

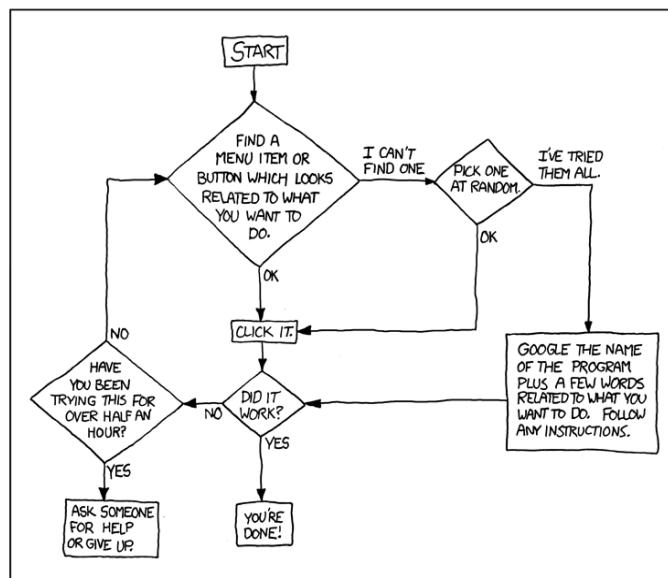
Introduction à l'analyse d'enquêtes avec

à partir d'un document original de
Julien Barnier
julien.barnier@ens-lyon.fr

complété par
Joseph Larmarange
joseph.larmarange@ceped.org

4 novembre 2013

WE DON'T MAGICALLY KNOW HOW TO DO EVERYTHING IN EVERY PROGRAM. WHEN WE HELP YOU, WE'RE USUALLY JUST DOING THIS:



<http://xkcd.com/627/>

Table des matières

1	Introduction	6
1.1	À propos de ce document	6
1.2	Licence	6
1.3	Remerciements	6
1.4	Conventions typographiques	6
1.5	Présentation de R	7
1.6	Philosophie de R	8
2	Prise en main	9
2.1	L'invite de commandes	9
2.2	Des objets	12
2.2.1	Objets simples	12
2.2.2	Vecteurs	13
2.3	Des fonctions	15
2.3.1	Arguments	16
2.3.2	Quelques fonctions utiles	17
2.3.3	Aide sur une fonction	17
2.4	Exercices	18
3	Premier travail avec des données	20
3.1	Regrouper les commandes dans des scripts	20
3.2	Ajouter des commentaires	21
3.3	Tableaux de données	22
3.4	Inspecter les données	23
3.4.1	Structure du tableau	23
3.4.2	Inspection visuelle	24
3.4.3	Accéder aux variables	25
3.5	Analyser une variable	26
3.5.1	Variable quantitative	26
3.5.2	Variable qualitative	34
3.6	Exercices	42
4	Import/export de données	43
4.1	Accès aux fichiers et répertoire de travail	43
4.2	Import de données depuis un tableur	44
4.2.1	Depuis Excel	45
4.2.2	Depuis OpenOffice ou LibreOffice	46
4.2.3	Autres sources / en cas de problèmes	46
4.3	Import depuis d'autres logiciels	47
4.3.1	SAS	47
4.3.2	SPSS	47
4.3.3	Stata	47

4.3.4 Fichiers <code>dbf</code>	48
4.4 Autres sources	48
4.5 Sauver ses données	48
4.6 Exporter des données	49
4.7 Exercices	49
5 Manipulation de données	50
5.1 Variables	50
5.1.1 Types de variables	50
5.1.2 Renommer des variables	51
5.1.3 Facteurs	52
5.2 Indexation	54
5.2.1 Indexation directe	54
5.2.2 Indexation par nom	56
5.2.3 Indexation par conditions	58
5.2.4 Indexation et assignation	64
5.3 Sous-populations	65
5.3.1 Par indexation	65
5.3.2 Fonction <code>subset</code>	66
5.3.3 Fonction <code>tapply</code>	67
5.4 Recodages	68
5.4.1 Convertir une variable	68
5.4.2 Découper une variable numérique en classes	69
5.4.3 Regrouper les modalités d'une variable	71
5.4.4 Variables calculées	73
5.4.5 Combiner plusieurs variables	73
5.4.6 Variables scores	74
5.4.7 Vérification des recodages	75
5.5 Tri de tables	75
5.6 Fusion de tables	77
5.7 Organiser ses scripts	80
5.8 Exercices	82
6 Statistique bivariée	84
6.1 Deux variables quantitatives	84
6.2 Une variable quantitative et une variable qualitative	89
6.3 Deux variables qualitatives	96
6.3.1 Tableau croisé	96
6.3.2 χ^2 et dérivés	98
6.3.3 Représentation graphique	99
7 Régression logistique	103
7.1 Préparation des données	103
7.2 Régression logistique binaire	106
7.3 Sélection de modèles	112
7.4 Régression logistique multinomiale	113
7.5 Exercices	117
8 Données pondérées	118
8.1 Options de certaines fonctions	118
8.2 Fonctions de l'extension <code>questionr</code>	118
8.3 Présentation de l'extension <code>survey</code>	119
8.4 Définir un plan d'échantillonage complexe avec <code>survey</code>	124
8.4.1 Différents types d'échantillonnage	124

8.4.2	Les options de <code>svydesign</code>	124
8.4.3	Extraire un sous-échantillon	126
8.5	Conclusion	126
9	Analyse des correspondances multiples (ACM)	127
9.1	Principe général	127
9.2	ACM avec <code>ade4</code>	128
9.3	ACM avec FactoMineR	132
10	Classification ascendante hiérarchique (CAH)	138
10.1	Calculer une matrice des distances	138
10.1.1	Distance de Gower	139
10.1.2	Distance du Φ^2	140
10.1.3	Exemple	140
10.2	Calcul du dendrogramme	140
10.3	Découper le dendrogramme	142
10.4	CAH avec l'extension FactoMineR	144
11	Analyse de séquences	148
11.1	L'analyse de séquences	148
11.2	Installer TraMineR et récupérer les données	149
11.3	Appariement optimal et classification	151
11.4	Représentations graphiques	154
11.5	Bibliographie	162
12	Analyse de survie	164
13	Exporter les résultats	165
13.1	Export manuel de tableaux	165
13.1.1	Copier/coller vers Excel et Word via le presse-papier	165
13.1.2	Export vers Word ou OpenOffice/LibreOffice via un fichier	166
13.2	Export de graphiques	166
13.2.1	Export via l'interface graphique (Windows ou Mac OS X)	166
13.2.2	Export avec les commandes de R	167
13.3	Génération automatique de documents avec OpenOffice ou LibreOffice	168
13.3.1	Prérequis	168
13.3.2	Exemple	169
13.3.3	Utilisation	170
13.4	Génération automatique de documents avec knitr	172
13.4.1	Exemple	173
13.4.2	Syntaxe	173
13.4.3	Aller plus loin	175
14	Où trouver de l'aide	176
14.1	Aide en ligne	176
14.1.1	Aide sur une fonction	176
14.1.2	Naviguer dans l'aide	177
14.2	Ressources sur le Web	177
14.2.1	Moteur de recherche	177
14.2.2	Aide en ligne	178
14.2.3	Ressources officielles	178
14.2.4	Revue	179
14.2.5	Ressources francophones	179
14.3	Où poser des questions	180
14.3.1	Liste R-soc	180

14.3.2 StackOverflow	180
14.3.3 Forum Web en français	180
14.3.4 Canaux IRC (chat)	181
14.3.5 Listes de discussion officielles	181
A Installer R	182
A.1 Installation de R sous Windows	182
A.2 Installation de R sous Mac OS X	182
A.3 Mise à jour de R sous Windows	182
A.4 Interfaces graphiques	183
A.5 RStudio	183
B Extensions	185
B.1 Présentation	185
B.2 Installation des extensions	185
B.3 L'extension questionr	186
B.3.1 Installation	186
B.3.2 Fonctions et utilisation	187
B.3.3 Le jeu de données hdv2003	187
B.3.4 Le jeu de données rp99	188
C Solutions des exercices	189
Table des figures	200
Index des fonctions	203

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 a été réalisé avec R version 3.0.2 (2013-09-25).

Ce document est basé sur l'*Introduction à R* écrite par Julien Barnier et accessible sur <http://alea.fr.eu.org/pages/intro-R>. Il a été complété de plusieurs chapitres par Joseph Larmarange. Cette version modifiée est téléchargeable à l'adresse :

<https://github.com/larmarange/intro-r/blob/CoursM2/intro.pdf?raw=true>

1.2 Licence

Ce document est diffusé sous licence *Creative Commons Attribution - Pas d'utilisation commerciale - Partage dans les mêmes conditions* :



<https://creativecommons.org/licenses/by-nc-sa/3.0/fr/>

1.3 Remerciements

Julien Barnier tient à remercier Mayeul Kauffmann, Julien Biaudet, Frédérique Giraud, Joël Gombin et Joseph Larmarange 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 tient à remercier Julien Barnier pour avoir mis son travail sous licence *Creative Commons* et Nicolas Robette pour avoir autorisé la reproduction de son introduction à l'analyse de séquences (chapitre 11 page 148).

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, Li-

nux, 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.
-2.3100	-0.6780	0.0120	0.0282	0.6610	2.4600

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

```
R> coo <- scatterutil.base(dfxxy = 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 libre¹, publié sous licence GNU GPL.

L'utilisation de R présente plusieurs avantages :

- c'est un logiciel *multiplateforme*, qui fonctionne aussi bien sur des systèmes 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'extensions² ;
- 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 Modalisa³. 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.
- comme R s'apparente davantage à un langage de programmation qu'à un logiciel proprement dit, la courbe d'apprentissage peut être un peu « raide », notamment pour ceux n'ayant jamais programmé auparavant.

1. Pour plus d'informations sur ce qu'est un logiciel libre, voir : <http://www.gnu.org/philosophy/free-sw.fr.html>

2. Il en existe actuellement plus de 4800, disponibles sur le *Comprehensive R Archive Network* (CRAN) : <http://cran.r-project.org/>

3. Certaines extensions ou logiciels proposent cependant des interfaces graphiques plus ou moins généralistes. Voir la section A.4, page 183

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.

Partie 2

Prise en main

L'installation du logiciel proprement dite n'est pas décrite ici mais indiquée dans l'annexe A, page 182. 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 183.

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 suit¹ :

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 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 suivante montre l'interface par défaut sous Windows.

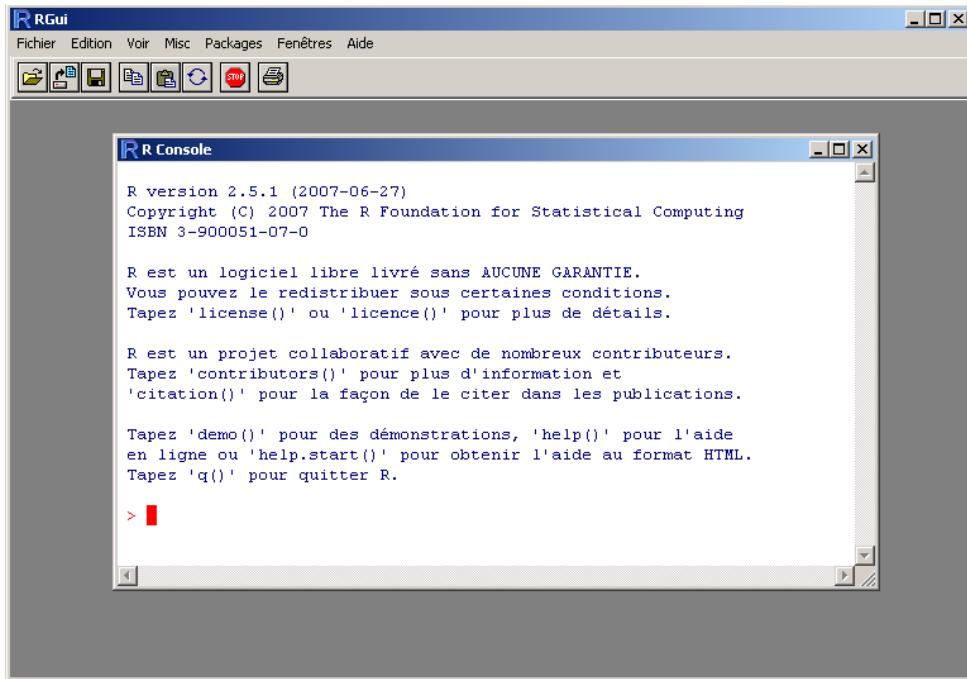


FIGURE 2.1 – L’interface de R sous Windows au démarrage

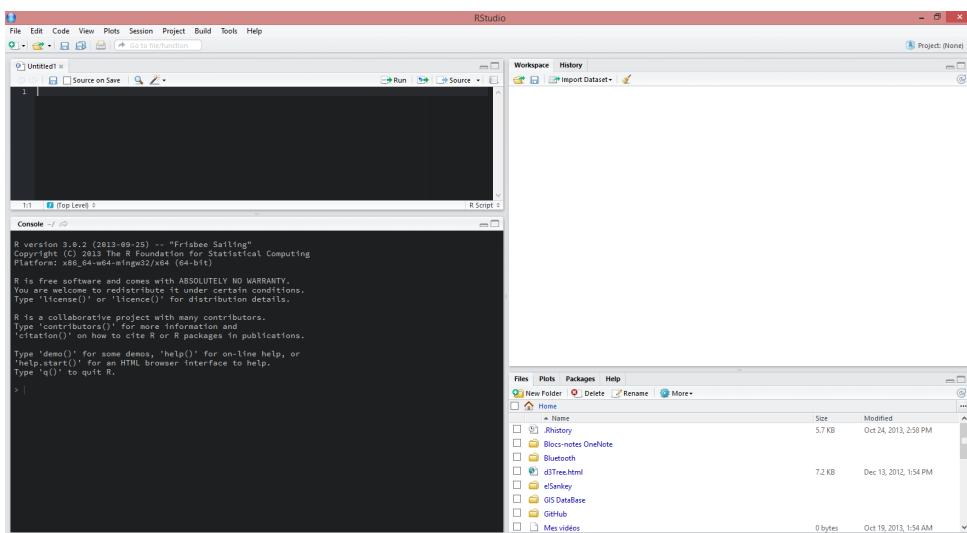


FIGURE 2.2 – L’interface de RStudio sous Windows au démarrage



L'interface de RStudio se présente différemment (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 chiffre². 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 14.

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 c³. 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'abréviation 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] 30.84 20.07 27.73 16.51 26.37 23.67 22.57 14.90 18.82 19.03
```



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 surprises⁴.

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 54.



Sous RStudio, vous avez du 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 `tailles`.

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] 76.71
```

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.

Reprendons 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] 171.3
```

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

On peut rajouter les fonctions de base suivantes :

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



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 14, page 176.



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 189*

Construire le vecteur suivant :

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

Exercice 2.2

▷ *Solution page 189*

Générez les vecteurs suivants chacun de deux manières différentes :

```
[1] 1 2 3 4
[1] 1 2 3 4 8 9 10 11
[1] 2 4 6 8
```

Exercice 2.3

▷ *Solution page 189*

On a demandé à 4 ménages le revenu du chef de ménage, celui de son conjoint, et le nombre de personnes du ménage :

```
R> chef <- c(1200, 1180, 1750, 2100)
R> conjoint <- 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 190*

Dans l'exercice précédent, calculez le revenu minimum et le revenu maximum parmi ceux du chef de ménage :

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

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

```
R> chef.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 *Édition*, et choisissez *Exécuter la ligne ou sélection*. Apparemment 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 *Édition* 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 > R Script*. Votre nouveau fichier apparaîtra 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 [186](#).

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.3.3](#), page [187](#).

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

```
R> library(questionr)
Loading required namespace: car
```

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 structure 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] "calso"         "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 ...
 $ calso         : 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 ...
```

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 **âge** 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 52.



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 tableau, *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$sex
[1] Femme Femme Homme Homme Femme Femme
[12] Homme Femme Femme Femme Femme Homme Femme Femme Femme Femme Femme Femme Femme Femme Femme Femme
[23] Femme Femme Femme Homme Femme Homme Homme Homme Homme Homme Homme Homme Homme Homme Homme
[34] Homme Femme Femme Homme Femme Femme
[45] Femme Homme Femme Femme
[56] Femme Femme Femme Homme Femme Femme
[67] Homme Homme Femme Femme Homme Femme Femme
[78] Femme Femme
[89] Homme Homme Homme Femme Homme Femme Femme
[100] Femme
[ reached getOption("max.print") -- omitted 1900 entries ]
Levels: Homme Femme
```

On constate alors que R a bien accédé au contenu de notre variable `sex` 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 poids
1 1 28 Femme Enseignement superieur y compris technique superieur 2634
2 2 23 Femme                                     <NA> 9738

  occup qualif freres.soeurs cleso
1 Exerce une profession Employe           8 Oui
2 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] 2.247

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

[1] 1.776

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
0.00    1.00    2.00    2.25    3.00   12.00      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 suivante.

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'histogramme⁴.

Deux exemples sont donnés figure 3.2 page 29 et figure 3.3 page 30.

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 31.

