159 185 160
```

On a bien notre suite de quarante tailles, mais on peut remarquer la présence de nombres entre crochets au début de chaque ligne (`[1]`, `[18]` et `[35]`). En fait ces nombres entre crochets indiquent la position du premier élément de la ligne dans notre vecteur. Ainsi, le 185 en début de deuxième ligne est le 18^{e} élément du vecteur, tandis que le 182 de la troisième ligne est à la 35^{e} position.

On en déduira d'ailleurs que lorsque l'on fait :

```
R> 2

[1] 2
```

R considère en fait le nombre 2 comme un vecteur à un seul élément.

On peut appliquer des opérations arithmétiques simples directement sur des vecteurs :

```
R> tailles <- c(167, 192, 173, 174, 172, 167, 171, 185, 163, 170)
R> tailles + 20

 [1] 187 212 193 194 192 187 191 205 183 190

R> tailles/100

 [1] 1.67 1.92 1.73 1.74 1.72 1.67 1.71 1.85 1.63 1.70

R> tailles^2

 [1] 27889 36864 29929 30276 29584 27889 29241 34225 26569 28900
```

On peut aussi combiner des vecteurs entre eux. L'exemple suivant calcule l'indice de masse corporelle à partir de la taille et du poids :

```
R> tailles <- c(167, 192, 173, 174, 172, 167, 171, 185, 163, 170)
R> poids <- c(86, 74, 83, 50, 78, 66, 66, 51, 50, 55)
R> tailles.m <- tailles/100
R> imc <- poids/(tailles.m^2)
R> imc

 [1] 30.83653 20.07378 27.73230 16.51473 26.36560 23.66524 22.57105
 [8] 14.90139 18.81892 19.03114
```

Quand on fait des opérations sur les vecteurs, il faut veiller à soit utiliser un vecteur et un chiffre (dans des opérations du type `v * 2` ou `v + 10`), soit à utiliser des vecteurs de même longueur (dans des opérations du type `u + v`).
Si on utilise des vecteurs de longueur différentes, on peut avoir quelques surprises[4].

On a vu jusque-là des vecteurs composés de nombres, mais on peut tout à fait créer des vecteurs composés de chaînes de caractères, représentant par exemple les réponses à une question ouverte ou fermée :

```
R> reponse <- c("Bac+2", "Bac", "CAP", "Bac", "Bac", "CAP", "BEP")
```

Enfin, notons que l'on peut accéder à un élément particulier du vecteur en faisant suivre le nom du vecteur de crochets contenant le numéro de l'élément désiré. Par exemple :

```
R> reponse <- c("Bac+2", "Bac", "CAP", "Bac", "Bac", "CAP", "BEP")
R> reponse[2]

[1] "Bac"
```

Cette opération s'appelle *l'indexation* d'un vecteur. Il s'agit ici de sa forme la plus simple, mais il en existe d'autres beaucoup plus complexes. L'indexation des vecteurs et des tableaux dans R est l'un des éléments particulièrement souples et puissants du langage (mais aussi l'un des plus délicats à comprendre et à maîtriser). Nous en reparlerons section 5.2 page 53.

Sous RStudio, vous avez dû remarquer que ce dernier effectue une coloration syntaxique. Lorsque vous tapez une commande, les valeurs numériques sont affichées dans une certaine couleur, les valeurs textuelles dans une autre et les noms des fonctions dans une troisième. De plus, si vous tapez une parenthèse ouvrante, RStudio va créer automatiquement après le curseur la parenthèse fermante correspondante (de même avec les guillements). De plus, si vous placez le curseur juste après une parenthèse fermante, la parenthèse ouvrante correspondante sera surlignée, ce qui sera bien pratique lors de la rédaction de commandes complexes.

2.3 Des fonctions

Nous savons désormais faire des opérations simples sur des nombres et des vecteurs, stocker ces données et résultats dans des objets pour les réutiliser par la suite.

Pour aller un peu plus loin nous allons aborder, après les *objets*, l'autre concept de base de R, à savoir les *fonctions*. Une fonction se caractérise de la manière suivante :

— elle a un nom ;
— elle accepte des arguments (qui peuvent avoir un nom ou pas) ;
— elle retourne un résultat et peut effectuer une action comme dessiner un graphique, lire un fichier, etc. ;

En fait rien de bien nouveau puisque nous avons déjà utilisé plusieurs fonctions jusqu'ici, dont la plus visible est la fonction `c`. Dans la ligne suivante :

```
R> reponse <- c("Bac+2", "Bac", "CAP", "Bac", "Bac", "CAP", "BEP")
```

on fait appel à la fonction nommée `c`, on lui passe en arguments (entre parenthèses et séparées par des virgules) une série de chaînes de caractères, et elle retourne comme résultat un vecteur de chaînes de caractères, que nous stockons dans l'objet `reponse`.

Prenons tout de suite d'autres exemples de fonctions courantes :

4. Quand R effectue une opération avec deux vecteurs de longueurs différentes, il recopie le vecteur le plus court de manière à lui donner la même taille que le plus long, ce qui s'appelle la *règle de recyclage* (*recycling rule*). Ainsi, `c(1,2) + c(4,5,6,7,8)` vaudra l'équivalent de `c(1,2,1,2,1) + c(4,5,6,7,8)`.

```
R> tailles <- c(167, 192, 173, 174, 172, 167, 171, 185, 163, 170)
R> length(tailles)

[1] 10

R> mean(tailles)

[1] 173.4

R> var(tailles)

[1] 76.71111
```

Ici, la fonction `length` nous renvoie le nombre d'éléments du vecteur, la fonction `mean` nous donne la moyenne des éléments du vecteur et la fonction `var` sa variance.

2.3.1 Arguments

Les arguments de la fonction lui sont indiqués entre parenthèses, juste après son nom. En général les premiers arguments passés à la fonction sont des données servant au calcul, et les suivants des paramètres influant sur ce calcul. Ceux-ci sont en général transmis sous la forme d'argument nommés.

Reprenons l'exemple des tailles précédent :

```
R> tailles <- c(167, 192, 173, 174, 172, 167, 171, 185, 163, 170)
```

Imaginons que le deuxième enquêté n'ait pas voulu nous répondre. Nous avons alors dans notre vecteur une valeur manquante. Celle-ci est symbolisée dans R par le code `NA` :

```
R> tailles <- c(167, NA, 173, 174, 172, 167, 171, 185, 163, 170)
```

Recalculons notre taille moyenne :

```
R> mean(tailles)

[1] NA
```

Et oui, par défaut, R renvoie `NA` pour un grand nombre de calculs (dont la moyenne) lorsque les données comportent une valeur manquante. On peut cependant modifier ce comportement en fournissant un paramètre supplémentaire à la fonction `mean`, nommé `na.rm` :

```
R> mean(tailles, na.rm = TRUE)

[1] 171.3333
```

Positionner le paramètre `na.rm` à `TRUE` (vrai) indique à la fonction `mean` de ne pas tenir compte des valeurs manquantes dans le calcul.

Lorsqu'on passe un argument à une fonction de cette manière, c'est-à-dire sous la forme `nom=valeur`, on parle d'*argument nommé*.

`NA` signifie *not available*. Cette valeur particulière peut être utilisée pour indiquer une valeur manquante pour tout type de liste (nombres, textes, valeurs logique, etc.).

2.3.2 Quelques fonctions utiles

Récapitulons la liste des fonctions que nous avons déjà rencontrées :

Fonction	Description
`c`	construit un vecteur à partir d'une série de valeurs
`length`	nombre d'éléments d'un vecteur
`mean`	moyenne d'un vecteur de type numérique
`var`	variance d'un vecteur de type numérique
`+`, `-`, `*`, `/`	opérateurs mathématiques de base
`^`	passage à la puissance

On peut rajouter les fonctions de base suivantes :

Fonction	Description
`min`	valeur minimale d'un vecteur numérique
`max`	valeur maximale d'un vecteur numérique
`sd`	écart-type d'un vecteur numérique
`:`	génère une séquence de nombres. `1:4` équivaut à `c(1,2,3,4)`

Autre outil bien utile de RStudio, l'auto-complétion. Tapez les premières lettres d'une fonction, par exemple `me` puis appuyez sur la touche `<Tabulation>`. RStudio affichera la liste des fonctions dont le nom commence par `me` ainsi qu'un court descriptif de chacune. Un appui sur la touche *Entrée* provoquera la saisie du nom complet de la fonction choisie. Vous pouvez également utiliser l'auto-complétion pour retrouver un objet que vous avez créé — par exemple, appuyez sur la touche `<Tabulation>` après avoir saisi `mean(t` — ou bien pour retrouver un argument nommé d'une fonction — par exemple, appuyez sur la touche `<Tabulation>` après avoir saisi `mean(taille,`.

2.3.3 Aide sur une fonction

Il est très fréquent de ne plus se rappeler quels sont les paramètres d'une fonction ou le type de résultat qu'elle retourne. Dans ce cas on peut très facilement accéder à l'aide décrivant une fonction particulière en tapant (remplacer `fonction` par le nom de la fonction) :

```
R> help("fonction")
```

Ou, de manière équivalente, `?fonction`[5].

Ces deux commandes affichent une page (en anglais) décrivant la fonction, ses paramètres, son résultat, le tout accompagné de diverses notes, références et exemples. Ces pages d'aide contiennent à peu près tout ce que vous pourrez chercher à savoir, mais elles ne sont pas toujours d'une lecture aisée.

5. L'utilisation du raccourci `?fonction` ne fonctionne pas pour certains opérateurs comme `*`. Dans ce cas on pourra utiliser `?'*'` ou bien simplement `help("*")`.

Un autre cas très courant dans R est de ne pas se souvenir ou de ne pas connaître le nom de la fonction effectuant une tâche donnée. Dans ce cas on se reportera aux différentes manières de trouver de l'aide décrites dans l'annexe 9, page 115.

Dans RStudio, les pages d'aide en ligne s'ouvriront dans le quadrant bas-droite sous l'onglet *Help*. Un clic sur l'icône en forme de maison vous affichera la page d'accueil de l'aide.

2.4 Exercices

Exercice 2.1

▷ *Solution page 129*

Construire le vecteur suivant :

```
[1] 120 134 256  12
```

Exercice 2.2

▷ *Solution page 129*

Construire le vecteur suivant :

```
[1] 1 2 3 4 5 6 7 8 9
```

À partir de ce premier vecteur, générer les deux vecteurs suivants :

```
[1] 101 102 103 104 105 106 107 108 109
[1]  2  4  6  8 10 12 14 16 18
```

Exercice 2.3

▷ *Solution page 129*

On a demandé à 4 ménages le revenu des deux conjoints, et le nombre de personnes du ménage :

```
R> conjoint1 <- c(1200, 1180, 1750, 2100)
R> conjoint2 <- c(1450, 1870, 1690, 0)
R> nb.personnes <- c(4, 2, 3, 2)
```

Calculez le revenu total par personne du ménage.

Exercice 2.4

▷ *Solution page 129*

Dans l'exercice précédent, calculez le revenu minimum et le revenu maximum parmi ceux du premier conjoint :

```
R> conjoint1 <- c(1200, 1180, 1750, 2100)
```

Recommencer avec les revenus suivants, parmi lesquels l'un des enquêtés n'a pas voulu répondre :

```
R> conjoint1.na <- c(1200, 1180, 1750, NA)
```

Partie 3

Premier travail avec des données

3.1 Regrouper les commandes dans des scripts

Jusqu'à maintenant nous avons utilisé uniquement la console pour communiquer avec R *via* l'invite de commandes. Le principal problème de ce mode d'interaction est qu'une fois qu'une commande est tapée, elle est pour ainsi dire « perdue », c'est-à-dire qu'on doit la saisir à nouveau si on veut l'exécuter une seconde fois. L'utilisation de la console est donc restreinte aux petites commandes « jetables », le plus souvent utilisées comme test.

La plupart du temps, les commandes seront stockées dans un fichier à part, que l'on pourra facilement ouvrir, éditer et exécuter en tout ou partie si besoin. On appelle en général ce type de fichier un *script*.

Pour comprendre comment cela fonctionne, dans le menu *Fichier*, sélectionnez l'entrée *Nouveau script*[1]. Une nouvelle fenêtre (vide) apparaît. Nous pouvons désormais y saisir des commandes. Par exemple, tapez sur la première ligne la commande suivante :

```
2+2
```

Ensuite, allez dans le menu *Éditon*, et choisissez *Exécuter la ligne ou sélection*. Apparement rien ne se passe, mais si vous jetez un œil à la fenêtre de la console, les lignes suivantes ont dû faire leur apparition :

```
R> 2+2

[1] 4
```

Voici donc comment soumettre rapidement à R les commandes saisies dans votre fichier. Vous pouvez désormais l'enregistrer, l'ouvrir plus tard, et en exécuter tout ou partie. À noter que vous avez plusieurs possibilités pour soumettre des commandes à R :

— vous pouvez exécuter la ligne sur laquelle se trouve votre curseur en sélectionnant *Éditon* puis *Exécuter la ligne ou sélection*, ou plus simplement en appuyant simultanément sur les touches `<Ctrl>` et `<R>`[2] ;

1. Les indications données ici concernent l'interface par défaut de R sous Windows. Elles sont très semblables sous Mac OS X.
2. Sous Mac OS X, on utilise les touches `<Pomme>` et `<Entrée>`.

— vous pouvez sélectionner plusieurs lignes contenant des commandes et les exécuter toutes en une seule fois exactement de la même manière ;
— vous pouvez exécuter d'un coup l'intégralité de votre fichier en choisissant *Édition* puis *Exécuter tout*.

La plupart du travail sous R consistera donc à éditer un ou plusieurs fichiers de commandes et à envoyer régulièrement les commandes saisies à R en utilisant les raccourcis clavier *ad hoc*.

Les commandes sont légèrement différentes avec RStudio mais le principe est le même. Pour créer un nouveau script R, faire *File > New file > R Script*. Votre nouveau fichier apparaitra dans le quadrant haut-gauche. Pour exécuter une ou plusieurs lignes de code, sélectionnez les lignes en question puis cliquez sur l'icône *Run* ou bien appuyez simultanément sur les touches `<Ctrl>` et `<Entrée>`.
Pour plus d'astuces (en anglais) : http://www.rstudio.com/ide/docs/using/source

3.2 Ajouter des commentaires

Un commentaire est une ligne ou une portion de ligne qui sera ignorée par R. Ceci signifie qu'on peut y écrire ce qu'on veut, et qu'on va les utiliser pour ajouter tout un tas de commentaires à notre code permettant de décrire les différentes étapes du travail, les choses à se rappeler, les questions en suspens, etc.

Un commentaire sous R commence par un ou plusieurs symboles `#` (qui s'obtient avec les touches `<Alt Gr>` et `<3>` sur les claviers de type PC). Tout ce qui suit ce symbole jusqu'à la fin de la ligne est considéré comme un commentaire. On peut créer une ligne entière de commentaire, par exemple en la faisant débuter par `##` :

```
## Tableau croisé de la CSP par le nombre de livres lus
## Attention au nombre de non réponses !
```

On peut aussi créer des commentaires pour une ligne en cours :

```
x <- 2  # On met 2 dans x, parce qu'il le vaut bien
```

Dans tous les cas, il est très important de documenter ses fichiers R au fur et à mesure, faute de quoi on risque de ne plus y comprendre grand chose si on les reprend ne serait-ce que quelques semaines plus tard.

Avec RStudio, vous pouvez également utiliser les commentaires pour créer des sections au sein de votre script et naviguer plus rapidement.
Voir (en anglais) : http://www.rstudio.com/ide/docs/using/code_folding

3.3 Tableaux de données

questionr

Dans cette partie nous allons utiliser un jeu de données inclus dans l'extension questionr. Cette extension et son installation sont décrites dans la partie B.3, page 124.

Le jeu de données en question est un extrait de l'enquête *Histoire de vie* réalisée par l'INSEE en 2003. Il contient 2000 individus et 20 variables. Le descriptif des variables est indiqué dans l'annexe B.3.4, page 127.

Pour pouvoir utiliser ces données, il faut d'abord charger l'extension questionr (après l'avoir installée, bien entendu) :

```
R> library(questionr)
```

Puis indiquer à R que nous souhaitons accéder au jeu de données à l'aide de la commande `data` :

```
R> data(hdv2003)
```

Bien. Et maintenant, elles sont où mes données ? Et bien elles se trouvent dans un objet nommé `hdv2003` désormais accessible directement. Essayons de taper son nom à l'invite de commande :

```
R> hdv2003
```

Le résultat (non reproduit ici) ne ressemble pas forcément à grand-chose... Il faut se rappeler que par défaut, lorsqu'on lui fournit seulement un nom d'objet, R essaye de l'afficher de la manière la meilleure (ou la moins pire) possible. La réponse à la commande `hdv2003` n'est donc rien moins que l'affichage des données brutes contenues dans cet objet.

Ce qui signifie donc que l'intégralité de notre jeu de données est inclus dans l'objet nommé `hdv2003` ! En effet, dans R, un objet peut très bien contenir un simple nombre, un vecteur ou bien le résultat d'une enquête tout entier. Dans ce cas, les objets sont appelés des *data frames*, ou tableaux de données. Ils peuvent être manipulés comme tout autre objet. Par exemple :

```
R> d <- hdv2003
```

va entraîner la copie de l'ensemble de nos données dans un nouvel objet nommé `d`, ce qui peut paraître parfaitement inutile mais a en fait l'avantage de fournir un objet avec un nom beaucoup plus court, ce qui diminuera la quantité de texte à saisir par la suite.

Résumons Comme nous avons désormais décidé de saisir nos commandes dans un script et non plus directement dans la console, les premières lignes de notre fichier de travail sur les données de l'enquête *Histoire de vie* pourraient donc ressembler à ceci :

```
## Chargement des extensions nécessaires
library(questionr)

## Jeu de données hdv2003
data(hdv2003)
d <- hdv2003
```

3.4 Inspecter les données

3.4.1 Structure du tableau

Avant de travailler sur les données, nous allons essayer de voir à quoi elles ressemblent. Dans notre cas il s'agit de se familiariser avec la stucture du fichier. Lors de l'import de données depuis un autre logiciel, il s'agira souvent de vérifier que l'importation s'est bien déroulée.

Les fonctions `nrow`, `ncol` et `dim` donnent respectivement le nombre de lignes, le nombre de colonnes et les dimensions de notre tableau. Nous pouvons donc d'ores et déjà vérifier que nous avons bien 2000 lignes et 20 colonnes :

```
R> nrow(d)

[1] 2000

R> ncol(d)

[1] 20

R> dim(d)

[1] 2000   20
```

La fonction `names` donne les noms des colonnes de notre tableau, c'est-à-dire les noms des variables :

```
R> names(d)

 [1] "id"           "age"          "sexe"         "nivetud"
 [5] "poids"        "occup"        "qualif"       "freres.soeurs"
 [9] "clso"         "relig"        "trav.imp"     "trav.satisf"
[13] "hard.rock"    "lecture.bd"   "peche.chasse" "cuisine"
[17] "bricol"       "cinema"       "sport"        "heures.tv"
```

La fonction `str` est plus complète. Elle liste les différentes variables, indique leur type et donne le cas échéant des informations supplémentaires ainsi qu'un échantillon des premières valeurs prises par cette variable :

```
R> str(d)

'data.frame': 2000 obs. of  20 variables:
 $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age          : int  28 23 59 34 71 35 60 47 20 28 ...
 $ sexe         : Factor w/ 2 levels "Homme","Femme": 2 2 1 1 2 2 2 1 2 1 ...
 $ nivetud      : Factor w/ 8 levels "N'a jamais fait d'etudes",..: 8 NA 3 8 3 6 3 6 NA 7 ...
 $ poids        : num  2634 9738 3994 5732 4329 ...
 $ occup        : Factor w/ 7 levels "Exerce une profession",..: 1 3 1 1 4 1 6 1 3 1 ...
 $ qualif       : Factor w/ 7 levels "Ouvrier specialise",..: 6 NA 3 3 6 6 2 2 NA 7 ...
 $ freres.soeurs: int  8 2 2 1 0 5 1 5 4 2 ...
 $ clso         : Factor w/ 3 levels "Oui","Non","Ne sait pas": 1 1 2 2 1 2 1 2 1 2 ...
 $ relig        : Factor w/ 6 levels "Pratiquant regulier",..: 4 4 4 3 1 4 3 4 3 2 ...
 $ trav.imp     : Factor w/ 4 levels "Le plus important",..: 4 NA 2 3 NA 1 NA 4 NA 3 ...
```

```
$ trav.satisf  : Factor w/ 3 levels "Satisfaction",..: 2 NA 3 1 NA 3 NA 2 NA 1 ...
$ hard.rock    : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 ...
$ lecture.bd   : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 ...
$ peche.chasse : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 2 2 1 ...
$ cuisine      : Factor w/ 2 levels "Non","Oui": 2 1 1 2 1 1 2 2 1 ...
$ bricol       : Factor w/ 2 levels "Non","Oui": 1 1 1 2 1 1 1 2 1 ...
$ cinema       : Factor w/ 2 levels "Non","Oui": 1 2 1 2 1 2 1 1 2 ...
$ sport        : Factor w/ 2 levels "Non","Oui": 1 2 2 2 1 2 1 1 1 ...
$ heures.tv    : num  0 1 0 2 3 2 2.9 1 2 2 ...
```

La première ligne nous informe qu'il s'agit bien d'un tableau de données avec 2000 observations et 20 variables. Vient ensuite la liste des variables. La première se nomme `id` et est de type *nombre entier* (`int`). La seconde se nomme `age` et est de type *numérique*. La troisième se nomme `sexe`, il s'agit d'un *facteur* (*factor*).

Un *facteur* et une variable pouvant prendre un nombre limité de modalités (*levels*). Ici notre variable a deux modalités possibles : `Homme` et `Femme`. Ce type de variable est décrit plus en détail section 5.1.3 page 50.

La fonction `str` peut s'appliquer à n'importe quel type d'objet. C'est un excellent moyen de connaître la structure d'un objet.

3.4.2 Inspection visuelle

La particularité de R par rapport à d'autres logiciels comme Modalisa ou SPSS est de ne pas proposer, par défaut, de vue des données sous forme de tableau. Ceci peut parfois être un peu déstabilisant dans les premiers temps d'utilisation, même si on perd vite l'habitude et qu'on finit par se rendre compte que « voir » les données n'est pas forcément un gage de productivité ou de rigueur dans le traitement.

Néanmoins, R propose une visualisation assez rudimentaire des données sous la forme d'une fenêtre de type tableur, *via* la fonction `edit` :

```
R> edit(d)
```

La fenêtre qui s'affiche permet de naviguer dans le tableau, et même d'éditer le contenu des cases et donc de modifier les données. Lorsque vous fermez la fenêtre, le contenu du tableau s'affiche dans la console : il s'agit en fait du tableau comportant les éventuelles modifications effectuées, `d` restant inchangé. Si vous souhaitez appliquer ces modifications, vous pouvez le faire en créant un nouveau tableau :

```
R> d.modif <- edit(d)
```

ou en remplaçant directement le contenu de `d`[3] :

```
R> d <- edit(d)
```

3. Dans ce cas on peut utiliser la fonction `fix` sous la forme `fix(d)`, qui est équivalente à `d <- edit(d)`.

La fonction `edit` peut être utile pour un avoir un aperçu visuel des données, par contre il est **très fortement** déconseillé de l'utiliser pour modifier les données. Si on souhaite effectuer des modifications, on remonte en général aux données originales (retouches ponctuelles dans un tableur par exemple) ou on les effectue à l'aide de commandes (qui seront du coup reproductibles).

Sous RStudio, la liste des objets en mémoire est affichée dans le quadrant haut-droite sous l'onglet *Workspace*. Un clic sur un tableau de données permet d'afficher son contenu sous un onglet dédié dans le quadrant haut-gauche. Cette manière de procéder est plus simple que le recours à la fonction `edit`.

3.4.3 Accéder aux variables

`d` représente donc l'ensemble de notre tableau de données. Nous avons vu que si l'on saisit simplement `d` à l'invite de commandes, on obtient un affichage du tableau en question. Mais comment accéder aux variables, c'est à dire aux colonnes de notre tableau ?

La réponse est simple : on utilise le nom de l'objet, suivi de l'opérateur `$`, suivi du nom de la variable, comme ceci :

```
R> d$sexe

 [1] Femme Femme Homme Homme Femme Femme Femme Homme Femme Homme Femme
[12] Homme Femme Femme Femme Femme Homme Femme Homme Femme Femme Homme
[23] Femme Femme Femme Homme Femme Homme Homme Homme Homme Homme Homme
[34] Homme Femme Femme Homme Femme Femme Homme Femme Homme Homme Femme
[45] Femme Homme Femme Femme Femme Femme Homme Femme Homme Femme Homme
[56] Femme Femme Femme Homme Femme Femme Homme Homme Homme Homme Femme
[67] Homme Homme Femme Femme
 [ reached getOption("max.print") -- omitted 1930 entries ]
Levels: Homme Femme
```

On constate alors que R a bien accédé au contenu de notre variable `sexe` du tableau `d` et a affiché son contenu, c'est-à-dire l'ensemble des valeurs prises par la variable.

Les fonctions `head` et `tail` permettent d'afficher seulement les premières (respectivement les dernières) valeurs prises par la variable. On peut leur passer en argument le nombre d'éléments à afficher :

```
R> head(d$sport)

[1] Non Oui Oui Oui Non Oui
Levels: Non Oui

R> tail(d$age, 10)

 [1] 52 42 50 41 46 45 46 24 24 66
```

À noter que ces fonctions marchent aussi pour afficher les lignes du tableau `d` :

```
R> head(d, 2)

  id age  sexe                                                nivetud
1  1  28 Femme Enseignement superieur y compris technique superieur
2  2  23 Femme                                                  <NA>
     poids                occup  qualif freres.soeurs clso
1 2634.398 Exerce une profession Employe             8  Oui
2 9738.396       Etudiant, eleve    <NA>             2  Oui
                       relig      trav.imp    trav.satisf hard.rock
1 Ni croyance ni appartenance Peu important Insatisfaction       Non
2 Ni croyance ni appartenance          <NA>           <NA>       Non
  lecture.bd peche.chasse cuisine bricol cinema sport heures.tv
1        Non          Non     Oui    Non    Non   Non         0
2        Non          Non     Non    Non    Oui   Oui         1
```

3.5 Analyser une variable

3.5.1 Variable quantitative

Principaux indicateurs

Comme la fonction `str` nous l'a indiqué, notre tableau `d` contient plusieurs valeurs numériques, dont la variable `heures.tv` qui représente le nombre moyen passé par les enquêtés à regarder la télévision quotidiennement. On peut essayer de déterminer quelques caractéristiques de cette variable, en utilisant des fonctions déjà vues précédemment :

```
R> mean(d$heures.tv)

[1] NA

R> mean(d$heures.tv, na.rm = TRUE)

[1] 2.246566

R> sd(d$heures.tv, na.rm = TRUE)

[1] 1.775853

R> min(d$heures.tv, na.rm = TRUE)

[1] 0

R> max(d$heures.tv, na.rm = TRUE)

[1] 12

R> range(d$heures.tv, na.rm = TRUE)

[1]  0 12
```

On peut lui ajouter la fonction `median`, qui donne la valeur médiane, et le très utile `summary` qui donne toutes ces informations ou presque en une seule fois, avec en plus les valeurs des premier et troisième quartiles et le nombre de valeurs manquantes (`NA`) :

```
R> median(d$heures.tv, na.rm = TRUE)

[1] 2

R> summary(d$heures.tv)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.000   2.000   2.247   3.000  12.000       5
```

La fonction `summary` peut-être utilisée sur tout type d'objet, y compris un tableau de données. Essayez donc `summary(d)`.

Histogramme

Tout cela est bien pratique, mais pour pouvoir observer la distribution des valeurs d'une variable quantitative, il n'y a quand même rien de mieux qu'un bon graphique.

On peut commencer par un histogramme de la répartition des valeurs. Celui-ci peut être généré très facilement avec la fonction `hist`, comme indiqué figure 3.1 page ci-contre.

Ici, les options `main`, `xlab` et `ylab` permettent de personnaliser le titre du graphique, ainsi que les étiquettes des axes. De nombreuses autres options existent pour personnaliser l'histogramme, parmi celles-ci on notera :

probability si elle vaut `TRUE`, l'histogramme indique la proportion des classes de valeurs au lieu des effectifs.

breaks permet de contrôler les classes de valeurs. On peut lui passer un chiffre, qui indiquera alors le nombre de classes, un vecteur, qui indique alors les limites des différentes classes, ou encore une chaîne de caractère ou une fonction indiquant comment les classes doivent être calculées.

col la couleur de l'histogramme [4].

Deux exemples sont donnés figure 3.2 page 28 et figure 3.3 page 29.

Voir la page d'aide de la fonction `hist` pour plus de détails sur les différentes options.

Boîtes à moustaches

Les boîtes à moustaches, ou `boxplot` en anglais, sont une autre représentation graphique de la répartition des valeurs d'une variable quantitative. Elles sont particulièrement utiles pour comparer les distributions de plusieurs variables ou d'une même variable entre différents groupes, mais peuvent aussi être utilisées pour représenter la dispersion d'une unique variable. La fonction qui produit ces graphiques est la fonction `boxplot`. On trouvera un exemple figure 3.4 page 30.

```
R> hist(d$heures.tv, main = "Nombre d'heures passées devant la télé par jour",
+     xlab = "Heures", ylab = "Effectif")
```

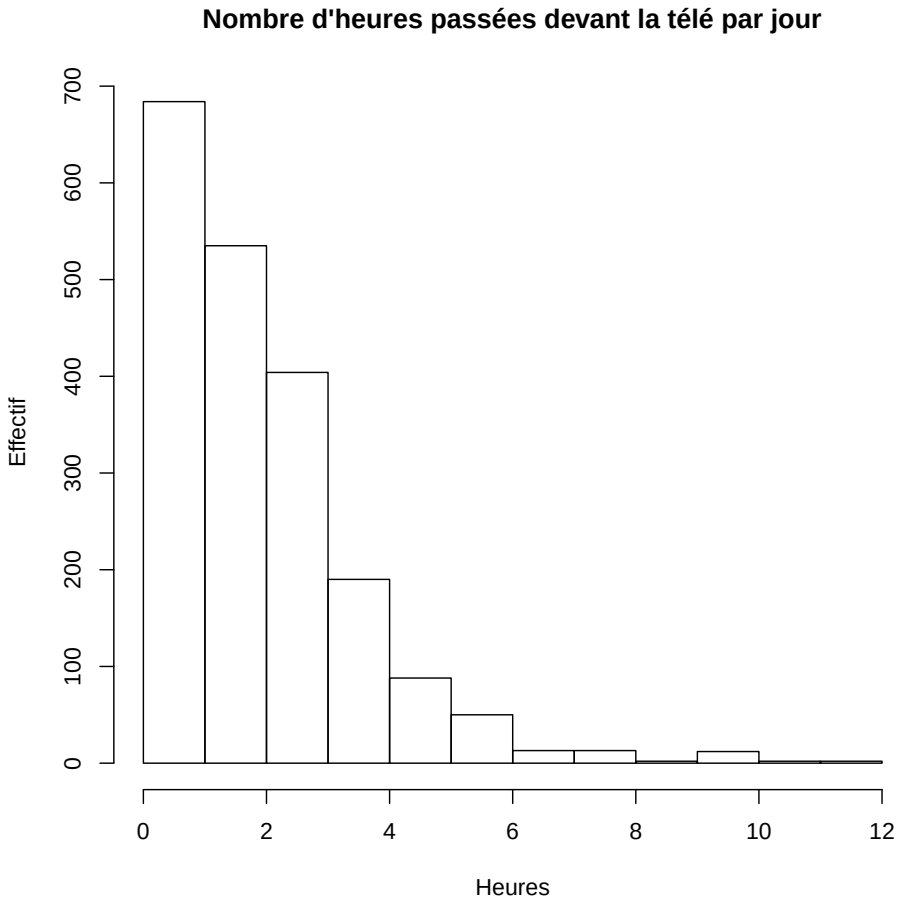


FIGURE 3.1 – Exemple d'histogramme

```
R> hist(d$heures.tv, main = "Heures de télé en 7 classes", breaks = 7, xlab = "Heures",
+     ylab = "Proportion", probability = TRUE, col = "orange")
```

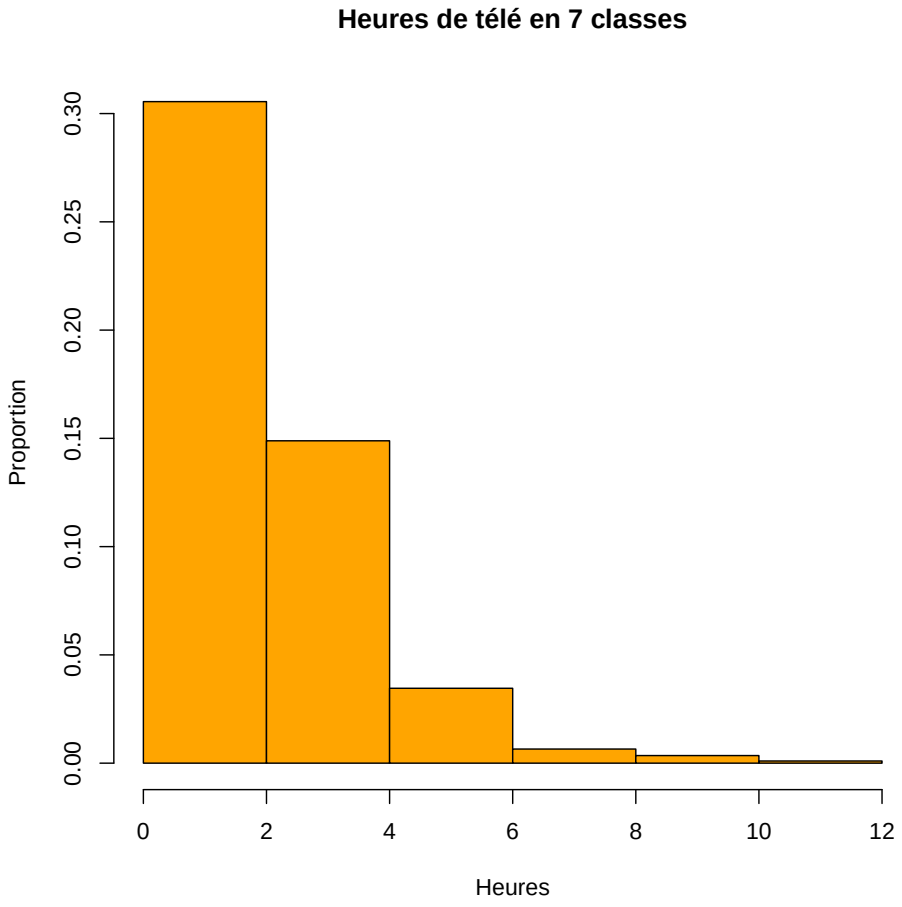


FIGURE 3.2 – Un autre exemple d'histogramme

```
R> hist(d$heures.tv, main = "Heures de télé avec classes spécifiées", breaks = c(0,
+     1, 4, 9, 12), xlab = "Heures", ylab = "Proportion", col = "red")
```

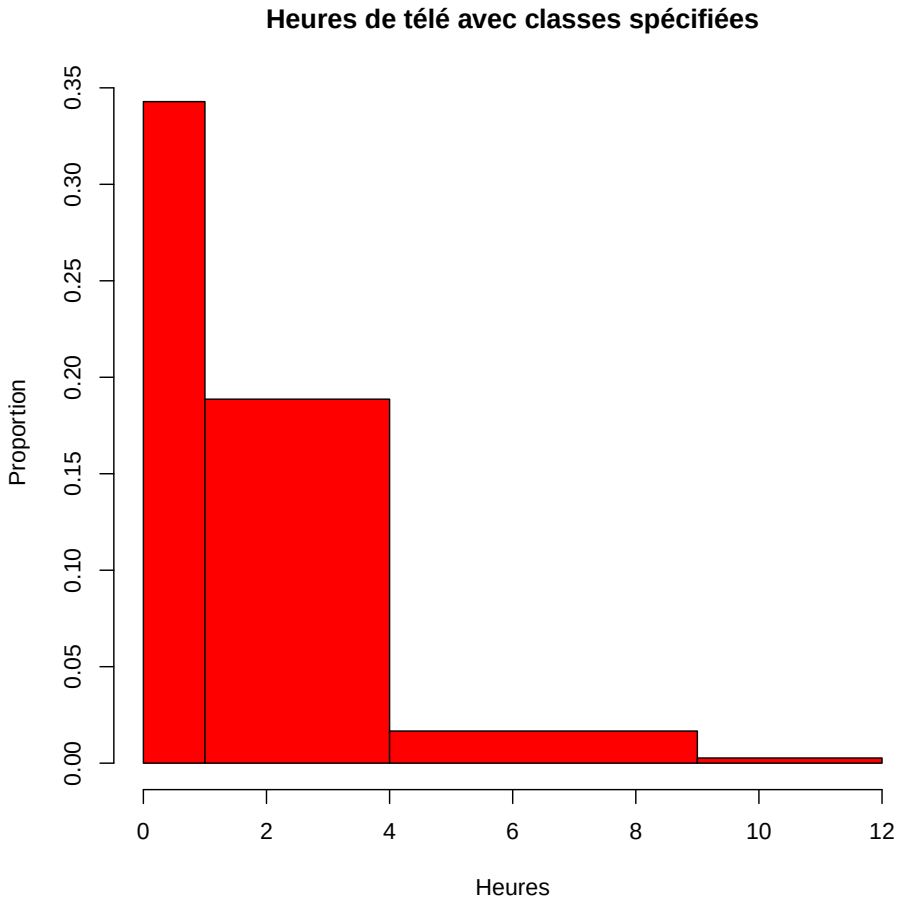


FIGURE 3.3 – Encore un autre exemple d'histogramme

```
R> boxplot(d$heures.tv, main = "Nombre d'heures passées devant la télé par jour",
+     ylab = "Heures")
```

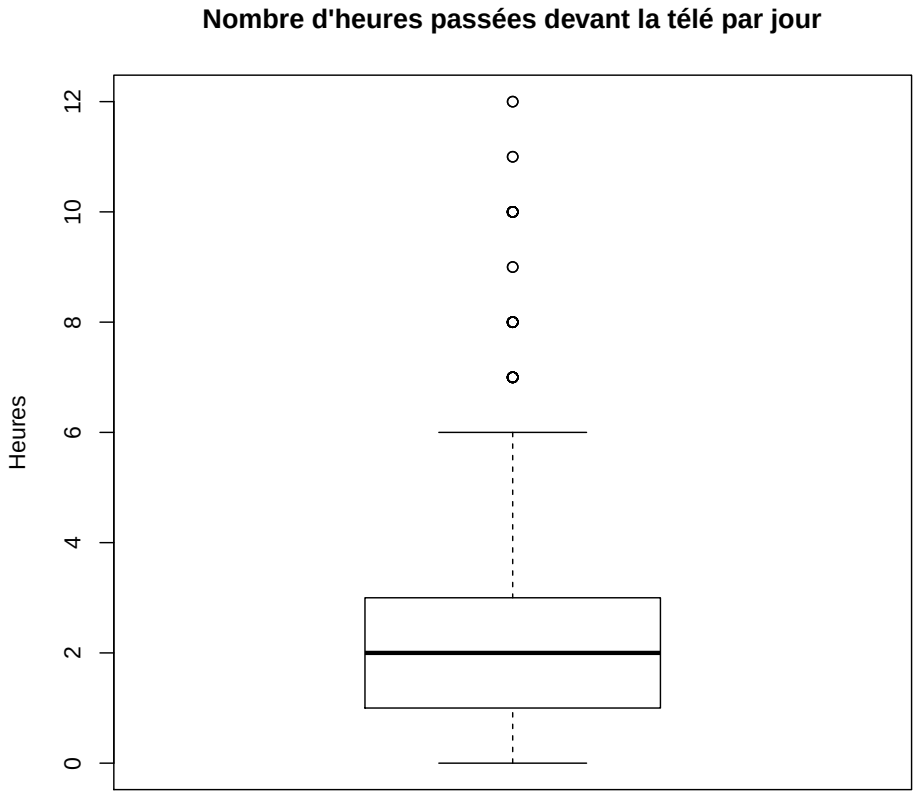


FIGURE 3.4 – Exemple de boîte à moustaches

```
R> boxplot(d$heures.tv, col = grey(0.8), main = "Nombre d'heures passées devant la télé par jour",
+     ylab = "Heures")
R> abline(h = median(d$heures.tv, na.rm = TRUE), col = "navy", lty = 2)
R> text(1.35, median(d$heures.tv, na.rm = TRUE) + 0.15, "Médiane", col = "navy")
R> Q1 <- quantile(d$heures.tv, probs = 0.25, na.rm = TRUE)
R> abline(h = Q1, col = "darkred")
R> text(1.35, Q1 + 0.15, "Q1 : premier quartile", col = "darkred", lty = 2)
R> Q3 <- quantile(d$heures.tv, probs = 0.75, na.rm = TRUE)
R> abline(h = Q3, col = "darkred")
R> text(1.35, Q3 + 0.15, "Q3 : troisième quartile", col = "darkred", lty = 2)
R> arrows(x0 = 0.7, y0 = quantile(d$heures.tv, probs = 0.75, na.rm = TRUE), x1 = 0.7,
+     y1 = quantile(d$heures.tv, probs = 0.25, na.rm = TRUE), length = 0.1, code = 3)
R> text(0.7, Q1 + (Q3 - Q1)/2 + 0.15, "h", pos = 2)
R> mtext("L'écart inter-quartile h contient 50 % des individus", side = 1)
R> abline(h = Q1 - 1.5 * (Q3 - Q1), col = "darkgreen")
R> text(1.35, Q1 - 1.5 * (Q3 - Q1) + 0.15, "Q1 -1.5 h", col = "darkgreen", lty = 2)
R> abline(h = Q3 + 1.5 * (Q3 - Q1), col = "darkgreen")
R> text(1.35, Q3 + 1.5 * (Q3 - Q1) + 0.15, "Q3 +1.5 h", col = "darkgreen", lty = 2)
```

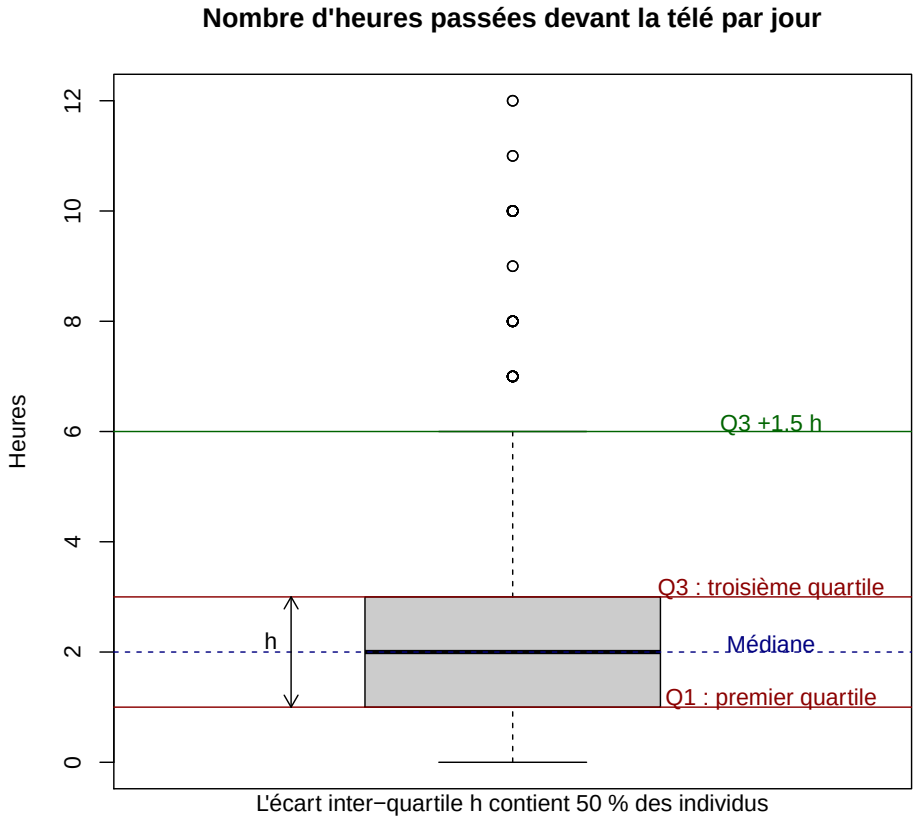


FIGURE 3.5 – Interprétation d'une boîte à moustaches

```
R> boxplot(d$heures.tv, main = "Nombre d'heures passées devant la télé par\njour",
+     ylab = "Heures")
R> rug(d$heures.tv, side = 2)
```

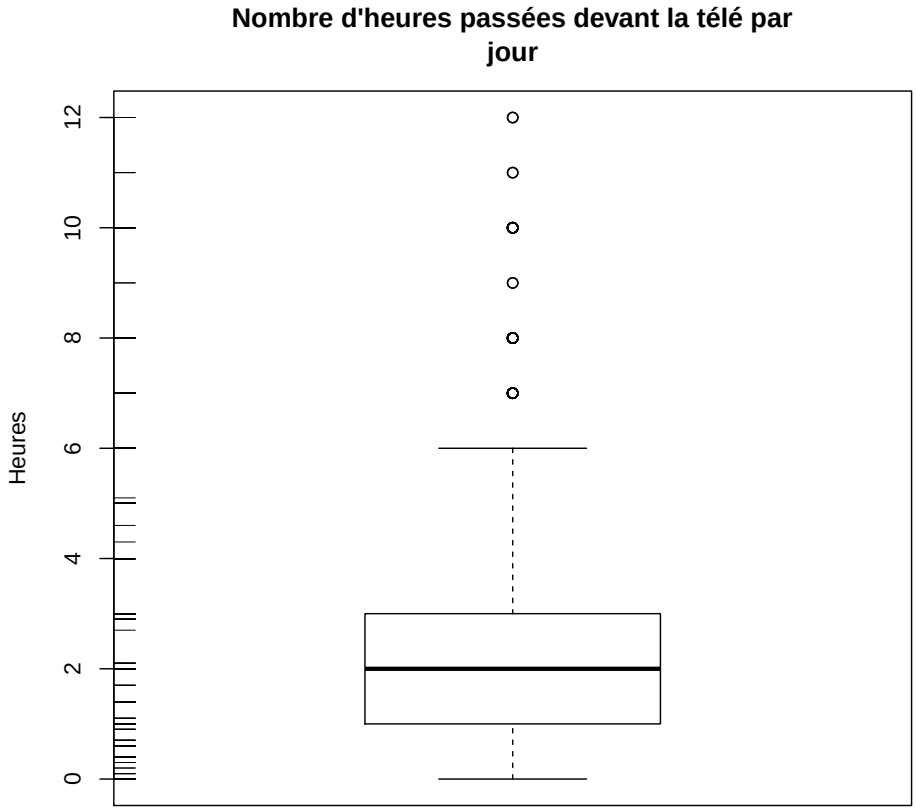


FIGURE 3.6 – Boîte à moustaches avec représentation des valeurs

Comment interpréter ce graphique ? On le comprendra mieux à partir de la figure 3.5 page précédente [5].

Le carré au centre du graphique est délimité par les premiers et troisième quartiles, avec la médiane représentée par une ligne plus sombre au milieu. Les « fourchettes » s'étendant de part et d'autres vont soit jusqu'à la valeur minimale ou maximale, soit jusqu'à une valeur approximativement égale au quartile le plus proche plus 1,5 fois l'écart inter-quartile. Les points se situant en-dehors de cette fourchette sont représentés par des petits ronds et sont généralement considérés comme des valeurs extrêmes, potentiellement aberrantes.

On peut ajouter la représentation des valeurs sur le graphique pour en faciliter la lecture avec des petits traits dessinés sur l'axe vertical (fonction `rug`), comme sur la figure 3.6 de la présente page.

4. Il existe un grand nombre de couleurs prédéfinies dans R. On peut récupérer leur liste en utilisant la fonction `colors` en tapant simplement `colors()` dans la console, ou en consultant le document suivant : http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

5. Le code ayant servi à générer cette figure est une copie quasi conforme de celui présenté dans l'excellent document de Jean Lobry sur les graphiques de base avec R, téléchargeable sur le site du Pôle bioinformatique lyonnais : http://pbil.univ-lyon1.fr/R/pdf/lang04.pdf.

3.5.2 Variable qualitative

Tris à plat

La fonction la plus utilisée pour le traitement et l'analyse des variables qualitatives (variable prenant ses valeurs dans un ensemble de modalités) est sans aucun doute la fonction `table`, qui donne les effectifs de chaque modalité de la variable.

```
R> table(d$sexe)


Homme Femme
  899  1101
```

La tableau précédent nous indique que parmi nos enquêtés on trouve 899 hommes et 1101 femmes.

Quand le nombre de modalités est élevé, on peut ordonner le tri à plat selon les effectifs à l'aide de la fonction `sort`.

```
R> table(d$occup)


Exerce une profession               Chomeur       Etudiant, eleve
                 1049                   134                    94
             Retraite   Retire des affaires              Au foyer
                  392                    77                   171
        Autre inactif
                   83

R> sort(table(d$occup))


  Retire des affaires         Autre inactif       Etudiant, eleve
                   77                    83                    94
              Chomeur              Au foyer              Retraite
                  134                   171                   392
Exerce une profession
                 1049

R> sort(table(d$occup), decreasing = TRUE)


Exerce une profession              Retraite              Au foyer
                 1049                   392                   171
              Chomeur       Etudiant, eleve         Autre inactif
                  134                    94                    83
  Retire des affaires
                   77
```

À noter que la fonction `table` exclut par défaut les non-réponses du tableau résultat. L'argument `useNA` de cette fonction permet de modifier ce comportement :

— avec `useNA="no"` (valeur par défaut), les valeurs manquantes ne sont jamais incluses dans le tri à plat ;

— avec `useNA="ifany"`, une colonne `NA` est ajoutée si des valeurs manquantes sont présentes dans les données ;
— avec `useNA="always"`, une colonne `NA` est toujours ajoutée, même s'il n'y a pas de valeurs manquantes dans les données.

On peut donc utiliser :

```
R> table(d$trav.satisf, useNA = "ifany")


  Satisfaction Insatisfaction      Equilibre           <NA>
           480            117            451            952
```

L'utilisation de `summary` permet également l'affichage du tri à plat et du nombre de non-réponses :

```
R> summary(d$trav.satisf)

  Satisfaction Insatisfaction      Equilibre           NA's
           480            117            451            952
```

questionr Pour obtenir un tableau avec la répartition en pourcentages, on peut utiliser la fonction `freq` de l'extension questionr [6].

```
R> freq(d$qualif)

                           n    % val%
Ouvrier specialise       203 10.2 12.3
Ouvrier qualifie         292 14.6 17.7
Technicien                86  4.3  5.2
Profession intermediaire 160  8.0  9.7
Cadre                    260 13.0 15.7
Employe                  594 29.7 35.9
Autre                     58  2.9  3.5
NA                       347 17.3   NA
```

La colonne `n` donne les effectifs bruts, et la colonne `%` la répartition en pourcentages. La fonction accepte plusieurs paramètres permettant d'afficher les totaux, les pourcentages cumulés, de trier selon les effectifs ou de contrôler l'affichage. Par exemple :

```
R> freq(d$qualif, cum = TRUE, total = TRUE, sort = "inc", digits = 2, exclude = NA)

                            n      %   %cum
Autre                      58   3.51   3.51
Technicien                 86   5.20   8.71
Profession intermediaire  160   9.68  18.39
Ouvrier specialise        203  12.28  30.67
Cadre                     260  15.73  46.40
Ouvrier qualifie          292  17.66  64.07
Employe                   594  35.93 100.00
Total                    1653 100.00 100.00
```

6. En l'absence de l'extension `questionr`, on pourra se rabattre sur la fonction `prop.table` avec la commande suivante : `prop.table(table(d$qualif))`.

La colonne `%cum` indique ici le pourcentage cumulé, ce qui est ici une très mauvaise idée puisque pour ce type de variable cela n'a aucun sens. Les lignes du tableau résultat ont été triés par effectifs croissants, les totaux ont été ajoutés, les non-réponses exclues, et les pourcentages arrondis à deux décimales.

Pour plus d'informations sur la commande `freq`, consultez sa page d'aide en ligne avec `?freq` ou `help("freq")`.

Représentation graphique

Pour représenter la répartition des effectifs parmi les modalités d'une variable qualitative, on a souvent tendance à utiliser des diagrammes en secteurs (camemberts). Ceci est possible sous R avec la fonction `pie`, mais la page d'aide de ladite fonction nous le déconseille assez vivement : les diagrammes en secteur sont en effet une mauvaise manière de présenter ce type d'information, car l'œil humain préfère comparer des longueurs plutôt que des surfaces [7].

On privilégiera donc d'autres formes de représentations, à savoir les diagrammes en bâtons et les diagrammes de Cleveland.

Les diagrammes en bâtons sont utilisés automatiquement par R lorsqu'on applique la fonction générique `plot` à un tri à plat obtenu avec `table`. On privilégiera cependant ce type de représentations pour les variables de type numérique comportant un nombre fini de valeurs. Le nombre de frères, sœurs, demi-frères et demi-sœurs est un bon exemple, indiqué figure 3.7 page suivante.

Pour les autres types de variables qualitatives, on privilégiera les diagrammes de Cleveland, obtenus avec la fonction `dotchart`. On doit appliquer cette fonction au tri à plat de la variable, obtenu avec la fonction `table`. Le résultat se trouve figure 3.8 page 37.

Quand la variable comprend un grand nombre de modalités, il est préférable d'ordonner le tri à plat obtenu à l'aide de la fonction `sort` (voir figure 3.9 page 38).

L'agument `pch`, qui est utilisé par la plupart des graphiques de type points, permet de spécifier le symbole à utiliser. Il peut prendre soit un nombre entier compris entre 0 et 25, soit un charactère textuel (voir figure 3.10 page 39).

3.6 Exercices

Exercice 3.5

▷ *Solution page 130*

Créer un script qui effectue les actions suvantes et exécutez-le :

— charger l'extension questionr
— charger le jeu de données `hdv2003`
— placer le jeu de données dans un objet nommé `df`
— afficher la liste des variables de `df` et leur type

Exercice 3.6

▷ *Solution page 131*

Des erreurs se sont produites lors de la saisie des données de l'enquête. En fait le premier individu du jeu de données n'a pas 42 ans mais seulement 24, et le second individu n'est pas un homme mais une femme. Corrigez les erreurs et stockez les données corrigées dans un objet nommé `df.ok`.

7. On trouvera des exemples illustrant cette idée dans le document de Jean Lobry cité précédemment.

```
R> plot(table(d$freres.soeurs), main = "Nombre de frères, soeurs, demi-frères et demi-soeurs",
+     ylab = "Effectif")
```

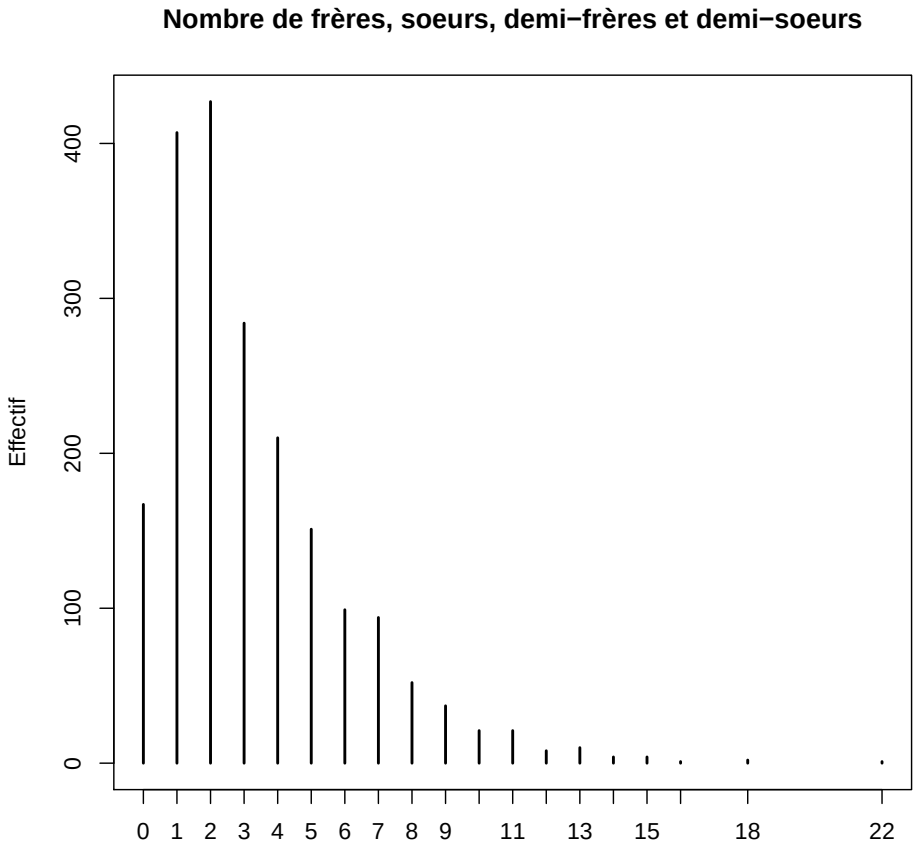


FIGURE 3.7 – Exemple de diagramme en bâtons

```
R> dotchart(table(d$clso), main = "Sentiment d'appartenance à une classe sociale",
+     pch = 19)
```

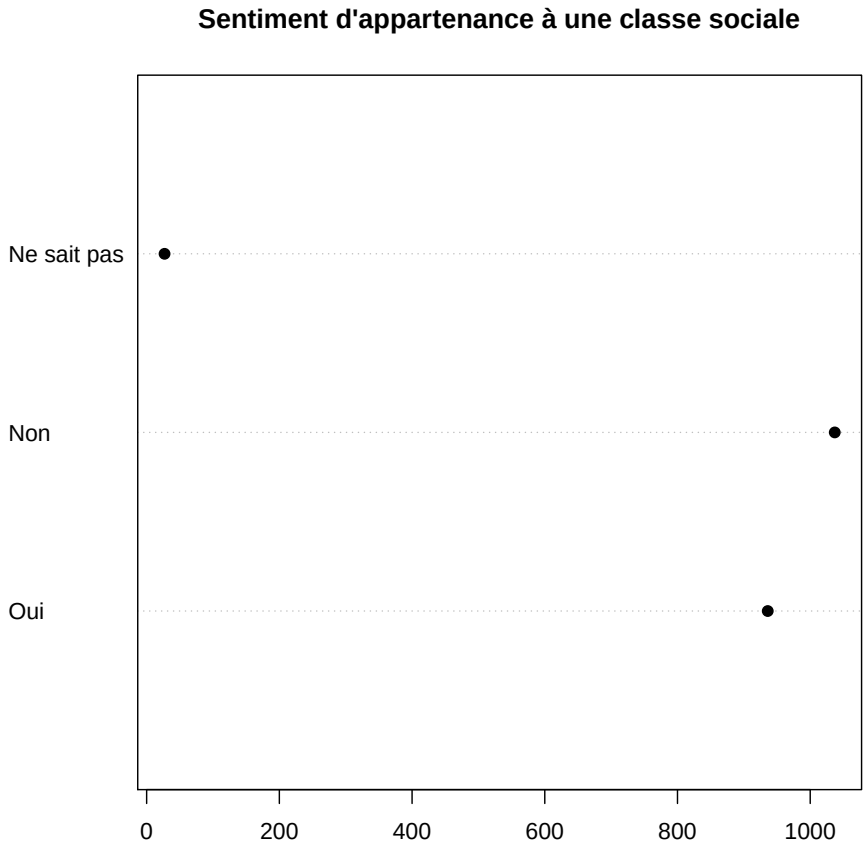


FIGURE 3.8 – Exemple de diagramme de Cleveland

```
R> dotchart(sort(table(d$qualif)), main = "Niveau de qualification")
```

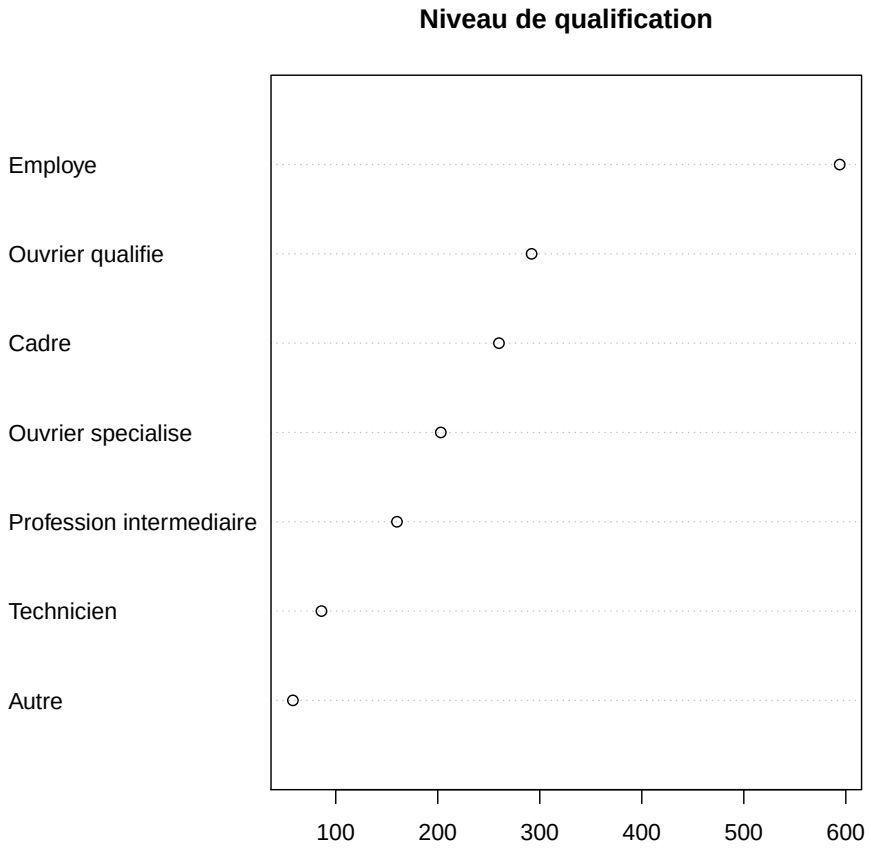


FIGURE 3.9 – Exemple de diagramme de Cleveland ordonné

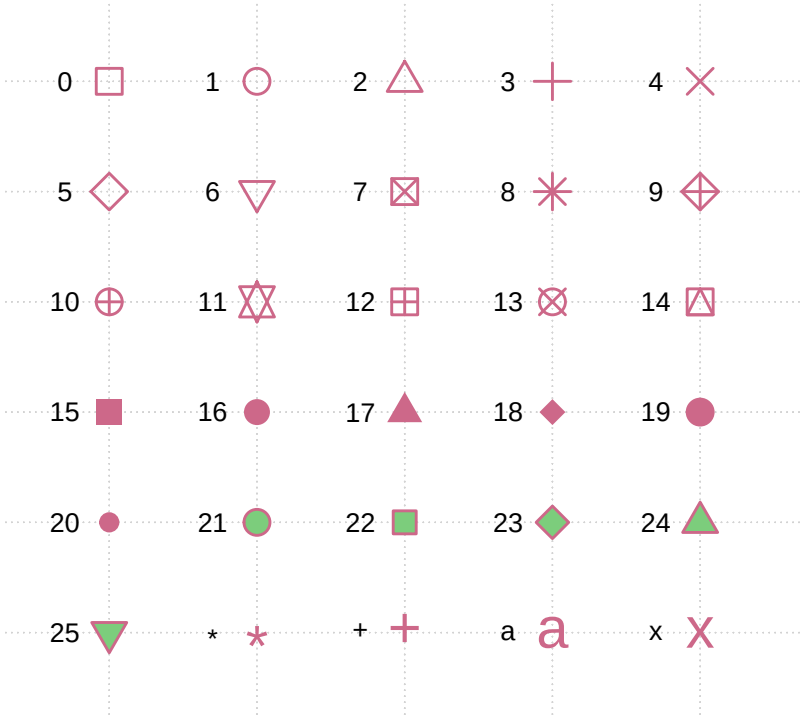


FIGURE 3.10 – Différentes valeurs possibles pour l'argument `pch`

Affichez ensuite les 4 premières lignes de `df.ok` pour vérifier que les modifications ont bien été prises en compte.

Exercice 3.7

▷ *Solution page 131*

Nous souhaitons étudier la répartition des âges des enquêtés (variable `age`). Pour cela, affichez les principaux indicateurs de cette variable. Représentez ensuite sa distribution par un histogramme en 10 classes, puis sous forme de boîte à moustache, et enfin sous la forme d'un diagramme en bâtons représentant les effectifs de chaque âge.

Exercice 3.8

▷ *Solution page 131*

On s'intéresse maintenant à l'importance accordée par les enquêtés à leur travail (variable `trav.imp`). Faites un tri à plat des effectifs des modalités de cette variable avec la commande `table`. Y'a-t-il des valeurs manquantes ?

Faites un tri à plat affichant à la fois les effectifs et les pourcentages de chaque modalité.

Représentez graphiquement les effectifs des modalités à l'aide d'un diagramme de Cleveland.

Partie 4

Import/export de données

L'import et l'export de données depuis ou vers d'autres applications est couvert en détail dans l'un des manuels officiels (en anglais) nommé *R Data Import/Export* et accessible, comme les autres manuels, à l'adresse suivante :

http://cran.r-project.org/manuals.html

Cette partie est très largement tirée de ce document, et on pourra s'y reporter pour plus de détails.

Importer des données est souvent l'une des première opérations que l'on effectue lorsque l'on débute sous R, et ce n'est pas la moins compliquée. En cas de problème il ne faut donc pas hésiter à demander de l'aide par les différents moyens disponibles (voir partie 9 page 115) avant de se décourager.

Un des points délicats pour l'importation de données dans R concerne le nom des variables. Pour être utilisables dans R ceux-ci doivent être à la fois courts et explicites, ce qui n'est pas le cas dans d'autres applications comme Modalisa par exemple. La plupart des fonctions d'importation s'occupent de convertir les noms de manières à ce qu'ils soient compatibles avec les règles de R (remplacement des espaces par des points par exemple), mais un renommage est souvent à prévoir, soit au sein de l'application d'origine, soit une fois les données importées dans R.

4.1 Accès aux fichiers et répertoire de travail

Dans ce qui suit, puisqu'il s'agit d'importer des données externes, nous allons avoir besoin d'accéder à des fichiers situés sur le disque dur de notre ordinateur.

Par exemple, la fonction `read.table`, très utilisée pour l'import de fichiers texte, prend comme premier argument le nom du fichier à importer, ici `fichier.txt` :

```
R> donnees <- read.table("fichier.txt")
```

Cependant, ceci ne fonctionnera que si le fichier se trouve dans le *répertoire de travail* de R. De quoi s'agit-il ? Tout simplement du répertoire dans lequel R est actuellement en train de s'exécuter. Pour savoir quel est le répertoire de travail actuel, on peut utiliser la fonction `getwd`[1] :

```
R> getwd()

[1] "/home/julien/r/doc/intro"
```

Si on veut modifier le répertoire de travail, on utilise `setwd` en lui indiquant le chemin complet. Par exemple sous Linux :