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

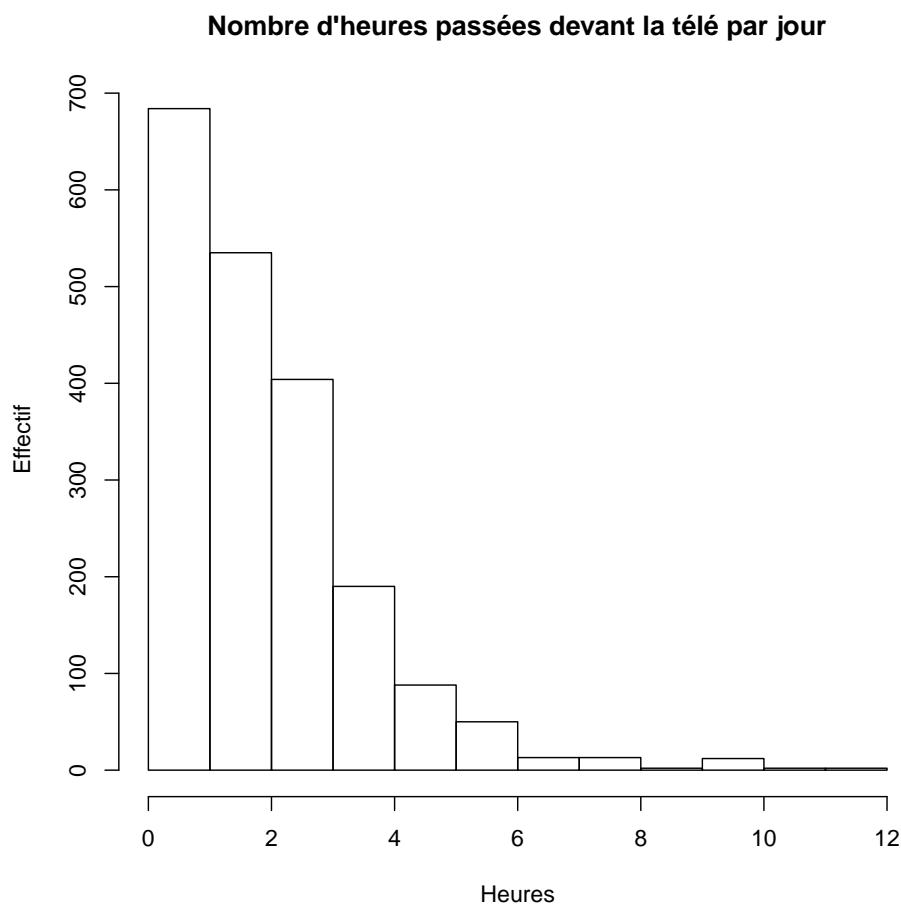


FIGURE 3.1 – Exemple d'histogramme

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

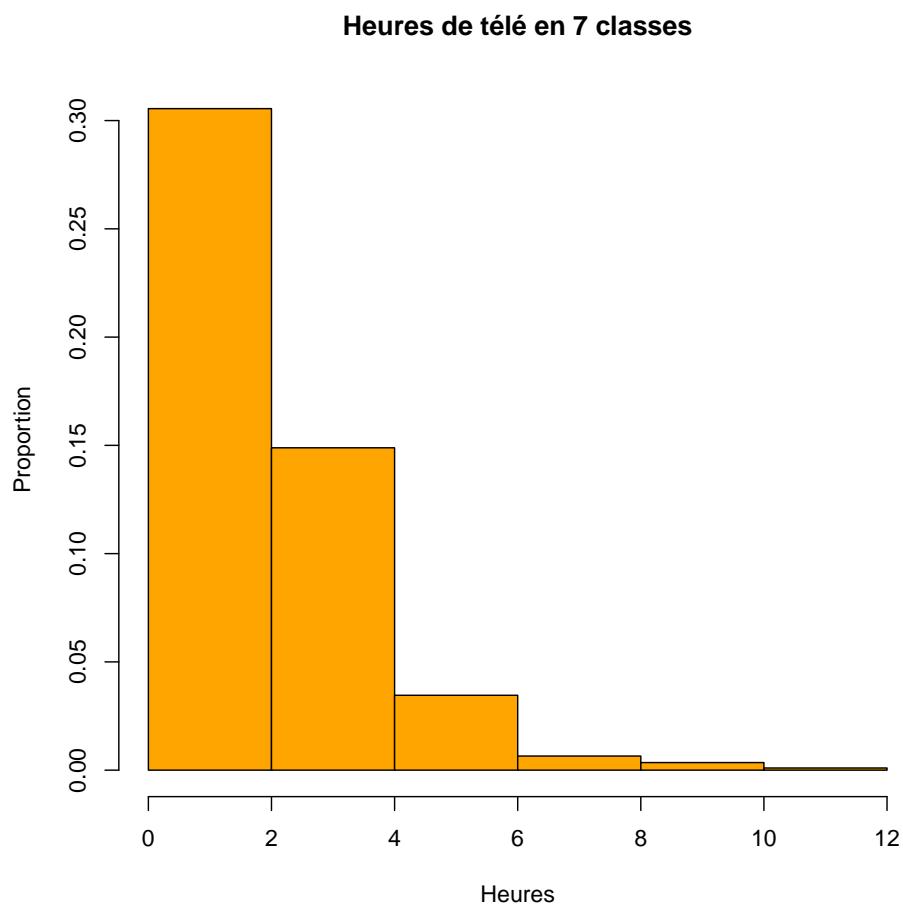


FIGURE 3.2 – Un autre exemple d’histogramme

```
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")
```

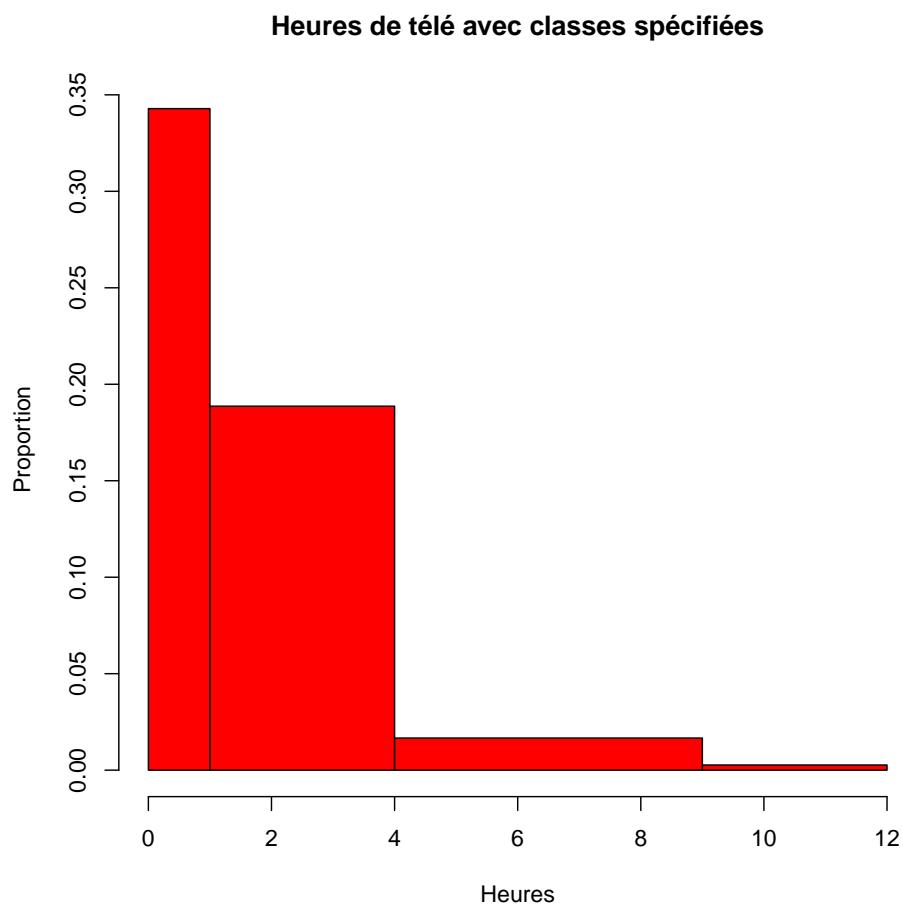


FIGURE 3.3 – Encore un autre exemple d’histogramme

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

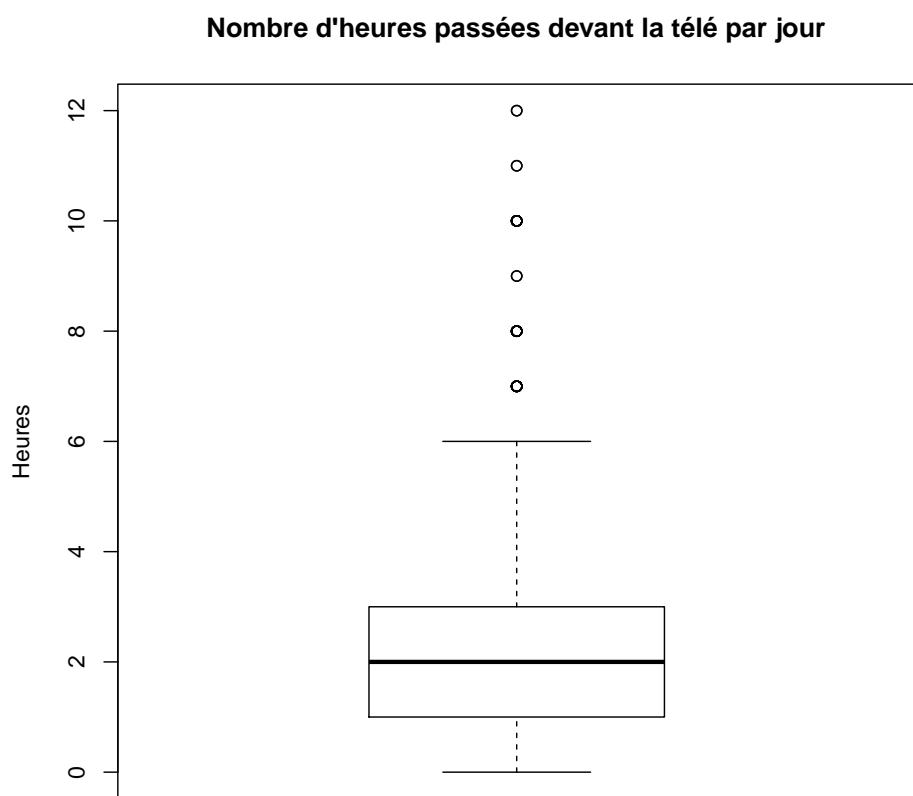


FIGURE 3.4 – Exemple de boîte à moustaches

```
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)
```

Nombre d'heures passées devant la télé par jour

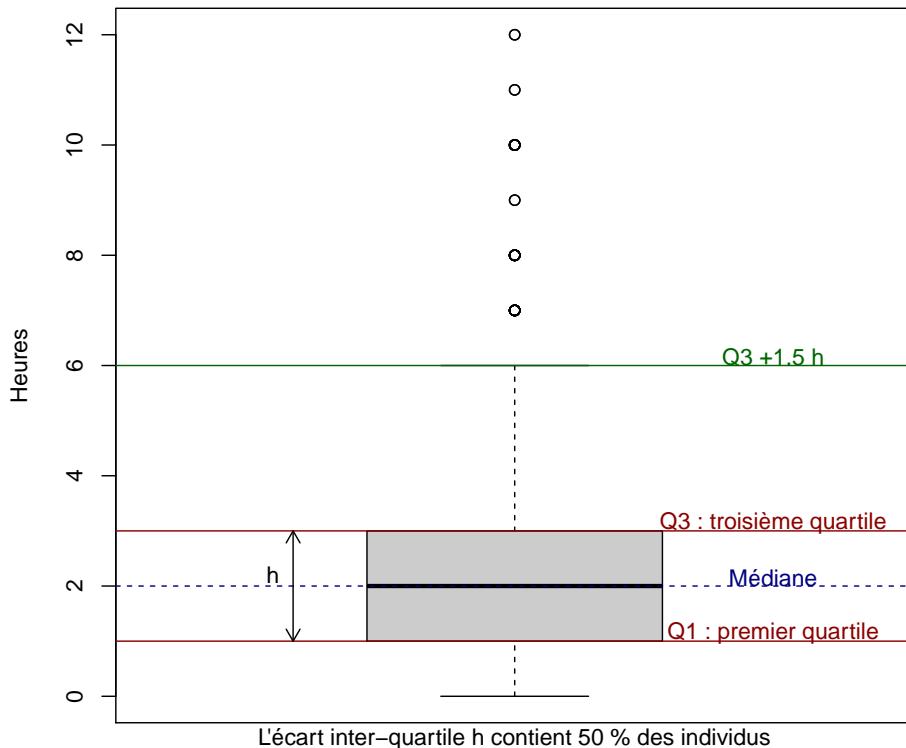


FIGURE 3.5 – Interprétation d'une boîte à moustaches

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

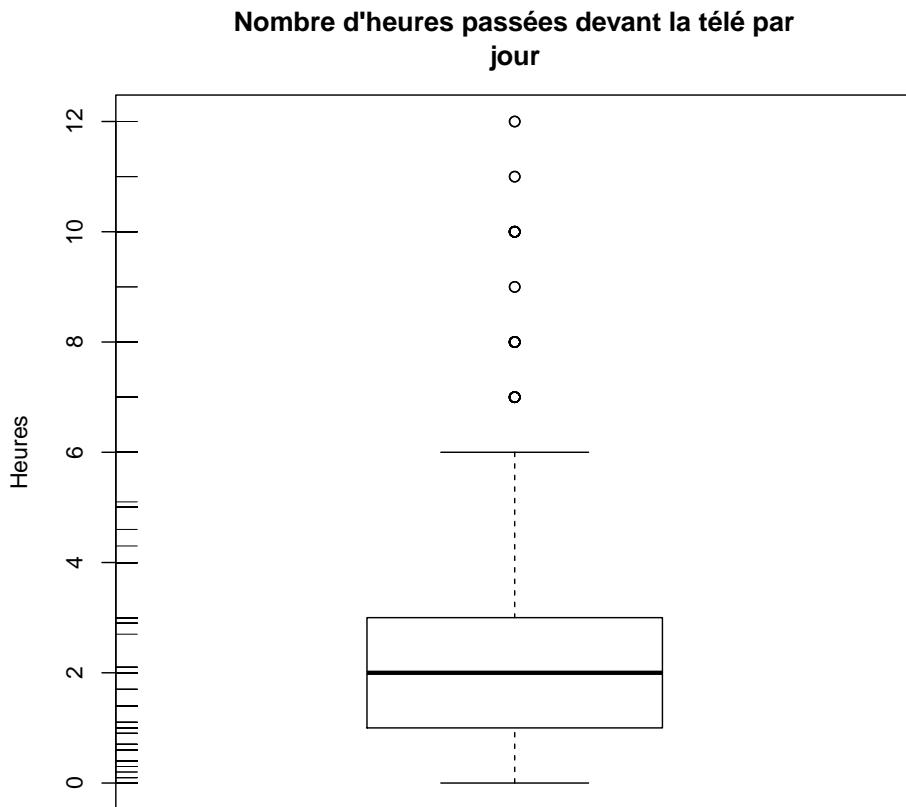


FIGURE 3.6 – Boîte à moustaches avec représentation des valeurs

Comment interpréter ce graphique ? On le comprendra mieux à partir de la figure 3.5 page précédente⁵.

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'autre 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/~tzhang/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>.

Intervalle de confiance

L'intervalle de confiance d'une moyenne peut être calculé avec la fonction `t.test` (fonction qui permet également de réaliser un test t de Student comme nous le verrons section 6.1 page 84) :

```
R> t.test(d$heures.tv)

One Sample t-test

data: d$heures.tv
t = 56.5, df = 1994, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.169 2.325
sample estimates:
mean of x
 2.247
```

Le niveau de confiance peut être précisé via l'argument `conf.level` :

```
R> t.test(d$heures.tv, conf.level = 0.9)

One Sample t-test

data: d$heures.tv
t = 56.5, df = 1994, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 2.181 2.312
sample estimates:
mean of x
 2.247
```

Le nombre d'heures moyennes à regarder la télévision parmi les enquêtés s'avère être de 2,2 heures, avec un intervalle de confiance à 95 % de [2,17 - 2,33] et un intervalle de confiance à 90 % de [2,18 - 2,31].

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

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	%
Ouvrier specialise	203	10.2
Ouvrier qualifie	292	14.6
Technicien	86	4.3
Profession intermediaire	160	8.0
Cadre	260	13.0
Employe	594	29.7
Autre	58	2.9
NA	347	17.3

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

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 triées 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 surfaces⁷.

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))`.

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")
```

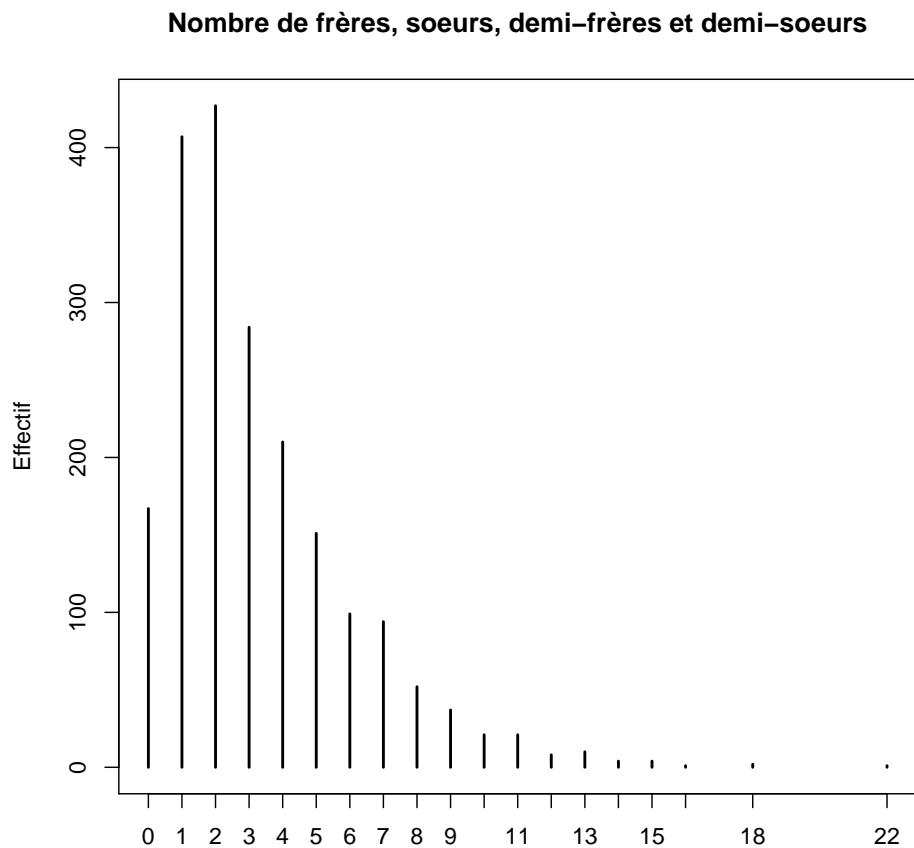


FIGURE 3.7 – Exemple de diagramme en bâtons

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 de la présente page.

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`⁸. Le résultat se trouve figure 3.8 page suivante.

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 39).

8. Pour des raisons liées au fonctionnement interne de la fonction `dotchart`, on doit l'appliquer à la transposition du tri à plat obtenu, d'où l'appel à la fonction `t`.