```
R> setwd("/home/julien/projets/R")
```

Sous Windows le chemin du répertoire est souvent un peu plus compliqué. Si vous utilisez l'interface graphique par défaut, vous pouvez utiliser la fonction *Changer le répertoire courant* du menu *Fichier*. Celle-ci vous permet de sélectionner le répertoire de travail de la session en cours en le sélectionnant via une boîte de dialogue.

Si vous utilisez RStudio, Vous pouvez utiliser une des commandes *set working directory* du menu *session* ou, mieux, utiliser les fonctionnalités de gestion de projet qui vous permettent de mémoriser, projet par projet, le répertoire de travail, la liste des fichiers ouverts ainsi que différents paramétrages spécifiques.

Une fois le répertoire de travail fixé, on pourra accéder aux fichiers qui s'y trouvent directement, en spécifiant seulement leur nom. On peut aussi créer des sous-répertoires dans le répertoire de travail ; une potentielle bonne pratique peut être de regrouper tous les fichiers de données dans un sous-répertoire nommé `donnees`. On pourra alors accéder aux fichiers qui s'y trouvent de la manière suivante :

```
R> donnees <- read.table("donnees/fichier.txt")
```

Dans ce qui suit on supposera que les fichiers à importer se trouvent directement dans le répertoire de travail, et on n'indiquera donc que le nom du fichier, sans indication de chemin ou de répertoire supplémentaire.

Si vous utilisez l'environnement de développement RStudio, vous pouvez vous débarasser du problème des répertoires de travail en utilisant sa fonctionnalité de gestion de *projets*. Les projets sont accessibles en haut à droite de l'écran. Un projet permet de centraliser tout ses fichiers dans un même répertoire. De plus, il est très facile de basculer très rapidement d'un projet à un autre, en retrouvrant sa session de travail dans l'était où elle était.

4.2 Import de données depuis un tableur

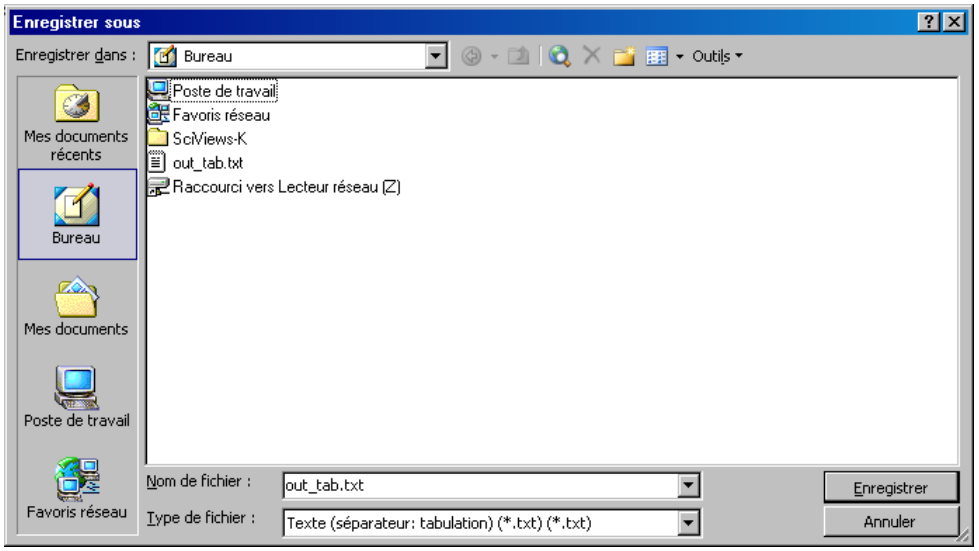
Il est assez courant de vouloir importer des données saisies ou traitées avec un tableur du type OpenOffice/LibreOffice ou Excel. En général les données prennent alors la forme d'un tableau avec les variables en colonne et les individus en ligne.

1. Le résultat indiqué ici correspond à un système Linux, sous Windows vous devriez avoir quelque chose de la forme `C:/Documents and Settings/` ...

	A	B	C	D
1	**Country or Area**	**Year**	**Educational levels**	**Value**
2	Afghanistan	2002	Primary level	3266737
3	Afghanistan	2001	Primary level	773623
4	Afghanistan	2001	Secondary level	362415
5	Afghanistan	2000	Primary level	500068
6	Afghanistan	1999	Primary level	957403
7	Afghanistan	1998	Primary level	1046338
8	Afghanistan	1995	Primary level	1312197
9	Afghanistan	1995	Secondary level	512851
10	Afghanistan	1994	Primary level	1161444
11	Afghanistan	1994	Secondary level	497762
12	Afghanistan	1993	Primary level	786532
13	Afghanistan	1993	Secondary level	332170
14	Afghanistan	1991	Primary level	627888
15	Afghanistan	1991	Secondary level	281928
16	Afghanistan	1990	Primary level	622513
17	Afghanistan	1990	Secondary level	182340

4.2.1 Depuis Excel

La démarche pour importer ces données dans R est d'abord de les enregistrer dans un format de type texte. Sous Excel, on peut ainsi sélectionner *Fichier*, *Enregistrer sous*, puis dans la zone *Type de fichier* choisir soit *Texte (séparateur tabulation)*, soit *CSV (séparateur : point-virgule)*.



Dans le premier cas, on peut importer le fichier en utilisant la fonction `read.delim2`, de la manière suivante :

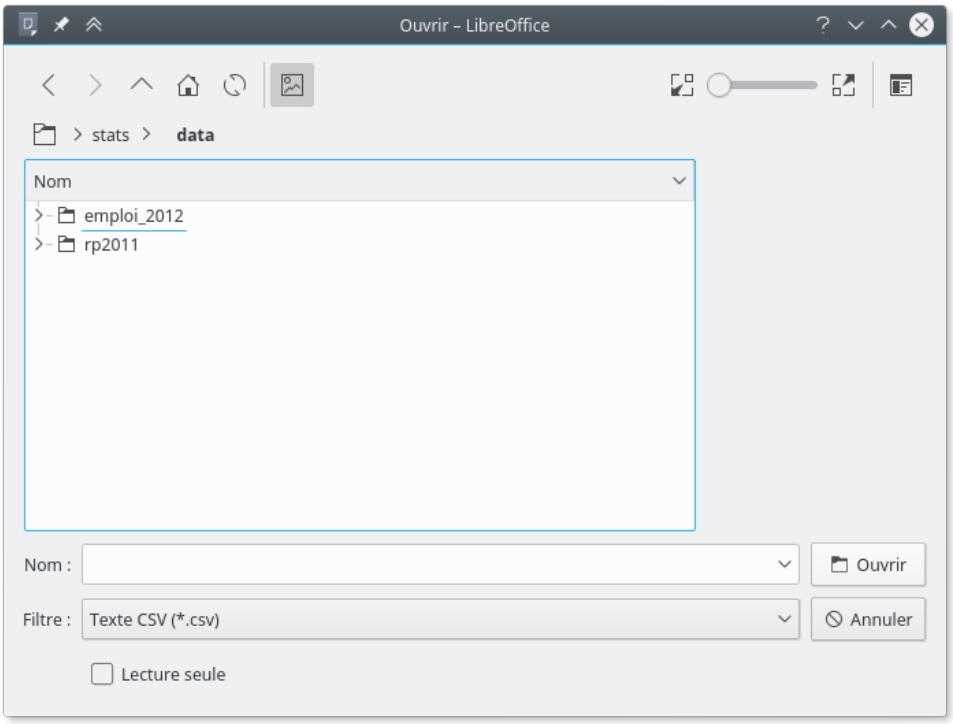
```
R> donnees <- read.delim2("fichier.txt")
```

Dans le second cas, on utilise `read.csv2`, de la même manière :

```
R> donnees <- read.csv2("fichier.csv")
```

4.2.2 Depuis OpenOffice ou LibreOffice

Depuis OpenOffice on procédera de la même manière, en sélectionnant le type de fichier *Texte CSV*.



On importe ensuite les données dans R à l'aide de la fonction `read.csv` :

```
R> read.csv("fichier.csv", dec = ",")
```

4.2.3 Autres sources / en cas de problèmes

Les fonctions `read.csv` et compagnie sont en fait des dérivées de la fonction plus générique `read.table`. Celle-ci contient de nombreuses options permettant d'adapter l'import au format du fichier texte. On pourra se reporter à la page d'aide de `read.table` si on recontre des problèmes ou si on souhaite importer des fichiers d'autres sources.

Parmi les options disponibles, on citera notamment :

`header` indique si la première ligne du fichier contient les noms des variables (valeur `TRUE`) ou non (valeur `FALSE`).

`sep` indique le caractère séparant les champs. En général soit une virgule, soit un point-virgule, soit une tabulation. Pour cette dernière l'option est `sep="\t"`.

`quote` indique le caractère utilisé pour délimiter les champs. En général on utilise soit des guillemets doubles (`quote="\""`) soit rien du tout (`quote=""`).

`dec` indique quel est le caractère utilisé pour séparer les nombres et leurs décimales. Il s'agit le plus souvent de la virgule lorsque les données sont en français (`dec=","`), et le point pour les données anglophones (`dec="."`).

D'autres options sont disponibles, pour gérer le format d'encodage du fichier source ou de nombreux autres paramètres d'importation. On se réfèrera alors à la page d'aide de `read.table` et à la section *Spreadsheet-like data* de *R Data Import/Export* :

http://cran.r-project.org/doc/manuals/R-data.html#Spreadsheet_002dlike-data

4.3 Import depuis d'autres logiciels

La plupart des fonctions permettant l'import de fichiers de données issus d'autres logiciels font partie d'une extension nommée foreign, présente à l'installation de R mais qu'il est nécessaire de charger en mémoire avant utilisation avec l'instruction :

```
R> library(foreign)
```

4.3.1 SAS

Les fichiers au format SAS se présentent en général sous deux format : format SAS export (extension `.xport` ou `.xpt`) ou format SAS natif (extension `.sas7bdat`).

R peut lire directement les fichiers au format export *via* la fonction `read.xport` de l'extension foreign.

Celle-ci s'utilise très simplement, en lui passant le nom du fichier en argument :

```
R> donnees <- read.xport("fichier.xpt")
```

En ce qui concerne les fichiers au format SAS natif, on pourra utiliser la fonction `read_sas` de l'extension `haven`. Celle-ci est capable de lire directement les fichiers au format `sas7bdat` et `sas7bcat` :

```
R> library(haven)
R> donnees <- read_sas("fichier.sas7bdat")
```

4.3.2 SPSS

Les fichiers générés par SPSS sont accessibles depuis R avec la fonction `read.spss` de l'extension `foreign`. Celle-ci peut lire aussi bien les fichiers sauvegardés avec la fonction *Enregistrer* que ceux générés par la fonction *Exporter*.

La syntaxe est également très simple :

```
R> donnees <- read.spss("fichier.sav", to.data.frame = TRUE)
```

Plusieurs options permettant de contrôler l'importation des données sont disponibles. On se reportera à la page d'aide de la fonction pour plus d'informations. Il est vivement recommandé d'utiliser systématiquement l'option `to.data.frame=TRUE`.

Une alternative est fournie par l'extension `haven` et ses fonctions `read_por` et `read_sav`.

4.3.3 Stata

Les fichiers générés par Stata sont accessibles depuis R avec la fonction `read.dta` de l'extension `foreign`.

La syntaxe est également très simple :

```
R> donnees <- read.data("fichier.dta", to.data.frame = TRUE)
```

L'importation des dates est parfois mal gérées. Dans ces cas là, l'opération suivante peut fonctionner. Sans garantie néanmoins, il est toujours vivement conseillé de vérifier le résultat obtenu !
`donnees$date <- as.Date(donnees$Date/(1000*3600*24), origin='1960-01-01')`

4.3.4 Fichiers `dbf`

L'Insee diffuse ses fichiers détails depuis son site Web au format dBase (extension `.dbf`). Ceux-ci sont directement lisibles dans R avec la fonction `read.dbf` de l'extension `foreign`.

```
R> donnees <- read.dbf("fichier.dbf")
```

La principale limitation des fichiers `dbf` est de ne pas gérer plus de 256 colonnes. Les tables des enquêtes de l'Insee sont donc parfois découpées en plusieurs fichiers `dbf` qu'il convient de fusionner avec la fonction `merge`. L'utilisation de cette fonction est détaillée dans la section 5.6 page 76.

4.4 Autres sources

R offre de très nombreuses autres possibilités pour accéder aux données. Il est ainsi possible d'importer des données depuis d'autres applications qui n'ont pas été évoquées (Epi Info, S-Plus, etc.), de se connecter à un système de base de données relationelle type MySql, de lire des données *via* `ODBC` ou des connexions réseau, etc.

Pour plus d'informations on consultera le manuel *R Data Import/Export* :

http://cran.r-project.org/manuals.html

4.5 Sauver ses données

R dispose également de son propre format pour sauvegarder et échanger des données. On peut sauver n'importe quel objet créé avec R et il est possible de sauver plusieurs objets dans un même fichier. L'usage est d'utiliser l'extension `.RData` pour les fichiers de données R. La fonction à utiliser s'appelle tout simplement `save`.

Par exemple, si l'on souhaite sauvegarder son tableau de données `d` ainsi que les objets `tailles` et `poids` dans un fichier `export.RData` :

```
R> save(d, tailles, poids, file = "export.RData")
```

À tout moment, il sera toujours possible de recharger ces données en mémoire à l'aide de la fonction `load` :

```
R> load("export.RData")
```

Si entre temps vous aviez modifié votre tableau `d`, vos modifications seront perdues. En effet, si lors du chargement de données, un objet du même nom existe en mémoire, ce dernier sera remplacé par l'objet importé.

La fonction `save.image` est un raccourci pour sauvergarder tous les objets de la session de travail dans le fichier `.RData` (un fichier un peu étrange car il n'a pas de nom mais juste une extension). Lors de la fermeture de R ou de RStudio, il vous sera demandé si vous souhaitez enregistrer votre session. Si vous répondez *Oui*, c'est cette fonction `save.image` qui sera appliquée.

```
R> save.image()
```

4.6 Exporter des données

R propose également différentes fonctions permettant d'exporter des données vers des formats variés.

— `write.table` est l'équivalent de `read.table` et permet d'enregistrer des tableaux de données au format texte, avec de nombreuses options ;
— `write.foreign`, de l'extension `foreign`, permet d'exporter des données aux formats SAS, SPSS ou Stata ;
— `write.dbf`, de l'extension `foreign`, permet d'exporter des données au format dBase ;

À nouveau, pour plus de détails on se référera aux pages d'aide de ces fonctions et au manuel *R Data Import/Export*.

4.7 Exercices

Exercice 4.9

▷ *Solution page 131*

Saisissez quelques données fictives dans une application de type tableur, enregistrez-les dans un format texte et importez-les dans R.

Vérifiez que l'importation s'est bien déroulée.

Exercice 4.10

▷ *Solution page 131*

L'adresse suivante permet de télécharger un fichier au format dBase contenant une partie des données de l'enquête *EPCV Vie associative* de l'INSEE (2002) :

http://telechargement.insee.fr/fichiersdetail/epcv1002/dbase/epcv1002_BENEVOLAT_dbase.zip

Téléchargez le fichier, décompressez-le et importez les données dans R.

Partie 5

Manipulation de données

Cette partie est un peu aride et pas forcément très intuitive. Elle aborde cependant la base de tous les traitements et manipulation de données sous R, et mérite donc qu'on s'y arrête un moment, ou qu'on y revienne un peu plus tard en cas de saturation...

5.1 Variables

Le type d'objet utilisé par R pour stocker des tableaux de données s'appelle un *data frame*. Celui-ci comporte des observations en ligne et des variables en colonnes. On accède aux variables d'un *data frame* avec l'opérateur `$`.

Dans ce qui suit on travaillera sur le jeu de données tiré de l'enquête *Histoire de vie*, fourni avec l'extension questionr et décrit dans l'annexe B.3.4, page 127.

```
R> library(questionr)
R> data(hdv2003)
R> d <- hdv2003
```

Mais aussi sur le jeu de données tiré du recensement 1999, décrit page 128 :

```
R> data(rp99)
```

5.1.1 Types de variables

On peut considérer qu'il existe quatre type de variables dans R :

- les variables **numériques**, ou quantitatives ;
- les **facteurs**, qui prennent leurs valeurs dans un ensemble défini de modalités. Elles correspondent en général aux questions fermées d'un questionnaire ;
- les variables **caractères**, qui contiennent des chaînes de caractères plus ou moins longues. On les utilise pour les questions ouvertes ou les champs libres ;
- les variables **booléennes**, qui ne peuvent prendre que la valeur *vrai* (`TRUE`) ou *faux* (`FALSE`). On les utilise dans R pour les calculs et les recodages.

Pour connaître le type d'une variable donnée, on peut utiliser la fonction `class`.

Résultat de `class`	Type de variable
`factor`	Facteur
`integer`	Numérique
`double`	Numérique
`numeric`	Numérique
`character`	Caractères
`logical`	Booléenne

```
R> class(d$age)

[1] "integer"

R> class(d$sexe)

[1] "factor"

R> class(c(TRUE, TRUE, FALSE))

[1] "logical"
```

La fonction `str` permet également d'avoir un listing de toutes les variables d'un tableau de données et indique le type de chacune d'elle.

5.1.2 Renommer des variables

Une opération courante lorsqu'on a importé des variables depuis une source de données externe consiste à renommer les variables importées. Sous R les noms de variables doivent être à la fois courts et explicites tout en obéissant à certaines règles décrites dans la remarque page 12.

On peut lister les noms des variables d'un *data frame* à l'aide de la fonction `names` :

```
R> names(d)

 [1] "id"            "age"           "sexe"          "nivetud"
 [5] "poids"         "occup"         "qualif"        "freres.soeurs"
 [9] "clso"          "relig"         "trav.imp"      "trav.satisf"
[13] "hard.rock"     "lecture.bd"    "peche.chasse"  "cuisine"
[17] "bricol"        "cinema"        "sport"         "heures.tv"
```

Cette fonction peut également être utilisée pour renommer l'ensemble des variables. Si par exemple on souhaitait passer les noms de toutes les variables en majuscules, on pourrait faire :

```
R> d.maj <- d
R> names(d.maj) <- c("ID", "AGE", "SEXE", "NIVETUD", "POIDS", "OCCUP", "QUALIF",
+     "FRERES.SOEURS", "CLSO", "RELIG", "TRAV.IMP", "TRAV.SATISF", "HARD.ROCK",
+     "LECTURE.BD", "PECHE.CHASSE", "CUISINE", "BRICOL", "CINEMA", "SPORT", "HEURES.TV")
R> summary(d.maj$SEXE)

Homme Femme
  899  1101
```

Ce type de renommage peut être utile lorsqu'on souhaite passer en revue tous les noms de variables d'un fichier importé pour les corriger le cas échéant. Pour faciliter un peu ce travail pas forcément passionant, on peut utiliser la fonction `dput` :

```
R> dput(names(d))

c("id", "age", "sexe", "nivetud", "poids", "occup", "qualif",
"freres.soeurs", "clso", "relig", "trav.imp", "trav.satisf",
"hard.rock", "lecture.bd", "peche.chasse", "cuisine", "bricol",
"cinema", "sport", "heures.tv")
```

On obtient en résultat la liste des variables sous forme de vecteur déclaré. On n'a plus alors qu'à copier/coller cette chaîne, rajouter `names(d) <-` devant, et modifier un à un les noms des variables.

questionr

Si on souhaite seulement modifier le nom d'une variable, on peut utiliser la fonction `rename.variable` de l'extension questionr. Celle-ci prend en argument le tableau de données, le nom actuel de la variable et le nouveau nom. Par exemple, si on veut renommer la variable `bricol` du tableau de données `d` en `bricolage` :

```
R> d <- rename.variable(d, "bricol", "bricolage")
R> table(d$bricolage)


 Non  Oui
1147  853
```

5.1.3 Facteurs

Parmi les différents types de variables, les *facteurs* (`factor`) sont à la fois à part et très utilisés, car ils vont correspondre à la plupart des variables issues d'une question fermée dans un questionnaire.

Les facteurs prennent leurs valeurs dans un ensemble de modalités prédéfinies, et ne peuvent en prendre d'autres. La liste des valeurs possibles est donnée par la fonction `levels` :

```
R> levels(d$sexe)

[1] "Homme" "Femme"
```

Si on veut modifier la valeur du sexe du premier individu de notre tableau de données avec une valeur différente, on obient un message d'erreur et une valeur manquante est utilisée à la place :

```
R> d$sexe[1] <- "Chihuahua"

Warning in `[<-.factor`(`*tmp*`, 1, value = structure(c(NA, 2L, 1L, 1L, : invalid
factor level, NA generated

R> d$sexe[1]

[1] <NA>
Levels: Homme Femme
```

On peut très facilement créer un facteur à partir d'une variable de type caractères avec la commande `factor` :

```
R> v <- factor(c("H", "H", "F", "H"))
R> v

[1] H H F H
Levels: F H
```

Par défaut, les niveaux d'un facteur nouvellement créés sont l'ensemble des valeurs de la variable caractères, ordonnées par ordre alphabétique. Cette ordre des niveaux est utilisé à chaque fois qu'on utilise des fonctions comme `table`, par exemple :

```
R> table(v)

v
F H
1 3
```

On peut modifier cet ordre au moment de la création du facteur en utilisant l'option `levels` :

```
R> v <- factor(c("H", "H", "F", "H"), levels = c("H", "F"))
R> table(v)

v
H F
3 1
```

On peut aussi modifier l'ordre des niveaux d'une variable déjà existante :

```
R> d$qualif <- factor(d$qualif, levels = c("Ouvrier specialise", "Ouvrier qualifie",
+     "Employe", "Technicien", "Profession intermediaire", "Cadre", "Autre"))
R> table(d$qualif)


      Ouvrier specialise         Ouvrier qualifie                  Employe
                     203                      292                      594
              Technicien Profession intermediaire                   Cadre
                      86                      160                      260
                  Autre
                      58
```

Interface interactive

questionr

L'extension questionr propose une interface interactive pour le réordonnancement des niveaux d'un facteur. Cette fonction, nommée `iorder`, vous permet de réordonner les modalités de manière graphique et de générer le code R correspondant.

Vous pouvez lancer `iorder` directement depuis un script ou la console. Dans l'exemple précédent, vous pouvez exécuter :

```
R> iorder(d$qualif)
```

Avec une version récente de RStudio, vous pouvez également sélectionner le menu *Addins*, puis choisir *Levels ordering*.

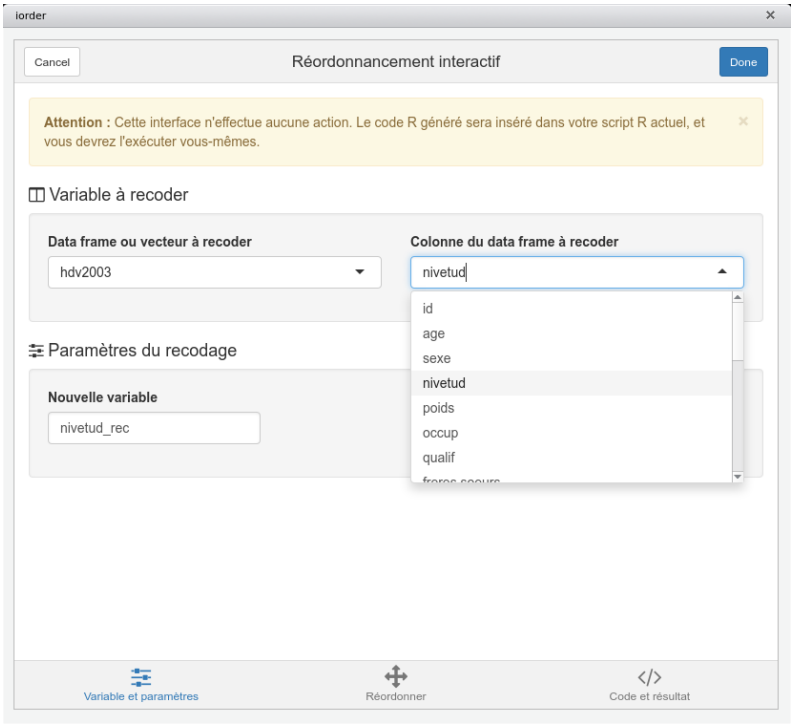


FIGURE 5.1 – Interface de la commande `iorder`

Dans les deux cas, une interface semblable à celle de la figure 5.1, de la présente page devrait s'afficher soit directement dans RStudio, soit dans votre navigateur Web.

Vous pouvez alors choisir la variable à réordonner dans l'onglet *Variable et paramètres*, déplacer les modalités par glisser-déposer dans l'onglet *Réordonner*, et vérifier le résultat dans l'onglet *Code et résultat*. Une fois le résultat satisfaisant, cliquez sur *Done* : si vous êtes sous RStudio le code généré sera directement inséré dans votre script actuel à l'emplacement du curseur. Sinon, ce code sera affiché dans la console et vous pourrez le copier/coller pour l'inclure dans votre script [1].

On peut également modifier les niveaux eux-mêmes. Imaginons que l'on souhaite créer une nouvelle variable `qualif.abr` contenant les noms abrégés des catégories socioprofessionnelles de `qualif`. On peut alors procéder comme suit :

```
R> d$qualif.abr <- factor(d$qualif, levels = c("Ouvrier specialise", "Ouvrier qualifie",
+     "Employe", "Technicien", "Profession intermediaire", "Cadre", "Autre"),
+     labels = c("OS", "OQ", "Empl", "Tech", "Interm", "Cadre", "Autre"))
R> table(d$qualif.abr)


   OS    OQ  Empl  Tech Interm Cadre Autre
  203   292   594    86   160   260    58
```

Dans ce qui précède, le paramètre `levels` de `factor` permet de spécifier quels sont les niveaux retenus dans le facteur résultat, ainsi que leur ordre. Le paramètre `labels`, lui, permet de modifier les noms de ces niveaux dans le facteur résultat. Il est donc capital d'indiquer les noms de `labels` exactement dans le même ordre que les niveaux de `levels`. Pour s'assurer de ne pas avoir commis d'erreur, il est recommandé d'effectuer un tableau croisé entre l'ancien et le nouveau facteur :

1. Le fonctionnement des interfaces interactives de questionr est décrit plus en détail section B.3.3, page 125.

```
R> table(d$qualif, d$qualif.abr)


                           OS  OQ Empl Tech Interm Cadre Autre
  Ouvrier specialise      203   0    0    0      0     0     0
  Ouvrier qualifie          0 292    0    0      0     0     0
  Employe                   0   0  594    0      0     0     0
  Technicien                0   0    0   86      0     0     0
  Profession intermediaire  0   0    0    0    160     0     0
  Cadre                     0   0    0    0      0   260     0
  Autre                     0   0    0    0      0     0    58
```

On a donc ici un premier moyen d'effectuer un recodage des modalités d'une variable de type facteur. D'autres méthodes existent, elles sont notamment détaillées section 5.4 page 66.

À noter que par défaut, les valeurs manquantes ne sont pas considérées comme un niveau de facteur. On peut cependant les transformer en niveau en utilisant la fonction `addNA`. Ceci signifie cependant qu'elle ne seront plus considérées comme manquantes par R :

```
R> summary(d$trav.satisf)

  Satisfaction Insatisfaction      Equilibre           NA's
           480            117            451            952

R> summary(addNA(d$trav.satisf))

  Satisfaction Insatisfaction      Equilibre           <NA>
           480            117            451            952
```

5.2 Indexation

L'indexation est l'une des fonctionnalités les plus puissantes mais aussi les plus difficiles à maîtriser de R. Il s'agit d'opérations permettant de sélectionner des sous-ensembles d'observations et/ou de variables en fonction de différents critères. L'indexation peut porter sur des vecteurs, des matrices ou des tableaux de données.

Le principe est toujours le même : on indique, entre crochets et à la suite du nom de l'objet à indexer, une série de conditions indiquant ce que l'on garde ou non. Ces conditions peuvent être de différents types.

5.2.1 Indexation directe

Le mode le plus simple d'indexation consiste à indiquer la position des éléments à conserver. Dans le cas d'un vecteur cela permet de sélectionner un ou plusieurs éléments de ce vecteur.

Soit le vecteur suivant :

```
R> v <- c("a", "b", "c", "d", "e", "f", "g")
```

Si on souhaite le premier élément du vecteur, on peut faire :

```
R> v[1]

[1] "a"
```

Si on souhaite les trois premiers éléments ou les éléments 2, 6 et 7 :

```
R> v[1:3]

[1] "a" "b" "c"

R> v[c(2, 6, 7)]

[1] "b" "f" "g"
```

Si on veut le dernier élément :

```
R> v[length(v)]

[1] "g"
```

Dans le cas de matrices ou de tableaux de données, l'indexation prend deux arguments séparés par une virgule : le premier concerne les *lignes* et le second les *colonnes*. Ainsi, si on veut l'élément correspondant à la troisième ligne et à la cinquième colonne du tableau de données `d` :

```
R> d[3, 5]

[1] 3994.102
```

On peut également indiquer des vecteurs :

```
R> d[1:3, 1:2]

  id age
1  1  28
2  2  23
3  3  59
```

Si on laisse l'un des deux critères vides, on sélectionne l'intégralité des lignes ou des colonnes. Ainsi si l'on veut seulement la cinquième colonne ou les deux premières lignes :

```
R> d[, 5]

 [1]  2634.3982  9738.3958  3994.1025  5731.6615  4329.0940  8674.6994
 [7]  6165.8035 12891.6408  7808.8721  2277.1605   704.3227  6697.8682
[13]  7118.4659   586.7714 11042.0774  9958.2287  4836.1393  1551.4846
[19]  3141.1572 27195.8378 14647.9983  8128.0603  1281.9156 11663.3383
[25]  8780.2614  1700.8437  6662.8375  3359.4690  8536.1101 10620.5259
[31]  5264.2953 14161.7597  1339.6196  9243.9153  4512.2959  7871.6452
[37]  1356.9621  7626.3300  1630.2746  2196.2485  5605.9846  8841.2960
[43]  9113.5378  2267.5912  7706.2944  2446.5111  8118.2639 10751.5037
```

```
[49]   831.8599  6591.6440  1936.8826   834.3845  3432.5286 11354.8932
[55]  9292.9762  6344.1227  4899.9404  4766.8652  3462.8121 23732.4853
[61]   833.8428  8529.4403  3190.3680  2423.1052  5945.9929 14991.8652
[67]  2062.1124  5702.0623 20604.2642  2634.4861
 [ reached getOption("max.print") -- omitted 1930 entries ]

R> d[1:2, ]

  id age  sexe                                                nivetud
1  1  28 Femme Enseignement superieur y compris technique superieur
2  2  23 Femme                                                   <NA>
     poids                occup  qualif freres.soeurs clso
1 2634.398 Exerce une profession Employe             8  Oui
2 9738.396       Etudiant, eleve    <NA>             2  Oui
                       relig      trav.imp    trav.satisf hard.rock
1 Ni croyance ni appartenance Peu important Insatisfaction       Non
2 Ni croyance ni appartenance          <NA>           <NA>       Non
  lecture.bd peche.chasse cuisine bricol cinema sport heures.tv qualif.abr
1        Non          Non     Oui    Non    Non   Non         0       Empl
2        Non          Non     Non    Non    Oui   Oui         1       <NA>
```

Enfin, si on préfixe les arguments avec le signe « - », ceci signifie « tous les éléments sauf ceux indiqués ». Si par exemple on veut tous les éléments de `v` sauf le premier :

```
R> v[-1]

[1] "b" "c" "d" "e" "f" "g"
```

Bien sûr, tous ces critères se combinent et on peut stocker le résultat dans un nouvel objet. Dans cet exemple `d2` contiendra les trois premières lignes de `d` mais sans les colonnes 2, 6 et 8.

```
R> d2 <- d[1:3, -c(2, 6, 8)]
```

5.2.2 Indexation par nom

Un autre mode d'indexation consiste à fournir non pas un numéro mais un nom sous forme de chaîne de caractères. On l'utilise couramment pour sélectionner les variables d'un tableau de données. Ainsi, les deux fonctions suivantes sont équivalentes [2] :

```
R> d$clso

 [1] Oui         Oui         Non         Non         Oui
 [6] Non         Oui         Non         Oui         Non
[11] Oui         Oui         Oui         Oui         Oui
[16] Non         Non         Non         Non         Non
[21] Oui         Oui         Non         Non         Non
[26] Oui         Non         Non         Non         Oui
[31] Non         Oui         Oui         Non         Non
[36] Oui         Oui         Non         Non         Oui
[41] Non         Non         Oui         Non         Non
```

```
[46] Non         Non         Oui         Oui         Non
[51] Non         Non         Oui         Non         Oui
[56] Oui         Non         Non         Oui         Non
[61] Non         Oui         Oui         Oui         Oui
[66] Non         Oui         Non         Non         Ne sait pas
 [ reached getOption("max.print") -- omitted 1930 entries ]
Levels: Oui Non Ne sait pas

R> d[, "clso"]

 [1] Oui         Oui         Non         Non         Oui
 [6] Non         Oui         Non         Oui         Non
[11] Oui         Oui         Oui         Oui         Oui
[16] Non         Non         Non         Non         Non
[21] Oui         Oui         Non         Non         Non
[26] Oui         Non         Non         Non         Oui
[31] Non         Oui         Oui         Non         Non
[36] Oui         Oui         Non         Non         Oui
[41] Non         Non         Oui         Non         Non
[46] Non         Non         Oui         Oui         Non
[51] Non         Non         Oui         Non         Oui
[56] Oui         Non         Non         Oui         Non
[61] Non         Oui         Oui         Oui         Oui
[66] Non         Oui         Non         Non         Ne sait pas
 [ reached getOption("max.print") -- omitted 1930 entries ]
Levels: Oui Non Ne sait pas
```

Là aussi on peut utiliser un vecteur pour sélectionner plusieurs noms et récupérer un « sous-tableau » de données :

```
R> d2 <- d[, c("id", "sexe", "age")]
```

Les noms peuvent également être utilisés pour les observations (lignes) d'un tableau de données si celles-ci ont été munies d'un nom avec la fonction `row.names`. Par défaut les noms de ligne sont leur numéro d'ordre, mais on peut leur assigner comme nom la valeur d'une variable d'identifiant. Ainsi, on peut assigner aux lignes du jeu de données `rp99` le nom des communes correspondantes :

```
R> row.names(rp99) <- rp99$nom
```

On peut alors accéder directement aux communes en donnant leur nom :

```
R> rp99[c("VILLEURBANNE", "OULLINS"), ]

                     nom  code pop.act pop.tot  pop15 nb.rp      agric
VILLEURBANNE VILLEURBANNE 69266   57252  124152 103157 55136 0.02095997
OULLINS           OULLINS 69149   11849   25186  20880 11091 0.10127437
                 artis   cadres   interm     empl     ouvr     retr
VILLEURBANNE 5.143925 13.13841 25.72312 31.41550 23.07343 36.65374
```

2. Une différence entre les deux est que `$` admet une correspondance partielle du nom de variable, si celle-ci est unique. Ainsi, `d$cls` renverra bien la variable `clso`, tandis que `d$c` renverra `NULL`, du fait que plusieurs variables de `d` commencent par `c`.

```
OULLINS      4.818972 10.20339 27.42847 31.53009 24.37336 41.54781
              tx.chom     etud dipl.sup dipl.aucun  proprio      hlm
VILLEURBANNE 14.82394 15.50646 9.744370   16.90045 37.61970 23.33684
OULLINS      10.64225 10.62739 7.624521   14.31513 51.51023 14.56136
             locataire    maison
VILLEURBANNE  32.76988  6.532937
OULLINS       29.91615 17.708052
```

Par contre il n'est pas possible d'utiliser directement l'opérateur « - » comme pour l'indexation directe. Pour exclure une colonne en fonction de son nom, on doit uiliser une autre forme d'indexation, *l'indexation par condition*, expliquée dans la section suivante. On peut ainsi faire :

```
R> d2 <- d[, names(d) != "qualif"]
```

Pour sélectionner toutes les colonnes sauf celle qui s'appelle `qualif`.

5.2.3 Indexation par conditions

Tests et conditions

Une condition est une expression logique dont le résultat est soit `TRUE` (vrai) soit `FALSE` (faux).

Une condition comprend la plupart du temps un opérateur de comparaison. Les plus courants sont les suivants :

Opérateur	Signification
==	égal à
!=	différent de
>	strictement supérieur à
<	strictement inférieur à
>=	supérieur ou égal à
<=	inférieur ou égal à

Voyons tout de suite un exemple :

```
R> d$sexe == "Homme"

 [1] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
[12]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE
[23] FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[34]  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE
[45] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE
[56] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
[67]  TRUE  TRUE FALSE FALSE
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Que s'est-il passé ? Nous avons fourni à R une condition qui signifie « la valeur de la variable `sexe` vaut "Homme" ». Et il nous a renvoyé un vecteur avec autant d'éléments qu'il y'a d'observations dans `d`, et dont la valeur est `TRUE` si l'observation correspond à un homme, et `FALSE` dans les autres cas.

Prenons un autre exemple. On n'affichera cette fois que les premiers éléments de notre variable d'intérêt à l'aide de la fonction `head` :

```
R> head(d$age)

[1] 28 23 59 34 71 35

R> head(d$age > 40)

[1] FALSE FALSE  TRUE FALSE  TRUE FALSE
```

On voit bien ici qu'à chaque élément du vecteur `d$age` dont la valeur est supérieure à 40 correspond un élément `TRUE` dans le résultat de la condition.

On peut combiner ou modifier des conditions à l'aide des opérateurs logiques habituels :

Opérateur	Signification
&	et logique
|	ou logique
!	négation logique

Comment les utilise-t-on ? Voyons tout de suite des exemples. Supposons que je veuille déterminer quels sont dans mon échantillon les hommes ouvriers spécialisés :

```
R> d$sexe == "Homme" & d$qualif == "Ouvrier specialise"

 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE    NA
[34]    NA FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[56] FALSE FALSE FALSE FALSE FALSE FALSE    NA FALSE    NA FALSE FALSE
[67] FALSE FALSE FALSE FALSE
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Si je souhaite identifier les personnes qui bricolent ou qui font la cuisine :

```
R> d$bricol == "Oui" | d$cuisine == "Oui"

 [1]  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE
[12]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE
[23]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[34]  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
[45] FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
[56]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
[67]  TRUE  TRUE  TRUE  TRUE
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Si je souhaite isoler les femmes qui ont entre 20 et 34 ans :

```
R> d$sexe == "Femme" & d$age >= 20 & d$age <= 34

 [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
[12] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
[23] FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
[34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[45] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
[56]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[67] FALSE FALSE FALSE FALSE
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Si je souhaite récupérer les enquêtés qui ne sont pas cadres, on peut utiliser l'une des deux formes suivantes :

```
R> d$qualif != "Cadre"

 [1]  TRUE    NA  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE    NA  TRUE  TRUE
[12]  TRUE  TRUE    NA    NA  TRUE FALSE    NA  TRUE  TRUE  TRUE  TRUE
[23]  TRUE    NA  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE    NA
[34]    NA  TRUE  TRUE  TRUE    NA  TRUE FALSE  TRUE  TRUE FALSE FALSE
[45]    NA  TRUE  TRUE    NA  TRUE FALSE  TRUE  TRUE  TRUE    NA  TRUE
[56]  TRUE    NA FALSE  TRUE    NA  TRUE    NA  TRUE    NA  TRUE  TRUE
[67]  TRUE  TRUE  TRUE    NA
 [ reached getOption("max.print") -- omitted 1930 entries ]

R> !(d$qualif == "Cadre")

 [1]  TRUE    NA  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE    NA  TRUE  TRUE
[12]  TRUE  TRUE    NA    NA  TRUE FALSE    NA  TRUE  TRUE  TRUE  TRUE
[23]  TRUE    NA  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE    NA
[34]    NA  TRUE  TRUE  TRUE    NA  TRUE FALSE  TRUE  TRUE FALSE FALSE
[45]    NA  TRUE  TRUE    NA  TRUE FALSE  TRUE  TRUE  TRUE    NA  TRUE
[56]  TRUE    NA FALSE  TRUE    NA  TRUE    NA  TRUE    NA  TRUE  TRUE
[67]  TRUE  TRUE  TRUE    NA
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Lorsqu'on mélange « et » et « ou » il est nécessaire d'utiliser des parenthèses pour différencier les blocs. La condition suivante identifie les femmes qui sont soit cadre, soit employée :

```
R> d$sexe == "Femme" & (d$qualif == "Employe" | d$qualif == "Cadre")

 [1]  TRUE    NA FALSE FALSE  TRUE  TRUE FALSE FALSE    NA FALSE  TRUE
[12] FALSE  TRUE    NA    NA  TRUE FALSE    NA FALSE FALSE  TRUE FALSE
[23]  TRUE    NA FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[34] FALSE FALSE  TRUE FALSE    NA  TRUE FALSE  TRUE FALSE FALSE  TRUE
[45]    NA FALSE FALSE    NA  TRUE  TRUE FALSE  TRUE FALSE    NA FALSE
[56] FALSE    NA  TRUE FALSE    NA  TRUE FALSE FALSE FALSE FALSE  TRUE
[67] FALSE FALSE  TRUE    NA
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

L'opérateur `%in%` peut être très utile : il teste si une valeur fait partie des éléments d'un vecteur. Ainsi on pourrait remplacer la condition précédente par :

```
R> d$sexe == "Femme" & d$qualif %in% c("Employe", "Cadre")

 [1]  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE
```

```
[12] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
[23]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[34] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE
[45] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
[56] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
[67] FALSE FALSE  TRUE FALSE
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Enfin, signalons qu'on peut utiliser les fonctions `table` ou `summary` pour avoir une idée du résultat de notre condition :

```
R> table(d$sexe)


Homme Femme
  899  1101

R> table(d$sexe == "Homme")


FALSE  TRUE
 1101   899

R> summary(d$sexe == "Homme")

   Mode   FALSE    TRUE    NA's
logical    1101     899       0
```

Utilisation pour l'indexation

L'utilisation des conditions pour l'indexation est assez simple : si on indexe un vecteur avec un vecteur booléen, seuls les éléments correspondant à `TRUE` seront conservés.

Ainsi, si on fait :

```
R> dh <- d[d$sexe == "Homme", ]
```

On obtiendra un nouveau tableau de données comportant l'ensemble des variables de `d`, mais seulement les observations pour lesquelles `d$sexe` vaut « Homme ».

La plupart du temps ce type d'indexation s'applique aux lignes, mais on peut aussi l'utiliser sur les colonnes d'un tableau de données. L'exemple suivant, un peu compliqué, sélectionne uniquement les variables dont le nom commence par `a` ou `s` :

```
R> d[, substr(names(d), 0, 1) %in% c("a", "s")]

     age  sexe sport
1     28 Femme   Non
2     23 Femme   Oui
3     59 Homme   Oui
4     34 Homme   Oui
```

```
5     71 Femme   Non
6     35 Femme   Oui
7     60 Femme   Non
8     47 Homme   Non
9     20 Femme   Non
10    28 Homme   Oui
11    65 Femme   Non
12    47 Homme   Oui
13    63 Femme   Non
14    67 Femme   Non
15    76 Femme   Non
16    49 Femme   Non
17    62 Homme   Oui
18    20 Femme   Oui
19    70 Homme   Non
20    39 Femme   Oui
21    30 Femme   Non
22    30 Homme   Non
23    37 Femme   Oui
 [ getOption("max.print") est atteint -- 1977 lignes omises ]
```

On peut évidemment combiner les différents type d'indexation. L'exemple suivant sélectionne les femmes de plus de 40 ans et ne conserve que les variables `qualif` et `bricol`.

```
R> d2 <- d[d$sexe == "Femme" & d$age > 40, c("qualif", "bricol")]
```

Valeurs manquantes dans les conditions

Une remarque importante : quand l'un des termes d'une condition comporte une valeur manquante (`NA`), le résultat de cette condition n'est pas toujours `TRUE` ou `FALSE`, il peut aussi être à son tour une valeur manquante.

```
R> v <- c(1:5, NA)
R> v

[1]  1  2  3  4  5 NA

R> v > 3

[1] FALSE FALSE FALSE  TRUE  TRUE    NA
```

On voit que le test `NA > 3` ne renvoie ni vrai ni faux, mais `NA`.

Le résultat d'une condition peut donc comporter un grand nombre de valeurs manquantes :

```
R> summary(d$trav.satisf == "Satisfaction")

   Mode   FALSE    TRUE    NA's
logical     568     480     952
```

Une autre conséquence importante de ce comportement est qu'on ne peut pas utiliser l'opérateur `== NA` pour tester la présence de valeurs manquantes. On utilisera à la place la fonction *ad hoc* `is.na`.

On comprendra mieux le problème avec l'exemple suivant :

```
R> v <- c(1, NA)
R> v

[1]  1 NA

R> v == NA

[1] NA NA

R> is.na(v)

[1] FALSE  TRUE
```

Pour compliquer encore un peu le tout, lorsqu'on utilise une condition pour l'indexation, si la condition renvoie `NA`, R ne sélectionne pas l'élément mais retourne quand même la valeur `NA`. Ceci aura donc des conséquences pour l'extraction de sous-populations, comme indiqué section 5.3.1 page suivante.

5.2.4 Indexation et assignation

Dans tous les exemples précédents, on a utilisé l'indexation pour extraire une partie d'un vecteur ou d'un tableau de données, en plaçant l'opération d'indexation à droite de l'opérateur `<-`.

Mais l'indexation peut également être placée à gauche de cet opérateur. Dans ce cas, les éléments sélectionnés par l'indexation sont alors remplacés par les valeurs indiquées à droite de l'opérateur `<-`.

Ceci est parfaitement incompréhensible. Prenons donc un exemple simple :

```
R> v <- 1:5
R> v

[1] 1 2 3 4 5

R> v[1] <- 3
R> v

[1] 3 2 3 4 5
```

Cette fois, au lieu d'utiliser quelque chose comme `x <- v[1]`, qui aurait placé la valeur du premier élément de `v` dans `x`, on a utilisé `v[1] <- 3`, ce qui a *mis à jour* le premier élément de `v` avec la valeur 3.

Ceci fonctionne également pour les tableaux de données et pour les différents types d'indexation évoqués précédemment :

```
R> d[257, "sexe"] <- "Homme"
```

Enfin on peut modifier plusieurs éléments d'un seul coup soit en fournissant un vecteur, soit en profitant du mécanisme de recyclage. Les deux commandes suivantes sont ainsi rigoureusement équivalentes :

```
R> d[c(257, 438, 889), "sexe"] <- c("Homme", "Homme", "Homme")
R> d[c(257, 438, 889), "sexe"] <- "Homme"
```

On commence à voir comment l'utilisation de l'indexation par conditions et de l'assignation va nous permettre de faire des recodages.

```
R> d$age[d$age >= 20 & d$age <= 30] <- "20-30 ans"
R> d$age[is.na(d$age)] <- "Inconnu"
```

5.3 Sous-populations

5.3.1 Par indexation

La première manière de construire des sous-populations est d'utiliser l'indexation par conditions. On peut ainsi facilement sélectionner une partie des observations suivant un ou plusieurs critères et placer le résultat dans un nouveau tableau de données.

Par exemple si on souhaite isoler les hommes et les femmes :

```
R> dh <- d[d$sexe == "Homme", ]
R> df <- d[d$sexe == "Femme", ]
R> table(d$sexe)


Homme Femme
  899  1101

R> dim(dh)

[1] 899  20

R> dim(df)

[1] 1101   20
```

On a à partir de là trois tableaux de données, `d` comportant la population totale, `dh` seulement les hommes et `df` seulement les femmes.

On peut évidemment combiner plusieurs critères :

```
R> dh.25 <- d[d$sexe == "Homme" & d$age <= 25, ]
R> dim(dh.25)

[1] 86 20
```

Si on utilise directement l'indexation, il convient cependant d'être extrêmement prudent avec les valeurs manquantes. Comme indiqué précédemment, la présence d'une valeur manquante dans une condition fait que celle-ci est évaluée en `NA` et qu'au final la ligne correspondante est conservée par l'indexation :

```
R> summary(d$trav.satisf)

  Satisfaction Insatisfaction      Equilibre           NA's
           480            117            451            952
```

```
R> d.satisf <- d[d$trav.satisf == "Satisfaction", ]
R> dim(d.satisf)

[1] 1432   20
```

Comme on le voit, ici `d.satisf` contient les individus ayant la modalité *Satisfaction* mais aussi ceux ayant une valeur manquante `NA`. C'est pourquoi il faut toujours soit vérifier au préalable qu'on n'a pas de valeurs manquantes dans les variables de la condition, soit exclure explicitement les `NA` de la manière suivante :

```
R> d.satisf <- d[d$trav.satisf == "Satisfaction" & !is.na(d$trav.satisf), ]
R> dim(d.satisf)

[1] 480  20
```

C'est notamment pour cette raison qu'on préfèrera le plus souvent utiliser la fonction `subset`.

5.3.2 Fonction `subset`

La fonction `subset` permet d'extraire des sous-populations de manière plus simple et un peu plus intuitive que l'indexation directe.

Celle-ci prend trois arguments principaux :

— le nom de l'objet de départ ;
— une condition sur les observations (`subset`) ;
— éventuellement une condition sur les colonnes (`select`).

Reprenons tout de suite un exemple déjà vu :

```
R> dh <- subset(d, sexe == "Homme")
R> df <- subset(d, sexe == "Femme")
```

L'utilisation de `subset` présente plusieurs avantages. Le premier est d'économiser quelques touches. On n'est en effet pas obligé de saisir le nom du tableau de données dans la condition sur les lignes. Ainsi les deux commandes suivantes sont équivalentes :

```
R> dh <- subset(d, d$sexe == "Homme")
R> dh <- subset(d, sexe == "Homme")
```

Le second avantage est que `subset` s'occupe du problème des valeurs manquantes évoquées précédemment et les exclut de lui-même, contrairement au comportement par défaut :

```
R> summary(d$trav.satisf)

  Satisfaction Insatisfaction      Equilibre           NA's
           480            117            451            952

R> d.satisf <- d[d$trav.satisf == "Satisfaction", ]
R> dim(d.satisf)

[1] 1432   20
```

```
R> d.satisf <- subset(d, trav.satisf == "Satisfaction")
R> dim(d.satisf)

[1] 480  20
```

Enfin, l'utilisation de l'argument `select` est simplifié pour l'expression de condition sur les colonnes. On peut ainsi spécifier les noms de variable sans guillemets et leur appliquer directement l'opérateur d'exclusion `-` :

```
R> d2 <- subset(d, select = c(sexe, sport))
R> d2 <- subset(d, age > 25, select = -c(id, age, bricol))
```

5.3.3 Fonction `tapply`

Cette section documente une fonction qui peut être très utile, mais pas forcément indispensable au départ. Vous pouvez donc passer directement à la suite si vous le souhaitez.

La fonction `tapply` n'est qu'indirectement liée à la notion de sous-population, mais peut permettre d'éviter d'avoir à créer ces sous-populations dans certains cas.

Son fonctionnement est assez simple, mais pas forcément intuitif. La fonction prend trois arguments : un vecteur, un facteur et une fonction. Elle applique ensuite la fonction aux éléments du vecteur correspondant à un même niveau du facteur. Vite, un exemple !

```
R> tapply(d$age, d$sexe, mean)

   Homme    Femme
48.16129 48.15350
```

Qu'est-ce que ça signifie ? Ici `tapply` a sélectionné toutes les observations correspondant à « Homme », puis appliqué la fonction `mean` aux valeurs de `age` correspondantes. Puis elle a fait de même pour les observations correspondant à « Femme ». On a donc ici la moyenne d'âge chez les hommes et chez les femmes.

On peut fournir à peu près n'importe quelle fonction à `tapply` :

```
R> tapply(d$bricol, d$sexe, freq)

$Homme
      n    % val%
Non 384 42.7 42.7
Oui 515 57.3 57.3
NA    0  0.0   NA

$Femme
      n    % val%
Non 763 69.3 69.3
Oui 338 30.7 30.7
NA    0  0.0   NA
```

Les arguments supplémentaires fournis à `tapply` sont en fait fournis directement à la fonction appelée.

```
R> tapply(d$bricol, d$sexe, freq, total = TRUE)

$Homme
        n     %  val%
Non   384  42.7  42.7
Oui   515  57.3  57.3
NA      0   0.0    NA
Total 899 100.0 100.0

$Femme
         n     %  val%
Non    763  69.3  69.3
Oui    338  30.7  30.7
NA       0   0.0    NA
Total 1101 100.0 100.0
```

À noter également, la fonction `by` est un équivalent (pour les tableaux de données) de `tapply`. La présentation des résultats diffère légèrement.

```
R> tapply(d$age, d$sexe, mean)

   Homme    Femme
48.16129 48.15350

R> by(d$age, d$sexe, mean)

d$sexe: Homme
[1] 48.16129
------------------------------------------------------------
d$sexe: Femme
[1] 48.1535
```

5.4 Recodages

Le recodage de variables est une opération extrêmement fréquente lors du traitement d'enquête. Celui-ci utilise soit l'une des formes d'indexation décrites précédemment, soit des fonctions *ad hoc* de R.

On passe ici en revue différents types de recodage parmi les plus courants. Les exemples s'appuient, comme précédemment, sur l'extrait de l'enquête *Histoire de vie* :

```
R> data(hdv2003)
R> d <- hdv2003
```

5.4.1 Convertir une variable

Il peut arriver qu'on veuille transformer une variable d'un type dans un autre.

Par exemple, on peut considérer que la variable numérique `freres.soeurs` est une « fausse » variable numérique et qu'une représentation sous forme de facteur serait plus adéquate. Dans ce cas il suffit de faire appel à la fonction `factor` :

```
R> d$fs.fac <- factor(d$freres.soeurs)
R> levels(d$fs.fac)

 [1] "0"  "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13"
[15] "14" "15" "16" "18" "22"
```

La conversion d'une variable caractères en facteur se fait de la même manière.

La conversion d'un facteur ou d'une variable numérique en variable caractères peut se faire à l'aide de la fonction `as.character` :

```
R> d$fs.char <- as.character(d$freres.soeurs)
R> d$qualif.char <- as.character(d$qualif)
```

La conversion d'un facteur en caractères est fréquemment utilisé lors des recodages du fait qu'il est impossible d'ajouter de nouvelles modalités à un facteur de cette manière. Par exemple, la première des commandes suivantes génère un message d'avertissement, tandis que les deux autres fonctionnent :

```
R> d$qualif[d$qualif == "Ouvrier specialise"] <- "Ouvrier"
R> d$qualif.char <- as.character(d$qualif)
R> d$qualif.char[d$qualif.char == "Ouvrier specialise"] <- "Ouvrier"
```

Dans le premier cas, le message d'avertissement indique que toutes les modalités « Ouvrier specialise » de notre variable `qualif` ont été remplacées par des valeurs manquantes `NA`.

Enfin, une variable de type caractères dont les valeurs seraient des nombres peut être convertie en variable numérique avec la fonction `as.numeric`. Si on souhaite convertir un facteur en variable numérique, il faut d'abord le convertir en variable de classe caractère :

```
R> d$fs.num <- as.numeric(as.character(d$fs.fac))
```

5.4.2 Découper une variable numérique en classes

Le premier type de recodage consiste à découper une variable de type numérique en un certain nombre de classes. On utilise pour cela la fonction `cut`.

Celle-ci prend, outre la variable à découper, un certain nombre d'arguments :
- `breaks` indique soit le nombre de classes souhaité, soit, si on lui fournit un vecteur, les limites des classes ;
- `labels` permet de modifier les noms de modalités attribués aux classes ;
- `include.lowest` et `right` influent sur la manière dont les valeurs situées à la frontière des classes seront inclues ou exclues ;
- `dig.lab` indique le nombre de chiffres après la virgule à conserver dans les noms de modalités.

Prenons tout de suite un exemple et tentons de découper notre variable `age` en cinq classes et de placer le résultat dans une nouvelle variable nommée `age5cl` :

```
R> d$age5cl <- cut(d$age, 5)
R> table(d$age5cl)


(17.9,33.8] (33.8,49.6] (49.6,65.4] (65.4,81.2] (81.2,97.1]
        454         628         556         319          43
```

Par défaut R nous a bien créé cinq classes d'amplitudes égales. La première classe va de 16,9 à 32,2 ans (en fait de 17 à 32), etc.

Les frontières de classe seraient plus présentables si elles utilisaient des nombres entiers. On va donc spécifier manuellement le découpage souhaité, par tranches de 20 ans :

```
R> d$age20 <- cut(d$age, c(0, 20, 40, 60, 80, 100))
R> table(d$age20)


  (0,20]  (20,40]  (40,60]  (60,80] (80,100]
      72      660      780      436       52
```

On aurait pu tenir compte des âges extrêmes pour la première et la dernière valeur :

```
R> range(d$age)

[1] 18 97

R> d$age20 <- cut(d$age, c(17, 20, 40, 60, 80, 93))
R> table(d$age20)


(17,20] (20,40] (40,60] (60,80] (80,93]
     72     660     780     436      50
```

Les symboles dans les noms attribués aux classes ont leur importance : `(` signifie que la frontière de la classe est exclue, tandis que `[` signifie qu'elle est incluse. Ainsi, `(20,40]` signifie « strictement supérieur à 20 et inférieur ou égal à 40 ».

On remarque que du coup, dans notre exemple précédent, la valeur minimale, 17, est exclue de notre première classe, et qu'une observation est donc absente de ce découpage. Pour résoudre ce problème on peut soit faire commencer la première classe à 16, soit utiliser l'option `include.lowest=TRUE` :

```
R> d$age20 <- cut(d$age, c(16, 20, 40, 60, 80, 93))
R> table(d$age20)


(16,20] (20,40] (40,60] (60,80] (80,93]
     72     660     780     436      50

R> d$age20 <- cut(d$age, c(17, 20, 40, 60, 80, 93), include.lowest = TRUE)
R> table(d$age20)


[17,20] (20,40] (40,60] (60,80] (80,93]
     72     660     780     436      50
```

On peut également modifier le sens des intervalles avec l'option `right=FALSE`, et indiquer manuellement les noms des modalités avec `labels` :

```
R> d$age20 <- cut(d$age, c(16, 20, 40, 60, 80, 93), right = FALSE, include.lowest = TRUE)
R> table(d$age20)
```

```
[16,20) [20,40) [40,60) [60,80) [80,93]
     48     643     793     454      60

R> d$age20 <- cut(d$age, c(17, 20, 40, 60, 80, 93), include.lowest = TRUE, labels = c("<20ans",
+     "21-40 ans", "41-60ans", "61-80ans", ">80ans"))
R> table(d$age20)


   <20ans 21-40 ans  41-60ans  61-80ans    >80ans
       72       660       780       436        50
```

Interface interactive

questionr propose une interface interactive à la fonction `cut`, nommée `icut`. Elle s'utilise soit dans RStudio via le menu *Addins*, puis *Variable cutting*, soit depuis la console ou un script de la manière suivante : *questionr*

```
R> icut(d$age)
```

R devrait alors afficher une interface semblable à celle de la figure 5.2, page suivante.

Vous pouvez alors choisir la variable à découper dans l'onglet *Variable et paramètres*, indiquer les limites de vos classes ainsi que quelques options complémentaires dans l'onglet *Découpage en classes*, et vérifier le résultat dans l'onglet *Code et résultat*. Une fois le résultat satisfaisant, cliquez sur *Done* : si vous êtes sous RStudio le code généré sera directement inséré dans votre script actuel à l'emplacement du curseur. Sinon, ce code sera affiché dans la console et vous pourrez le copier/coller pour l'inclure dans votre script [3].

quant.cut

Enfin, l'extension questionr propose une fonction `quant.cut` permettant de découper une variable numérique en un nombre de classes donné ayant des efffectifs semblables. Il suffit de lui passer le nombre de classes en argument : *questionr*

```
R> d$age6cl <- quant.cut(d$age, 6)
R> table(d$age6cl)


    [18,30)     [30,39)     [39,48) [48,55.667) [55.667,66)     [66,97]
        302         337         350         344         305         362
```

`quant.cut` admet les mêmes autres options que `cut` (`include.lowest, right, labels...`).

5.4.3 Regrouper les modalités d'une variable

Pour regrouper les modalités d'une variable qualitative (d'un facteur le plus souvent), on peut utiliser directement l'indexation.

Ainsi, si on veut recoder la variable `qualif` dans une variable `qualif.reg` plus « compacte », on peut utiliser :

3. Le fonctionnement des interfaces interactives de questionr est décrit plus en détail section B.3.3, page 125.

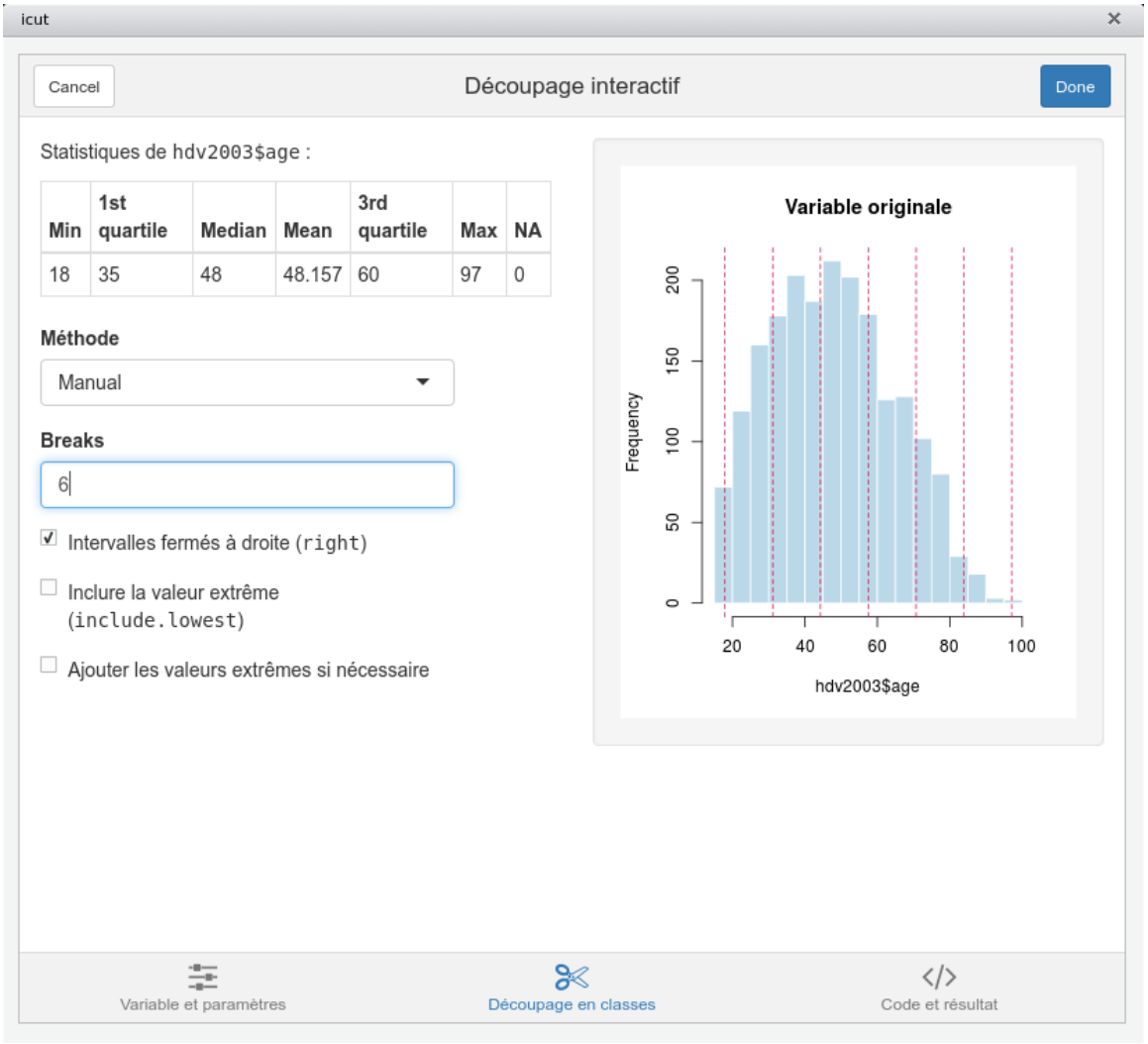


FIGURE 5.2 – Interface de la commande `icut`

```
R> table(d$qualif)


      Ouvrier specialise         Ouvrier qualifie                Technicien
                     203                      292                        86
Profession intermediaire                   Cadre                   Employe
                     160                      260                       594
                  Autre
                      58

R> d$qualif.reg[d$qualif == "Ouvrier specialise"] <- "Ouvrier"
R> d$qualif.reg[d$qualif == "Ouvrier qualifie"] <- "Ouvrier"
R> d$qualif.reg[d$qualif == "Employe"] <- "Employe"
R> d$qualif.reg[d$qualif == "Profession intermediaire"] <- "Intermediaire"
R> d$qualif.reg[d$qualif == "Technicien"] <- "Intermediaire"
R> d$qualif.reg[d$qualif == "Cadre"] <- "Cadre"
R> d$qualif.reg[d$qualif == "Autre"] <- "Autre"
R> table(d$qualif.reg)


        Autre         Cadre       Employe Intermediaire       Ouvrier
           58           260           594           246           495
```

On aurait pu représenter ce recodage de manière plus compacte, notamment en commençant par copier le contenu de `qualif` dans `qualif.reg`, ce qui permet de ne pas s'occuper de ce qui ne change pas. Il est cependant nécessaire de ne pas copier `qualif` sous forme de facteur, sinon on ne pourrait ajouter de nouvelles modalités. On copie donc la version *caractères* de `qualif` grâce à la fonction `as.character` :

```
R> d$qualif.reg <- as.character(d$qualif)
R> d$qualif.reg[d$qualif == "Ouvrier specialise"] <- "Ouvrier"
R> d$qualif.reg[d$qualif == "Ouvrier qualifie"] <- "Ouvrier"
R> d$qualif.reg[d$qualif == "Profession intermediaire"] <- "Intermediaire"
R> d$qualif.reg[d$qualif == "Technicien"] <- "Intermediaire"
R> table(d$qualif.reg)


        Autre         Cadre       Employe Intermediaire       Ouvrier
           58           260           594           246           495
```

On peut faire une version encore plus compacte en utilisant l'opérateur logique *ou* (`|`) :

```
R> d$qualif.reg <- as.character(d$qualif)
R> d$qualif.reg[d$qualif == "Ouvrier specialise" | d$qualif == "Ouvrier qualifie"] <- "Ouvrier"
R> d$qualif.reg[d$qualif == "Profession intermediaire" | d$qualif == "Technicien"] <- "Intermediaire"
R> table(d$qualif.reg)


        Autre         Cadre       Employe Intermediaire       Ouvrier
           58           260           594           246           495
```

Enfin, pour terminer ce petit tour d'horizon, on peut également remplacer l'opérateur `|` par `%in%`, qui peut parfois être plus lisible :

```
R> d$qualif.reg <- as.character(d$qualif)
R> d$qualif.reg[d$qualif %in% c("Ouvrier specialise", "Ouvrier qualifie")] <- "Ouvrier"
R> d$qualif.reg[d$qualif %in% c("Profession intermediaire", "Technicien")] <- "Intermediaire"
R> table(d$qualif.reg)


        Autre         Cadre       Employe Intermediaire       Ouvrier
           58           260           594           246           495
```

Dans tous les cas le résultat obtenu est une variable de type *caractère*. On pourra la convertir en *facteur* par un simple :

```
R> d$qualif.reg <- factor(d$qualif.reg)
```

Si on souhaite recoder les valeurs manquantes, il suffit de faire appel à la fonction `is.na` :

```
R> table(d$trav.satisf)


  Satisfaction Insatisfaction      Equilibre
           480            117            451

R> d$trav.satisf.reg <- as.character(d$trav.satisf)
R> d$trav.satisf.reg[is.na(d$trav.satisf)] <- "Valeur manquante"
R> table(d$trav.satisf.reg)


       Equilibre   Insatisfaction     Satisfaction Valeur manquante
             451              117              480              952
```

Interface interactive

questionr

questionr propose une interface interactive pour le recodage d'une variable qualitative (renommage et regroupement de modalités). Cette fonction, nommée `irec`, s'utilise depuis RStudio via le menu *Addins*, puis *Levels recoding*, soit depuis la console de la manière suivante :

```
R> irec(d$qualif)
```

Dans les deux cas, une interface semblable à celle de la figure B.1, page 126 devrait s'afficher soit directement dans RStudio, soit dans votre navigateur Web.