```
R> dotchart(t(table(d$c1so)), main = "Sentiment d'appartenance à une classe sociale",
+           pch = 19)
```

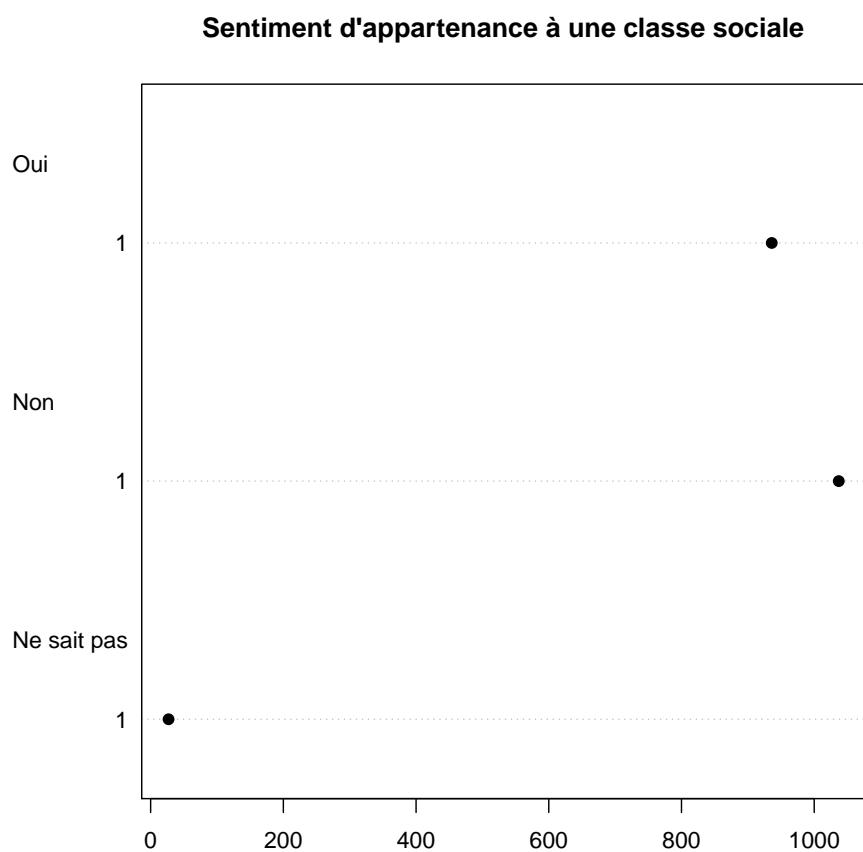


FIGURE 3.8 – Exemple de diagramme de Cleveland

```
R> dotchart(sort(table(d$qualif)), main = "Niveau de qualification")
Warning: 'x' is neither a vector nor a matrix: using as.numeric(x)
```

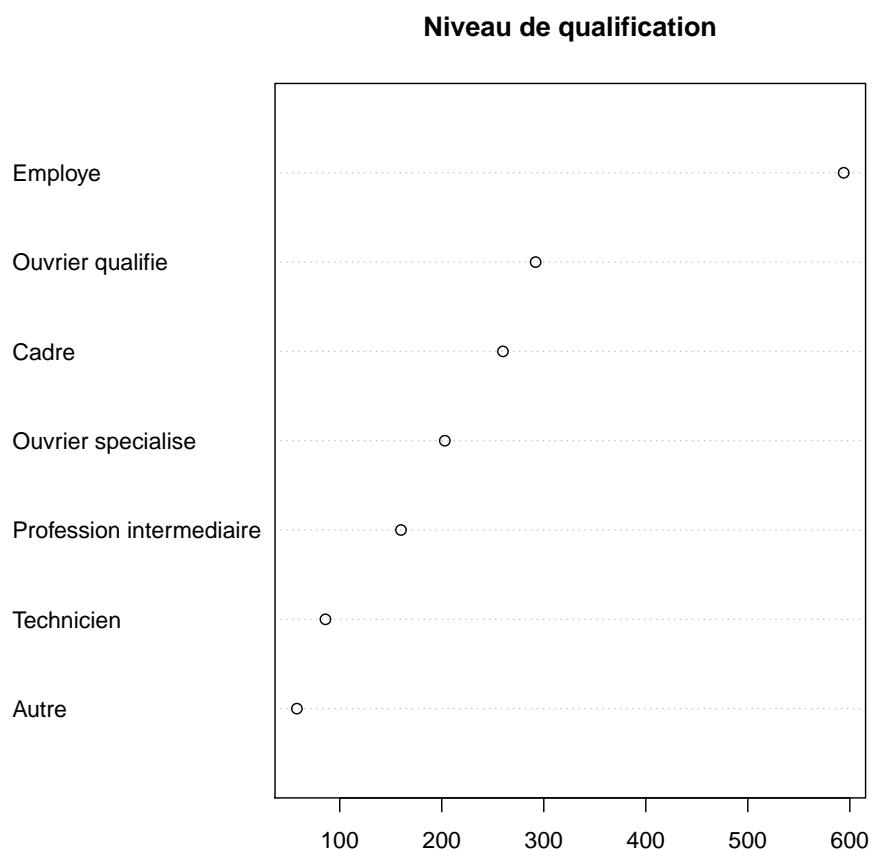


FIGURE 3.9 – Exemple de diagramme de Cleveland ordonné

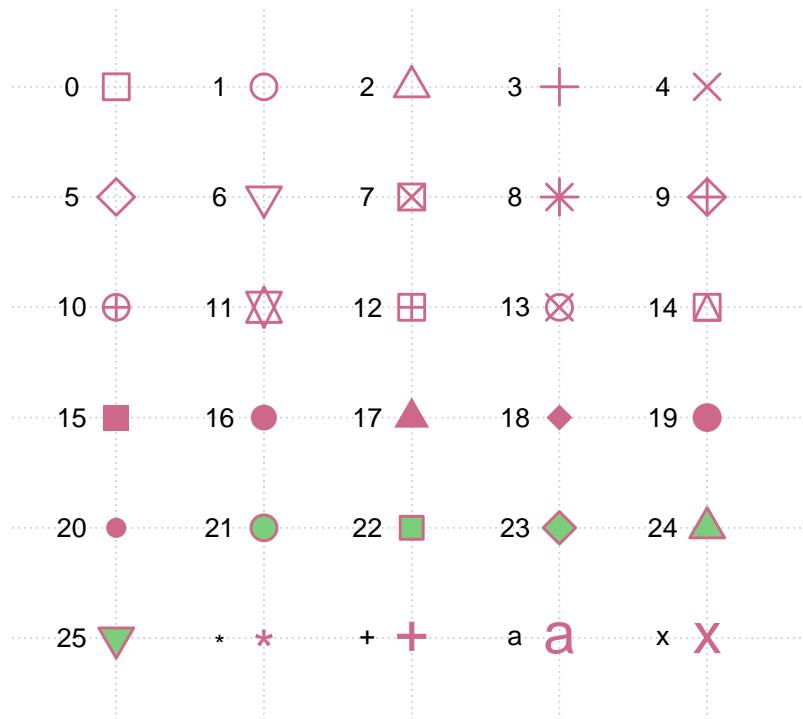


FIGURE 3.10 – Différentes valeurs possibles pour l’argument pch



L’argument `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 caractère textuel (voir figure 3.10 de la présente page).

Intervalle de confiance

La fonction `prop.test` permet de calculer l’intervalle de confiance d’une proportion. Une première possibilité consiste à lui transmettre une table à une dimension et deux entrées. Par exemple, si l’on s’intéresse à la proportion de personnes ayant pratiqué une activité physique au cours des douze derniers mois :

```
R> freq(d$sport)
```

	n	%
Non	1277	63.8
Oui	723	36.1

```
NA      0  0.0

R> prop.test(table(d$sport))

1-sample proportions test with continuity correction

data: table(d$sport), null probability 0.5
X-squared = 152.9, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.6169 0.6595
sample estimates:
    p
0.6385
```

On remarquera que la fonction a calculé l'intervalle de confiance correspondant à la première entrée du tableau, autrement dit celui de la proportion d'enquêtés n'ayant pas pratiqué une activité sportive.

Or, nous sommes intéressé par la proportion complémentaire, à savoir celle d'enquêtés ayant pratiqué une activité sportive. On peut dès lors modifier l'ordre de la table en indiquant notre modalité d'intérêt avec la fonction `relevel` ou bien indiquer à `prop.test` d'abord le nombre de succès puis l'effectif total :

```
R> prop.test(table(relevel(d$sport, "Oui")))

1-sample proportions test with continuity correction

data: table(relevel(d$sport, "Oui")), null probability 0.5
X-squared = 152.9, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3405 0.3831
sample estimates:
    p
0.3615

R> prop.test(sum(d$sport == "Oui"), length(d$sport))

1-sample proportions test with continuity correction

data: sum(d$sport == "Oui") out of length(d$sport), null probability 0.5
X-squared = 152.9, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3405 0.3831
sample estimates:
    p
0.3615
```

Enfin, le niveau de confiance peut être modifié via l'argument `conf.level` :

```
R> prop.test(table(relevel(d$sport, "Oui")), conf.level = 0.9)

1-sample proportions test with continuity correction

data: table(relevel(d$sport, "Oui")), null probability 0.5
X-squared = 152.9, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
0.3438 0.3796
sample estimates:
p
0.3615
```

3.6 Exercices

Exercice 3.5

▷ *Solution page 190*

Créer un script qui effectue les actions suivantes 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 191*

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`.

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 191*

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. Quel est l'intervalle de confiance à 95 % de l'âge moyen des enquêtés ?

Exercice 3.8

▷ *Solution page 191*

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 premières 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 14 page 176) 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] "C:/Users/Joseph/intro-r"
```

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 **données**. 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 débarrasser 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 retrouvant sa session de travail dans l'était où elle était.

4.2 Import de données depuis un tableur

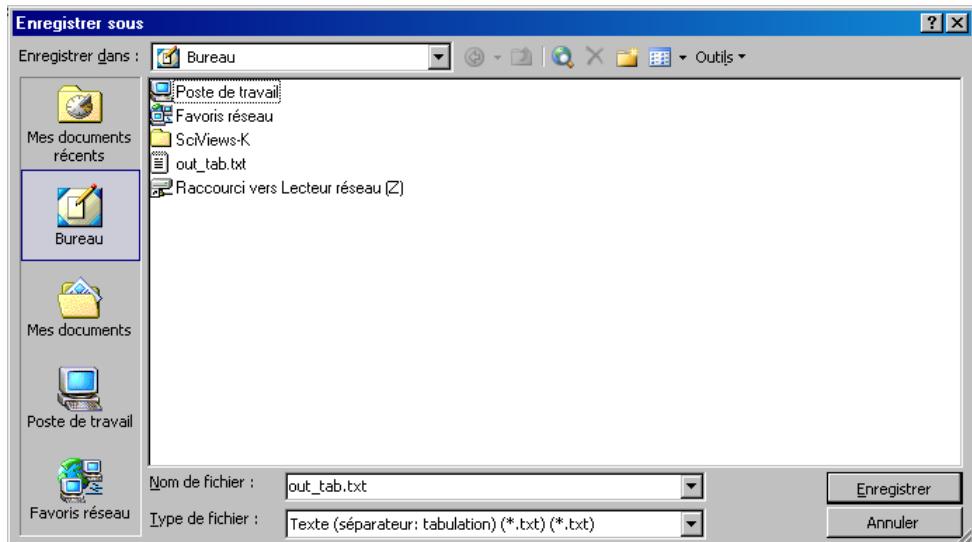
Il est assez courant de vouloir importer des données saisies ou traitées avec un tableur du type OpenOffice/LibreOffice ou Excel. En général les données prennent alors la forme d'un tableau avec les variables en colonne et les individus en ligne.

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		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 :

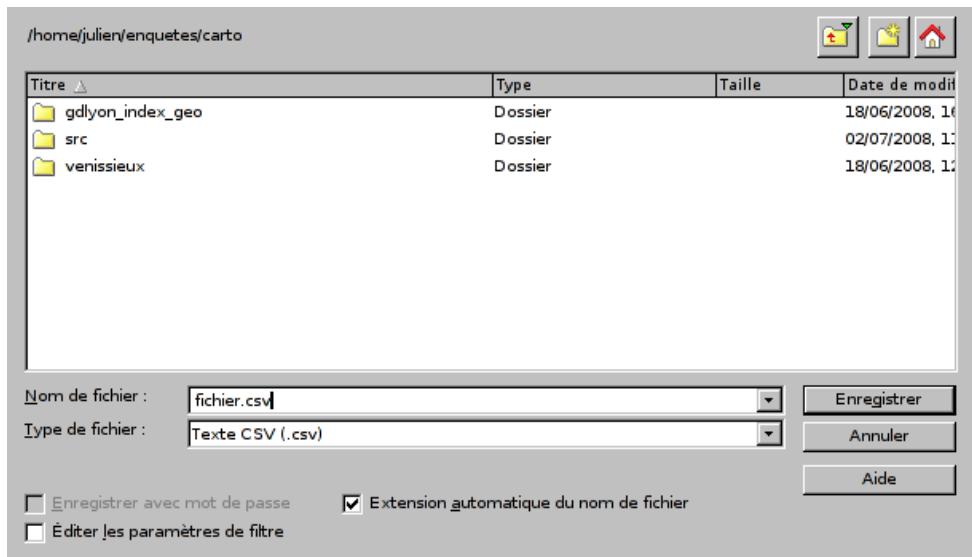
```
R> donnees <- read.delim2("fichier.txt")
```

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 rencontre 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 référera 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, il existe des fonctions permettant de les importer, mais elles nécessitent d'avoir une installation de SAS fonctionnelle sur sa machine (il s'agit des fonctions `read.ssd` de l'extension `foreign`, et `sas.get` de l'extension `Hmisc`).

Si on ne dispose que des fichiers au format SAS natif, le plus simple est d'utiliser l'application SAS System Viewer, qui permet de lire des fichiers SAS natif, de les visualiser et de les enregistrer dans un format texte. Cette application est téléchargeable gratuitement, mais ne fonctionne que sous Windows² :

<http://www.sas.com/apps/demosdownloads/setupcat.jsp?cat=SAS+System+Viewer>

Une fois le fichier de données au format SAS natif ouvert on peut l'enregistrer au format texte tabulé. L'import dans R se fait alors avec la commande suivante :

```
R> donnees <- read.delim("fichier.txt", na.strings = ".")
```

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`.

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 :

2. Ou sous Linux et Mac OS X avec wine.

```
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 !

```
R> 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 77.

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 relationnelle 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 perdus. 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 191*

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 192*

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.3.3, page 187.

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

Mais aussi sur le jeu de données tiré du recensement 1999, décrit page 188 :

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

5.1.1 Types de variables

On peut considérer qu'il existe quatre types 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 <code>class</code>	Type de variable
<code>factor</code>	Facteur
<code>integer</code>	Numérique
<code>double</code>	Numérique
<code>numeric</code>	Numérique
<code>character</code>	Caractères
<code>logical</code>	Booléenne

```
R> class(d$age)
[1] "integer"

R> class(d$sex)
[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 13.

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

```
R> names(d)
[1] "id"          "age"         "sex"        "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 passionnant, 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.

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 obtient un message d'erreur et une valeur manquante est utilisée à la place :

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

Warning: 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
```

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 :

```
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 68.