Vous pouvez alors sélectionner différentes options, et pour chaque ancienne modalité, indiquer la nouvelle valeur correspondante. Pour regrouper des modalités, il suffit de leur assigner des nouvelles valeurs identiques. Dans tous les cas n'hésitez pas à expérimenter, l'interface se contente de générer du code R mais ne l'exécute pas, et ne modifie donc jamais vos données ![4]

5.4.4 Variables calculées

La création d'une variable numérique à partir de calculs sur une ou plusieurs autres variables numériques se fait très simplement.

4. Le fonctionnement des interfaces interactives de questionr est décrit plus en détail section B.3.3, page 125.

Supposons que l'on souhaite calculer une variable indiquant l'écart entre le nombre d'heures passées à regarder la télévision et la moyenne globale de cette variable. On pourrait alors faire :

```
R> range(d$heures.tv, na.rm = TRUE)

[1]  0 12

R> mean(d$heures.tv, na.rm = TRUE)

[1] 2.246566

R> d$ecart.heures.tv <- d$heures.tv - mean(d$heures.tv, na.rm = TRUE)
R> range(d$ecart.heures.tv, na.rm = TRUE)

[1] -2.246566  9.753434

R> mean(d$ecart.heures.tv, na.rm = TRUE)

[1] 4.714578e-17
```

Autre exemple tiré du jeu de données `rp99` : si on souhaite calculer le pourcentage d'actifs dans chaque commune, on peut diviser la population active `pop.act` par la population totale `pop.tot`.

```
R> rp99$part.actifs <- rp99$pop.act/rp99$pop.tot * 100
```

5.4.5 Combiner plusieurs variables

La combinaison de plusieurs variables se fait à l'aide des techniques d'indexation déjà décrites précédemment. Le plus compliqué est d'arriver à formuler des conditions parfois complexes de manière rigoureuse.

On peut ainsi vouloir combiner plusieurs variables qualitatives en une seule :

```
R> d$act.manuelles <- NA
R> d$act.manuelles[d$cuisine == "Oui" & d$bricol == "Oui"] <- "Cuisine et Bricolage"
R> d$act.manuelles[d$cuisine == "Oui" & d$bricol == "Non"] <- "Cuisine seulement"
R> d$act.manuelles[d$cuisine == "Non" & d$bricol == "Oui"] <- "Bricolage seulement"
R> d$act.manuelles[d$cuisine == "Non" & d$bricol == "Non"] <- "Ni cuisine ni bricolage"
R> table(d$act.manuelles)


    Bricolage seulement    Cuisine et Bricolage       Cuisine seulement
                    437                     416                     465
Ni cuisine ni bricolage
                    682
```

On peut également combiner variables qualitatives et variables quantitatives :

```
R> d$age.sexe <- NA
R> d$age.sexe[d$sexe == "Homme" & d$age < 40] <- "Homme moins de 40 ans"
R> d$age.sexe[d$sexe == "Homme" & d$age >= 40] <- "Homme plus de 40 ans"
```

```
R> d$age.sexe[d$sexe == "Femme" & d$age < 40] <- "Femme moins de 40 ans"
R> d$age.sexe[d$sexe == "Femme" & d$age >= 40] <- "Femme plus de 40 ans"
R> table(d$age.sexe)


Femme moins de 40 ans  Femme plus de 40 ans Homme moins de 40 ans
                  376                   725                   315
 Homme plus de 40 ans
                  584
```

Les combinaisons de variables un peu complexes nécessitent parfois un petit travail de réflexion. En particulier, l'ordre des commandes de recodage a parfois une influence dans le résultat final.

5.4.6 Variables scores

Une variable score est une variable calculée en additionnant des poids accordés aux modalités d'une série de variables qualitatives.

Pour prendre un exemple tout à fait arbitraire, imaginons que nous souhaitons calculer un score d'activités extérieures. Dans ce score on considère que le fait d'aller au cinéma « pèse » 10, celui de pêcher ou chasser vaut 30 et celui de faire du sport vaut 20. On pourrait alors calculer notre score de la manière suivante :

```
R> d$score.ext <- 0
R> d$score.ext[d$cinema == "Oui"] <- d$score.ext[d$cinema == "Oui"] + 10
R> d$score.ext[d$peche.chasse == "Oui"] <- d$score.ext[d$peche.chasse == "Oui"] +
+     30
R> d$score.ext[d$sport == "Oui"] <- d$score.ext[d$sport == "Oui"] + 20
R> table(d$score.ext)


  0  10  20  30  40  50  60
800 342 229 509  31  41  48
```

Cette notation étant un peu lourde, on peut l'alléger un peu en utilisant la fonction `ifelse`. Celle-ci prend en argument une condition et deux valeurs. Si la condition est vraie elle retourne la première valeur, sinon elle retourne la seconde.

```
R> d$score.ext <- 0
R> d$score.ext <- ifelse(d$cinema == "Oui", 10, 0) + ifelse(d$peche.chasse == "Oui",
+     30, 0) + ifelse(d$sport == "Oui", 20, 0)
R> table(d$score.ext)


  0  10  20  30  40  50  60
800 342 229 509  31  41  48
```

5.4.7 Vérification des recodages

Il est très important de vérifier, notamment après les recodages les plus complexes, qu'on a bien obtenu le résultat escompté. Les deux points les plus sensibles étant les valeurs manquantes et les erreurs dans les conditions.

Pour vérifier tout cela le plus simple est sans doute de faire des tableaux croisés entre la variable recodée et celles ayant servi au recodage, à l'aide de la fonction `table`, et de vérifier le nombre de valeurs manquantes dans la variable recodée avec `summary`, `freq` ou `table`.

Par exemple :

```
R> d$act.manuelles <- NA
R> d$act.manuelles[d$cuisine == "Oui" & d$bricol == "Oui"] <- "Cuisine et Bricolage"
R> d$act.manuelles[d$cuisine == "Oui" & d$bricol == "Non"] <- "Cuisine seulement"
R> d$act.manuelles[d$cuisine == "Non" & d$bricol == "Oui"] <- "Bricolage seulement"
R> d$act.manuelles[d$cuisine == "Non" & d$bricol == "Non"] <- "Ni cuisine ni bricolage"
R> table(d$act.manuelles, d$cuisine)


                         Non Oui
  Bricolage seulement     437   0
  Cuisine et Bricolage      0 416
  Cuisine seulement         0 465
  Ni cuisine ni bricolage 682   0

R> table(d$act.manuelles, d$bricol)


                         Non Oui
  Bricolage seulement       0 437
  Cuisine et Bricolage      0 416
  Cuisine seulement       465   0
  Ni cuisine ni bricolage 682   0
```

5.5 Tri de tables

On a déjà évoqué l'existence de la fonction `sort`, qui permet de trier les éléments d'un vecteur.

```
R> sort(c(2, 5, 6, 1, 8))

[1] 1 2 5 6 8
```

On peut appliquer cette fonction à une variable, mais celle-ci ne permet que d'ordonner les valeurs de cette variable, et pas l'ensemble du tableau de données dont elle fait partie. Pour cela nous avons besoin d'une autre fonction, nommée `order`. Celle-ci ne renvoie pas les valeurs du vecteur triées, mais les emplacements de ces valeurs.

Un exemple pour comprendre :

```
R> order(c(15, 20, 10))

[1] 3 1 2
```

Le résultat renvoyé signifie que la plus petite valeur est la valeur située en 3ème position, suivie de celle en 1ère position et de celle en 2ème position. Tout cela ne paraît pas passionnant à première vue, mais si on mélange ce résultat avec un peu d'indexation directe, ça devient intéressant...

```
R> order(d$age)

 [1]  162  215  346  377  511  646  852  916 1211 1213 1261 1333 1395 1447
[15] 1600 1774 1937   38  100  134  196  204  256  257  349  395  407  427
[29]  453  578  726  969 1052 1056 1077 1177 1234 1250 1342 1377 1381 1382
[43] 1540 1559 1607 1634 1689 1983    9   18   25  231  335  347  358  488
[57]  496  642  826  922 1023 1042 1156 1175 1290 1384 1464 1467 1608 1661
 [ reached getOption("max.print") -- omitted 1930 entries ]
```

Ce que cette fonction renvoie, c'est l'ordre dans lequel on doit placer les éléments de `age`, et donc par extension les lignes de `d`, pour que la variable soit triée par ordre croissant. Par conséquent, si on fait :

```
R> d.tri <- d[order(d$age), ]
```

Alors on a trié les lignes de `d` par ordre d'âge croissant ! Et si on fait un petit :

```
R> head(d.tri, 3)

     id age  sexe nivetud    poids           occup qualif freres.soeurs
162 162  18 Homme    <NA> 4982.964 Etudiant, eleve   <NA>             2
215 215  18 Homme    <NA> 4631.188 Etudiant, eleve   <NA>             2
    clso                       relig trav.imp trav.satisf hard.rock
162  Non  Appartenance sans pratique     <NA>        <NA>       Non
215  Oui Ni croyance ni appartenance     <NA>        <NA>       Non
    lecture.bd peche.chasse cuisine bricol cinema sport heures.tv fs.fac
162        Non          Non     Non    Non    Non   Oui         3      2
215        Non          Non     Oui    Non    Oui   Oui         2      2
    fs.char qualif.char fs.num      age5cl  age20  age6cl qualif.reg
162       2        <NA>      2 (17.9,33.8] <20ans [18,30)       <NA>
215       2        <NA>      2 (17.9,33.8] <20ans [18,30)       <NA>
     trav.satisf.reg ecart.heures.tv           act.manuelles
162 Valeur manquante       0.7534336 Ni cuisine ni bricolage
215 Valeur manquante      -0.2465664       Cuisine seulement
                age.sexe score.ext
162 Homme moins de 40 ans        20
215 Homme moins de 40 ans        30
 [ getOption("max.print") est atteint -- ligne 1 omise ]
```

On a les caractéristiques des trois enquêtés les plus jeunes.

On peut évidemment trier par ordre décroissant en utilisant l'option `decreasing=TRUE`. On peut donc afficher les caractéristiques des trois individus les plus âgés avec :

```
R> head(d[order(d$age, decreasing = TRUE), ], 3)
```

5.6 Fusion de tables

Lorsqu'on traite de grosses enquêtes, notamment les enquêtes de l'INSEE, on a souvent à gérer des données réparties dans plusieurs tables, soit du fait de la construction du questionnaire, soit du fait de contraintes techniques (fichiers `dbf` ou Excel limités à 256 colonnes, par exemple).

Une opération relativement courante consiste à *fusionner* plusieurs tables pour regrouper tout ou partie des données dans un unique tableau.

Nous allons simuler artificiellement une telle situation en créant deux tables à partir de l'extrait de l'enquête *Histoire de vie* :

```
R> data(hdv2003)
R> d <- hdv2003
R> dim(d)

[1] 2000   20

R> d1 <- subset(d, select = c("id", "age", "sexe"))
R> dim(d1)

[1] 2000    3

R> d2 <- subset(d, select = c("id", "clso"))
R> dim(d2)

[1] 2000    2
```

On a donc deux tableaux de données, `d1` et `d2`, comportant chacun 2000 lignes et respectivement 3 et 2 colonnes. Comment les rassembler pour n'en former qu'un ?

Intuitivement, cela paraît simple. Il suffit de « coller » `d2` à la droite de `d1`, comme dans l'exemple suivant.

Id	V1	V2
1	H	12
2	H	17
3	F	41
4	F	9
⋮	⋮	⋮

$+$

Id	V3
1	Rouge
2	Bleu
3	Bleu
4	Rouge
⋮	⋮

$=$

Id	V1	V2	V3
1	H	12	Rouge
2	H	17	Bleu
3	F	41	Bleu
4	F	9	Rouge
⋮	⋮	⋮	⋮

Cela semble fonctionner. La fonction qui permet d'effectuer cette opération sous R s'appelle `cbind`, elle « colle » des tableaux côte à côte en regroupant leurs colonnes [5].

```
R> cbind(d1, d2)

       id age  sexe   id        clso
1       1  28 Femme    1         Oui
2       2  23 Femme    2         Oui
3       3  59 Homme    3         Non
4       4  34 Homme    4         Non
5       5  71 Femme    5         Oui
6       6  35 Femme    6         Non
7       7  60 Femme    7         Oui
8       8  47 Homme    8         Non
9       9  20 Femme    9         Oui
10     10  28 Homme   10         Non
```

5. L'équivalent de `cbind` pour les lignes s'appelle `rbind`.

```
11     11  65 Femme   11         Oui
12     12  47 Homme   12         Oui
13     13  63 Femme   13         Oui
14     14  67 Femme   14         Oui
 [ getOption("max.print") est atteint -- 1986 lignes omises ]
```

À part le fait qu'on a une colonne `id` en double, le résultat semble satisfaisant. À première vue seulement. Imaginons maintenant que nous avons travaillé sur `d1` et `d2`, et que nous avons ordonné les lignes de `d1` selon l'âge des enquêtés :

```
R> d1 <- d1[order(d1$age), ]
```

Répétons l'opération de collage :

```
R> cbind(d1, d2)

       id age  sexe   id        clso
162   162  18 Homme    1         Oui
215   215  18 Homme    2         Oui
346   346  18 Femme    3         Non
377   377  18 Homme    4         Non
511   511  18 Homme    5         Oui
646   646  18 Homme    6         Non
852   852  18 Femme    7         Oui
916   916  18 Femme    8         Non
1211 1211  18 Homme    9         Oui
1213 1213  18 Femme   10         Non
1261 1261  18 Homme   11         Oui
1333 1333  18 Femme   12         Oui
1395 1395  18 Homme   13         Oui
1447 1447  18 Femme   14         Oui
 [ getOption("max.print") est atteint -- 1986 lignes omises ]
```

Que constate-t-on ? La présence de la variable `id` en double nous permet de voir que les identifiants ne coïncident plus ! En regroupant nos colonnes nous avons donc attribué à des individus les réponses d'autres individus.

La commande `cbind` ne peut en effet fonctionner que si les deux tableaux ont exactement le même nombre de lignes, et dans le même ordre, ce qui n'est pas le cas ici.

On va donc être obligé de pocéder à une *fusion* des deux tableaux, qui va permettre de rendre à chaque ligne ce qui lui appartient. Pour cela nous avons besoin d'un identifiant qui permet d'identifier chaque ligne de manière unique et qui doit être présent dans tous les tableaux. Dans notre cas, c'est plutôt rapide, il s'agit de la variable `id`.

Une fois l'identifiant identifié[6], on peut utiliser la commande `merge`. Celle-ci va fusionner les deux tableaux en supprimant les colonnes en double et en regroupant les lignes selon leurs identifiants :

```
R> d.complet <- merge(d1, d2, by = "id")
R> head(d.complet)

  id age  sexe clso
```

6. Si vous me passez l'expression...

```
1  1  28 Femme  Oui
2  2  23 Femme  Oui
3  3  59 Homme  Non
4  4  34 Homme  Non
5  5  71 Femme  Oui
6  6  35 Femme  Non
```

Ici l'utilisation de la fonction est plutôt simple car nous sommes dans le cas de figure idéal : les lignes correspondent parfaitement et l'identifiant est clairement identifié. Parfois les choses peuvent être un peu plus compliquées :

— parfois les identifiants n'ont pas le même nom dans les deux tableaux. On peut alors les spécifier par les options `by.x` et `by.y` ;
— parfois les deux tableaux comportent des colonnes (hors identifiants) ayant le même nom. `merge` conserve dans ce cas ces deux colonnes mais les renomme en les suffixant par `.x` pour celles provenant du premier tableau, et `.y` pour celles du second ;
— parfois on n'a pas d'identifiant unique préétabli, mais on en construit un à partir de plusieurs variables. On peut alors donner un vecteur en paramètres de l'option `by`, par exemple `by=c("nom","prenom","date.naissance")`.

Une subtilité supplémentaire intervient lorsque les deux tableaux fusionnés n'ont pas exactement les mêmes lignes. Par défaut, `merge` ne conserve que les lignes présentes dans les deux tableaux :

Id	V1
1	H
2	H
3	F

+

Id	V2
1	10
2	15
5	31

=

Id	V1	V2
1	H	10
2	H	15

On peut cependant modifier ce comportement avec les options `all.x=TRUE` et `all.y=TRUE`. La première option indique de conserver toutes les lignes du premier tableau. Dans ce cas `merge` donne une valeur `NA` pour ces lignes aux colonnes provenant du second tableau. Ce qui donnerait :

Id	V1
1	H
2	H
3	F

+

Id	V2
1	10
2	15
5	31

=

Id	V1	V2
1	H	10
2	H	15
3	F	NA

`all.y` fait la même chose en conservant toutes les lignes du second tableau. On peut enfin décider toutes les lignes des deux tableaux en utilisant à la fois `all.x=TRUE` et `all.y=TRUE`, ce qui donne :

Id	V1
1	H
2	H
3	F

+

Id	V2
1	10
2	15
5	31

=

Id	V1	V2
1	H	10
2	H	15
3	F	NA
5	NA	31

Parfois, l'un des identifiants est présent à plusieurs reprises dans l'un des tableaux (par exemple lorsque l'une des tables est un ensemble de ménages et que l'autre décrit l'ensemble des individus de ces ménages). Dans ce cas les lignes de l'autre table sont dupliquées autant de fois que nécessaires :

Id	V1
1	H
2	H
3	F

+

Id	V2
1	10
1	18
1	21
2	11
3	31

=

Id	V1	V2
1	H	10
1	H	18
1	H	21
2	H	11
3	F	31

5.7 Organiser ses scripts

Il ne s'agit pas ici de manipulation de données à proprement parler, mais plutôt d'une conséquence de ce qui a été vu précédemment : à mesure que recodages et traitements divers s'accumulent, votre script R risque de devenir rapidement très long et pas très pratique à éditer.

Il est très courant de répartir son travail entre différents fichiers, ce qui est rendu très simple par la fonction `source`. Celle-ci permet de lire le contenu d'un fichier de script et d'exécuter son contenu.

Prenons tout de suite un exemple. La plupart des scripts R commencent par charger les extensions utiles, importer les données, effectuer manipulations, traitements et recodages, puis à mettre en oeuvre les analyses. Prenons le fichier fictif suivant :

```
library(questionr)
library(foreign)

## IMPORT DES DONNÉES

d1 <- read.dbf("data/tab1.dbf")
d2 <- read.dbf("data/tab2.dbf")

d <- merge(d1, d2, by = "id")

## RECODAGES

d$tx.chomage <- as.numeric(d$tx.chomage)

d$pcs[d$pcs == "Ouvrier qualifie"] <- "Ouvrier"
d$pcs[d$pcs == "Ouvrier specialise"] <- "Ouvrier"

d$age5cl <- cut(d$age, 5)

## ANALYSES

tab <- table(d$tx.chomage, d$age5cl)
tab
chisq.test(tab)
```

Une manière d'organiser notre script [7] pourrait être de placer les opérations d'import des données et celles de recodage dans deux fichiers scripts séparés. Créons alors un fichier nommé `import.R` dans notre répertoire de travail et copions les lignes suivantes :

```
## IMPORT DES DONNÉES

d1 <- read.dbf("data/tab1.dbf")
d2 <- read.dbf("data/tab2.dbf")

d <- merge(d1, d2, by = "id")
```

Créons également un fichier `recodages.R` avec le contenu suivant :

7. Ceci n'est qu'une suggestion, la manière d'organiser (ou non) son travail étant bien évidemment très hautement subjective.

```
## RECODAGES

d$tx.chomage <- as.numeric(d$tx.chomage)

d$pcs[d$pcs == "Ouvrier qualifie"] <- "Ouvrier"
d$pcs[d$pcs == "Ouvrier specialise"] <- "Ouvrier"

d$age5cl <- cut(d$age, 5)
```

Dès lors, si nous rajoutons les appels à la fonction `source` qui vont bien, le fichier suivant sera strictement équivalent à notre fichier de départ :

```
library(questionr)
library(foreign)

source("import.R")
source("recodages.R")

## ANALYSES

tab <- table(d$tx.chomage, d$age5cl)
tab
chisq.test(tab)
```

Au fur et à mesure du travail sur les données, on placera les recodages que l'on souhaite conserver dans le fichier `recodages.R`.

Cette méthode présente plusieurs avantages :

— bien souvent, lorsqu'on effectue des recodages on se retrouve avec des variables recodées qu'on ne souhaite pas conserver. Si on prend l'habitude de placer les recodages intéressants dans le fichier `recodages.R`, alors il suffit d'exécuter les cinq premières lignes du fichier pour se retrouver avec un tableau de données `d` propre et complet.

— on peut répartir ses analyses dans différents scripts. Il suffit alors de copier les cinq premières lignes du fichier précédent dans chacun des scripts, et on aura l'assurance de travailler sur exactement les mêmes données.

Le premier point illustre l'une des caractéristiques de R : il est rare que l'on stocke les données modifiées. En général on repart toujours du fichier source original, et les recodages sont conservés sous forme de scripts et recalculés à chaque fois qu'on recommence à travailler. Ceci offre une traçabilité parfaite du traitement effectué sur les données.

5.8 Exercices

Exercice 5.11

▷ *Solution page 131*

Renommer la variable `clso` du jeu de données `hdv2003` en `classes.sociales`, puis la renommer en `clso`.

Exercice 5.12

▷ *Solution page 132*

Réordonner les niveaux du facteur `clso` pour que son tri à plat s'affiche de la manière suivante :

```
tmp
       Non Ne sait pas         Oui
      1037          27         936
```

Exercice 5.13

▷ *Solution page 132*

Affichez :
- les 3 premiers éléments de la variable `cinema`
- les éléments 12 à 30 de la variable `lecture.bd`
- les colonnes 4 et 8 des lignes 5 et 12 du jeu de données `hdv2003`
- les 4 derniers éléments de la variable `age`

Exercice 5.14

▷ *Solution page 132*

Construisez les sous-tableaux suivants avec la fonction `subset` :
- âge et sexe des lecteurs de BD
- ensemble des personnes n'étant pas chômeur (variable `occup`), sans la variable `cinema`
- identifiants des personnes de plus de 45 ans écoutant du hard rock
- femmes entre 25 et 40 ans n'ayant pas fait de sport dans les douze derniers mois
- hommes ayant entre 2 et 4 frères et sœurs et faisant la cuisine ou du bricolage

Exercice 5.15

▷ *Solution page 133*

Calculez le nombre moyen d'heures passées devant la télévision chez les lecteurs de BD, d'abord en construisant les sous-populations, puis avec la fonction `tapply`.

Exercice 5.16

▷ *Solution page 133*

Convertissez la variable `freres.soeurs` en variable de type caractères. Convertissez cette nouvelle variable en facteur. Puis convertissez à nouveau ce facteur en variable numérique. Vérifiez que votre variable finale est identique à la variable de départ.

Exercice 5.17

▷ *Solution page 133*

Découpez la variable `freres.soeurs` :
- en cinq classes d'amplitude égale
- en catégories « de 0 à 2 », « de 2 à 4 », « plus de 4 », avec les étiquettes correspondantes
- en quatre classes d'effectif équivalent
- d'où vient la différence d'effectifs entre les deux découpages précédents ?

Exercice 5.18

▷ *Solution page 134*

Recodez la variable `trav.imp` en `trav.imp2cl` pour obtenir les modalités « Le plus ou aussi important » et « moins ou peu important ». Vérifiez avec des tris à plat et un tableau croisé.

Recodez la variable `relig` en `relig.4cl` en regroupant les modalités « Pratiquant regulier » et « Pratiquant occasionnel » en une seule modalité « Pratiquant », et en remplaçant la modalité « NSP ou NVPR » par des valeurs manquantes. Vérifiez avec un tri croisé.

Exercice 5.19

▷ *Solution page 135*

Créez une variable ayant les modalités suivantes :
- Homme de plus de 40 ans lecteur de BD
- Homme de plus de 30 ans
- Femme faisant du bricolage
- Autre

Vérifier avec des tris croisés.

Exercice 5.20

▷ *Solution page 136*

Ordonner le tableau de données selon le nombre de frères et soeurs croissant. Afficher le sexe des 10 individus regardant le plus la télévision.

Partie 6

Statistique bivariée

On entend par statistique bivariée l'étude des relations entre deux variables, celles-ci pouvant être quantitatives ou qualitatives.

questionr

Comme dans la partie précédente, on travaillera sur les jeux de données fournis avec l'extension questionr et tiré de l'enquête *Histoire de vie* et du recensement 1999 :

```
R> data(hdv2003)
R> d <- hdv2003
R> data(rp99)
```

6.1 Deux variables quantitatives

La comparaison de deux variables quantitatives se fait en premier lieu graphiquement, en représentant l'ensemble des couples de valeurs. On peut ainsi représenter les valeurs du nombre d'heures passées devant la télévision selon l'âge (figure 6.1 page ci-contre).

Le fait que des points sont superposés ne facilite pas la lecture du graphique. On peut utiliser une représentation avec des points semi-transparents (figure 6.2 page 86).

Plus sophistiqué, on peut faire une estimation locale de densité et représenter le résultat sous forme de « carte ». Pour cela on commence par isoler les deux variables, supprimer les observations ayant au moins une valeur manquante à l'aide de la fonction `complete.cases`, estimer la densité locale à l'aide de la fonction `kde2d` de l'extension MASS [1] et représenter le tout à l'aide d'une des fonctions `image`, `contour` ou `filled.contour`... Le résultat est donné figure 6.3 page 87.

Dans tous les cas, il n'y a pas de structure très nette qui semble se dégager. On peut tester ceci mathématiquement en calculant le cœfficient de corrélation entre les deux variables à l'aide de la fonction `cor` :

```
R> cor(d$age, d$heures.tv, use = "complete.obs")

[1] 0.1776249
```

L'option `use` permet d'éliminer les observations pour lesquelles l'une des deux valeurs est manquante. Le cœfficient de corrélation est très faible.

1. MASS est installée par défaut avec la version de base de R.

```
R> plot(d$age, d$heures.tv)
```

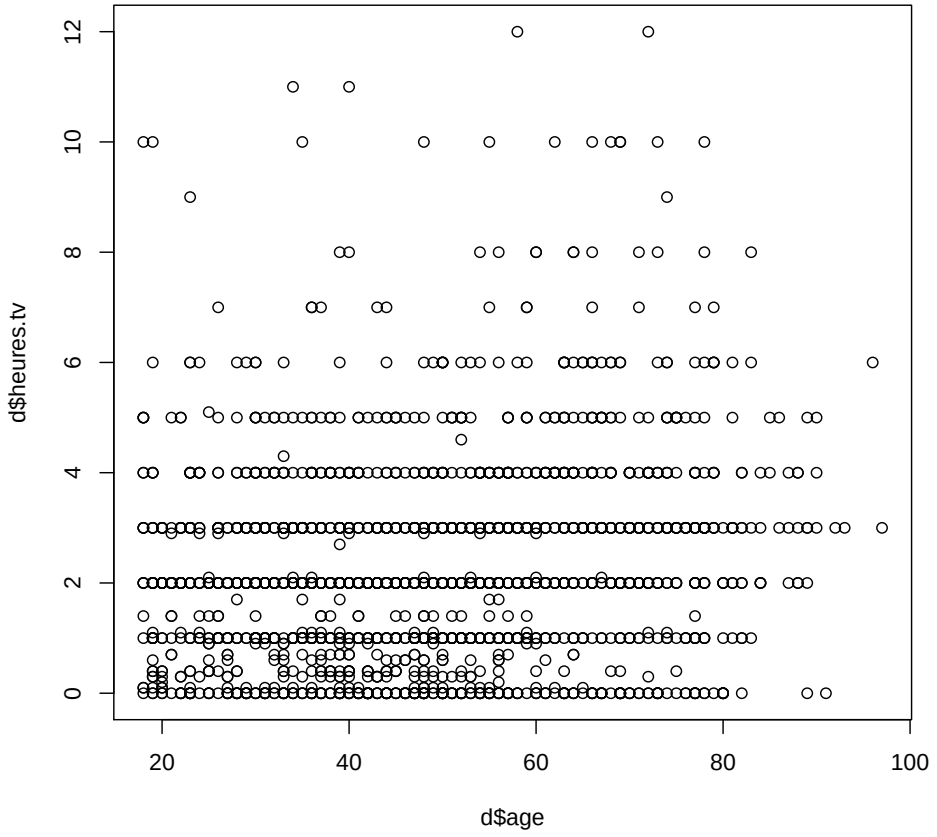


FIGURE 6.1 – Nombre d'heures de télévision selon l'âge

```
R> plot(d$age, d$heures.tv, pch = 19, col = rgb(1, 0, 0, 0.1))
```

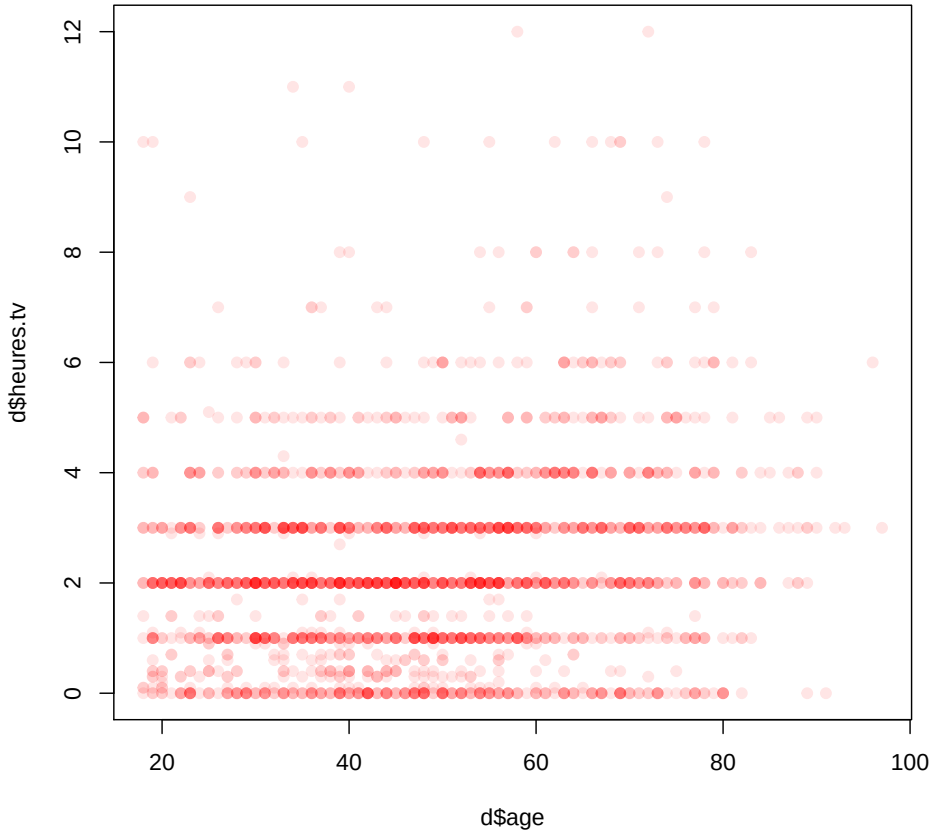


FIGURE 6.2 – Nombre d'heures de télévision selon l'âge avec semi-transparence

```
R> library(MASS)
R> tmp <- d[, c("age", "heures.tv")]
R> tmp <- tmp[complete.cases(tmp), ]
R> filled.contour(kde2d(tmp$age, tmp$heures.tv), color = terrain.colors)
```

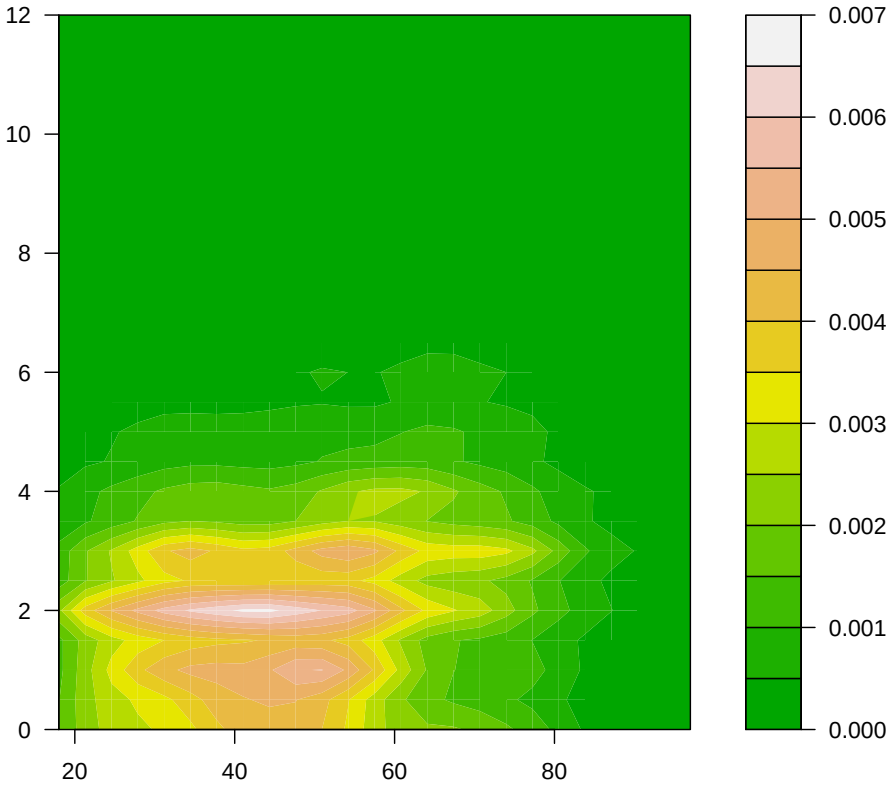


FIGURE 6.3 – Représentation de l'estimation de densité locale

```
R> plot(rp99$dipl.sup, rp99$cadres, ylab = "Part des cadres", xlab = "Part des diplomês du supérieur"
```

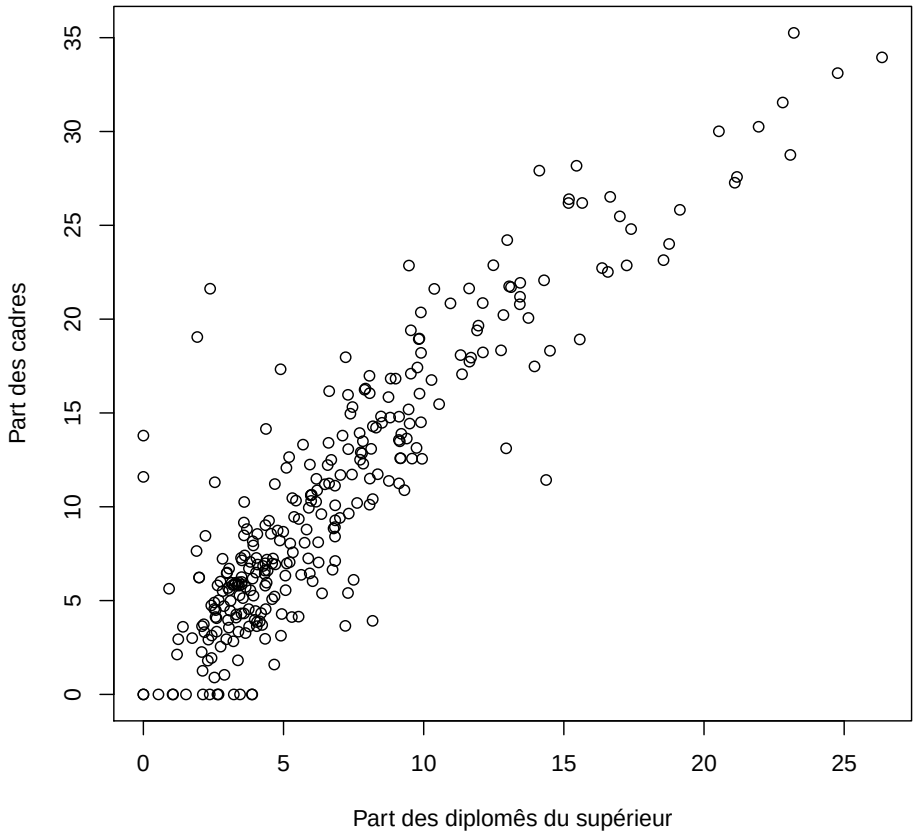


FIGURE 6.4 – Proportion de cadres et proportion de diplômés du supérieur

On va donc s'intéresser plutôt à deux variables présentes dans le jeu de données `rp99`, la part de diplômés du supérieur et la proportion de cadres dans les communes du Rhône en 1999.