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
```

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]	2634.4	9738.4	3994.1	5731.7	4329.1	8674.7	6165.8	12891.6
[9]	7808.9	2277.2	704.3	6697.9	7118.5	586.8	11042.1	9958.2
[17]	4836.1	1551.5	3141.2	27195.8	14648.0	8128.1	1281.9	11663.3
[25]	8780.3	1700.8	6662.8	3359.5	8536.1	10620.5	5264.3	14161.8

```
[33] 1339.6 9243.9 4512.3 7871.6 1357.0 7626.3 1630.3 2196.2
[41] 5606.0 8841.3 9113.5 2267.6 7706.3 2446.5 8118.3 10751.5
[49] 831.9 6591.6 1936.9 834.4 3432.5 11354.9 9293.0 6344.1
[57] 4899.9 4766.9 3462.8 23732.5 833.8 8529.4 3190.4 2423.1
[65] 5946.0 14991.9 2062.1 5702.1 20604.3 2634.5 13544.6 4748.8
[73] 2348.4 3718.8 3850.6 6571.6 3238.3 17333.3 6168.5 7357.9
[81] 3909.6 1514.1 2916.9 3156.1 12404.6 2388.0 6590.6 2590.6
[89] 2786.3 7457.7 11076.9 2841.5 2422.1 4454.3 5581.6 5875.0
[97] 2064.1 827.7 13405.1 2186.5
[ reached getOption("max.print") -- omitted 1900 entries ]
```

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

	id	age	sexe	niveau	etud	poids		
1	1	28	Femme	Enseignement superieur	y compris technique superieur	2634		
2	2	23	Femme		<NA>	9738		
	occup	qualif	freres.soeurs	clso				
1	Exerce une profession	Employe		8	Oui			
2	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 équivalentes¹ :

```
R> d$clos
```

	Oui	Oui	Non	Non	Oui
[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
[71] Non     Non     Oui     Non     Oui
[76] Non     Oui      Non     Oui     Non
[81] Non     Non     Non     Oui     Oui
[86] Non     Oui      Oui     Non     Oui
[91] Oui      Oui     Non     Oui     Oui
[96] Non     Oui      Non     Oui     Oui
[ reached getOption("max.print") -- omitted 1900 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
[71] Non     Non     Oui     Non     Oui
[76] Non     Oui      Non     Oui     Non
[81] Non     Non     Non     Oui     Oui
[86] Non     Oui      Oui     Non     Oui
[91] Oui     Oui      Non     Oui     Oui
[96] Non     Oui      Non     Oui     Oui
[ reached getOption("max.print") -- omitted 1900 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")]
```

1. 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`.

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	artis
VILLEURBANNE	VILLEURBANNE	69266	57252	124152	103157	55136	0.02096	5.144
OULLINS	OULLINS	69149	11849	25186	20880	11091	0.10127	4.819
	cadres	interm	empl	ouvr	retr	tx.chom	etud	dipl.sup
VILLEURBANNE	13.14	25.72	31.42	23.07	36.65	14.82	15.51	9.744
OULLINS	10.20	27.43	31.53	24.37	41.55	10.64	10.63	7.625
	dipl.aucun	proprio	hlm	locataire	maison			
VILLEURBANNE	16.90	37.62	23.34		32.77	6.533		
OULLINS	14.32	51.51	14.56		29.92	17.708		

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 utiliser 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
<code>==</code>	égal à
<code>!=</code>	different de
<code>></code>	strictement supérieur à
<code><</code>	strictement inférieur à
<code>>=</code>	supérieur ou égal à
<code><=</code>	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 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 TRUE TRUE FALSE FALSE FALSE FALSE TRUE  
[78] FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE TRUE TRUE  
[89] TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE  
[100] FALSE  
[ reached getOption("max.print") -- omitted 1900 entries ]
```

Que s'est-il passé ? Nous avons fourni à R une condition qui signifie « la valeur de la variable `sex` 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
[12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE NA
[34] NA FALSE FALSE
[45] FALSE FALSE
[56] FALSE FALSE FALSE FALSE FALSE FALSE NA FALSE NA FALSE FALSE FALSE FALSE
[67] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE NA FALSE FALSE FALSE FALSE
[89] FALSE FALSE
[100] FALSE
[ reached getOption("max.print") -- omitted 1900 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 FALSE TRUE TRUE
[67] TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
[78] FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE
[89] TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
[100] TRUE

[ reached getOption("max.print") -- omitted 1900 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 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
[45] 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 FALSE FALSE FALSE FALSE TRUE FALSE FALSE
[78] FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
[89] FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[100] FALSE

[ reached getOption("max.print") -- omitted 1900 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 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 TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[78] NA TRUE NA TRUE TRUE TRUE FALSE NA FALSE TRUE FALSE
[89] TRUE FALSE FALSE FALSE TRUE NA NA TRUE TRUE TRUE TRUE
[100] TRUE

[ reached getOption("max.print") -- omitted 1900 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  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[78]    NA  TRUE    NA  TRUE  TRUE  TRUE FALSE    NA FALSE  TRUE FALSE
[89]  TRUE FALSE FALSE FALSE  TRUE    NA    NA  TRUE  TRUE  TRUE  TRUE
[100]   TRUE
[ reached getOption("max.print") -- omitted 1900 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$sex == "Femme" & (d$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 FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE
[78]    NA  TRUE    NA  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE
[89] FALSE FALSE FALSE  TRUE FALSE    NA    NA  TRUE FALSE FALSE FALSE
[100]   TRUE
[ reached getOption("max.print") -- omitted 1900 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$sex == "Femme" & d$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 FALSE  TRUE FALSE
[23]  TRUE 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 FALSE  TRUE
[67] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE
[78] FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE
[89] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
[100]   TRUE
[ reached getOption("max.print") -- omitted 1900 entries ]
```

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

```
R> table(d$sex)
```

	Homme	Femme
899	1101	

```
R> table(d$sex == "Homme")

FALSE  TRUE
1101   899

R> summary(d$sex == "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$sex == "Homme", ]
```

On obtiendra un nouveau tableau de données comportant l'ensemble des variables de `d`, mais seulement les observations pour lesquelles `d$sex` 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
24   79 Femme   Non
25   20 Femme   Oui
26   74 Homme  Non
27   31 Femme   Non
28   35 Homme  Non
29   35 Homme  Non
30   30 Homme  Oui
31   54 Homme  Non
32   29 Homme  Non
33   49 Homme  Non
[ reached getOption("max.print") -- omitted 1967 rows ]

```

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

```
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$sex == "Homme" & d$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.

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
48.16 48.15
```

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    %
Non 384 42.7
Oui 515 57.3
NA   0  0.0
```

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

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    %
Non 384 42.7
```

```
Oui    515  57.3
NA      0   0.0
Total  899 100.0
```

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



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

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

Homme Femme
48.16 48.15

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

d$sex: Homme
[1] 48.16
-----
d$sex: Femme
[1] 48.15
```

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 incluses 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

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

```
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 :

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

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

```
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

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.

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] 2.247

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] -2.247 9.753

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

[1] 4.715e-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$age.sex <- NA
R> d$age.sex[d$sex == "Homme" & d$age < 40] <- "Homme moins de 40 ans"
R> d$age.sex[d$sex == "Homme" & d$age >= 40] <- "Homme plus de 40 ans"
R> d$age.sex[d$sex == "Femme" & d$age < 40] <- "Femme moins de 40 ans"
R> d$age.sex[d$sex == "Femme" & d$age >= 40] <- "Femme plus de 40 ans"
R> table(d$age.sex)
```

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
[71] 1795 1971 262 444 704 744 1105 1109 1164 1243 1309 1357 1403 1627
[85] 1640 1645 1660 1737 1783 1847 79 205 217 312 367 378 393 412
[99] 544 576
[ reached getOption("max.print") -- omitted 1900 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 ciso
162 162 18 Homme    <NA> 4983 Etudiant, eleve <NA>           2 Non
215 215 18 Homme    <NA> 4631 Etudiant, eleve <NA>           2 Oui
346 346 18 Femme    <NA> 1725 Etudiant, eleve <NA>           9 Non
                                relig trav.imp trav.satisf hard.rock lecture.bd
162 Appartenance sans pratique    <NA>      <NA>       Non       Non
215 Ni croyance ni appartenance <NA>      <NA>       Non       Non
346 Pratiquant regulier        <NA>      <NA>       Non       Non
                                peche.chasse cuisine bricol cinema sport heures.tv fs.fac fs.char
162           Non   Non   Non   Non   Oui     3     2     2
215           Non   Oui   Non   Oui   Oui     2     2     2
346           Non   Non   Non   Oui   Non     2     9     9
                                qualif.char fs.num      age5cl age20 age6cl qualif.reg
162           <NA>      2 (17.9,33.8] <20ans [18,30)      <NA>
215           <NA>      2 (17.9,33.8] <20ans [18,30)      <NA>
346           <NA>      9 (17.9,33.8] <20ans [18,30)      <NA>
                                trav.satisf.reg ecart.heures.tv      act.manuelles
```

```

162 Valeur manquante      0.7534 Ni cuisine ni bricolage
215 Valeur manquante      -0.2466      Cuisine seulement
346 Valeur manquante      -0.2466 Ni cuisine ni bricolage
                    age.sex score.ext
162 Homme moins de 40 ans   20
215 Homme moins de 40 ans   30
346 Femme moins de 40 ans   10

```

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		Id	V3		Id	V1	V2	V3
1	H	12		1	Rouge		1	H	12	Rouge
2	H	17	+	2	Bleu	=	2	H	17	Bleu
3	F	41		3	Bleu		3	F	41	Bleu
4	F	9		4	Rouge		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 côté en regroupant leurs colonnes².

```
R> cbind(d1, d2)
```

	id	age	sexe	id	cuso
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
11	11	65	Femme	11	Oui
12	12	47	Homme	12	Oui
13	13	63	Femme	13	Oui
14	14	67	Femme	14	Oui
15	15	76	Femme	15	Oui
16	16	49	Femme	16	Non
17	17	62	Homme	17	Non
18	18	20	Femme	18	Non
19	19	70	Homme	19	Non
20	20	39	Femme	20	Non

```
[ reached getOption("max.print") -- omitted 1980 rows ]
```

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	cuso
162	162	18	Homme	1	Oui
215	215	18	Homme	2	Oui
346	346	18	Femme	3	Non
377	377	18	Homme	4	Non

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

```

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
1600 1600  18 Femme  15     Oui
1774 1774  18 Homme  16     Non
1937 1937  18 Homme  17     Non
38    38    19 Femme  18     Non
100   100   19 Femme  19     Non
134   134   19 Femme  20     Non
[ reached getOption("max.print") -- omitted 1980 rows ]

```

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 procéder 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 identifié³, 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
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 ;

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

- 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		Id	V2	=	Id	V1	V2
1	H	+	1	10		1	H	10
2	H		2	15		2	H	15
3	F		5	31				

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

Id	V1		Id	V2	=	Id	V1	V2
1	H	+	1	10		1	H	10
2	H		1	18		1	H	18
3	F		1	21		1	H	21

Id	V1		Id	V2	=	Id	V1	V2
1	H	+	1	10		1	H	10
2	H		1	18		1	H	18
3	F		1	21		1	H	21

Id	V1		Id	V2	=	Id	V1	V2
1	H	+	2	11		2	H	11
2	H		3	31		3	F	31

Id	V1		Id	V2	=	Id	V1	V2
1	H	+	2	11		2	H	11
2	H		3	31		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, par définir le répertoire de travail à l'aide de `setwd`, à importer les données, à effectuer manipulations, traitements et recodages, puis à mettre en oeuvre les analyses. Prenons le fichier fictif suivant :

```

library(questionr)
library(foreign)
setwd("/home/julien/r/projet")
## IMPORT DES DONNÉES
d1 <- read.dbf("tab1.dbf")
d2 <- read.dbf("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 script⁴ 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("tab1.dbf")
d2 <- read.dbf("tab2.dbf")
d <- merge(d1, d2, by = "id")

```

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

```

## 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)
setwd("/home/julien/r/projet")
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 :

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

- 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 à 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 192*

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

Exercice 5.12

▷ *Solution page 192*

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 192*

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 192*

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 193*

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 193*

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 193*

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 194*

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 régulier » 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 195*

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 196*

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

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 MASS¹ 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 coefficient de corrélation entre les deux variables à l'aide de la fonction `cor` :

```
R> cor(d$age, d$heures.tv, use = "complete.obs")
[1] 0.1776
```

L'option `use` permet d'éliminer les observations pour lesquelles l'une des deux valeurs est manquante. Le coefficient 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)
```

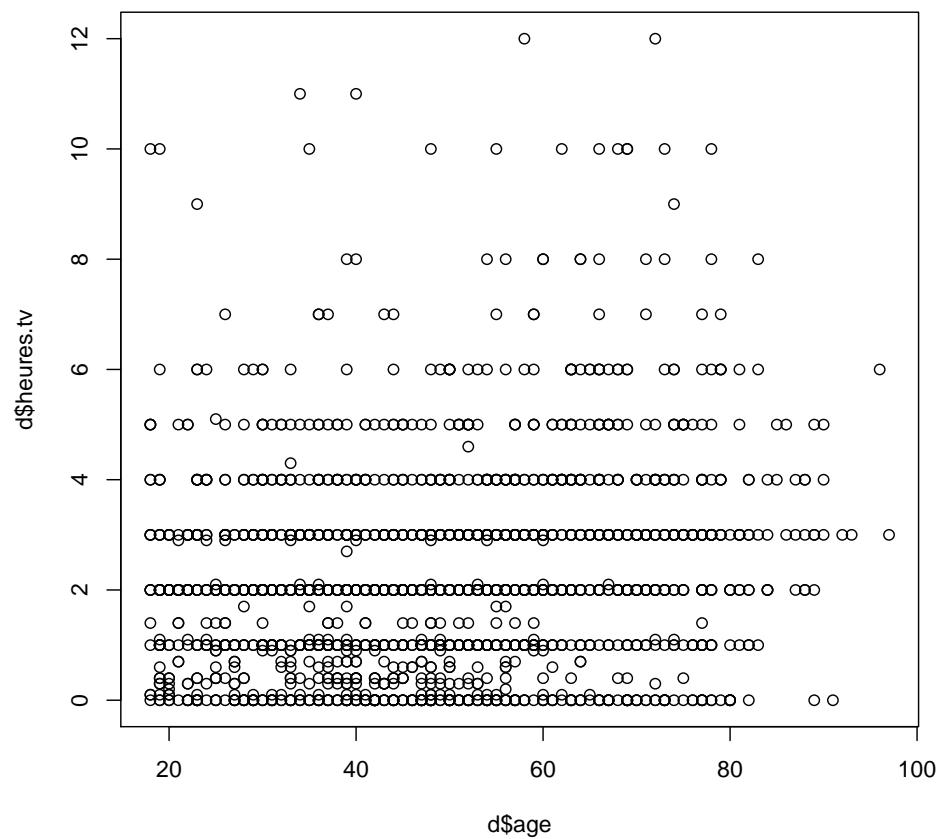


FIGURE 6.1 – Nombre d’heures de télévision selon l’âge

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

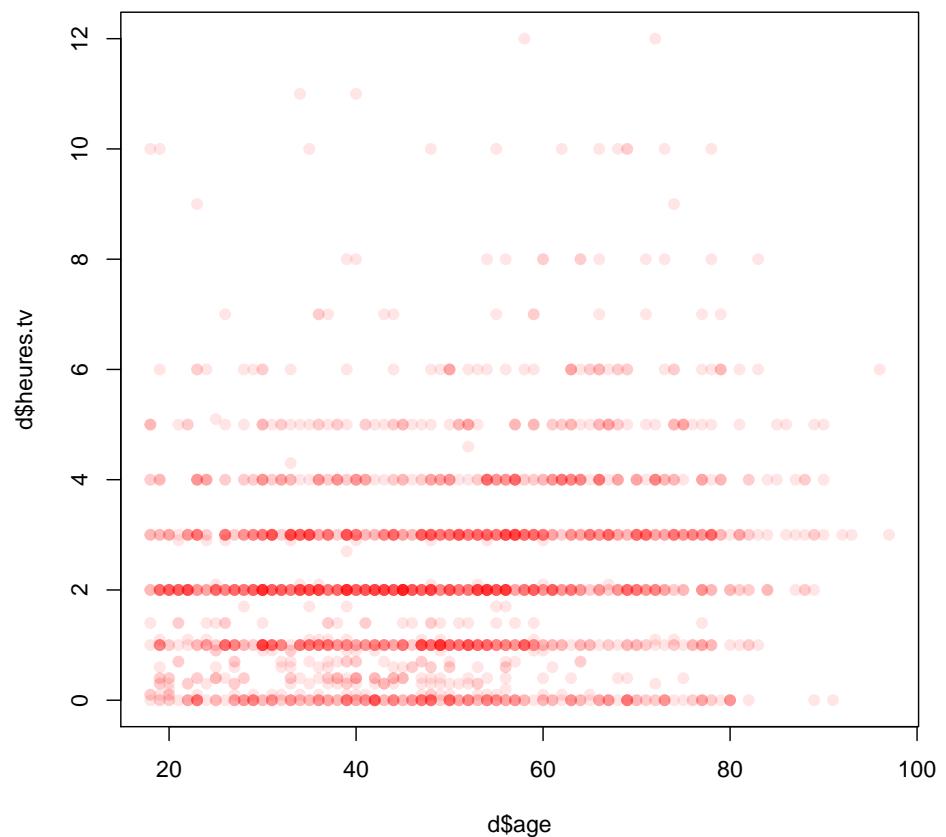


FIGURE 6.2 – Nombre d’heures de télévision selon l’âge avec semi-transparence

```
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)
```