À nouveau, commençons par représenter les deux variables (figure 6.4 de la présente page). Ça ressemble déjà beaucoup plus à une relation de type linéaire.

Calculons le coefficient de corrélation :

```
R> cor(rp99$dipl.sup, rp99$cadres)

[1] 0.8975282
```

C'est beaucoup plus proche de 1. On peut alors effectuer une régression linéaire complète en utilisant la fonction `lm` :

```
R> reg <- lm(cadres ~ dipl.sup, data = rp99)
R> summary(reg)
```

```
Call:
lm(formula = cadres ~ dipl.sup, data = rp99)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6905 -1.9010 -0.1823  1.4913 17.0866

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.24088    0.32988   3.762 0.000203 ***
dipl.sup     1.38352    0.03931  35.196  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.281 on 299 degrees of freedom
Multiple R-squared:  0.8056,Adjusted R-squared:  0.8049
F-statistic:  1239 on 1 and 299 DF,  p-value: < 2.2e-16
```

Le résultat montre que les cœfficients sont significativement différents de 0. La part de cadres augmente donc avec celle de diplômés du supérieur (ô surprise). On peut très facilement représenter la droite de régression à l'aide de la fonction `abline` (figure 6.5 page suivante).

On remarquera que le premier argument passé à la fonction `lm` a une syntaxe un peu particulière. Il s'agit d'une *formule*, utilisée de manière générale dans les modèles statistiques. On indique la variable d'intérêt à gauche et la variable explicative à droite, les deux étant séparées par un tilde ~ (obtenu sous Windows en appuyant simultanément sur les touches `<Alt Gr>` et `<2>`). On remarquera que les noms des colonnes de notre tableau de données ont été écrites sans guillemets. Dans le cas présent, nous avons calculé une régression linéaire simple entre deux variables, d'où l'écriture `cadres ~ dipl.sup`. Si nous avions voulu expliquer une variable `z` par deux variables `x` et `y`, nous aurions écrit `z ~ x + y`. Il est possible de spécifier des modèles encore plus complexes. Pour un aperçu de la syntaxe des formules sous R, voir http://ww2.coastal.edu/kingw/statistics/R-tutorials/formulae.html.

6.2 Une variable quantitative et une variable qualitative

Quand on parle de comparaison entre une variable quantitative et une variable qualitative, on veut en général savoir si la distribution des valeurs de la variable quantitative est la même selon les modalités de la variable qualitative. En clair : est ce que l'âge de ceux qui écoutent du hard rock est différent de l'âge de ceux qui n'en écoutent pas ?

Là encore, l'idéal est de commencer par une représentation graphique. Les boîtes à moustaches sont parfaitement adaptées pour cela.

Si on a construit des sous-populations d'individus écoutant ou non du hard rock, on peut utiliser la fonction `boxplot` comme indiqué figure 6.6 page 91.

Mais construire les sous-populations n'est pas nécessaire. On peut utiliser directement la version de `boxplot` prenant une *formule* en argument (figure 6.7 page 92).

À première vue, ô surprise, la population écoutant du hard rock a l'air sensiblement plus jeune. Peut-on le tester mathématiquement ? On peut calculer la moyenne d'âge des deux groupes en utilisant la fonction `tapply` [2] :

2. Fonction décrite page 65.

```
R> plot(rp99$dipl.sup, rp99$cadres, ylab = "Part des cadres", xlab = "Part des diplômés du supérieur"
R> abline(reg, col = "red")
```

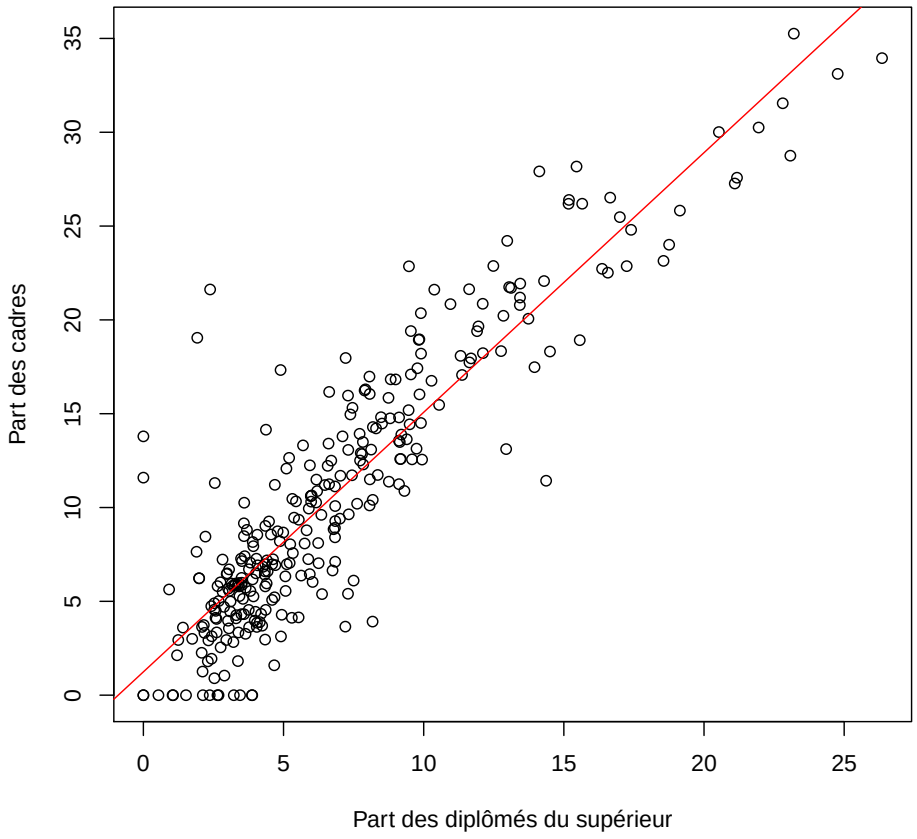


FIGURE 6.5 – Régression de la proportion de cadres par celle de diplômés du supérieur

```
R> d.hard <- subset(d, hard.rock == "Oui")
R> d.non.hard <- subset(d, hard.rock == "Non")
R> boxplot(d.hard$age, d.non.hard$age)
```

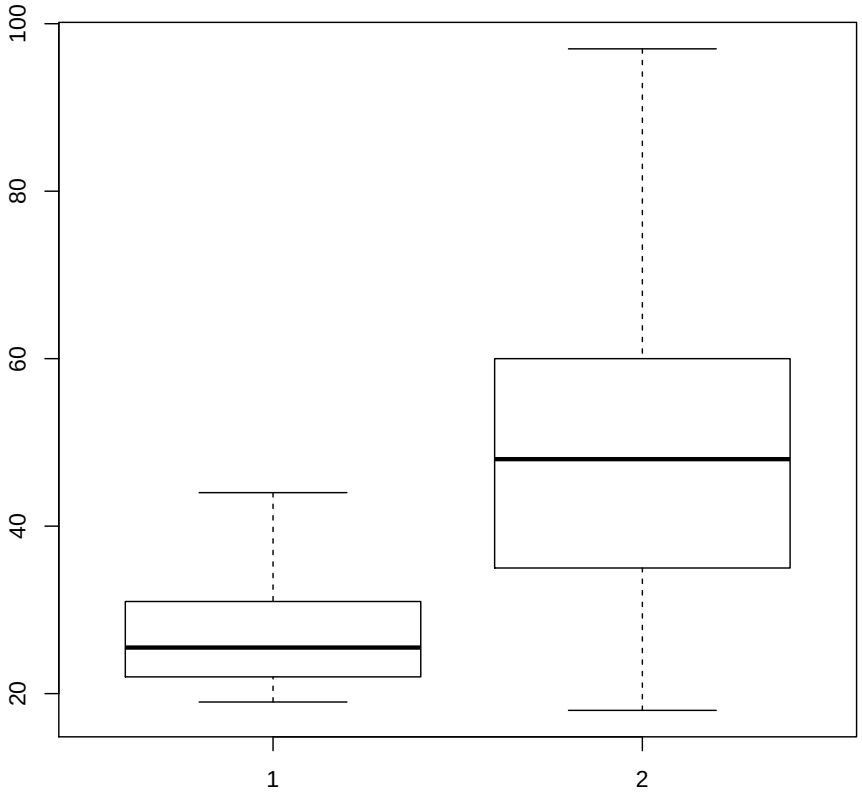


FIGURE 6.6 – *Boxplot* de la répartition des âges (sous-populations)

```
R> boxplot(age ~ hard.rock, data = d)
```

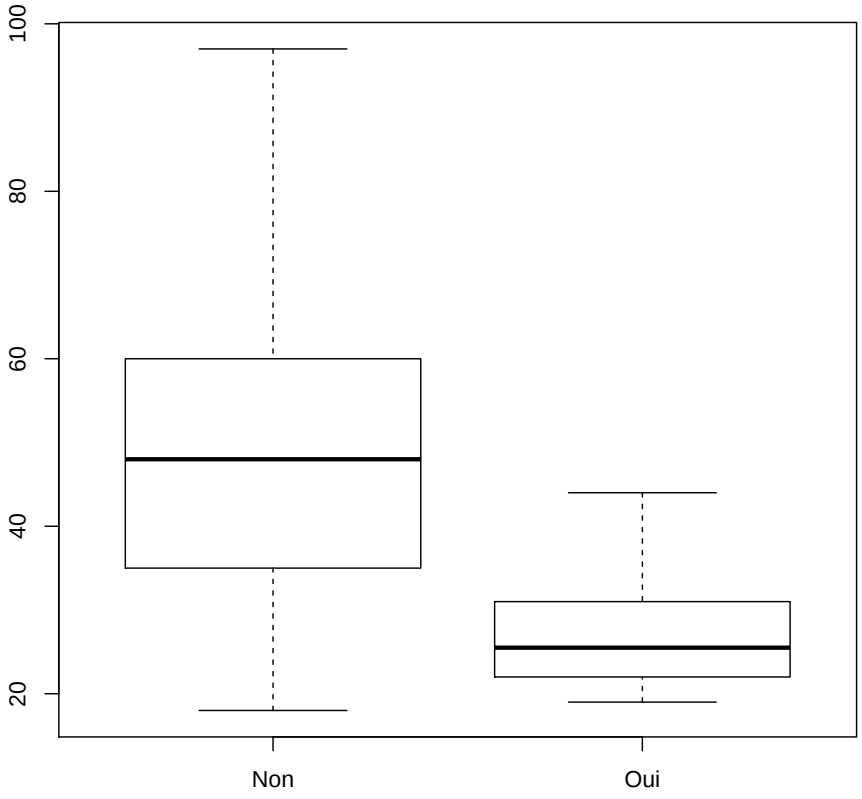


FIGURE 6.7 – *Boxplot* de la répartition des âges (formule)

```
R> tapply(d$age, d$hard.rock, mean)

     Non      Oui
48.30211 27.57143
```

L'écart est très important. Est-il statistiquement significatif ? Pour cela on peut faire un test t de comparaison de moyennes à l'aide de la fonction `t.test` :

```
R> t.test(d$age ~ d$hard.rock)


Welch Two Sample t-test

data:  d$age by d$hard.rock
t = 9.6404, df = 13.848, p-value = 1.611e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 16.11379 25.34758
sample estimates:
mean in group Non mean in group Oui
         48.30211          27.57143
```

Le test est extrêmement significatif. L'intervalle de confiance à 95 % de la différence entre les deux moyennes va de 14,5 ans à 21,8 ans.

La valeur affichée pour p est de `1.611e-07`. Cette valeur peut paraître étrange pour les non avertis. Cela signifie tout simplement 1,611 multiplié par 10 à la puissance -7, autrement dit 0,0000001611. Cette manière de représenter un nombre est couramment appelée *notation scientifique*.

Nous sommes cependant allés un peu vite en besogne, car nous avons négligé une hypothèse fondamentale du test t : les ensembles de valeur comparés doivent suivre approximativement une loi normale et être de même variance [3]. Comment le vérifier ?

D'abord avec un petit graphique, comme sur la figure 6.8 page suivante.

Ça a l'air à peu près bon pour les « Sans hard rock », mais un peu plus limite pour les fans de *Metallica*, dont les effectifs sont d'ailleurs assez faibles. Si on veut en avoir le cœur net on peut utiliser le test de normalité de Shapiro-Wilk avec la fonction `shapiro.test` :

```
R> shapiro.test(d$age[d$hard.rock == "Oui"])


Shapiro-Wilk normality test

data:  d$age[d$hard.rock == "Oui"]
W = 0.86931, p-value = 0.04104

R> shapiro.test(d$age[d$hard.rock == "Non"])
```

3. Concernant cette seconde condition, R propose une option nommée `var.equal` qui permet d'utiliser une approximation dans le cas où les variances ne sont pas égales

```
R> par(mfrow = c(1, 2))
R> hist(d$age[d$hard.rock == "Oui"], main = "Hard rock", col = "red")
R> hist(d$age[d$hard.rock == "Non"], main = "Sans hard rock", col = "red")
```

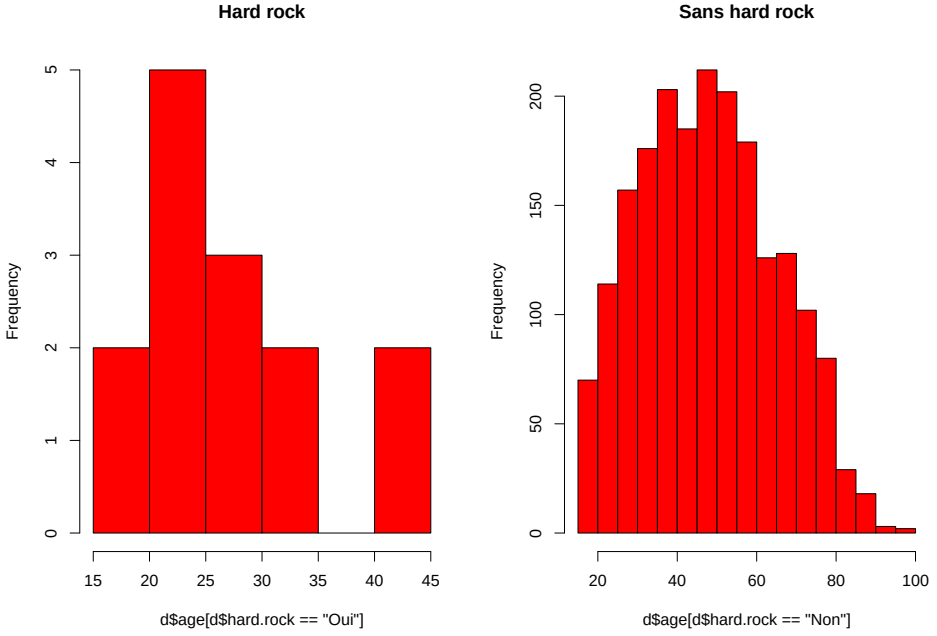


FIGURE 6.8 – Distribution des âges pour appréciation de la normalité

```
Shapiro-Wilk normality test

data:  d$age[d$hard.rock == "Non"]
W = 0.98141, p-value = 2.079e-15
```

Visiblement, le test estime que les distributions ne sont pas suffisamment proches de la normalité dans les deux cas.

Et concernant l'égalité des variances ?

```
R> tapply(d$age, d$hard.rock, var)

      Non       Oui
285.62858  62.72527
```

L'écart n'a pas l'air négligeable. On peut le vérifier avec le test fourni par la fonction `var.test` :

```
R> var.test(d$age ~ d$hard.rock)


F test to compare two variances

data:  d$age by d$hard.rock
F = 4.5536, num df = 1985, denom df = 13, p-value = 0.003217
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.751826 8.694405
sample estimates:
ratio of variances
          4.553644
```

La différence est très significative. En toute rigueur le test t n'aurait donc pas pu être utilisé.

Damned ! Ces maudits tests statistiques vont-ils nous empêcher de faire connaître au monde entier notre fabuleuse découverte sur l'âge des fans de *Sepultura* ? Non ! Car voici qu'approche à l'horizon un nouveau test, connu sous le nom de *Wilcoxon/Mann-Whitney*. Celui-ci a l'avantage d'être *non-paramétrique*, c'est à dire de ne faire aucune hypothèse sur la distribution des échantillons comparés. Par contre il ne compare pas des différences de moyennes mais des différences de médianes :

```
R> wilcox.test(d$age ~ d$hard.rock)


Wilcoxon rank sum test with continuity correction

data:  d$age by d$hard.rock
W = 23980, p-value = 2.856e-06
alternative hypothesis: true location shift is not equal to 0
```

Ouf ! La différence est hautement significative[4]. Nous allons donc pouvoir entamer la rédaction de notre article pour la *Revue française de sociologie*.

4. Ce test peut également fournir un intervalle de confiance avec l'option `conf.int=TRUE`.

6.3 Deux variables qualitatives

La comparaison de deux variables qualitatives s'appelle en général un *tableau croisé*. C'est sans doute l'une des analyses les plus fréquentes lors du traitement d'enquêtes en sciences sociales.

6.3.1 Tableau croisé

La manière la plus simple d'obtenir un tableau croisé est d'utiliser la fonction `table` en lui donnant en paramètres les deux variables à croiser. En l'occurrence nous allons croiser un recodage du niveau de qualification regroupé avec le fait de pratiquer un sport.

On commence par calculer la variable recodée et par afficher le tri à plat des deux variables :

```
R> d$qualreg <- as.character(d$qualif)
R> d$qualreg[d$qualif %in% c("Ouvrier specialise", "Ouvrier qualifie")] <- "Ouvrier"
R> d$qualreg[d$qualif %in% c("Profession intermediaire", "Technicien")] <- "Intermediaire"
R> table(d$qualreg)


        Autre         Cadre       Employe Intermediaire       Ouvrier
           58           260           594           246           495

R> table(d$sport)


 Non  Oui
1277  723
```

Le tableau croisé des deux variables s'obtient de la manière suivante :

```
R> table(d$sport, d$qualreg)


      Autre Cadre Employe Intermediaire Ouvrier
  Non    38   117     401           127     381
  Oui    20   143     193           119     114
```

Il est d'ailleurs tout à fait possible de croiser trois variables ou plus. Par exemple :

```
R> table(d$sport, d$cuisine, d$sexe)

, ,  = Homme


      Non Oui
  Non 401 129
  Oui 228 141

, ,  = Femme


      Non Oui
```

```
  Non 358 389
  Oui 132 222
```

On n'a cependant que les effectifs, ce qui rend difficile les comparaisons. L'extension questionr fournit des fonctions permettant de calculer les pourcentages lignes, colonnes et totaux d'un tableau croisé. *questionr*

Les pourcentages lignes s'obtiennent avec la fonction `lprop`. Celle-ci s'applique au tableau croisé généré par `table` :

```
R> tab <- table(d$sport, d$qualreg)
R> lprop(tab)


           Autre Cadre Employe Intermediaire Ouvrier Total
  Non        3.6  11.0  37.7    11.9          35.8   100.0
  Oui        3.4  24.3  32.8    20.2          19.4   100.0
  Ensemble   3.5  15.7  35.9    14.9          29.9   100.0
```

Les pourcentages ligne ne nous intéressent guère ici. On ne cherche pas à voir quelle est la proportion de cadres parmi ceux qui pratiquent un sport, mais plutôt quelle est la proportion de sportifs chez les cadres. Il nous faut donc des pourcentages colonnes, que l'on obtient avec la fonction `cprop` :

```
R> cprop(tab)


        Autre Cadre Employe Intermediaire Ouvrier Ensemble
  Non    65.5  45.0  67.5    51.6          77.0    64.4
  Oui    34.5  55.0  32.5    48.4          23.0    35.6
  Total 100.0 100.0 100.0   100.0         100.0   100.0
```

Dans l'ensemble, le pourcentage de personnes ayant pratiqué un sport est de 35,6 %. Mais cette proportion varie fortement d'une catégorie professionnelle à l'autre : 55,0 % chez les cadres contre 23,0 % chez les ouvriers.

À noter qu'on peut personnaliser l'affichage de ces tableaux de pourcentages à l'aide de différentes options, dont `digits`, qui règle le nombre de décimales à afficher, et `percent`, qui indique si on souhaite ou non rajouter un symbole % dans chaque case du tableau. Cette personnalisation peut se faire directement au moment de la génération du tableau, et dans ce cas elle sera utilisée par défaut :

```
R> ctab <- cprop(tab, digits = 2, percent = TRUE)
R> ctab


        Autre   Cadre   Employe Intermediaire Ouvrier Ensemble
  Non    65.52%  45.00%  67.51%  51.63%        76.97%  64.37%
  Oui    34.48%  55.00%  32.49%  48.37%        23.03%  35.63%
  Total 100.00% 100.00% 100.00% 100.00%       100.00% 100.00%
```

Ou bien ponctuellement en passant les mêmes arguments aux fonctions `print` (pour affichage dans R) ou `clipcopy` (pour export vers un logiciel externe) :

```
R> ctab <- cprop(tab)
R> print(ctab, percent = TRUE)
```

```
       Autre  Cadre  Employe Intermediaire Ouvrier Ensemble
  Non   65.5%  45.0%  67.5%   51.6%         77.0%   64.4%
  Oui   34.5%  55.0%  32.5%   48.4%         23.0%   35.6%
  Total 100.0% 100.0% 100.0%  100.0%        100.0%  100.0%
```

6.3.2 χ^2 et dérivés

Pour tester l'existence d'un lien entre les modalités des deux variables, on va utiliser le très classique test du χ^2 [5]. Celui-ci s'obtient grâce à la fonction `chisq.test`, appliquée au tableau croisé obtenu avec `table` [6] :

```
R> chisq.test(tab)


Pearson's Chi-squared test

data:  tab
X-squared = 96.798, df = 4, p-value < 2.2e-16
```

Le test est hautement significatif, on ne peut pas considérer qu'il y a indépendance entre les lignes et les colonnes du tableau.

On peut affiner l'interprétation du test en déterminant dans quelle case l'écart à l'indépendance est le plus significatif en utilisant les *résidus* du test. Ceux-ci sont notamment affichables avec la fonction `chisq.residuals` de questionr :

questionr

```
R> chisq.residuals(tab)


      Autre Cadre Employe Intermediaire Ouvrier
  Non  0.11 -3.89    0.95         -2.49    3.49
  Oui -0.15  5.23   -1.28          3.35   -4.70
```

Les cases pour lesquelles l'écart à l'indépendance est significatif ont un résidu dont la valeur est supérieure à 2 ou inférieure à -2. Ici on constate que la pratique d'un sport est sur-représentée parmi les cadres et, à un niveau un peu moindre, parmi les professions intermédiaires, tandis qu'elle est sous-représentée chez les ouvriers.

Enfin, on peut calculer le cœfficient de contingence de Cramer du tableau, qui peut nous permettre de le comparer par la suite à d'autres tableaux croisés. On peut pour cela utiliser la fonction `cramer.v`de questionr :

questionr

```
R> cramer.v(tab)

[1] 0.24199
```

Et pour un tableau à 2×2 entrées, il est possible de calculer le test exact de Fisher avec la fonction `fisher.test`. On peut soit lui passer le résultat de `table`, soit directement les deux variables à croiser.

5. On ne donnera pas plus d'indications sur le test du χ^2 ici. Les personnes désirant une présentation plus détaillée pourront se reporter (attention, séance d'autopromotion !) à la page suivante : http://alea.fr.eu.org/pages/khi2.

6. On peut aussi appliquer directement le test en spécifiant les deux variables à croiser *via* `chisq.test(d$qualreg, d$sport)`

```
R> lprop(table(d$sexe, d$cuisine))


           Non   Oui   Total
  Homme     70.0  30.0 100.0
  Femme     44.5  55.5 100.0
  Ensemble  56.0  44.0 100.0

R> fisher.test(table(d$sexe, d$cuisine))


Fisher's Exact Test for Count Data

data:  table(d$sexe, d$cuisine)
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.402598 3.513723
sample estimates:
odds ratio
  2.903253
```

6.3.3 Représentation graphique

Enfin, on peut obtenir une représentation graphique synthétisant l'ensemble des résultats obtenus sous la forme d'un graphique en mosaïque, grâce à la fonction `mosaicplot`. Le résultat est indiqué figure 6.9 page suivante.

Comment interpréter ce graphique haut en couleurs [7] ? Chaque rectangle représente une case de tableau. Sa largeur correspond au pourcentage des modalités en colonnes (il y'a beaucoup d'employés et d'ouvriers et très peu d'« autres »). Sa hauteur correspond aux pourcentages-colonnes : la proportion de sportifs chez les cadres est plus élevée que chez les employés. Enfin, la couleur de la case correspond au résidu du test du χ^2 correspondant : les cases en rouge sont sous-représentées, les cases en bleu sur-représentées, et les cases blanches sont statistiquement proches de l'hypothèse d'indépendance.

Lorsque l'on s'intéresse principalement aux variations d'une variable selon une autre, par exemple ici à la pratique du sport selon le niveau de qualification, il peut être intéressant de présenter les pourcentages en colonne sous la forme de barres cumulées. Voir figure 6.10 page 101.

7. Sauf s'il est imprimé en noir et blanc...

```
R> mosaicplot(qualreg ~ sport, data = d, shade = TRUE, main = "Graphe en mosaïque")
```

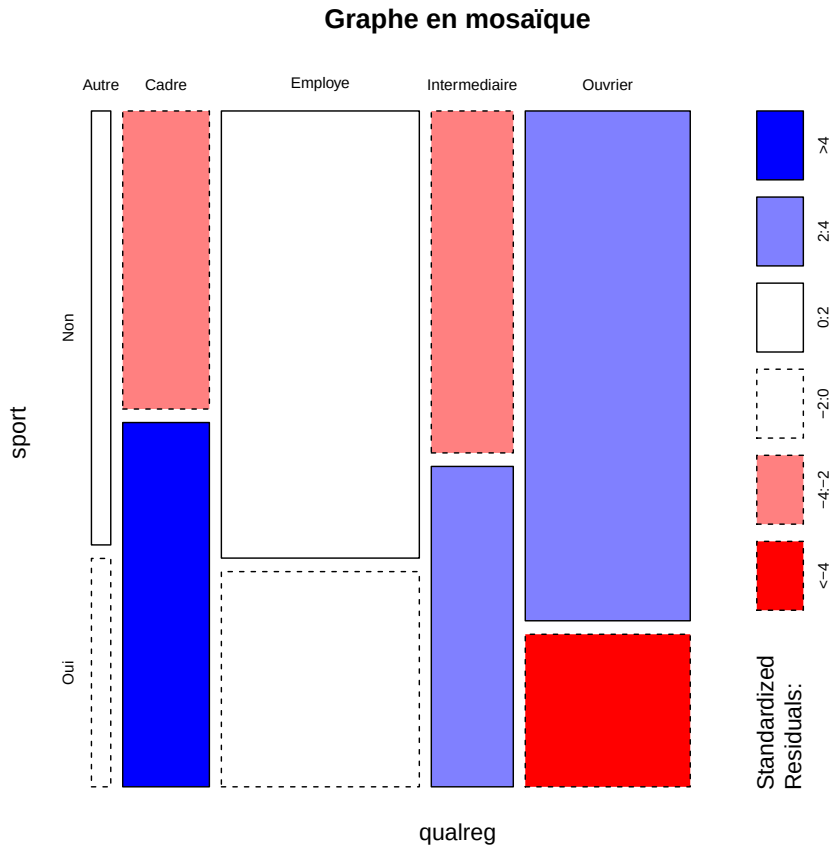


FIGURE 6.9 – Exemple de graphe en mosaïque

```
R> barplot(cprop(tab, total = FALSE), main = "Pratique du sport selon le niveau de qualification")
```

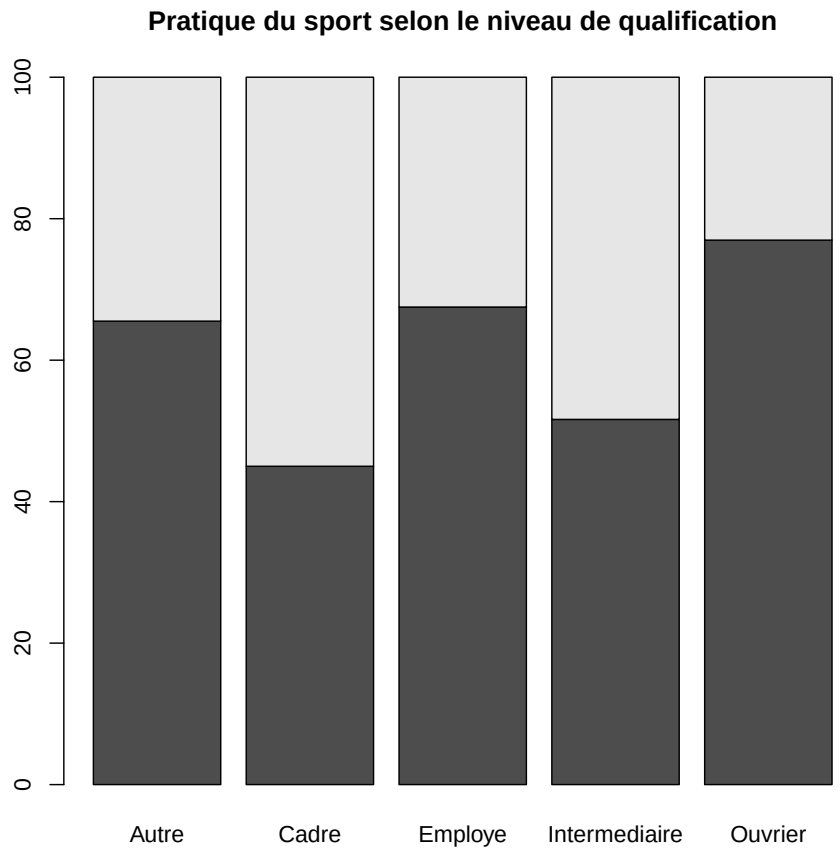


FIGURE 6.10 – Exemple de barres cumulées

Partie 7

Données pondérées

S'il est tout à fait possible de travailler avec des données pondérées sous R, cette fonctionnalité n'est pas aussi bien intégrée que dans la plupart des autres logiciels de traitement statistique. En particulier, il y a plusieurs manières possibles de gérer la pondération.

Dans ce qui suit, on utilisera le jeu de données tiré de l'enquête *Histoire de vie* et notamment sa variable de pondération `poids`[1].

```
R> data(hdv2003)
R> d <- hdv2003
R> range(d$poids)

[1]    78.07834 31092.14132
```

7.1 Options de certaines fonctions

Tout d'abord, certaines fonctions de R acceptent en argument un vecteur permettant de pondérer les observations (l'option est en général nommée `weights` ou `row.w`). C'est le cas par exemple des méthodes d'estimation de modèles linéaires (`lm`) ou de modèles linéaires généralisés (`glm`), ou dans les analyses de correspondances des extensions ade4 (`dudi.acm`) ou FactoMineR (`MCA`).

Par contre cette option n'est pas présente dans les fonctions de base comme `mean`, `var`, `table` ou `chisq.test`.

7.2 Fonctions de l'extension questionr

L'extension questionr propose quelques fonctions permettant de calculer des statistiques simples pondérées[2] :

`wtd.mean` moyenne pondérée

`wtd.var` variance pondérée

`wtd.table` tris à plat et tris croisés pondérés

1. On notera que cette variable est utilisée à titre purement illustratif. Le jeu de données étant un extrait d'enquête et la variable de pondération n'ayant pas été recalculée, elle n'a ici à proprement parler aucun sens.
2. Les fonctions `wtd.mean` et `wtd.var` sont des copies conformes des fonctions du même nom de l'extension Hmisc de Frank Harrel. Hmisc étant une extension « de taille », on a préféré recopié les fonctions pour limiter le poids des dépendances.