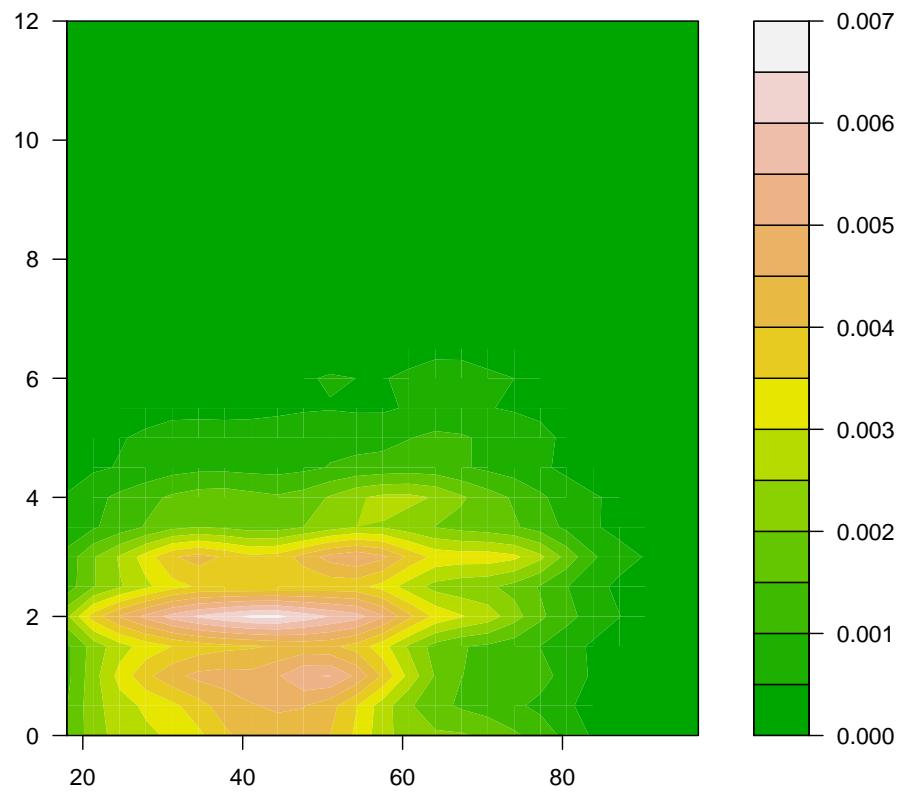


FIGURE 6.3 – Représentation de l'estimation de densité locale

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

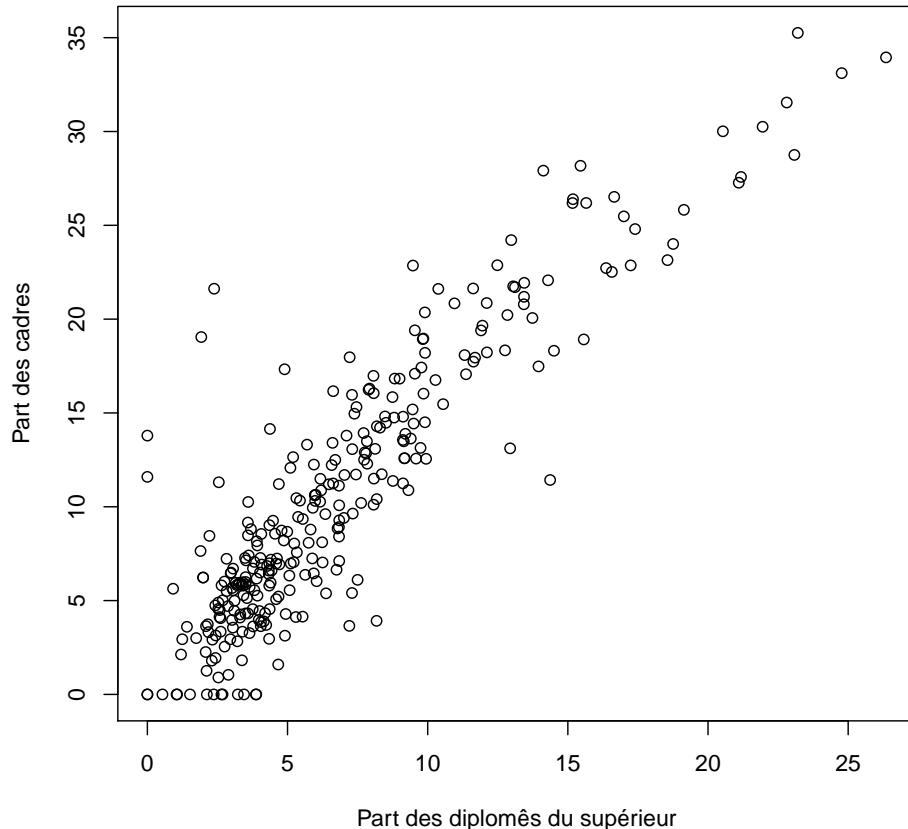


FIGURE 6.4 – Proportion de cadres et proportion de diplômés du supérieur

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] 0.8975
```

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 
-9.691 -1.901 -0.182  1.491 17.087 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.2409     0.3299   3.76   2e-04 ***  
dipl.sup     1.3835     0.0393  35.20  <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.28 on 299 degrees of freedom
Multiple R-squared:  0.806, Adjusted R-squared:  0.805 
F-statistic: 1.24e+03 on 1 and 299 DF,  p-value: <2e-16

```

Le résultat montre que les coefficients 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 67.

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

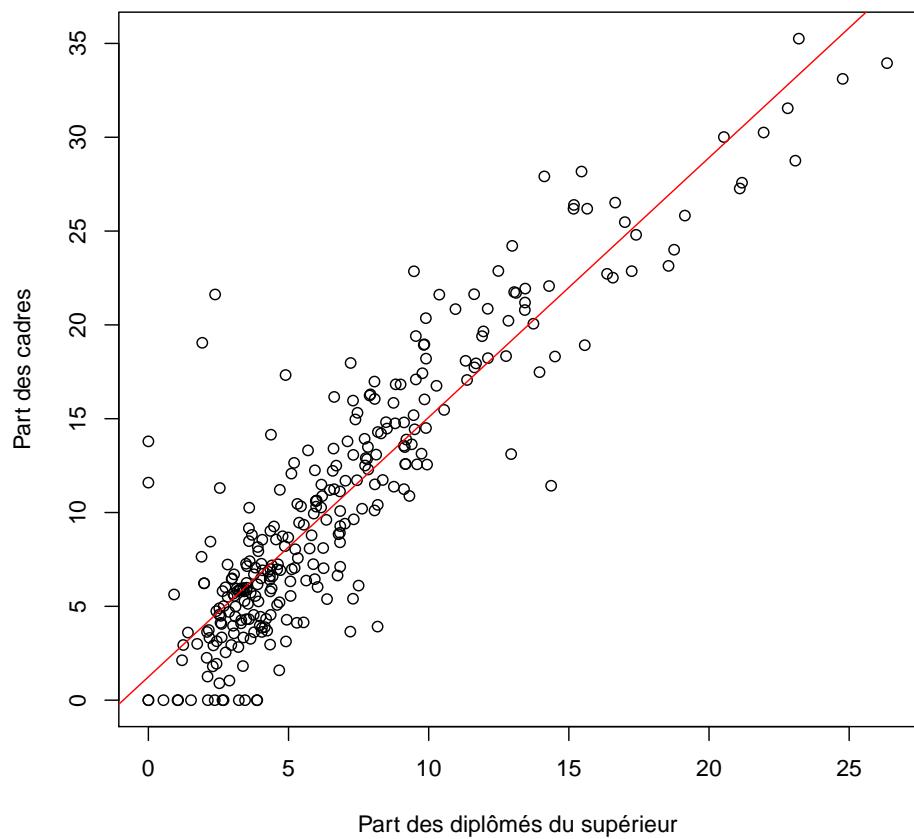


FIGURE 6.5 – Régression de la proportion de cadres par celle de diplômés du supérieur

```
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)
```

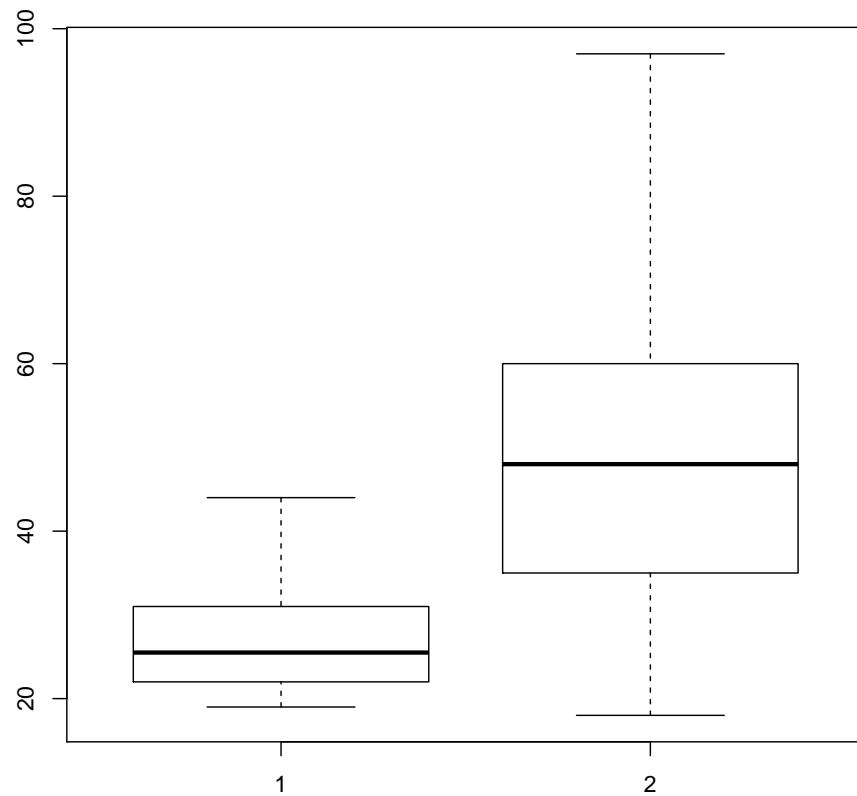


FIGURE 6.6 – Boxplot de la répartition des âges (sous-populations)

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

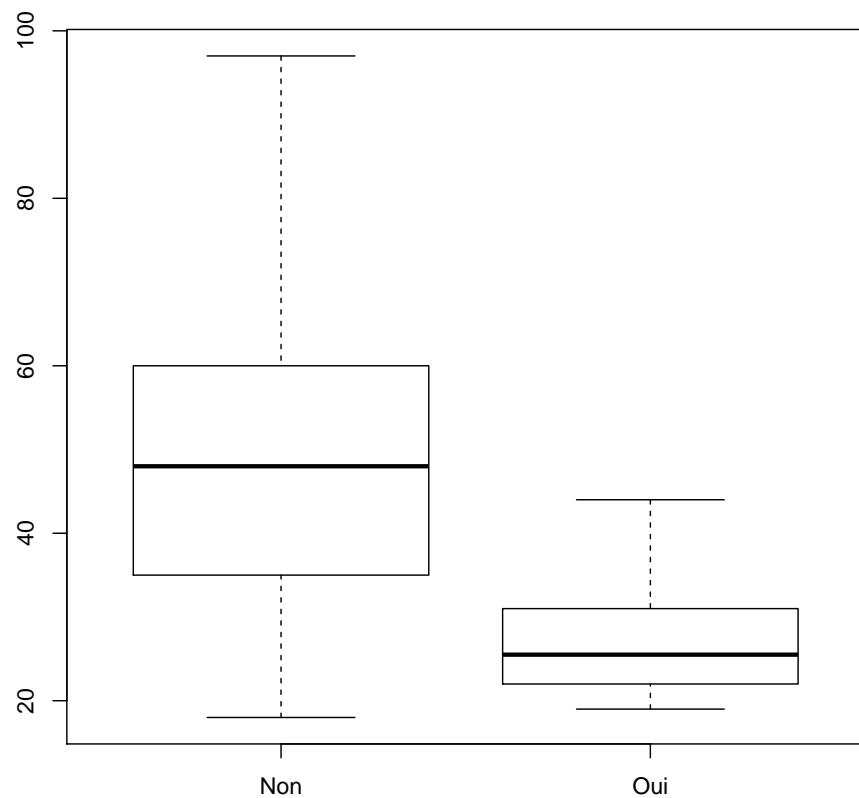


FIGURE 6.7 – *Boxplot* de la répartition des âges (formule)

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

Non    Oui
48.30 27.57
```

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 = 9.64, df = 13.85, p-value = 1.611e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
16.11 25.35
sample estimates:
mean in group Non mean in group Oui
48.30          27.57
```

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*. Voir aussi http://fr.wikipedia.org/wiki/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 variance³. 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 = 0.8693, 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")
```

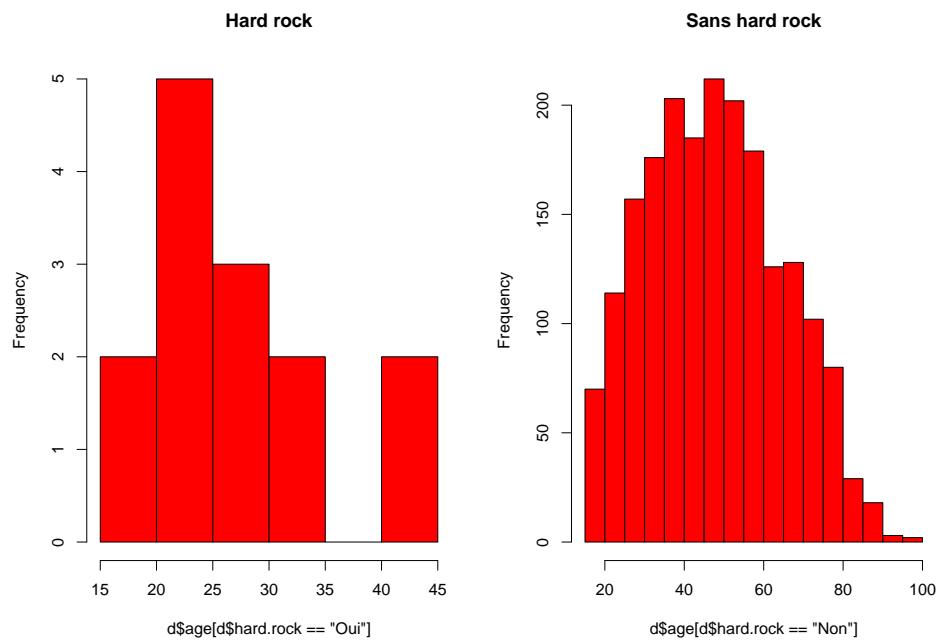


FIGURE 6.8 – Distribution des âges pour appréciation de la normalité

```
Shapiro-Wilk normality test
```

```
data: d$age[d$hard.rock == "Non"]
W = 0.9814, 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
285.63	62.73	

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 = 4.554, 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:
 1.752 8.694
sample estimates:
ratio of variances
        4.554
```

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 significative⁴. 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 tout à fait possible de croiser trois variables ou plus. Par exemple :

```
R> table(d$sport, d$cuisine, d$sex)
, ,   = 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.

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 `copy` (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 = 96.8, 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` :

```
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

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)`

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 coefficient 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.vde` `questionr` :

```
R> cramer.v(tab)
```

```
[1] 0.242
```



Pour un tableau à 2x2 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.

```
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:
 2.403 3.514
sample estimates:
odds ratio
 2.903
```

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 couleurs⁷? 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.

7. Sauf s'il est imprimé en noir et blanc...

```
R> mosaicplot(qualreg ~ sport, data = d, shade = TRUE, main = "Graphe en mosaïque")
```

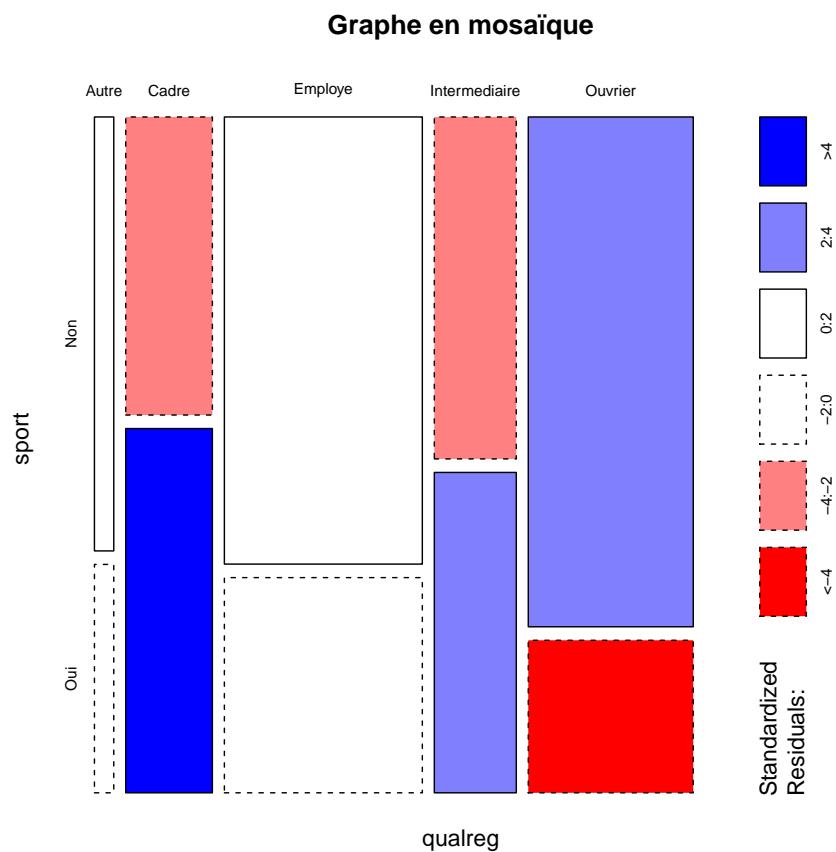


FIGURE 6.9 – Exemple de graphe en mosaïque