On les utilise de la manière suivante :

```
R> library(questionr)
R> mean(d$age)

[1] 48.157

R> wtd.mean(d$age, weights = d$poids)

[1] 46.34726

R> wtd.var(d$age, weights = d$poids)

[1] 325.2658
```

Pour les tris à plat, on utilise la fonction `wtd.table` à laquelle on passe la variable en paramètre :

```
R> wtd.table(d$sexe, weights = d$poids)

  Homme   Femme
5149382 5921844
```

Pour un tri croisé, il suffit de passer deux variables en paramètres :

```
R> wtd.table(d$sexe, d$hard.rock, weights = d$poids)

              Non        Oui
Homme 5109366.41   40016.02
Femme 5872596.42   49247.49
```

Ces fonctions admettent notamment les deux options suivantes :

`na.rm` si `TRUE`, on ne conserve que les observations sans valeur manquante

`normwt` si `TRUE`, on normalise les poids pour que les effectifs totaux pondérés soient les mêmes que les effectifs initiaux. Il faut utiliser cette option, notamment si on souhaite appliquer un test sensible aux effectifs comme le χ^2.

Ces fonctions rendent possibles l'utilisation des statistiques descriptives les plus simples et le traitement des tableaux croisés (les fonctions `lprop`, `cprop` ou `chisq.test` peuvent être appliquées au résultat d'un `wtd.table`) mais restent limitées en termes de tests statistiques ou de graphiques...

7.3 L'extension survey

L'extension survey est spécialement dédiée au traitement d'enquêtes ayant des techniques d'échantillonnage et de pondération potentiellement très complexes. L'extension s'installe comme la plupart des autres :

```
R> install.packages("survey", dep = TRUE)
```

Le site officiel (en anglais) comporte beaucoup d'informations, mais pas forcément très accessibles :

http://faculty.washington.edu/tlumley/survey/

Pour utiliser les fonctionnalités de l'extension, on doit d'abord définir un *design* de notre enquête. C'est-à-dire indiquer quel type de pondération nous souhaitons lui appliquer. Dans notre cas nous utilisons le *design* le plus simple, avec une variable de pondération déjà calculée. Ceci se fait à l'aide de la fonction `svydesign` :

```
R> library(survey)
R> dw <- svydesign(ids = ~1, data = d, weights = ~d$poids)
```

Cette fonction crée un nouvel objet, que nous avons nommé `dw`. Cet objet n'est pas à proprement parler un tableau de données, mais plutôt un tableau de données *plus* une méthode de pondération. `dw` et `d` sont des objets distincts, les opérations effectuées sur l'un n'ont pas d'influence sur l'autre. On peut cependant retrouver le contenu de `d` depuis `dw` en utilisant `dw$variables` :

```
R> mean(d$age)

[1] 48.157

R> mean(dw$variables$age)

[1] 48.157
```

Lorsque notre *design* est déclaré, on peut lui appliquer une série de fonctions permettant d'effectuer diverses opérations statistiques en tenant compte de la pondération. On citera notamment :

`svymean`, `svyvar`, `svytotal` statistiques univariées
`svytable` tableaux croisés
`svychisq` test du χ^2
`svyby` statistiques selon un facteur
`svyglm` modèles linéaires généralisés
`svyplot`, `svyhist`, `svyboxplot` fonctions graphiques

D'autres fonctions sont disponibles, comme `svyratio`, mais elles ne seront pas abordées ici.

Pour ne rien arranger, ces fonctions prennent leurs arguments sous forme de formules, c'est-à-dire pas de la manière habituelle. En général l'appel de fonction se fait en spécifiant d'abord les variables d'intérêt sous forme de formule, puis l'objet *design*.

Voyons tout de suite quelques exemples :

```
R> svymean(~age, dw)

      mean     SE
age 46.347 0.5284

R> svyvar(~heures.tv, dw, na.rm = TRUE)

          variance     SE
heures.tv   2.9886 0.1836

R> svytable(~sexe, dw)

sexe
  Homme   Femme
5149382 5921844
```

```
R> svytable(~sexe + clso, dw)

       clso
sexe           Oui        Non Ne sait pas
  Homme 2658744.04 2418187.64    72450.75
  Femme 2602031.76 3242389.36    77422.79
```

En particulier, les tris à plat se déclarent en passant comme argument le nom de la variable précédé d'un symbole `~`, tandis que les tableaux croisés utilisent les noms des deux variables séparés par un `+` et précédés par un `~`.

On peut récupérer le tableau issu de `svytable` dans un objet et le réutiliser ensuite comme n'importe quel tableau croisé :

```
R> tab <- svytable(~sexe + clso, dw)
R> tab

       clso
sexe           Oui        Non Ne sait pas
  Homme 2658744.04 2418187.64    72450.75
  Femme 2602031.76 3242389.36    77422.79

R> lprop(tab)

          clso
sexe       Oui   Non   Ne sait pas Total
  Homme     51.6  47.0   1.4         100.0
  Femme     43.9  54.8   1.3         100.0
  Ensemble  47.5  51.1   1.4         100.0

R> svychisq(~sexe + clso, dw)


Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~sexe + clso, dw)
F = 3.3331, ndf = 1.9734, ddf = 3944.9000, p-value = 0.03641
```

Les fonctions `lprop` et `cprop` de questionr sont donc tout à fait compatibles avec l'utilisation de survey. La fonction `freq` peut également être utilisée si on lui passe en argument non pas la variable elle-même, mais son tri à plat obtenu avec `svytable` :

questionr

```
R> tab <- svytable(~peche.chasse, dw)
R> freq(tab, total = TRUE)

             n     %  val%
Non    9716683  87.8  87.8
Oui    1354544  12.2  12.2
Total 11071226 100.0 100.0
```

Par contre, il **ne faut pas** utiliser `chisq.test` sur un tableau généré par `svytable`. Les effectifs étant extrapolés à partir de la pondération, les résultats du test seraient complètement faussés. Si on veut faire un test du χ^2 sur un tableau croisé pondéré, il faut utiliser `svychisq` :

```
R> svychisq(~sexe + clso, dw)


Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~sexe + clso, dw)
F = 3.3331, ndf = 1.9734, ddf = 3944.9000, p-value = 0.03641
```

Le principe de la fonction `svyby` est similaire à celui de `tapply` (voir section 5.3.3 page 65). Elle permet de calculer des statistiques selon plusieurs sous-groupes définis par un facteur. Par exemple :

```
R> svyby(~age, ~sexe, dw, svymean)

       sexe      age        se
Homme Homme 45.20200 0.7419450
Femme Femme 47.34313 0.7420836
```

Enfin, survey est également capable de produire des graphiques à partir des données pondérées. Des exemples sont donnés figure 7.1 page ci-contre.

7.4 Conclusion

En attendant mieux, la gestion de la pondération sous R n'est sans doute pas ce qui se fait de plus pratique et de plus simple. On pourra quand même donner les conseils suivants :

- utiliser les options de pondération des fonctions usuelles ou les fonctions d'extensions comme questionr pour les cas les plus simples ;
- si on utilise survey, effectuer tous les recodages et manipulations sur les données non pondérées autant que possible ;
- une fois les recodages effectués, on déclare le *design* et on fait les analyses en tenant compte de la pondération ;
- surtout ne jamais modifier les variables du *design*. Toujours effectuer recodages et manipulations sur les données non pondérées, puis redéclarer le *design* pour que les mises à jour effectuées soient disponibles pour l'analyse ;

```
R> par(mfrow = c(2, 2))
R> svyplot(~age + heures.tv, dw, col = "red", main = "Bubble plot")
R> svyhist(~heures.tv, dw, col = "peachpuff", main = "Histogramme")
R> svyboxplot(age ~ 1, dw, main = "Boxplot simple", ylab = "Âge")
R> svyboxplot(age ~ sexe, dw, main = "Boxplot double", ylab = "Âge", xlab = "Sexe")
```

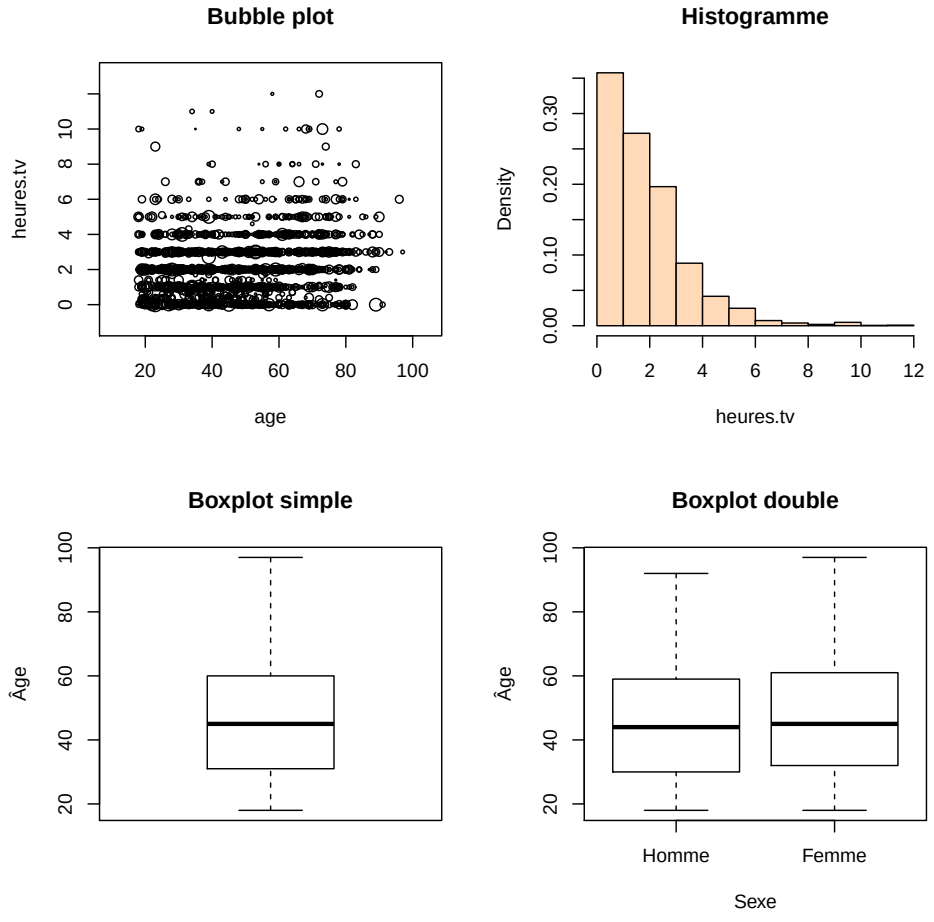


FIGURE 7.1 – Fonctions graphiques de l'extension survey

Partie 8

Exporter les résultats

Cette partie décrit comment, une fois les analyses réalisées, on peut exporter les résultats (tableaux et graphiques) dans un traitement de texte ou une application externe.

8.1 Export manuel de tableaux

questionr

Les tableaux générés par R (et plus largement, tous les types d'objets) peuvent être exportés pour inclusion dans un traitement de texte à l'aide de la fonction `clipcopy` de l'extension questionr [1].

Il suffit pour cela de lui passer en argument le tableau ou l'objet qu'on souhaite exporter. Dans ce qui suit on utilisera le tableau suivant, placé dans un objet nommé `tab` :

```
R> data(hdv2003)
R> tab <- table(hdv2003$sexe, hdv2003$bricol)
R> tab



        Non Oui
  Homme 384 515
  Femme 763 338
```

8.1.1 Copier/coller vers Excel et Word *via* le presse-papier

La première possibilité est d'utiliser les options par défaut de `clipcopy`. Celle-ci va alors transformer le tableau (ou l'objet) en `HTML` et placer le résultat dans le presse papier du système. Ceci ne fonctionne malheureusement que sous Windows [2].

```
R> clipcopy(tab)
```

On peut ensuite récupérer le résultat dans une feuille Excel en effectuant un simple *Coller*.

1. Celle-ci nécessite que l'extension `R2HTML` soit également installée sur le système *via* `install.packages("R2HTML",dep=TRUE)`.
2. En fait cela fonctionne aussi sous Linux si le programme `xclip` est installé et accessible. Cela fonctionne peut-être aussi sous Mac OS X mais n'a pas pu être testé.

	A	B	C
1		**Non**	**Oui**
2	Homme	383	511
3	Femme	771	335

On peut ensuite sélectionner le tableau sous Excel, le copier et le coller dans Word :

	Non	Oui
Homme	383	511
Femme	771	335

8.1.2 Export vers Word ou OpenOffice/LibreOffice *via* un fichier

L'autre possibilité ne nécessite pas de passer par Excel, et fonctionne sous Word, OpenOffice et LibreOffice sur toutes les plateformes.

Elle nécessite de passer à la fonction `clipcopy` l'option `file=TRUE` qui enregistre le contenu de l'objet dans un fichier plutôt que de le placer dans le presse-papier :

```
R> clipcopy(tab, file = TRUE)
```

Par défaut le résultat est placé dans un fichier nommé `temp.html` dans le répertoire courant, mais on peut modifier le nom et l'emplacement avec l'option `filename` :

```
R> clipcopy(tab, file = TRUE, filename = "exports/tab1.html")
```

On peut ensuite l'intégrer directement dans Word ou dans OpenOffice en utilisant le menu *Insertion* puis *Fichier* et en sélectionnant le fichier de sortie généré précédemment.

	Non	Oui
Homme	383	511
Femme	771	335

8.2 Export de graphiques

8.2.1 Export *via* l'interface graphique (Windows ou Mac OS X)

L'export de graphiques est très simple si on utilise l'interface graphique sous Windows. En effet, les fenêtres graphiques possèdent un menu *Fichier* qui comporte une entrée *Sauver sous* et une entrée *Copier dans le presse papier*.

L'option *Sauver sous* donne le choix entre plusieurs formats de sortie, vectoriels (Metafile, Postscript) ou bitmaps (jpeg, png, tiff, etc.). Une fois l'image enregistrée on peut ensuite l'inclure dans n'importe quel document ou la retravailler avec un logiciel externe.

Une image *bitmap* est une image stockée sous forme de points, typiquement une photographie. Une image *vectorielle* est une image enregistrée dans un langage de description, typiquement un schéma ou une figure. Le second format présente l'avantage d'être en général beaucoup plus léger et d'être redimensionnable à l'infini sans perte de qualité. Pour plus d'informations voir http://fr.wikipedia.org/wiki/Image_matricielle et http://fr.wikipedia.org/wiki/Image_vectorielle.

L'option *Copier dans le presse papier* permet de placer le contenu de la fenêtre dans le presse-papier soit dans un format vectoriel soit dans un format bitmap. On peut ensuite récupérer le résultat dans un traitement de texte ou autre avec un simple *Coller*.

Des possibilités similaires sont offertes par l'interface sous Mac OS X, mais avec des formats proposés un peu différents.

Avec RStudio, les commandes d'export sont situées dans le menu *Plots* qui comporte les entrées *Save Plot as image* et *Save Plot as PDF*. Ces mêmes commandes sont accessibles via le bouton *Export* situé au dessus du graphique dans le quadrant bas-droit. Les options d'export sont plus importantes que celle de l'interface graphique de base, avec notamment le support du format SVG ou encore la possibilité de modifier la taille du graphique exporté.

8.2.2 Export avec les commandes de R

On peut également exporter les graphiques dans des fichiers de différents formats directement avec des commandes R. Ceci a l'avantage de fonctionner sur toutes les plateformes, et de faciliter la mise à jour du graphique exporté (on n'a qu'à relancer les commandes concernées pour que le fichier externe soit mis à jour).

La première possibilité est d'exporter le contenu d'une fenêtre déjà existante à l'aide de la fonction `dev.`. On doit fournir à celle-ci le format de l'export (option `device`) et le nom du fichier (option `file`). Par exemple :

```
R> boxplot(rnorm(100))
R> dev.print(device = png, file = "export.png", width = 600)
```

Les formats de sortie possibles varient selon les plateformes, mais on retrouve partout les formats bitmap bmp, jpeg, png, tiff, et les formats vectoriels postscript ou pdf. La liste complète disponible pour votre installation de R est disponible dans la page d'aide de `Devices` :

```
R> ?Devices
```

L'autre possibilité est de rediriger directement la sortie graphique dans un fichier, avant d'exécuter la commande générant la figure. On doit pour cela faire appel à l'une des commandes permettant cette redirection. Les plus courantes sont `bmp`, `png`, `jpeg` et `tiff` pour les formats bitmap, `postscript`, `pdf`, `svg`[3] et `win.metafile`[4] pour les formats vectoriels.

Les formats vectoriels ont l'avantage de pouvoir être redimensionnés à volonté sans perte de qualité, et produisent des fichiers en général de plus petite taille. On pourra donc privilégier le format SVG, par exemple, si on utilise LibreOffice ou OpenOffice.

3. Ne fonctionne pas sous Word.
4. Ne fonctionne que sous Word.

Ces fonctions prennent différentes options permettant de personnaliser la sortie graphique. Les plus courantes sont `width` et `height` qui donnent la largeur et la hauteur de l'image générée (en pixels pour les images bitmap, en pouces pour les images vectorielles), et `pointsize` qui donne la taille de base des polices de caractère utilisées.

```
R> png(file = "out.png", width = 800, height = 700)
R> plot(rnorm(100))
R> dev.off()
R>
R> pdf(file = "out.pdf", width = 9, height = 9, pointsize = 10)
R> plot(rnorm(150))
R> dev.off()
```

Il est nécessaire de faire un appel à la fonction `dev.off` après génération du graphique pour que le résultat soit bien écrit dans le fichier de sortie (dans le cas contraire on se retrouve avec un fichier vide).

8.3 Génération automatique de documents avec RMarkdown

R Markdown est une extension R, développée par RStudio, qui permet de mélanger du code R dans des documents au format `Markdown` et de produire en retour des documents comportant, à la place du code en question, le résultat de son exécution (texte, tableaux, graphiques, etc.).

Site officiel de l'extension :

http://rmarkdown.rstudio.com/

R Markdown est extrêmement versatile. À partir d'un format de balisage léger, il permet d'inclure du code R et de générer des documents dans différents formats, notamment HTML, PDF [5] ou traitement de texte. À noter que RStudio [6] propose une interface pratique à R Markdown, mais qu'on peut aussi parfaitement l'utiliser en ligne de commande sans passer par RStudio.

8.3.1 Exemple

Voyons tout de suite un exemple. Dans RStudio, choisissez le menu *File*, puis *New file*, puis *R Markdown...*. Une boîte dialogue vous permet alors d'indiquer un titre, un auteur et un format de sortie [7]. Un nouveau document s'ouvre. Effacez son contenu et remplacez le par quelque chose comme :

5. Sous réserve d'avoir une installation fonctionnelle de LaTeX
6. Pour plus d'informations sur RStudio, voir section A.5 page 122.
7. Le format de sortie peut être modifié par la suite à tout moment

```
---
title: "Test"
output: html_document
---

# Exemple de titre

Ceci est un paragraphe avec du texte en *italique*, en **gras** et en
`police à chasse fixe`.

Ensuite vient un bloc de code R qui affiche du texte :

```{r exemple1}
data(iris)
mean(iris$Sepal.Width)
```

Grâce à `kable`, on peut aussi afficher des tableaux avec le code suivant :

```{r exempletab}
library(knitr)
tab <- table(iris$Species, cut(iris$Sepal.Width, breaks=3))
kable(tab)
```

Et on peut, enfin, inclure des graphiques directement :

```{r examplegraph, echo=FALSE}
plot(iris$Sepal.Width, iris$Sepal.Length, col="red")
```
```

Enregistrez le fichier avec un nom de votre choix suivi de l'extension `.Rmd`, puis cliquez sur le bouton *Knit HTML*. Vous devriez voir apparaître une fenêtre ressemblant à la figure 8.1 page ci-contre.

Vous avez ensuite la possibilité d'enregistrer ce fichier HTML, ou même, *via* le bouton *Publish*, de le mettre en ligne sur le site *Rpubs* (http://rpubs.com) pour pouvoir le partager facilement.

Si vous n'utilisez pas RStudio, vous pouvez compiler votre fichier R Markdown en lançant R dans le même répertoire et en utilisant le code suivant :

```
R> library(rmarkdown)
R> render("test.Rmd")
```

Le résultat se trouvera dans le fichier `test.html` du même répertoire.

8.3.2 Syntaxe

Dans l'exemple précédent, il faut bien différencier ce qui relève de la syntaxe de *Markdown* et ce qui relève de la syntaxe de knitr.

Markdown est un langage de balisage permettant de mettre en forme du texte en désignant des niveaux de titre, du gras, des listes à puce, etc. Ainsi, du texte placé entre deux astérisques sera mis en italique, une ligne précédée d'un dièse `#` sera transformée en titre de niveau 1, etc. Dans RStudio, choisissez le

Test

Exemple de titre

Ceci est un paragraphe avec du texte en *italique*, en **gras** et en `police à chasse fixe`.

Ensuite vient un bloc de code R qui affiche du texte :

```
data(iris)
mean(iris$Sepal.Width)
```

```
## [1] 3.057333
```

Grâce à `xtable`, on peut aussi afficher des tableaux avec le code suivant :

```
library(knitr)
tab <- table(iris$Species, cut(iris$Sepal.Width, breaks=3))
kable(tab)
```

	(2,2.8]	(2.8,3.6]	(3.6,4.4]
setosa	1	36	13
versicolor	27	23	0
virginica	19	29	2

Et on peut, enfin, inclure des graphiques directement :

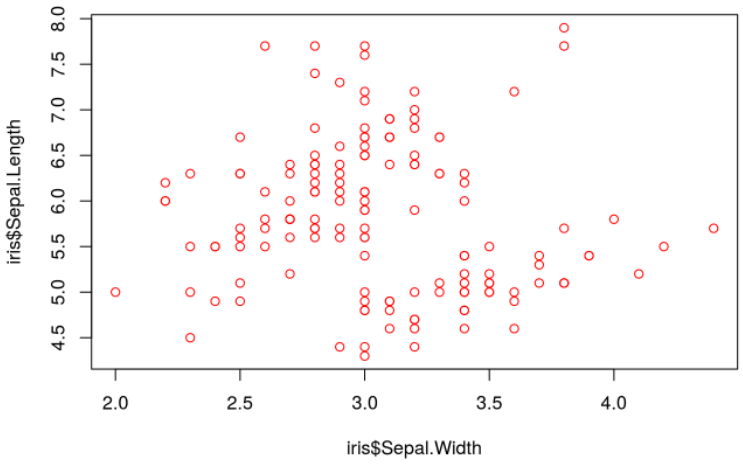


FIGURE 8.1 – Résultat de la génération d'un document HTML par R Markdown

menu *Help* puis *Markdown Quick Reference* pour afficher un aperçu des différentes possibilités de mise en forme.

Ensuite, le document contient plusieurs blocs de code R. Ceux-ci sont délimités par la syntaxe suivante :

```
```{r nom_du_bloc, options}
```
```

Le bloc commence et se termine par trois quotes inverses suivies, entre accolades, du langage utilisé dans le bloc (ici toujours `r`), du nom du bloc (ce qui permet de l'identifier facilement en cas d'erreur), et d'une liste d'options éventuelles séparées par des virgules.

Ces options permettent de modifier le comportement du bloc. Par exemple, spécifier `echo=FALSE` fera que le code R ne sera pas affiché dans le document final, `fig.width=8` modifiera la largeur des images générées pour un graphique, etc. Un aperçu des principales options peut être trouvé à l'adresse suivante :

http://rpubs.com/gallery/options

Et la liste exhaustive se trouve ici :

http://yihui.name/knitr/options

8.3.3 Aller plus loin

L'objectif ici était de présenter un aperçu de l'intérêt et des possibilités de R Markdown. Grâce à ce système, le code R peut être intégré directement aux analyses, et le document final contient le résultat de l'exécution de ce code. La mise à jour de l'ensemble de ces résultats (en cas de modification des données par exemple) peut alors se faire d'un simple clic ou d'une seule commande. L'intérêt en terme de reproductibilité des recherches est également énorme.

Pour aller au-delà de l'exemple donné ici, on trouvera de nombreuses ressources en ligne sur les possibilités et l'utilisation de knitr. L'extension propose même une démonstration permettant de modifier un fichier *R Markdown* et de voir le résultat juste en appuyant sur la touche `F4`. Pour lancer cette démonstration :

```
R> if (!require("shiny")) install.packages("shiny")
R> demo("notebook", package = "knitr")
```

Partie 9

Où trouver de l'aide

9.1 Aide en ligne

R dispose d'une aide en ligne très complète, mais dont l'usage n'est pas forcément très simple. D'une part car elle est intégralement en anglais, d'autre part car son organisation prend un certain temps à être maîtrisée.

9.1.1 Aide sur une fonction

La fonction la plus utile est sans doute celle qui permet d'afficher la page d'aide liée à une ou plusieurs fonctions. Celle-ci permet de lister les arguments de la fonction, d'avoir des informations détaillées sur son fonctionnement, les résultats qu'elle retourne, etc.

Pour accéder à l'aide de la fonction `mean`, par exemple, il vous suffit de saisir directement :

```
R> help("mean")
```

Ou sa forme abrégée `?mean`.

Chaque page d'aide comprend plusieurs sections, en particulier :

Description donne un résumé en une phrase de ce que fait la fonction

Usage indique la ou les manières de l'utiliser

Arguments détaille tous les arguments possibles et leur signification

Value indique la forme du résultat renvoyé par la fonction

Details apporte des précisions sur le fonctionnement de la fonction

Note pour des remarques éventuelles

References pour des références bibliographiques ou des URL associées

See Also *très utile*, renvoie vers d'autres fonctions semblables ou liées, ce qui peut être très utile pour découvrir ou retrouver une fonction dont on a oublié le nom

Examples série d'exemples d'utilisation

Les exemples peuvent être directement exécutés en utilisant la fonction `example` :

```
R> example(mean)
```

```
meanR> x <- c(0:10, 50)

meanR> xm <- mean(x)

meanR> c(xm, mean(x, trim = 0.10))
[1] 8.75 5.50
```

9.1.2 Naviguer dans l'aide

La fonction `help.start` permet d'afficher le contenu de l'aide en ligne au format HTML dans votre navigateur Web. Pour comprendre ce que cela signifie, saisissez simplement :

```
R> help.start()
```

Ceci devrait lancer votre navigateur favori et afficher une page vous permettant alors de naviguer parmi les différentes extensions installées, d'afficher les pages d'aide des fonctions, de consulter les manuels, d'effectuer des recherches, etc.

À noter qu'à partir du moment où vous avez lancé `help.start()`, les pages d'aide demandées avec `help("lm")` ou `?plot` s'afficheront désormais dans votre navigateur.

Si vous souhaitez rechercher quelque chose dans le contenu de l'aide directement dans la console, vous pouvez utiliser la fonction `help.search` (ou `??` qui est équivalente), qui renvoie une liste des pages d'aide contenant les termes recherchés. Par exemple :

```
R> help.search("logistic")  # equivalent a ??logistic
```

Dans RStudio, les pages d'aide en ligne s'ouvriront dans le quadrant bas-droite sous l'onglet *Help*. Un clic sur l'icône en forme de maison vous affichera la page d'accueil de l'aide.

9.2 Ressources sur le Web

De nombreuses ressources existent en ligne, mais la plupart sont en anglais.

9.2.1 Aide en ligne

Le site R documentation propose un accès clair et rapide à la documentation de R et des extensions hébergées sur le CRAN. Il permet notamment de rechercher et naviguer facilement entre les pages des différentes fonctions :

http://www.rdocumentation.org/

9.2.2 Ressources officielles

La documentation officielle de R est accessible en ligne depuis le site du projet :

http://www.r-project.org/

Les liens de l'entrée *Documentation* du menu de gauche vous permettent d'accéder à différentes ressources.

Les manuels sont des documents complets de présentation de certains aspects de R. Ils sont accessibles en ligne, ou téléchargeables au format PDF :

`http://cran.r-project.org/manuals.html`

On notera plus particulièrement *An introduction to R*, normalement destiné aux débutants, mais qui nécessite quand même un minimum d'aisance en informatique et en statistiques :

`http://cran.r-project.org/doc/manuals/R-intro.html`

R Data Import/Export explique notamment comment importer des données depuis d'autres logiciels :

`http://cran.r-project.org/doc/manuals/R-data.html`

Les FAQ regroupent des questions fréquemment posées et leurs réponses. À lire donc ou au moins à parcourir avant toute chose :

`http://cran.r-project.org/faqs.html`

La FAQ la plus utile est la FAQ généraliste sur R :

`http://cran.r-project.org/doc/FAQ/R-FAQ.html`

Mais il existe également une FAQ dédiée aux questions liées à Windows, et une autre à la plateforme Mac OS X.

Les manuels et les FAQ sont accessibles même si vous n'avez pas d'accès à Internet en utilisant la fonction `help.start()` décrite précédemment.

R-announce est la liste de diffusion électronique officielle du projet. Elle ne comporte qu'un nombre réduit de messages (quelques-uns par mois tout au plus) et diffuse les annonces concernant de nouvelles versions de R ou d'autres informations particulièrement importantes. On peut s'y abonner à l'adresse suivante :

`https://stat.ethz.ch/mailman/listinfo/r-announce`

R Journal est la revue officielle du projet R, qui a succédé début 2009 à la lettre de nouvelles *R News*. Elle paraît entre deux et cinq fois par an et contient des informations sur les nouvelles versions du logiciel, des articles présentant des extensions, des exemples d'analyse... Les parutions sont annoncées sur la liste de diffusion *R-announce*, et les numéros sont téléchargeables à l'adresse suivante :

`http://journal.r-project.org/`

Autres documents On trouvera de nombreux documents dans différentes langues, en général au format PDF, dans le répertoire suivant :

`http://cran.r-project.org/doc/contrib/`

Parmi ceux-ci, les cartes de référence peuvent être très utiles, ce sont des aides-mémoire recensant les fonctions les plus courantes :

`http://cran.r-project.org/doc/contrib/Short-refcard.pdf`

On notera également un document d'introduction en anglais progressif et s'appuyant sur des méthodes statistiques relativement simples :

`http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`

9.2.3 Revue

La revue *Journal of Statistical Software* est une revue électronique anglophone, dont les articles sont en accès libre, et qui traite de l'utilisation de logiciels d'analyse de données dans un grand nombre de domaines. De nombreux articles (la majorité) sont consacrés à R et à la présentation d'extensions plus ou moins spécialisées.

Les articles qui y sont publiés prennent souvent la forme de tutoriels plus ou moins accessibles mais qui fournissent souvent une bonne introduction et une ressource riche en informations et en liens.

Adresse de la revue :

http://www.jstatsoft.org/

9.2.4 Ressources francophones

Il existe des ressources en français sur l'utilisation de R, mais peu sont réellement destinées aux débutants, elles nécessitent en général des bases à la fois en informatique et en statistique.

Le document le plus abordable et le plus complet est sans doute *R pour les débutants*, d'Emmanuel Paradis, accessible au format PDF :

http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf

La somme de documentation en français la plus importante liée à R est sans nulle doute celle mise à disposition par le *Pôle bioinformatique lyonnais*. Leur site propose des cours complets de statistique utilisant R :

http://pbil.univ-lyon1.fr/R/enseignement.html

La plupart des documents sont assez pointus niveau mathématique et plutôt orientés biostatistique, mais on trouvera des documents plus introductifs ici :

http://pbil.univ-lyon1.fr/R/html/cours1

Dans tous les cas la somme de travail et de connaissances mise à disposition librement est impressionnante...

Enfin, le site de Vincent Zoonekynd comprend de nombreuses notes prises au cours de sa découverte du logiciel. On notera cependant que l'auteur est normalien et docteur en mathématiques...

http://zoonek2.free.fr/UNIX/48_R_2004/all.html

9.2.5 Cours en ligne

De nombreux cours en ligne existent désormais pour l'apprentissage de R ou de méthodes statistiques en utilisant R.

Le site *France Université Numérique* propose plusieurs cours en français, dont un sur l'analyse de données multidimensionnelles :

https://www.fun-mooc.fr/

Le site *Datacamp* propose une formation en ligne en français gratuite, basée sur la pratique :

https://www.datacamp.com/courses/introduction-a-r/

En anglais, on trouvera des ressources sur R sur les principales plateformes de MOOCs. On pourra citer par exemple les cours de la spécialisation *Data science* sur Coursera :

https://www.coursera.org/specializations/jhu-data-science

Par ailleurs, *Swirl* est une extension qui propose des cours « interactifs » directement intégrés à R :

http://swirlstats.com/students.html

9.3 Où poser des questions

La communauté des utilisateurs de R est très active et en général très contente de pouvoir répondre aux questions (nombreuses) des débutants et à celles (tout aussi nombreuses) des utilisateurs plus expérimentés.

Dans tous les cas, les règles de base à respecter avant de poser une question sont toujours les mêmes : avoir cherché soi-même la réponse auparavant, notamment dans les FAQ et dans l'aide en ligne, et poser sa question de la manière la plus claire possible, de préférence avec un exemple de code posant problème.

9.3.1 Liste `R-soc`

Une liste de discussion a été créée spécialement pour permettre aide et échanges autour de l'utilisation de R en sciences sociales. Elle est hébergée par le CRU et on peut s'y abonner à l'adresse suivante :

https://groupes.renater.fr/sympa/subscribe/r-soc

Grâce aux services offerts par le site `gmane.org`, la liste est également disponible sous d'autres formes (forum Web, blog, `NNTP`, fils RSS) permettant de lire et de poster sans avoir à s'inscrire et à recevoir les messages sous forme de courrier électronique.