```
R> barplot(cprop(tab, total = FALSE), main = "Pratique du sport selon le niveau de qualification")
```

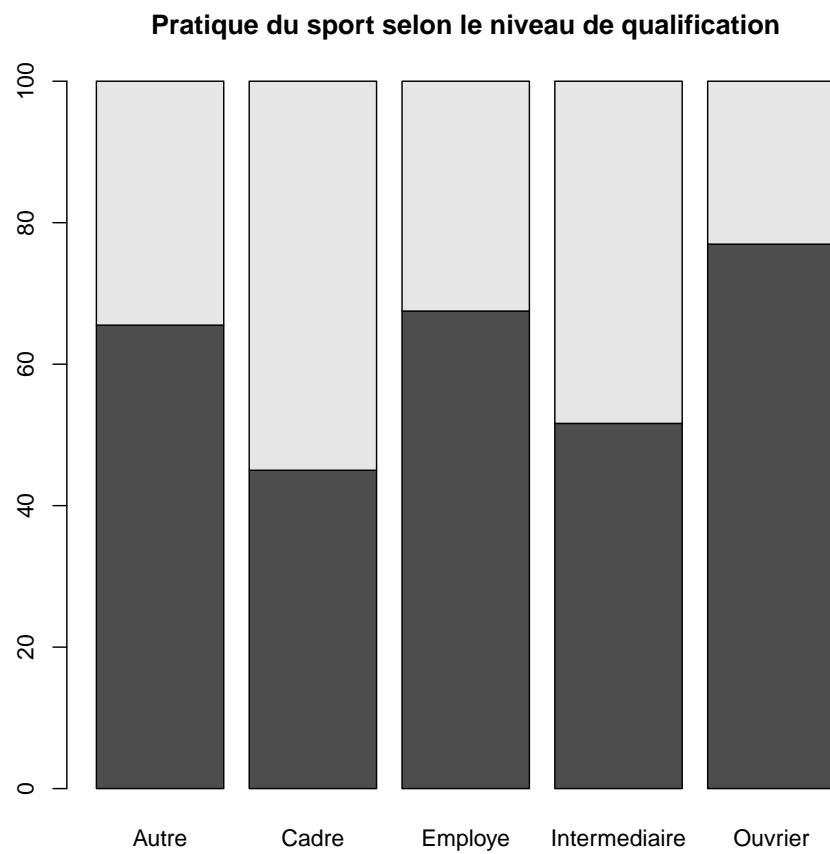


FIGURE 6.10 – Exemple de barres cumulées

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 précédente.

Partie 7

Régression logistique

La régression logistique est fréquemment utilisée en sciences sociales car elle permet d'effectuer un raisonnement dit *toutes choses étant égales par ailleurs*. Plus précisément, la régression logistique a pour but d'isoler les effets de chaque variable, c'est-à-dire d'identifier les effets résiduels d'une *variable explicative* sur une *variable d'intérêt*, une fois pris en compte les autres variables explicatives introduites dans le modèle. La régression logistique est ainsi prisée en épidémiologie pour identifier les facteurs associés à telle ou telle pathologie.

La régression logistique ordinaire ou régression logistique binaire vise à expliquer une variable d'intérêt binaire (c'est-à-dire de type « Oui/Non »). Les variables explicatives qui seront introduites dans le modèle peuvent être quantitatives ou qualitatives.

7.1 Préparation des données

Dans ce chapitre, nous allons encore une fois utiliser les données de l'enquête *Histoire de vie*, fournies avec l'extension `questionr` et décrites dans l'annexe B.3.3, page 187.

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

À titre d'exemple, nous allons étudier l'effet de l'âge, du sexe, du niveau d'étude, de la pratique religieuse et du nombre moyen d'heures passées à regarder la télévision par jour.

En premier lieu, il importe de vérifier que notre variable d'intérêt (ici `sport`) est correctement codée. Une possibilité consiste à créer une variable booléenne(vrai / faux) selon que l'individu a pratiqué du sport ou non :

```
R> d$sport2 <- FALSE
R> d$sport2[d$sport == "Oui"] <- TRUE
```

Dans le cas présent, cette variable n'a pas de valeur manquante. Mais le cas échéant il faut bien renseigner NA pour les valeurs manquantes, les individus en question étant alors exclu de l'analyse.

Il n'est pas forcément nécessaire de transformer notre variable d'intérêt en variable booléenne. En effet, R accepte sans problème une variable de type facteur. Cependant, l'ordre des valeurs d'un facteur a de l'importance. En effet, R considère toujours la première modalité comme étant la *modalité de référence*. Dans le cas de la variable d'intérêt, la modalité de référence correspond au fait de ne pas remplir le critère étudié, dans notre exemple au fait de ne pas avoir eu d'activité sportive au cours des douze derniers mois.

Pour connaître l'ordre des modalités d'une variable de type facteur, on peut utiliser la fonction `levels` ou bien encore tout simplement la fonction `freq` :

```
R> levels(d$sport)
```

```
[1] "Non" "Oui"
```

```
R> freq(d$sport)
```

	n	%
Non	1277	63.8
Oui	723	36.1
NA	0	0.0

Dans notre exemple, la modalité « Non » est déjà la première modalité. Il n'y a donc pas besoin de modifier notre variable. Si ce n'est pas le cas, il faudra modifier la modalité de référence avec la fonction `relevel` comme nous allons le voir un peu plus loin.



Il est possible d'indiquer un facteur à plus de deux modalités. Dans une telle situation, R considérera que tous les modalités, sauf la modalité de référence, est une réalisation de la variable d'intérêt. Cela serait correct par exemple notre variable `sport` était codée ainsi : « Non », « Oui, toutes les semaines », « Oui, au moins une fois par mois », « Oui, moins d'une fois par mois ». Cependant, afin d'éviter tout risque d'erreur ou de mauvaise interprétation, il est vivement conseillé de recoder au préalable sa variable d'intérêt en un facteur à deux modalités.

La notion de modalité de référence s'applique également aux variables explicatives qualitatives. En effet, dans un modèle, tous les coefficients sont calculés par rapport à la modalité de référence. Il importe de choisir une modalité de référence qui fasse sens afin de faciliter l'interprétation. Par ailleurs, ce choix peut également dépendre de la manière dont on souhaite présenter les résultats. De manière générale on évitera de choisir comme référence une modalité peu représentée dans l'échantillon ou bien une modalité correspondant à une situation atypique.

Prenons l'exemple de la variable `sexe`. Souhaite-t-on connaître l'effet d'être une femme par rapport au fait d'être un homme ou bien l'effet d'être un homme par rapport au fait d'être une femme ? Si l'on opte pour le second, alors notre modalité de référence sera le sexe féminin. Comme est codée cette variable ?

```
R> freq(d$sexe)
```

	n	%
Homme	899	45
Femme	1101	55
NA	0	0

La modalité `Femme` s'avère ne pas être la première modalité. Nous devons appliquer la fonction `relevel` :

```
R> d$sexe <- relevel(d$sexe, "Femme")
R> freq(d$sexe)
```

	n	%
--	---	---

```
Femme 1101 55
Homme 899 45
NA      0  0
```

Les variables `age` et `heures.tv` sont des variables quantitatives. Il importe de vérifier qu'elles sont bien enregistrées en tant que variables numériques. En effet, il arrive parfois que dans le fichier source les variables quantitatives soient renseignées sous forme de valeur textuelle et non sous forme numérique.

```
R> str(d$age)
int [1:2000] 28 23 59 34 71 35 60 47 20 28 ...
R> str(d$heures.tv)
num [1:2000] 0 1 0 2 3 2 2.9 1 2 2 ...
```

Nos deux variables sont bien renseignées sous forme numérique.

Cependant, l'effet de l'âge est rarement linéaire. Un exemple trivial est par exemple le fait d'occuper un emploi qui sera moins fréquent aux jeunes âges et aux âges élevés. Dès lors, on pourra transformer la variable « âge » en groupe d'âges(voir section 5.4.2 page 69) :

```
R> d$grpage <- cut(d$age, c(16, 25, 45, 65, 93), right = FALSE, include.lowest = TRUE)
R> freq(d$grpae)

      n    %
[16,25) 169  8.5
[25,45) 706 35.3
[45,65) 745 37.2
[65,93] 378 18.9
NA        2   0.1
```

Jetons maintenant un oeil à la variable `nivetud` :

```
R> freq(d$nivetud)

      n    %
N'a jamais fait d'etudes          39  2.0
A arrete ses etudes, avant la dernière année d'etudes primaires 86  4.3
Derniere année d'etudes primaires 341 17.1
1er cycle                         204 10.2
2eme cycle                        183  9.2
Enseignement technique ou professionnel court 463 23.2
Enseignement technique ou professionnel long   131  6.6
Enseignement superieur y compris technique superieur 441 22.1
NA                                112  5.6
```

En premier lieu, cette variable est détaillée en pas moins de huit modalités dont certaines sont peu représentées (seulement 39 individus soit 2 % n'ont jamais fait d'études par exemple). Afin d'améliorer notre modèle logistique, il peut être pertinent de regrouper certaines modalités (voir section 5.4.3 page 71) :

```
R> d$etud <- d$ivetud
R> levels(d$etud) <- c("Primaire", "Primaire", "Primaire", "Secondaire", "Secondaire",
+   "Technique/Professionnel", "Technique/Professionnel", "Supérieur")
R> freq(d$etud)

          n      %
Primaire     466  23.3
Secondaire    387 19.4
Technique/Professionnel 594 29.7
Supérieur     441 22.1
NA             112  5.6
```

Notre variable comporte également 112 individus avec une valeur manquante. Si nous conservons cette valeur manquante, ces 112 individus seront, par défaut, exclus de l'analyse. Ces valeurs manquantes n'étant pas négligeable (5,6 %), nous pouvons également faire le choix de considérer ces valeurs manquantes comme une modalité supplémentaire. Auquel cas, nous utiliserons la fonction addNA :

```
R> levels(d$etud)
[1] "Primaire"                 "Secondaire"
[3] "Technique/Professionnel" "Supérieur"

R> d$etud <- addNA(d$etud)
R> levels(d$etud)
[1] "Primaire"                 "Secondaire"
[3] "Technique/Professionnel" "Supérieur"
[5] NA
```

7.2 Régression logistique binaire

La fonction `glm` (pour *generalized linear models*) permet de calculer une grande variété de modèles statistiques. La régression logistique ordinaire correspond au modèle *logit* de la famille des modèles binomiaux, ce que l'on indique à `glm` avec l'argument `family=binomial(logit)`.

Le modèle proprement dit sera renseigné sous la forme d'une *formule* (que nous avons déjà rencontrée page 89). On indiquera d'abord la variable d'intérêt, suivie du signe ~ puis de la liste des variables explicatives séparées par un signe +. Enfin, l'argument **data** permettra d'indiquer notre tableau de données.

```

    grp[65,93]           etudSecondaire
    -1.37927             0.94831
etudTechnique/Professionnel etudSupérieur
    1.04716              1.88962
    etudNA                religPratiquant occasionnel
    2.14854              -0.02060
religAppartenance sans pratique religNi croyance ni appartenance
    -0.00618              -0.21407
    religRejet            religNSP ou NVPR
    -0.38274              -0.08336
    heures.tv
    -0.12072

Degrees of Freedom: 1992 Total (i.e. Null); 1978 Residual
(7 observations deleted due to missingness)
Null Deviance: 2610
Residual Deviance: 2210 AIC: 2240

```



Il est possible de spécifier des modèles plus complexes. Par exemple, `x:y` permet d'indiquer l'interaction entre les variables `x` et `y`. `x * y` sera équivalent à `x + y + x:y`. Pour aller plus loin, voir <http://www.coastal.edu/kingw/statistics/R-tutorials/formulae.html>.

Une présentation plus complète des résultats est obtenue avec `summary` :

```
R> summary(reg)

Call:
glm(formula = sport ~ sexe + grp[25,45) + grp[45,65) + grp[65,93] + etudSecondaire + etudSupérieur + religPratiquant occasionnel + religNi croyance ni appartenance, family = binomial(logit), data = d)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-1.878  -0.886  -0.481   1.003   2.420 

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         -0.79736   0.32383  -2.46   0.01381 *  
sexeHomme                            0.43900   0.10606   4.14   3.5e-05 *** 
grp[25,45)                          -0.42031   0.22804  -1.84   0.06531 .  
grp[45,65)                           -1.08546   0.23770  -4.57   5.0e-06 *** 
grp[65,93]                           -1.37927   0.27379  -5.04   4.7e-07 *** 
etudSecondaire                       0.94831   0.19743   4.80   1.6e-06 *** 
etudTechnique/Professionnel        1.04716   0.18978   5.52   3.4e-08 *** 
etudSupérieur                        1.88962   0.19519   9.68   < 2e-16 *** 
etudNA                               2.14854   0.33020   6.51   7.7e-11 *** 
religPratiquant occasionnel       -0.02060   0.18920  -0.11   0.91331  
religAppartenance sans pratique  -0.00618   0.17471  -0.04   0.97180  
religNi croyance ni appartenance -0.21407   0.19310  -1.11   0.26760 
```

```

religRejet           -0.38274   0.28588  -1.34  0.18062
religNSP ou NVPR    -0.08336   0.41097  -0.20  0.83926
heures.tv           -0.12072   0.03359  -3.59  0.00033 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2607.4 on 1992 degrees of freedom
Residual deviance: 2205.9 on 1978 degrees of freedom
(7 observations deleted due to missingness)
AIC: 2236

Number of Fisher Scoring iterations: 4

```

Dans le cadre d'un modèle logistique, généralement on ne présente pas les coefficients du modèle mais leur valeur exponentielle, cette dernière correspondant en effet à des *odds ratio*, également appelé *rappor des cotes*. L'odds ratio diffère du *risque relatif*. Cependant son interprétation est similaire. Un odds ratio de 1 signifie l'absence d'effet. Un odds ratio largement supérieur à 1 correspond à une augmentation du phénomène étudié et un odds ratio largement inférieur à 1 correspond à une diminution du phénomène étudié¹.

La fonction `coef` permet d'obtenir les coefficients d'un modèle, `confint` leurs intervalles de confiance et `exp` de calculer l'exponentiel. Les odds ratio et leurs intervalles de confiance s'obtiennent ainsi :

```
R> exp(coef(reg))

(Intercept)                      sexeHomme
                           0.4505          1.5512
grpage[25,45]                   grpage [45,65]
                           0.6568          0.3377
grpage[65,93]                   etudSecondaire
                           0.2518          2.5813
etudTechnique/Professionnel     etudSupérieur
                           2.8495          6.6169
etudNA      religPratiquant occasionnel
                           8.5724          0.9796
religAppartenance sans pratique religNi croyance ni appartenance
                           0.9938          0.8073
religRejet           religNSP ou NVPR
                           0.6820          0.9200
heures.tv            0.8863
```

```
R> exp(confint(reg))
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	0.2380	0.8481
sexeHomme	1.2606	1.9107
grpage[25,45]	0.4195	1.0276

1. Pour plus de détails, voir <http://www.spc.univ-lyon1.fr/polycop/odds%20ratio.htm>

grpage[45,65)	0.2115	0.5381
grpage[65,93]	0.1467	0.4297
etudSecondaire	1.7613	3.8240
etudTechnique/Professionnel	1.9764	4.1640
etudSupérieur	4.5427	9.7745
etudNA	4.5265	16.5563
religPratiquant occasionnel	0.6768	1.4217
religAppartenance sans pratique	0.7067	1.4027
religNi croyance ni appartenance	0.5531	1.1798
religRejet	0.3872	1.1899
religNSP ou NVPR	0.3999	2.0222
heures.tv	0.8292	0.9459

On pourra faciliter la lecture en combinant les deux :

```
R> exp(cbind(coef(reg), confint(reg)))
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	0.4505	0.2380
sexeHomme	1.5512	1.2606
grpage[25,45)	0.6568	0.4195
grpage[45,65)	0.3377	0.2115
grpage[65,93]	0.2518	0.1467
etudSecondaire	2.5813	1.7613
etudTechnique/Professionnel	2.8495	1.9764
etudSupérieur	6.6169	4.5427
etudNA	8.5724	4.5265
religPratiquant occasionnel	0.9796	0.6768
religAppartenance sans pratique	0.9938	0.7067
religNi croyance ni appartenance	0.8073	0.5531
religRejet	0.6820	0.3872
religNSP ou NVPR	0.9200	0.3999
heures.tv	0.8863	0.8292

Pour savoir si un odds ratio diffère significativement de 1 (ce qui est identique au fait que le coefficient soit différent de 0), on pourra se référer à la colonne `Pr(>|z|)` obtenue avec `summary`.

Il y a également une petite fonction bien pratique appelée `odds.ratio` et disponible à cette adresse : <http://joseph.lamarange.net/?Calculer-les-Odds-Ratio-d-une>. En premier lieu, on va copier le code de cette fonction et l'exécuter dans R :

```
R> odds.ratio <- function(reg, level = 0.95, digits = 3) {
+   if ("glm" %in% class(reg)) {
+     if (reg$family$family == "binomial") {
+       r <- cbind(exp(coef(reg)), exp(confint(reg, level = level)), summary(reg)$coefficients
+                   [4])
+       r[, 1:3] <- round(r[, 1:3], digits = digits)
+       colnames(r)[1] <- "OR"
+       colnames(r)[4] <- "p"
+       printCoefmat(r, signif.stars = TRUE, has.Pvalue = TRUE)
+     } else {
```

```

+         stop("reg should be a glm with family=binomial or the result of multinom.")
+
+     }
+ } else if ("multinom" %in% class(reg)) {
+   coef <- summary(reg)$coefficients
+   ci <- confint(reg, level = level)
+   # From http://www.ats.ucla.edu/stat/r/dae/mlogit.htm
+   z <- summary(reg)$coefficients/summary(reg)$standard.errors
+   p <- p <- (1 - pnorm(abs(z), 0, 1)) * 2
+   d <- dim(ci)
+   r <- array(NA, c(d[1] * d[3], d[2] + 2))
+   dimnames(r)[[1]] <- rep("", d[1] * d[3])
+   for (i in 1:d[3]) {
+     fl <- (i - 1) * d[1] + 1 #first line
+     ll <- i * d[1] #last line
+     r[fl:ll, ] <- cbind(coef[i, ], ci[, , i], p[i, ])
+     rownames(r)[fl:ll] <- paste0(rownames(coef)[i], "/", colnames(coef))
+   }
+   r[, 1:3] <- round(r[, 1:3], digits = digits)
+   colnames(r) <- c("OR", dimnames(ci)[[2]], "p")
+   printCoefmat(r, signif.stars = TRUE, has.Pvalue = TRUE)
+ } else stop("reg should be a glm with family=binomial or the result of multinom.")
+
}

```

Ensuite, il n'y a plus qu'à l'exécuter :