Pour plus d'informations :

http://dir.gmane.org/gmane.comp.lang.r.user.french

9.3.2 StackOverflow

Le site *StackOverflow* (qui fait partie de la famille des sites *StackExchange*) comprend une section (anglophone) dédiée à R qui permet de poser des questions et en général d'obtenir des réponses assez rapidement :

http://stackoverflow.com/questions/tagged/r

La première chose à faire, évidemment, est de vérifier que sa question n'a pas déjà été posée.

9.3.3 Forum Web en français

Le Cirad a mis en ligne un forum dédié aux utilisateurs de R, très actif :

http://forums.cirad.fr/logiciel-R/index.php

Les questions diverses et variées peuvent être posées dans la rubrique *Questions en cours* :

http://forums.cirad.fr/logiciel-R/viewforum.php?f=3

Il est tout de même conseillé de faire une recherche rapide sur le forum avant de poser une question, pour voir si la réponse ne s'y trouverait pas déjà.

9.3.4 Canal IRC (chat)

L'IRC, ou *Internet Relay Chat* est le vénérable ancêtre toujours très actif des messageries instantanées actuelles. Un canal (en anglais) est notamment dédié aux échanges autour de R (`#R`).

Si vous avez déjà l'habitude d'utiliser IRC, il vous suffit de pointer votre client préféré sur Freenode (`irc.freenode.net`) puis de rejoindre le canal en question.

Sinon, le plus simple est certainement d'utiliser l'interface Web de Mibbit, accessible à l'adresse :

http://www.mibbit.com/

Dans le champ *Connect to IRC*, sélectionnez *Freenode.net*, puis saisissez un pseudonyme dans le champ *Nick* et `#R` dans le champ *Channel*. Vous pourrez alors discuter directement avec les personnes présentes.

Le canal `#R` est normalement peuplé de personnes qui seront très heureuses de répondre à toutes les questions, et en général l'ambiance y est très bonne. Une fois votre question posée, n'hésitez pas à être patient et à attendre quelques minutes, voire quelques heures, le temps qu'un des habitués vienne y faire un tour.

9.3.5 Listes de discussion officielles

La liste de discussion d'entraide (par courrier électronique) officielle du logiciel R s'appelle `R-help`. On peut s'y abonner à l'adresse suivante, mais il s'agit d'une liste avec de nombreux messages :

`https://stat.ethz.ch/mailman/listinfo/r-help`

Pour une consultation ou un envoi ponctuels, le mieux est sans doute d'utiliser les interfaces Web fournies par gmane :

`http://blog.gmane.org/gmane.comp.lang.r.general`

`R-help` est une liste avec de nombreux messages, suivie par des spécialistes de R, dont certains des développeurs principaux. Elle est cependant à réserver aux questions particulièrement techniques qui n'ont pas trouvé de réponses par d'autres biais.

Dans tous les cas, il est nécessaire avant de poster sur cette liste de bien avoir pris connaissance du *posting guide* correspondant :

`http://www.r-project.org/posting-guide.html`

Plusieurs autres listes plus spécialisées existent également, elles sont listées à l'adresse suivante :

`http://www.r-project.org/mail.html`

Annexe A

Installer R

A.1 Installation de R sous Windows

Nous ne couvrons ici que l'installation de R sous Windows. Rappelons qu'en tant que logiciel libre, R est librement et gratuitement installable par quiconque.

La première chose à faire est de télécharger la dernière version du logiciel. Pour cela il suffit de se rendre à l'adresse suivante :

http://cran.r-project.org/bin/windows/base/release.htm

Vous allez alors vous voir proposer le téléchargement d'un fichier nommé `R-3.X.X-win.exe` (les `X` étant remplacés par les numéros de la dernière version disponible). Une fois ce fichier sauvegardé sur votre poste, exécutez-le et procédez à l'installation du logiciel : celle-ci s'effectue de manière tout à fait classique, c'est-à-dire en cliquant un certain nombre de fois [1] sur le bouton *Suivant*.

Une fois l'installation terminée, vous devriez avoir à la fois une magnifique icône R sur votre bureau ainsi qu'une non moins magnifique entrée R dans les programmes de votre menu *Démarrer*. Il ne vous reste donc plus qu'à lancer le logiciel pour voir à quoi il ressemble.

A.2 Installation de R sous Mac OS X

L'installation est très simple :

1. Se rendre à la page suivante : http://cran.r-project.org/bin/macosx/
2. Télécharger le fichier nommé `R-3.X.Y.pkg`
3. Procéder à l'installation.

A.3 Mise à jour de R sous Windows

La méthode conseillée pour mettre à jour R sur les plateformes Windows est la suivante [2] :

1. Désinstaller R. Pour cela on pourra utiliser l'entrée *Uninstall R* présente dans le groupe R du menu *Démarrer*.

1. Voire un nombre de fois certain. Vous pouvez laisser les options par défaut à chaque étape de l'installation.
2. Méthode conseillée dans l'entrée correspondante de la FAQ de R pour Windows : http://cran.r-project.org/bin/windows/rw-FAQ.html#What_0027s-the-best-way-to-upgrade_003f

2. Installer la nouvelle version comme décrit précédemment.
3. Se rendre dans le répertoire d'installation de R, en général `C:\Program Files\R`. Sélectionner le répertoire de l'ancienne installation de R et copier le contenu du dossier nommé `library` dans le dossier du même nom de la nouvelle installation. En clair, si vous mettez à jour de R 3.0.0 vers R 3.1.0, copiez tout le contenu du répertoire `C:\Program Files\R\R-3.0.0\library` dans `C:\Program Files\R\R-3.1.0\library`.
4. Lancez la nouvelle version de R et exécuter la commande `update.packages` pour mettre à jour les extensions.

A.4 Interfaces graphiques

L'interface par défaut sous Windows est celle présentée figure 2.1 page 9. Il en existe d'autres, plus ou moins sophistiquées, qui vont de la simple coloration syntaxique à des interfaces plus complètes se rapprochant de modèles du type SPSS.

On pourra notamment regarder l'extension R Commander, qui propose une interface graphique intégrée à R pour certaines fonctions de traitement et d'analyse de données :

http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/

Ou encore RKWard, un logiciel multiplateforme proposant une interface assez complète et des fonctions d'export de résultats :

http://rkward.sf.net/

Cependant, le projet RStudio tend à s'imposer comme l'environnement de développement de référence pour R, d'autant qu'il a l'avantage d'être libre, gratuit et multiplateforme. Son installation est décrite section A.5 de la présente page.

Au final, ce document se basant toujours sur une utilisation de R basée sur la saisie de commandes textuelles, l'interface choisie importe peu. Celles-ci ne diffèrent que par le niveau de confort ou d'efficacité supplémentaires qu'elles apportent.

A.5 RStudio

RStudio est un environnement de développement intégré libre, gratuit, et qui fonctionne sour Windows, Mac OS X et Linux. Il fournit un éditeur de script avec coloration syntaxique, des fonctionnalités pratiques d'édition et d'exécution du code, un affichage simultané du code, de la console R, des fichiers, graphiques et pages d'aide, une gestion des extensions, une intégration avec des systèmes de contrôle de versions comme git, etc.

Il est en développement actif et de nouvelles fonctionnalités sont ajoutées régulièrement. Son seul défaut est d'avoir une interface uniquement anglophone.

Pour avoir un aperçu de l'interface de RStudio, on pourra se référer à la page *Screenshots* du site du projet :

http://www.rstudio.com/ide/screenshots/

L'installation de RStudio est très simple, il suffit de se rendre sur la page de téléchargement et de sélectionner le fichier correspondant à son système d'exploitation :

http://www.rstudio.com/ide/download/

L'installation s'effectue ensuite de manière tout à fait classique.

NB : il est préférable d'installer d'abord R avant de procéder à l'installation de RStudio.

La documentation de RStudio (en anglais) est disponible en ligne à :

http://www.rstudio.com/ide/docs/

Annexe B

Extensions

B.1 Présentation

L'installation par défaut du logiciel R contient le cœur du programme ainsi qu'un ensemble de fonctions de base fournissant un grand nombre d'outils de traitement de données et d'analyse statistiques.

R étant un logiciel libre, il bénéficie d'une forte communauté d'utilisateurs qui peuvent librement contribuer au développement du logiciel en lui ajoutant des fonctionnalités supplémentaires. Ces contributions prennent la forme d'extensions (*packages*) pouvant être installées par l'utilisateur et fournissant alors diverses fonctions supplémentaires.

Il existe un très grand nombre d'extensions (environ 1500 à ce jour), qui sont diffusées par un réseau baptisé CRAN (*Comprehensive R Archive Network*).

La liste de toutes les extensions disponibles sur le CRAN est disponible ici :

`http://cran.r-project.org/web/packages/`

Pour faciliter un peu le repérage des extensions, il existe un ensemble de regroupements thématiques (économétrie, finance, génétique, données spatiales...) baptisés *Task views* :

`http://cran.r-project.org/web/views/`

On y trouve notamment une *Task view* dédiée aux sciences sociales, listant de nombreuses extensions potentiellement utiles pour les analyses statistiques dans ce champ disciplinaire :

`http://cran.r-project.org/web/views/SocialSciences.html`

B.2 Installation des extensions

Les interfaces graphiques sous Windows ou Mac OS X permettent la gestion des extensions par le biais de boîtes de dialogues (entrées du menu *Packages* sous Windows par exemple). Nous nous contenterons ici de décrire cette gestion *via* la console.

On notera cependant que l'installation et la mise à jour des extensions nécessite d'être connecté à l'Internet.

L'installation d'une extension se fait par la fonction `install.packages`, à qui on fournit le nom de l'extension. Ici on souhaite installer l'extension ade4 :

```
R> install.packages("ade4", dep = TRUE)
```

L'option `dep=TRUE` indique à R de télécharger et d'installer également toutes les extensions dont l'extension choisie dépend pour son fonctionnement.

En général R va alors vous demander de choisir un *miroir* depuis lequel récupérer les données nécessaires. Le plus simple est de sélectionner le premier miroir de la liste, baptisé `0-cloud`, qui est une redirection automatique fournie par les éditeurs de RStudio.

Une fois l'extension installée, elle peut être appelée depuis la console ou un fichier script avec la commande :

```
R> library(ade4)
```

À partir de là, on peut utiliser les fonctions de l'extension, consulter leur page d'aide en ligne, accéder aux jeux de données qu'elle contient, etc.

Pour mettre à jour l'ensemble des extensions installées, une seule commande suffit :

```
R> update.packages()
```

Si on souhaite désinstaller une extension précédemment installée, on peut utiliser la fonction `remove.packages` :

```
R> remove.packages("ade4")
```

Il est important de bien comprendre la différence entre `install.packages` et `library`. La première va chercher les extensions sur l'Internet et les installe en local sur le disque dur de l'ordinateur. On n'a besoin d'effectuer cette opération qu'une seule fois. La seconde lit les informations de l'extension sur le disque dur et les met à disposition de R. On a besoin de l'exécuter à chaque début de session ou de script.

B.3 L'extension questionr

questionr est une extension pour R comprenant quelques fonctions potentiellement utiles pour l'utilisation du logiciel en sciences sociales, ainsi que différents jeux de données. Elle est développée en collaboration avec François Briatte et Joseph Larmarange.

B.3.1 Installation

L'installation nécessite d'avoir une connexion active à Internet. L'extension est hébergée sur le CRAN (*Comprehensive R Archive Network*), le réseau officiel de diffusion des extensions de R. Elle est donc installable de manière très simple, comme n'importe quelle autre extension, par un simple :

```
R> install.packages("questionr", dep = TRUE)
```

Si vous souhaitez utiliser la toute dernière version en ligne sur Github, vous pouvez utiliser la fonction `install_github`, de l'extension devtools :

```
R> install_github("juba/questionr")
```

L'extension s'utilise alors de manière classique grâce à l'instruction `library` en début de session ou de fichier R :

```
R> library(questionr)
```

B.3.2 Fonctions et utilisation

Pour plus de détails sur la liste des fonctions de l'extension et son utilisation, on pourra se reporter aux pages Web de l'extension, hébergées sur Github :

https://github.com/juba/questionr

Un document PDF (en anglais) regroupant les pages d'aide en ligne de l'extension est disponible sur le CRAN :

http://cran.r-project.org/web/packages/questionr/questionr.pdf

Le même document est également accessible en ligne sur R documentation :

http://www.rdocumentation.org/packages/questionr

B.3.3 Interfaces interactives

questionr introduit depuis la version 0.3 des *interfaces interactives*, c'est-à-dire des interfaces dynamiques permettant de générer le code R correspondant à des opérations courantes. Ces interfaces sont en fait des applications Web [1] qui s'ouvrent soit directement dans RStudio, soit dans le navigateur Web de l'utilisateur.

L'exemple suivant permettra de mieux comprendre ce dont il s'agit : après installation de questionr, exécutez le code suivant dans une session R :

```
R> library(questionr)
R> data(hdv2003)
```

À partir de là, vous pouvez soit aller dans le menu *Addins* et choisir *Levels recoding*, soit exécuter le code suivant dans la console :

```
R> irec(hdv2003$sexe)
```

Une interface semblable à celle de la figure B.1, page suivante devrait alors s'afficher soit directement dans RStudio, soit dans votre navigateur Web.

La fonction `irec` a pour objet de faciliter le recodage d'une variable qualitative. Ici, nous souhaitons recoder la variable `sexe` du jeu de données `hdv2003`.

L'interface se divise en trois onglet :

— dans l'onglet *Variable et paramètres*, vous pouvez sélectionner la variable à recoder, choisir le nom de la nouvelle variable, le style de recodage, et indiquer si vous souhaitez que cette nouvelle variable soit convertie en facteur.
— dans l'onglet *Recodages*, des champs de formulaires listent les valeurs de la variable et vous permettent de les modifier.

1. Applications Web développées grâce au *framework* Shiny : http://www.rstudio.com/shiny/.

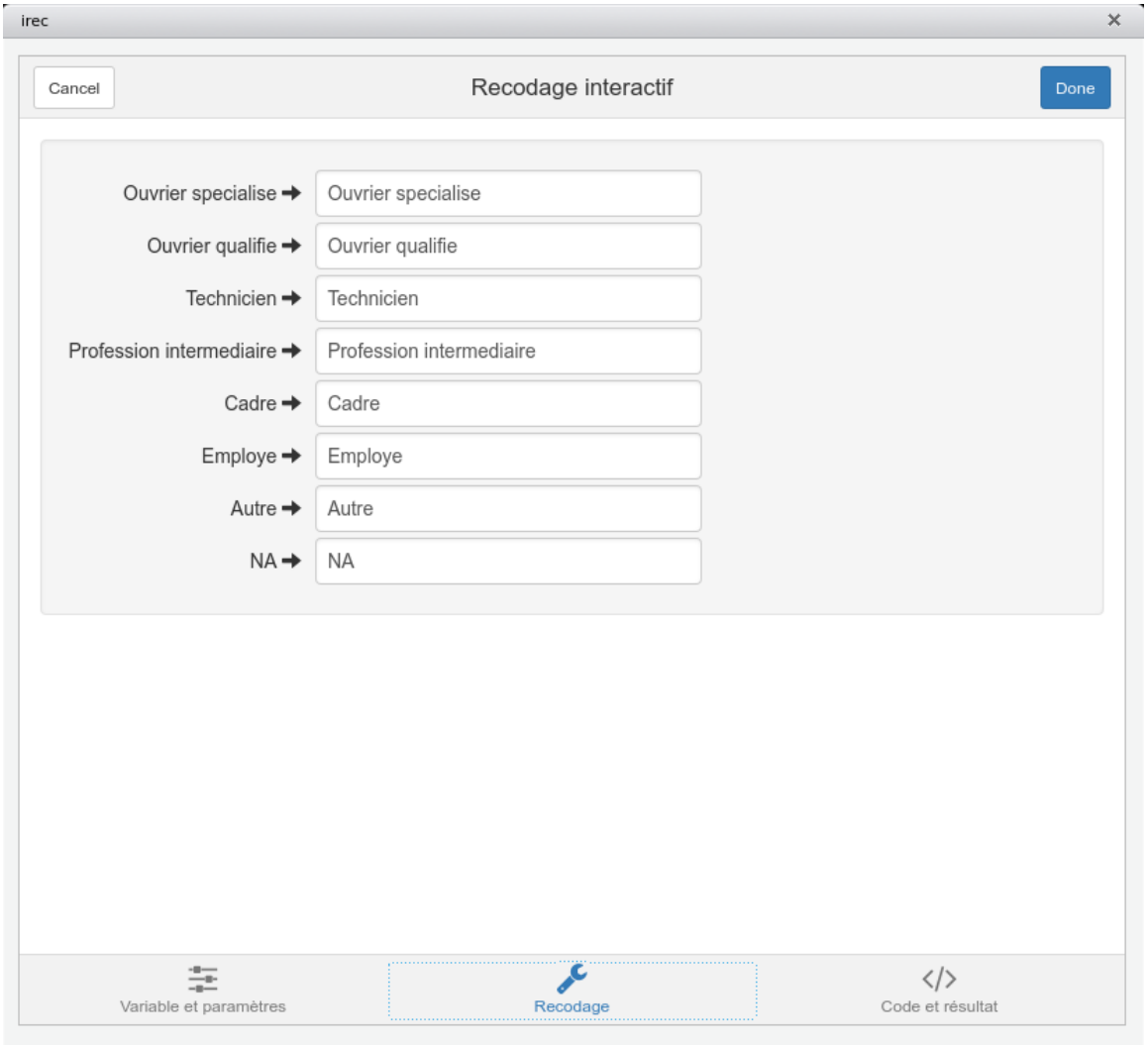


FIGURE B.1 – Interface de la commande `irec`

— enfin, dans l'onglet *Code et résultat*, vous trouverez à la fois le code R correspondant aux données remplies dans le formulaire précédent, ainsi qu'un tableau croisé de l'ancienne et de la nouvelle variable vous permettant de vérifier que le résultat est bien celui attendu.

Vous remarquerez que le code généré se met à jour automatiquement, au fur et à mesure que vous changez les options ou que vous modifiez des valeurs.

Il est important de bien comprendre que cette interface ne modifie jamais vos données et n'exécute pas le code qu'elle génère ! Vous pouvez donc expérimenter librement, par contre c'est-à-vous de copier le code généré, de l'insérer dans votre script et de l'exécuter.

Une fois votre recodage terminé, cliquez sur le bouton *Done*. Si vous êtes sous RStudio, le code R généré est alors directement inséré dans votre script, à l'emplacement actuel du curseur. Sinon, il est affiché dans la console et vous pouvez donc le copier/coller pour l'inclure dans votre script. À noter que vous devez alors exécuter le code vous-mêmes si vous souhaitez que votre recodage soit appliqué.

Le fonctionnement est le même pour les autres fonctions proposant une interface interactive, notamment `iorder` pour réordonner les niveaux d'un facteur ou `icut` pour découper une variable numérique en classes.

B.3.4 Le jeu de données `hdv2003`

L'extension questionr contient plusieurs jeux de données (*dataset*) destinés à l'apprentissage de R.

`hdv2003` est un extrait comportant 2000 individus et 20 variables provenant de l'enquête *Histoire de Vie* réalisée par l'INSEE en 2003.

L'extrait est tiré du fichier détail mis à disposition librement (ainsi que de nombreux autres) par l'INSEE à l'adresse suivante :

http://www.insee.fr/fr/themes/detail.asp?ref_id=fd-HDV03

Les variables retenues ont été parfois partiellement recodées. La liste des variables est la suivante :

Variable	Description
`id`	Identifiant (numéro de ligne)
`poids`	Variable de pondération [2]
`age`	Âge
`sexe`	Sexe
`nivetud`	Niveau d'études atteint
`occup`	Occupation actuelle
`qualif`	Qualification de l'emploi actuel
`freres.soeurs`	Nombre total de frères, sœurs, demi-frères et demi-sœurs
`clso`	Sentiment d'appartenance à une classe sociale
`relig`	Pratique et croyance religieuse
`trav.imp`	Importance accordée au travail
`trav.satisf`	Satisfaction ou insatisfaction au travail
`hard.rock`	Ecoute du Hard rock ou assimilés
`lecture.bd`	Lecture de bandes dessinées
`peche.chasse`	Pêche ou chasse pour le plaisir au cours des 12 derniers mois
`cuisine`	Cuisine pour le plaisir au cours des 12 derniers mois
`bricol`	Bricolage ou mécanique pour le plaisir au cours des 12 derniers mois
`cinema`	Cinéma au cours des 12 derniers mois
`sport`	Sport ou activité physique pour le plaisir au cours des 12 derniers mois
`heures.tv`	Nombre moyen d'heures passées àregarder la télévision par jour

B.3.5 Le jeu de données `rp99`

`rp99` est issu du recensement de la population de 1999 de l'INSEE. Il comporte une petite partie des résultats pour l'ensemble des communes du Rhône, soit 301 lignes et 21 colonnes

La liste des variables est la suivante :

Variable	Description
`nom`	nom de la commune
`code`	Code de la commune
`pop.act`	Population active
`pop.tot`	Population totale
`pop15`	Population des 15 ans et plus
`nb.rp`	Nombre de résidences principales
`agric`	Part des agriculteurs dans la population active
`artis`	Part des artisans, commerçants et chefs d'entreprises
`cadres`	Part des cadres
`interm`	Part des professions intermédiaires
`empl`	Part des employés
`ouvr`	Part des ouvriers
`retr`	Part des retraités
`tx.chom`	Part des chômeurs
`etud`	Part des étudiants
`dipl.sup`	Part des diplômés du supérieur
`dipl.aucun`	Part des personnes sans diplôme
`proprio`	Part des propriétaires parmi les résidences principales
`hlm`	Part des logements HLM parmi les résidences principales
`locataire`	Part des locataires parmi les résidences principales
`maison`	Part des maisons parmi les résidences principales

2. Comme il s'agit d'un extrait du fichier, cette variable de pondération n'a en toute rigueur aucune valeur statistique. Elle a été tout de même incluse à des fins « pédagogiques ».

Annexe C

Solutions des exercices

Exercice 2.1, page 17

```
R> c(120, 134, 256, 12)

[1] 120 134 256  12
```

Exercice 2.2, page 17

```
R> 1:9

[1] 1 2 3 4 5 6 7 8 9

R> 1:9 + 100

[1] 101 102 103 104 105 106 107 108 109

R> 1:9 * 2

[1]  2  4  6  8 10 12 14 16 18
```

Exercice 2.3, page 17

```
R> conjoint1 <- c(1200, 1180, 1750, 2100)
R> conjoint2 <- c(1450, 1870, 1690, 0)
R> nb.personnes <- c(4, 2, 3, 2)
R> (conjoint1 + conjoint2)/nb.personnes

[1]  662.500 1525.000 1146.667 1050.000
```

Exercice 2.4, page 17

```
R> conjoint1 <- c(1200, 1180, 1750, 2100)
R> min(conjoint1)

[1] 1180

R> max(conjoint1)

[1] 2100

R> conjoint1.na <- c(1200, 1180, 1750, NA)
R> min(conjoint1.na)

[1] NA

R> max(conjoint1.na)

[1] NA

R> min(conjoint1.na, na.rm = TRUE)

[1] 1180

R> max(conjoint1.na, na.rm = TRUE)

[1] 1750
```

Exercice 3.5, page 35

```
R> library(questionr)
R> data(hdv2003)
R> df <- hdv2003
R> str(df)

'data.frame': 2000 obs. of  20 variables:
 $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age          : int  28 23 59 34 71 35 60 47 20 28 ...
 $ sexe         : Factor w/ 2 levels "Homme","Femme": 2 2 1 1 2 2 2 1 2 1 ...
 $ nivetud      : Factor w/ 8 levels "N'a jamais fait d'etudes",..: 8 NA 3 8 3 6 3 6 NA 7 ...
 $ poids        : num  2634 9738 3994 5732 4329 ...
 $ occup        : Factor w/ 7 levels "Exerce une profession",..: 1 3 1 1 4 1 6 1 3 1 ...
 $ qualif       : Factor w/ 7 levels "Ouvrier specialise",..: 6 NA 3 3 6 6 2 2 NA 7 ...
 $ freres.soeurs: int  8 2 2 1 0 5 1 5 4 2 ...
 $ clso         : Factor w/ 3 levels "Oui","Non","Ne sait pas": 1 1 2 2 1 2 1 2 1 2 ...
 $ relig        : Factor w/ 6 levels "Pratiquant regulier",..: 4 4 4 3 1 4 3 4 3 2 ...
 $ trav.imp     : Factor w/ 4 levels "Le plus important",..: 4 NA 2 3 NA 1 NA 4 NA 3 ...
 $ trav.satisf  : Factor w/ 3 levels "Satisfaction",..: 2 NA 3 1 NA 3 NA 2 NA 1 ...
 $ hard.rock    : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 1 ...
 $ lecture.bd   : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 1 ...
 $ peche.chasse : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 2 2 1 1 ...
```

```
 $ cuisine      : Factor w/ 2 levels "Non","Oui": 2 1 1 2 1 1 2 2 1 1 ...
 $ bricol       : Factor w/ 2 levels "Non","Oui": 1 1 1 2 1 1 1 2 1 1 ...
 $ cinema       : Factor w/ 2 levels "Non","Oui": 1 2 1 2 1 2 1 1 2 2 ...
 $ sport        : Factor w/ 2 levels "Non","Oui": 1 2 2 2 1 2 1 1 1 2 ...
 $ heures.tv    : num  0 1 0 2 3 2 2.9 1 2 2 ...
```

Exercice 3.6, page 35

Utilisez la fonction suivante et corrigez manuellement les erreurs :

```
R> df.ok <- edit(df)
```

Attention à ne pas utiliser la fonction `fix` dans ce cas, celle-ci modifierait directement le contenu de `df`.

Puis utilisez la fonction `head` :

```
R> head(df.ok, 4)
```

Exercice 3.7, page 40

```
R> summary(df$age)
R> hist(df$age, breaks = 10, main = "Répartition des âges", xlab = "Âge", ylab = "Effectif")
R> boxplot(df$age)
R> plot(table(df$age), main = "Répartition des âges", xlab = "Âge", ylab = "Effectif")
```

Exercice 3.8, page 40

```
R> table(df$trav.imp)
R> summary(df$trav.imp)
R> freq(df$trav.imp)
R> dotchart(table(df$trav.imp))
```

Exercice 4.9, page 47

Utilisez la fonction `read.table` ou l'un de ses dérivés, en fonction du tableur utilisé et du format d'enregistrement.

Pour vérifier que l'importation s'est bien passée, on peut utiliser les fonctions `str`, `dim`, éventuellement `edit` et faire quelques tris à plat.

Exercice 4.10, page 47

Utilisez la fonction `read.dbf` de l'extension foreign.

Exercice 5.11, page 81

```
R> library(questionr)
R> data(hdv2003)
R> d <- hdv2003
R> d <- rename.variable(d, "clso", "classes.sociales")
R> d <- rename.variable(d, "classes.sociales", "clso")
```

Exercice 5.12, page 81

```
R> d$clso <- factor(d$clso, levels = c("Non", "Ne sait pas", "Oui"))
R> table(d$clso)


        Non Ne sait pas         Oui
       1037          27         936
```

Exercice 5.13, page 82

```
R> d$cinema[1:3]

[1] Non Oui Non
Levels: Non Oui

R> d$lecture.bd[12:30]

 [1] Non Non Non Non Non Non Non Non Non Non Non Non Non Non Non Non Non Non Non
[18] Non Non
Levels: Non Oui

R> d[c(5, 12), c(4, 8)]

                             nivetud freres.soeurs
5  Derniere annee d'etudes primaires             0
12                        2eme cycle             4

R> longueur <- length(d$age)
R> tail(d$age, 4)

[1] 46 24 24 66
```

Exercice 5.14, page 82

```
R> subset(d, lecture.bd == "Oui", select = c(age, sexe))
R> subset(d, occup != "Chômeur", select = -cinema)
R> subset(d, age >= 45 & hard.rock == "Oui", select = id)
R> subset(d, sexe == "Femme" & age >= 25 & age <= 40 & sport == "Non")
R> subset(d, sexe == "Homme" & freres.soeurs >= 2 & freres.soeurs <= 4 & (cuisine ==
+     "Oui" | bricol == "Oui"))
```

Exercice 5.15, page 82

```
R> d.bd.oui <- subset(d, lecture.bd == "Oui")
R> d.bd.non <- subset(d, lecture.bd == "Non")
R> mean(d.bd.oui$heures.tv)

[1] 1.76383

R> mean(d.bd.non$heures.tv, na.rm = TRUE)

[1] 2.258214

R> tapply(d$heures.tv, d$lecture.bd, mean, na.rm = TRUE)

     Non      Oui
2.258214 1.763830
```

Exercice 5.16, page 82

```
R> d$fs.char <- as.character(d$freres.soeurs)
R> d$fs.fac <- factor(d$fs.char)
R> d$fs.num <- as.numeric(as.character(d$fs.char))
R> table(d$fs.num == d$freres.soeurs)


TRUE
2000
```

Exercice 5.17, page 82

```
R> d$fs1 <- cut(d$freres.soeurs, 5)
R> table(d$fs1)


(-0.022,4.4]    (4.4,8.8]   (8.8,13.2]  (13.2,17.6]    (17.6,22]
        1495          396           97            9            3

R> d$fs2 <- cut(d$freres.soeurs, breaks = c(0, 2, 4, 19), include.lowest = TRUE,
+     labels = c("de 0 à 2", "de 2 à 4", "plus de 4"))
R> table(d$fs2)


 de 0 à 2  de 2 à 4 plus de 4
     1001       494       504

R> d$fs3 <- quant.cut(d$freres.soeurs, 3)
R> table(d$fs3)


 [0,2)  [2,4) [4,22]
   574    711    715
```

Exercice 5.18, page 82

```
R> d$trav.imp2cl[d$trav.imp == "Le plus important" | d$trav.imp == "Aussi important que le reste"] <- "Le plus ou aussi impo
R> d$trav.imp2cl[d$trav.imp == "Moins important que le reste" | d$trav.imp == "Peu important"] <- "moins ou peu important"
R> table(d$trav.imp)


           Le plus important Aussi important que le reste
                          29                         259
Moins important que le reste               Peu important
                         708                          52


R> table(d$trav.imp2cl)


Le plus ou aussi important     moins ou peu important
                       288                        760


R> table(d$trav.imp, d$trav.imp2cl)


                             Le plus ou aussi important
  Le plus important                                  29
  Aussi important que le reste                      259
  Moins important que le reste                        0
  Peu important                                       0

                             moins ou peu important
  Le plus important                                0
  Aussi important que le reste                     0
  Moins important que le reste                   708
  Peu important                                   52
```

```
R> d$relig.4cl <- as.character(d$relig)
R> d$relig.4cl[d$relig == "Pratiquant regulier" | d$relig == "Pratiquant occasionnel"] <- "Pratiquant"
R> d$relig.4cl[d$relig == "NSP ou NVPR"] <- NA
R> table(d$relig.4cl, d$relig, exclude = NULL)


                           Pratiquant regulier Pratiquant occasionnel
  Appartenance sans pratique                   0                     0
  Ni croyance ni appartenance                  0                     0
  Pratiquant                                 266                   442
  Rejet                                        0                     0
  <NA>                                         0                     0

                           Appartenance sans pratique
  Appartenance sans pratique                        760
  Ni croyance ni appartenance                         0
  Pratiquant                                          0
  Rejet                                               0
  <NA>                                                0

                           Ni croyance ni appartenance Rejet
  Appartenance sans pratique                           0     0
  Ni croyance ni appartenance                        399     0
  Pratiquant                                           0     0
  Rejet                                                0    93
  <NA>                                                 0     0

                           NSP ou NVPR <NA>
  Appartenance sans pratique           0    0
  Ni croyance ni appartenance          0    0
```

```
  Pratiquant                          0    0
  Rejet                               0    0
  <NA>                               40    0
```

Exercice 5.19, page 83

Attention, l'ordre des opérations a toute son importance !

```
R> d$var <- "Autre"
R> d$var[d$sexe == "Femme" & d$bricol == "Oui"] <- "Femme faisant du bricolage"
R> d$var[d$sexe == "Homme" & d$age > 30] <- "Homme de plus de 30 ans"
R> d$var[d$sexe == "Homme" & d$age > 40 & d$lecture.bd == "Oui"] <- "Homme de plus de 40 ans lecteur de BD"
R> table(d$var)


                              Autre
                                925
          Femme faisant du bricolage
                                338
             Homme de plus de 30 ans
                                728
Homme de plus de 40 ans lecteur de BD
                                  9

R> table(d$var, d$sexe)


                                    Homme Femme
  Autre                               162   763
  Femme faisant du bricolage            0   338
  Homme de plus de 30 ans             728     0
  Homme de plus de 40 ans lecteur de BD     9     0

R> table(d$var, d$bricol)


                                    Non Oui
  Autre                             847  78
  Femme faisant du bricolage          0 338
  Homme de plus de 30 ans           298 430
  Homme de plus de 40 ans lecteur de BD   2   7

R> table(d$var, d$lecture.bd)


                                    Non Oui
  Autre                             905  20
  Femme faisant du bricolage        324  14
  Homme de plus de 30 ans           724   4
  Homme de plus de 40 ans lecteur de BD   0   9

R> table(d$var, d$age > 30)


                                    FALSE TRUE
  Autre                               283  642
  Femme faisant du bricolage           68  270
  Homme de plus de 30 ans               0  728
  Homme de plus de 40 ans lecteur de BD     0    9

R> table(d$var, d$age > 40)
```

```
                                  FALSE TRUE
  Autre                             417  508
  Femme faisant du bricolage        152  186
  Homme de plus de 30 ans           163  565
  Homme de plus de 40 ans lecteur de BD   0    9
```

Exercice 5.20, page 83

```
R> d.ord <- d[order(d$freres.soeurs), ]
R> d.ord <- d[order(d$heures.tv, decreasing = TRUE), c("sexe", "heures.tv")]
R> head(d.ord, 10)

      sexe heures.tv
288  Femme        12
391  Femme        12
1324 Homme        11
1761 Femme        11
100  Femme        10
236  Femme        10
421  Homme        10
426  Femme        10
841  Femme        10
1075 Homme        10
```

Table des figures

Index des fonctions