```
R> odds.ratio(reg)
```

Waiting for profiling to be done...

	OR	2.5 %	97.5 %	p							
(Intercept)	0.451	0.238	0.85	0.01381 *							
sexeHomme	1.551	1.261	1.91	3.5e-05 ***							
grpage[25,45]	0.657	0.420	1.03	0.06531 .							
grpage[45,65]	0.338	0.212	0.54	5.0e-06 ***							
grpage[65,93]	0.252	0.147	0.43	4.7e-07 ***							
etudSecondaire	2.581	1.761	3.82	1.6e-06 ***							
etudTechnique/Professionnel	2.850	1.976	4.16	3.4e-08 ***							
etudSupérieur	6.617	4.543	9.78	< 2e-16 ***							
etudNA	8.572	4.527	16.56	7.7e-11 ***							
religPratiquant occasionnel	0.980	0.677	1.42	0.91331							
religAppartenance sans pratique	0.994	0.707	1.40	0.97180							
religNi croyance ni appartenance	0.807	0.553	1.18	0.26760							
religRejet	0.682	0.387	1.19	0.18062							
religNSP ou NVPR	0.920	0.400	2.02	0.83926							
heures.tv	0.886	0.829	0.95	0.00033 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

L'extension effects propose une représentation graphique résumant les effets de chaque variable du modèle. Attention : il y a un bug dans les versions antérieures à la 2.3-0. Il est recommandé d'installer la dernière version à partir de R-Forge en utilisant la commande suivante :

```
R> library(effects)

Loading required package: lattice
Loading required package: grid
Loading required package: colorspace

R> plot(allEffects(reg))
```

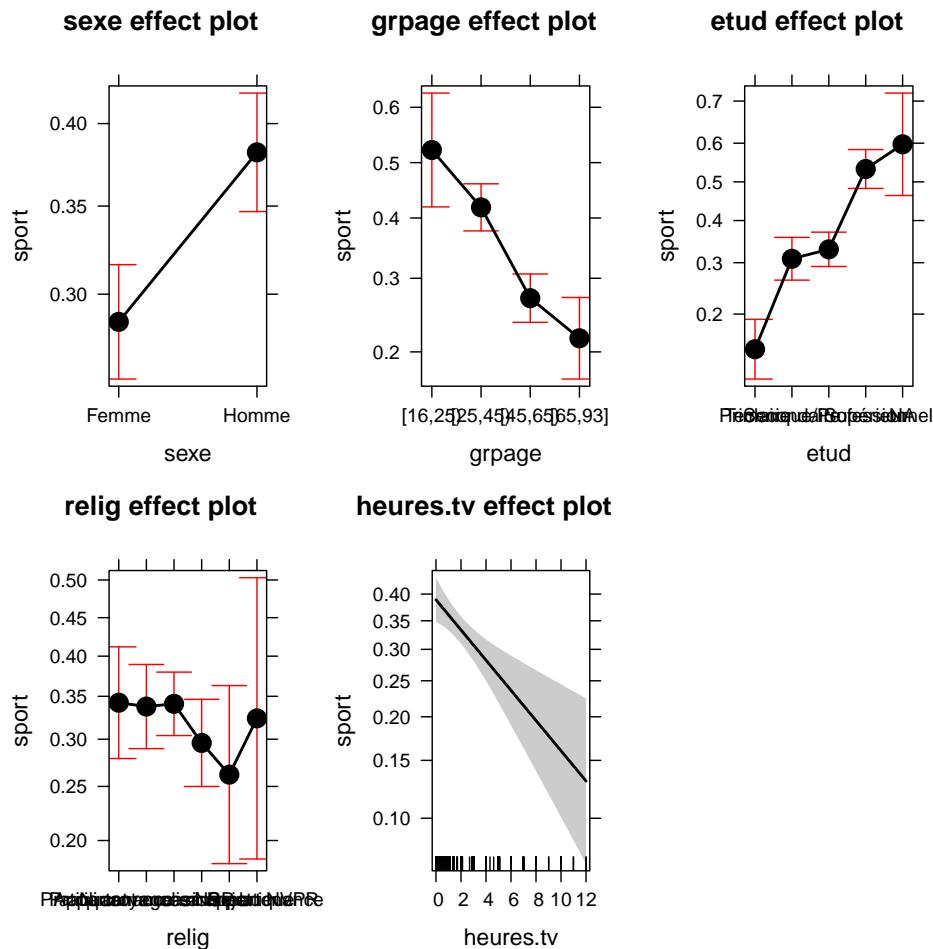


FIGURE 7.1 – Représentation graphique de l'effet de chaque variable du modèle logistique

```
R> install.packages("effects", repos = "http://R-Forge.R-project.org")
```

Nous allons appliquer la fonction `plot` au résultat de la fonction `allEffects`. Nous obtenons alors la figure 7.1 de la présente page.

Une manière de tester la qualité d'un modèle est le calcul d'une *matrice de confusion*, c'est-à-dire le tableau croisé des valeurs observées et celles des valeurs prédites en appliquant le modèle aux données d'origine.

La fonction `predict` avec l'argument `type="response"` permet d'appliquer notre modèle logistique à un tableau de données et renvoie pour chaque individu la probabilité qu'il ait vécu le phénomène étudié.

```
R> sport.pred <- predict(reg, type = "response", newdata = d)
R> head(d$sport.pred)

NULL
```

Or notre variable étudiée est de type binaire. Nous devons donc transformer nos probabilités prédictes en une variable du type « oui/non ». Usuellement, les probabilités prédictes seront réunies en deux groupes selon qu'elles soient supérieures ou inférieures à la moitié. La matrice de confusion est alors égale à :

```
R> table(sport.pred > 0.5, d$sport)
```

	Non	Oui
FALSE	1074	384
TRUE	199	336

Nous avons donc 583 (384+199) prédictions incorrectes sur un total de 1993, soit un taux de mauvais classement de 29,3 %.

7.3 Sélection de modèles

Il est toujours tentant lorsque l'on recherche les facteurs associés à un phénomène d'inclure un nombre important de variables explicatives potentielles dans un modèle logistique. Cependant, un tel modèle ne sera forcément le plus efficace et certaines variables n'auront probablement pas d'effet significatif sur la variable d'intérêt.

La technique de *sélection descendante pas à pas* est une approche visant à améliorer son modèle explicatif². On réalise un premier modèle avec toutes les variables spécifiées, puis on regarde s'il est possible d'améliorer le modèle en supprimant une des variables du modèle. Si plusieurs variables permettent d'améliorer le modèle, on supprimera la variable dont la suppression améliorera le plus le modèle. Puis on recommence le même procédé pour voir si la suppression d'une seconde variable peut encore améliorer le modèle et ainsi de suite. Lorsque le modèle ne peut plus être amélioré par la suppression d'une variable, on s'arrête.

Il faut également définir un critère pour déterminer la qualité d'un modèle. L'un des plus utilisés est le *Akaike information criterion* ou AIC. Plus l'AIC sera faible, meilleure sera le modèle.

La fonction `step` permet justement de sélectionner le meilleur modèle par une procédure pas à pas descendante basée sur la minimisation de l'AIC. La fonction affiche à l'écran les différentes étapes de la sélection et renvoie le modèle final.

```
R> reg2 <- step(reg)

Start:  AIC=2236
sport ~ sexe + grpage + etud + relig + heures.tv

          Df Deviance AIC
- relig      5     2210 2230
<none>           2206 2236
- heures.tv  1     2219 2247
- sexe       1     2223 2251
```

2. Il existe également des méthodes de *sélection ascendante pas à pas*, mais nous les aborderons pas ici.

```

- grpage     3      2259 2283
- etud       4      2330 2352

Step:  AIC=2230
sport ~ sexe + grpae + etud + heures.tv

          Df Deviance  AIC
<none>            2210 2230
- heures.tv   1      2224 2242
- sexe        1      2226 2244
- grpae       3      2260 2274
- etud        4      2334 2346

```

Le modèle initial a un AIC de 2114,8. À la première étape, il apparait que la suppression de la variable religion permet diminuer l'AIC à 2109,1. Lors de la seconde étape, toute suppression d'une autre variable ferait augmenter l'AIC. La procédure s'arrête donc.

7.4 Régression logistique multinomiale

La régression logistique multinomiale est une extension de la régression logistique aux variables qualitatives à trois modalités ou plus. Dans ce cas de figure, chaque modalité de la variable d'intérêt sera comparée à la modalité de référence. Les odds ratio seront donc exprimés par rapport à cette dernière.

Nous allons prendre pour exemple la variable `trav.satisf`, à savoir la satisfaction ou l'insatisfaction au travail.

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

      n    %
Satisfaction 480 24.0
Insatisfaction 117 5.9
Equilibre     451 22.6
NA           952 47.6
```

Nous allons choisir comme modalité de référence la position intermédiaire, à savoir l'« équilibre ». De plus, nous n'allons conserver pour l'analyse

```
R> d$trav.satisf <- relevel(d$trav.satisf, "Equilibre")
```

Enfin, nous allons aussi en profiter pour raccourcir les étiquettes de la variable `trav.imp` :

```
R> levels(d$trav.imp) <- c("Le plus", "Aussi", "Moins", "Peu")
```

Pour calculer un modèle logistique multinomial, nous allons utiliser la fonction `multinom` de l'extension `nnet`³. Si l'extension n'est pas disponible, vous devrez l'installer (voir section B.2 page 185). La syntaxe de `multinom` est similaire à celle de `glm`, le paramètre `family` en moins.

```
R> library(nnet)
R> regm <- multinom(trav.satisf ~ sexe + etud + grpae + trav.imp, data = d)
```

3. Une alternative est d'avoir recours à l'extension `mlogit` que nous n'aborderons pas ici.

```
# weights: 39 (24 variable)
initial value 1151.345679
iter 10 value 977.348901
iter 20 value 969.849189
iter 30 value 969.522965
final value 969.521855
converged
```

Comme pour la régression logistique, il est possible de réaliser une sélection pas à pas descendante :

```
R> regm2 <- step(regm)

Start: AIC=1987
trav.satisf ~ sexe + etud + grpage + trav.imp

trying - sexe
# weights: 36 (22 variable)
initial value 1151.345679
iter 10 value 978.538886
iter 20 value 970.453555
iter 30 value 970.294459
final value 970.293988
converged
trying - etud
# weights: 27 (16 variable)
initial value 1151.345679
iter 10 value 987.907714
iter 20 value 981.785467
iter 30 value 981.762800
final value 981.762781
converged
trying - grpae
# weights: 30 (18 variable)
initial value 1151.345679
iter 10 value 979.485430
iter 20 value 973.175923
final value 973.172389
converged
trying - trav.imp
# weights: 30 (18 variable)
initial value 1151.345679
iter 10 value 998.803976
iter 20 value 994.417973
iter 30 value 994.378914
final value 994.378869
converged
      Df  AIC
- grpae  18 1982
- sexe   22 1985
<none>  24 1987
- etud   16 1996
- trav.imp 18 2025
# weights: 30 (18 variable)
```

```
initial value 1151.345679
iter 10 value 979.485430
iter 20 value 973.175923
final value 973.172389
converged

Step: AIC=1982
trav.satisf ~ sexe + etud + trav.imp

trying - sexe
# weights: 27 (16 variable)
initial value 1151.345679
iter 10 value 976.669670
iter 20 value 973.928385
iter 20 value 973.928377
iter 20 value 973.928377
final value 973.928377
converged
trying - etud
# weights: 18 (10 variable)
initial value 1151.345679
iter 10 value 988.413720
final value 985.085797
converged
trying - trav.imp
# weights: 21 (12 variable)
initial value 1151.345679
iter 10 value 1001.517287
final value 998.204280
converged
      Df  AIC
- sexe     16 1980
<none>    18 1982
- etud     10 1990
- trav.imp 12 2020
# weights: 27 (16 variable)
initial value 1151.345679
iter 10 value 976.669670
iter 20 value 973.928385
iter 20 value 973.928377
iter 20 value 973.928377
final value 973.928377
converged

Step: AIC=1980
trav.satisf ~ etud + trav.imp

trying - etud
# weights: 15 (8 variable)
initial value 1151.345679
iter 10 value 986.124104
final value 986.034023
converged
```

```

trying - trav.imp
# weights: 18 (10 variable)
initial value 1151.345679
iter 10 value 1000.225356
final value 998.395273
converged
      Df AIC
<none> 16 1980
- etud     8 1988
- trav.imp 10 2017

```

La plupart des fonctions vues précédemment fonctionnent⁴, de même que la représentation graphique des effets (figure 7.2 page ci-contre).

```

R> summary(regm2)

Call:
multinom(formula = trav.satisf ~ etud + trav.imp, data = d)

Coefficients:
              (Intercept) etudSecondaire etudTechnique/Professionnel
Satisfaction    -0.1111        0.04916          0.07793
Insatisfaction   -1.1214       -0.09738          0.08393
                  etudSupérieur etudNA trav.impAussi trav.impMoins
Satisfaction      0.69950  -0.53842        0.2579       -0.1756
Insatisfaction    0.07755  -0.04364       -0.2280       -0.5330
                  trav.impPeu
Satisfaction      -0.5995
Insatisfaction     1.3402

Std. Errors:
              (Intercept) etudSecondaire etudTechnique/Professionnel
Satisfaction      0.4521        0.2636          0.2408
Insatisfaction    0.6517        0.4000          0.3580
                  etudSupérieur etudNA trav.impAussi trav.impMoins
Satisfaction      0.2473  0.5911        0.4261       0.4116
Insatisfaction    0.3831  0.8408        0.6214       0.5942
                  trav.impPeu
Satisfaction      0.5580
Insatisfaction     0.6587

Residual Deviance: 1948
AIC: 1980

```

```
R> odds.ratio(regm2)
```

	OR	2.5 %	97.5 %	p
Satisfaction/(Intercept)	-0.111	-0.997	0.78	0.8059
Satisfaction/etudSecondaire	0.049	-0.467	0.57	0.8520
Satisfaction/etudTechnique/Professionnel	0.078	-0.394	0.55	0.7463
Satisfaction/etudSupérieur	0.700	0.215	1.18	0.0047 **

4. Voir <http://www.ats.ucla.edu/stat/r/dae/mlogit.htm> (en anglais) pour plus de détails.

```
R> library(effects)
R> plot(allEffects(regm2))

Error: subscript out of bounds
```

FIGURE 7.2 – Représentation graphique de l'effet de chaque variable du modèle logistique

```
Satisfaction/etudNA           -0.538 -1.697  0.62 0.3624
Satisfaction/trav.impAussi    0.258 -0.577  1.09 0.5450
Satisfaction/trav.impMoins   -0.176 -0.982  0.63 0.6696
Satisfaction/trav.impPeu     -0.600 -1.693  0.49 0.2827
Insatisfaction/(Intercept)   -1.121 -2.399  0.16 0.0853 .
Insatisfaction/etudSecondaire -0.097 -0.881  0.69 0.8077
Insatisfaction/etudTechnique/Professionnel 0.084 -0.618  0.79 0.8146
Insatisfaction/etudSupérieur   0.078 -0.673  0.83 0.8396
Insatisfaction/etudNA          -0.044 -1.691  1.60 0.9586
Insatisfaction/trav.impAussi   -0.228 -1.446  0.99 0.7137
Insatisfaction/trav.impMoins   -0.533 -1.698  0.63 0.3697
Insatisfaction/trav.impPeu     1.340  0.049  2.63 0.0419 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De même, il est possible de calculer la matrice de confusion :

```
R> table(predict(regm2, newdata = d), d$trav.satisf)
```

	Equilibre	Satisfaction	Insatisfaction
Equilibre	262	211	49
Satisfaction	171	258	45
Insatisfaction	18	11	23

7.5 Exercices

Exercice 7.21

▷ *Solution page 196*

Nous allons utiliser le fichier de données `Aids2` fourni par l'extension MASS. Chargez cette extension en mémoire, puis utilisez la commande `data(Aids2)` pour charger ce fichier de données. Un descriptif (en anglais) de ce tableau de données est disponible via la commande `?Aids2`. Calculer un modèle de régression logistique évaluant l'effet du sexe, de la région (variable `state`), de l'âge et de la catégorie de transmission sur la probabilité d'être toujours en vie. Représentez graphiquement l'effet des variables explicatives. Calculez des groupes d'âges et refaites le modèle ainsi que le graphique. Calculez les odds ratio. Enfin, réalisez une sélection descendante pas à pas.

Partie 8

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] 78.08 31092.14
```

8.1 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`.

8.2 Fonctions de l'extension `questionr`

L'extension `questionr` propose quelques fonctions permettant de calculer des statistiques simples pondérées² :

`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> mean(d$age)
[1] 48.16

R> library(questionr)
R> wtd.mean(d$age, weights = d$poids)
[1] 46.35

R> wtd.var(d$age, weights = d$poids)
[1] 325.3
```

Pour les tris à plat, on utilise la fonction `wtd.table` à laquelle on passe la variable en paramètre :

```
R> wtd.table(d$sex, weights = d$poids)

Homme    Femme
5149382 5921844
```

Pour un tri croisé, il suffit de passer deux variables en paramètres :

```
R> wtd.table(d$sex, d$hard.rock, weights = d$poids)

Non      Oui
Homme 5109366 40016
Femme 5872596 49247
```

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...

8.3 Présentation de 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* ou *plan d'échantillonnage* le plus simple, avec une variable de pondération déjà calculée. Ceci se fait à l'aide de la fonction **svydesign** :

```
R> library(survey)

Attaching package: 'survey'
L'objet suivant est masqué from 'package:graphics':
  dotchart

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] 48.16

R> mean(dw$variables$age)
[1] 48.16
```

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, svyquantile** 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*. L'intervalle de confiance d'une moyenne s'obtient avec **confint** et celui d'une proportion avec **svyciprop**.

Voyons tout de suite quelques exemples³ :

```
R> svymean(~age, dw)

  mean    SE
age 46.3 0.53
```

3. Pour d'autres exemples, voir http://www.ats.ucla.edu/stat/r/faq/svy_r_oscluster.htm (en anglais).

```
R> confint(svymean(~age, dw)) # Intervalle de confiance
  2.5 % 97.5 %
age 45.31 47.38

R> svyquantile(~age, dw, quantile = c(0.25, 0.5, 0.75), ci = TRUE)

$quantiles
  0.25 0.5 0.75
age   31   45   60

$CIs
, , age
  0.25 0.5 0.75
(lower   30   43   58
upper) 32   47   62

R> svyvar(~heures.tv, dw, na.rm = TRUE)

      variance     SE
heures.tv    2.99 0.18

R> svytable(~sexe, dw)

sexe
  Homme Femme
5149382 5921844

R> svyciprop(~sexe, dw) # Intervalle de confiance
  2.5% 97.5%
sexe 0.535 0.507 0.56

R> svytable(~sexe + clso, dw)

  clso
sexe      Oui     Non Ne sait pas
  Homme 2658744 2418188        72451
  Femme 2602032 3242389        77423
```

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
```

```

sexes       Oui      Non Ne sait pas
Homme 2658744 2418188          72451
Femme 2602032 3242389          77423

R> lprop(tab)

clso
sexes       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 = 3.333, ndf = 1.973, ddf = 3944.902, p-value = 0.03641

```

questionr

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** :

```

R> tab <- svytable(~peche.chasse, dw)
R> freq(tab, total = TRUE)

      n      %
Non  9716683 87.8
Oui 1354544 12.2
Total 11071226 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 = 3.333, ndf = 1.973, ddf = 3944.902, p-value = 0.03641

```

Le principe de la fonction **svyby** est similaire à celui de **tapply** (voir section 5.3.3 page 67). 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 45.20 0.7419
Femme Femme 47.34 0.7421

```

```
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")
```

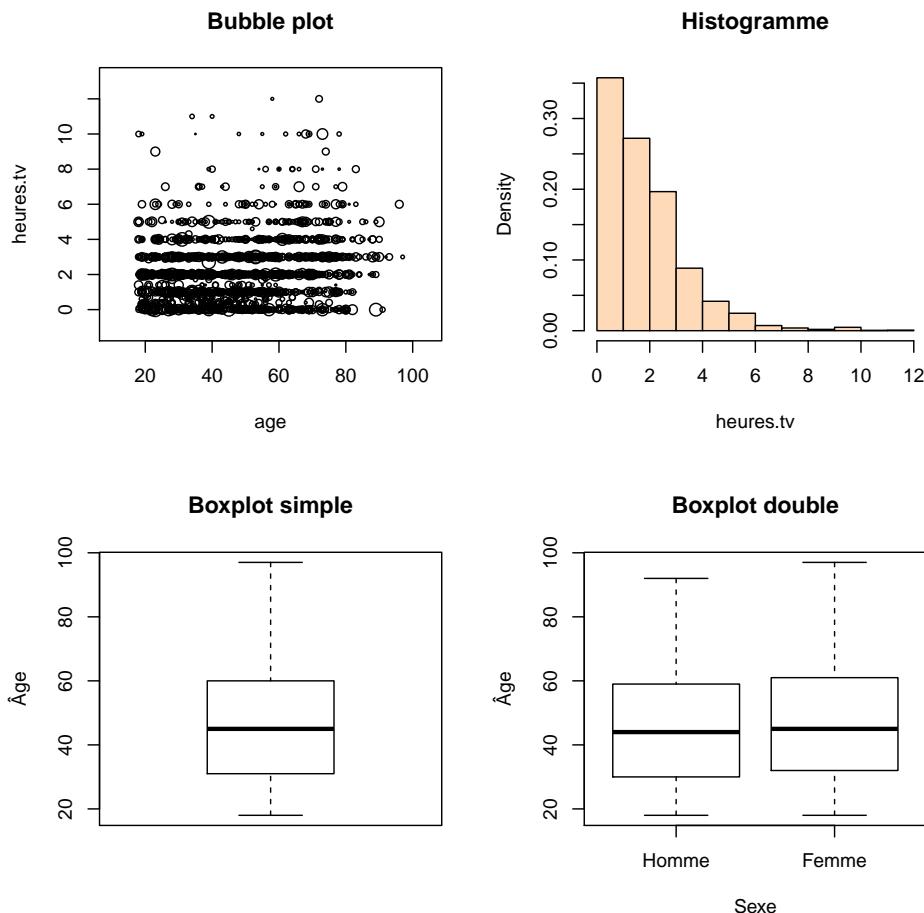


FIGURE 8.1 – Fonctions graphiques de l'extension survey

Enfin, **survey** est également capable de produire des graphiques à partir des données pondérées. Des exemples sont donnés figure 8.1 de la présente page.

Enfin, **survey** fournit une fonction **svyglm** permettant de calculer un modèle statistique tout en prenant en compte le plan d'échantillonnage spécifié. La syntaxe de **svyglm** est proche de celle **glm** :

```
R> reg <- svyglm(sport ~ sexe + age + relig + heures.tv, dw, family = binomial(logit))
Warning: non-integer #successes in a binomial glm!
```

Le résultat obtenu est similaire à celui de **glm** et l'on peut utiliser sans problème les fonctions **coef**, **confint**, **odds.ratio** ou **predict** abordées au chapitre 7.

Par contre, la sélection descendante pas à pas d'un modèle par minimisation de l'AIC avec **step** ne fonctionnera pas. En effet, il semble qu'il n'existe pas encore d'analogue de l'AIC dans le contexte

d'un plan d'échantillonage complexe⁴. On pourra se rabattre néanmoins sur une sélection raisonnée des variables à conserver ne prenant par exemple en compte que celles ayant un effet significatif à 5 % ou 10 %.

Par ailleurs, la fonction `allEffects` est elle aussi incompatible avec `svyglm`⁵.

8.4 Définir un plan d'échantillonage complexe avec survey

L'extension `survey` ne permet pas seulement d'indiquer une variable de pondération mais également de prendre les spécificités du plan d'échantillonnage (strates, grappes, ...). Le plan d'échantillonnage ne joue pas seulement sur la pondération des données, mais influence le calcul des variances et par ricochet tous les tests statistiques. Deux échantillons identiques avec la même variable de pondération mais des designs différents produiront les mêmes moyennes et proportions mais des intervalles de confiance différents.

8.4.1 Différents types d'échantillonnage

L'*échantillonnage aléatoire simple* ou *échantillonnage equiprobable* est une méthode pour laquelle tous les échantillons possibles (de même taille) ont la même probabilité d'être choisis et tous les éléments de la population ont une chance égale de faire partie de l'échantillon. C'est l'échantillonnage le plus simple : chaque individu à la même probabilité d'être sélectionné.

L'*échantillonnage stratifié* est une méthode qui consiste d'abord à subdiviser la population en groupes homogènes (strates) pour ensuite extraire un échantillon aléatoire de chaque strate. Cette méthode suppose la connaissance de la structure de la population. Pour estimer les paramètres, les résultats doivent être pondérés par l'importance relative de chaque strate dans la population.

L'*échantillonnage par grappes* est une méthode qui consiste à choisir un échantillon aléatoire d'unités qui sont elles-mêmes des sous-ensembles de la population (« grappes »). Cette méthode suppose que les unités de chaque grappe sont représentatives. Elle possède l'avantage d'être souvent plus économique.

Il est possible de combiner plusieurs de ces approches. Par exemple, les *enquêtes démographiques et de santé*⁶ (EDS) sont des enquêtes stratifiées en grappes à deux degrés. Dans un premier temps, la population est divisée en strates par région et milieu de résidence. Dans chaque strate, des zones d'enquêtes, correspondant à des unités de recensement, sont tirées au sort avec une probabilité proportionnelle au nombre de ménages de chaque zone au dernier recensement de population. Enfin, au sein de chaque zone d'enquête sélectionnée, un recensement de l'ensemble des ménages est effectué puis un nombre identique de ménages par zone d'enquête est tiré au sort de manière aléatoire simple.

8.4.2 Les options de svydesign

La fonction `svydesign` accepte plusieurs arguments décrits sur sa page d'aide (obtenue avec la commande `?svydesign`).

L'argument `data` permet de spécifier le tableau de données contenant les observations.

L'argument `ids` est obligatoire et spécifie sous la forme d'une formule les identifiants des différents niveaux d'un tirage en grappe. S'il s'agit d'un échantillon aléatoire simple, on entrera `ids=~1`. Autre situation : supposons une étude portant sur la population française. Dans un premier temps, on a tiré au sort un certain nombre de départements français. Dans un second temps, on tire au sort dans chaque département des communes. Dans chaque commune sélectionnée, on tire au sort des quartiers. Enfin, on interroge de manière exhaustive toutes les personnes habitant les quartiers enquêtés. Notre fichier de données devra

4. Voir cette discussion <https://groups.google.com/forum/#!topic/r-help-archive/XcrSW9s7kwI>.

5. Compatibilité qui pourra éventuellement être introduite dans une future version de l'extention `effects`.

6. Vaste programme d'enquêtes réalisées à intervalles réguliers dans les pays en développement, disponibles sur <http://www.measuredhs.com/>.

donc comporter pour chaque observation les variables `id_departement`, `id_commune` et `id_quartier`. On écrira alors pour l'argument `ids` la valeur suivante : `ids=~id_departement+id_commune+id_quartier`.

Si l'échantillon est stratifié, on spécifiera les strates à l'aide de l'argument `strata` en spécifiant la variable contenant l'identifiant des strates. Par exemple : `strata=~id_strate`.

Il faut encore spécifier les probabilités de tirage de chaque cluster ou bien la pondération des individus. Si l'on dispose de la probabilité de chaque observation d'être sélectionnée, on utilisera l'argument `probs`. Si, par contre, on connaît la pondération de chaque observation (qui doit être proportionnelle à l'inverse de cette probabilité), on utilisera l'argument `weights`.

Si l'échantillon est stratifié, qu'au sein de chaque strate les individus ont été tirés au sort de manière aléatoire et que l'on connaît la taille de chaque strate, il est possible de ne pas avoir à spécifier la probabilité de tirage ou la pondération de chaque observation. Il est préférable de fournir une variable contenant la taille de chaque strate à l'argument `fpc`. De plus, dans ce cas-là, une petite correction sera appliquée au modèle pour prendre en compte la taille finie de chaque strate.

Quelques exemples :

```
R> # Échantillonnage aléatoire simple
R> plan <- svydesign(ids = ~1, data = donnees)
R>
R> # Échantillonnage stratifié à un seul niveau (la taille de chaque strate est
R> # connue)
R> plan <- svydesign(ids = ~1, data = donnees, fpc = ~taille)
R>
R> # Échantillonnage en grappes avec tirages à quatre degrés (departement,
R> # commune, quartier, individus). La probabilité de tirage de chaque niveau
R> # de cluster est connue.
R> plan <- svydesign(ids = ~id_departement + id_commune + id_quartier, data = donnees,
+     Probs = ~proba_departement + proba_commune + proba_quartier)
R>
R> # Échantillonnage stratifié avec tirage à deux degrés (clusters et
R> # individus). Le poids statistiques de chaque observation est connu.
R> plan <- svydesign(ids = ~id_cluster, data = donnees, strata = ~id_strate, weights = ~poids)
```

Prenons l'exemple d'une enquête démographique et de santé. Le nom des différentes variables est standardisé et commun quelle que soit l'enquête. Nous supposerons que vous avez importé le fichier *individus* dans un tableau de données nommés `eds`. Le poids statistique de chaque individu est fourni par la variable `V005` qui doit au préalable être divisée par un million. Les grappes d'échantillonnage au premier degré sont fournies par la variable `V021` (*primary sample unit*). Si elle n'est pas renseignée par le numéro de grappe `V001` Enfin, le milieu de résidence (urbain / rural) est fourni par `V025` et la région par `V024`. Pour rappel, l'échantillon a été stratifié à la fois par région et par milieu de résidence. Certaines enquêtes fournissent directement un numéro de strate via `V022`. Si tel est le cas, on pourra préciser le plan d'échantillonnage ainsi :

```
R> eds$poids <- eds$V005/1e+06
R> design.eds <- svydesign(ids = ~V021, data = eds, strata = ~V022, weights = ~poids)
```

Si `V022` n'est pas fourni mais que l'enquête a bien été stratifié par région et milieu de résidence (vérifiez toujours le premier chapitre du rapport d'enquête), on pourra créer une variable `strata` ainsi⁷ :

7. L'astuce consiste à utiliser `as.integer` pour obtenir le code des facteurs et non leur valeur textuelle. L'addition des deux valeurs après multiplication du code de la région par 10 permet d'obtenir une valeur unique pour chaque combinaison des deux variables. On retrouve le résultat en facteurs puis on modifie les étiquettes des modalités.

```
R> eds$strate <- as.factor(as.integer(eds$V024) * 10 + as.integer(eds$V025))
R> levels(eds$strate) <- c(paste(levels(eds$V024), "Urbain"), paste(levels(eds$V024),
+ "Rural"))
R> design.eds <- svydesign(ids = ~V021, data = eds, strata = ~strate, weights = ~poids)
```

8.4.3 Extraire un sous-échantillon

Si l'on souhaite travailler sur un sous-échantillon tout en gardant les informations d'échantillonnage, on utilisera la fonction `subset` présentée en détail section 5.3.2 page 66.

```
R> sous <- subset(dw, sexe == "Femme" & age >= 40)
```

8.5 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 ;

Partie 9

Analyse des correspondances multiples (ACM)

Il existe plusieurs techniques d'*analyse factorielle* dont les plus courantes sont l'*analyse en composante principale* (ACP) porte sur des variables quantitatives, l'*analyse factorielle des correspondances* (AFC) porte sur deux variables qualitatives et l'*analyse des correspondances multiples* (ACM) sur plusieurs variables qualitatives (il s'agit d'une extension de l'AFC). Pour combiner des variables à la fois quantitatives et qualitatives, on pourra avoir recours à l'analyse mixte de Hill et Smith.

Bien que ces techniques soient disponibles dans les extensions standards de R, il est souvent préférable d'avoir recours à deux autres extensions plus complètes, `ade4` et `FactoMineR`, chacune ayant ses avantages et des possibilités différentes. Voici les fonctions les plus fréquentes :

Analyse	Variables	Fonction standard	Fonction ade4	Fonction FactoMineR
ACP	plusieurs variables quantitatives	<code>princomp</code> (<code>stats</code>)	<code>dudi.pca</code>	PCA
AFC	deux variables qualitatives	<code>corresp</code> (<code>MASS</code>)	<code>dudi.coa</code>	CA
ACM	plusieurs variables qualitatives	<code>mca</code> (<code>MASS</code>)	<code>dudi.acm</code>	MCA
Analyse mixte de Hill et Smith	plusieurs variables quantitatives et/ou qualitatives	—	<code>dudi.mix</code>	—

Dans la suite de ce chapitre, nous n'arboderons que l'analyse des correspondances multiples (ACM).



On trouvera également de nombreux supports de cours en français sur l'analyse factorielle sur le site de François Gilles Carpentier : <http://geai.univ-brest.fr/~carpenti/>.

9.1 Principe général

L'analyse des correspondances multiples est une technique descriptive visant à résumer l'information contenu dans un grand nombre de variables afin de faciliter l'interprétation des corrélations existantes

entre ces différentes variables. On cherche à savoir quelles sont les modalités corrélées entre elles.

L'idée générale est la suivante¹. L'ensemble des individus peut être représenté dans un espace à plusieurs dimensions où chaque axe représente les différentes variables utilisées pour décrire chaque individu. Plus précisément, pour chaque variable qualitative, il y a autant d'axes que de modalités moins un. Ainsi il faut trois axes pour décrire une variable à quatre modalités. Un tel nuage de points est aussi difficile à interpréter que de lire directement le fichier de données. On ne voit pas les corrélations qu'il peut y avoir entre modalités, par exemple qu'aller au cinéma est plus fréquent chez les personnes habitant en milieu urbain. Afin de mieux représenter ce nuage de points, on va procéder à un changement de systèmes de coordonnées. Les individus seront dès lors projetés et représentés sur un nouveau système d'axe. Ce nouveau système d'axes est choisi de telle manière que la majorité des variations soit concentrées sur les premiers axes. Les deux-trois premiers axes permettront d'expliquer la majorité des différences observées dans l'échantillon, les autres axes n'apportant qu'une faible part additionnelle d'information. Dès lors, l'analyse pourra se concentrer sur ses premiers axes qui constitueront un bon résumé des variations observables dans l'échantillon.

Avant toute ACM, il est indispensable de réaliser une analyse préliminaire de chaque variable, afin de voir si toutes les classes sont aussi bien représentées ou s'il existe un déséquilibre. L'ACM est sensible aux effectifs faibles, aussi regrouper les classes quand cela est nécessaire.

9.2 ACM avec ade4

Si l'extension `ade4` n'est pas présente sur votre PC, il vous faut l'installer :

```
R> install.packages("ade4", dep = TRUE)
```

Comme précédemment, nous utiliserons le fichier de données `hdv2003` fourni avec l'extension `questionr`.

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

En premier lieu, comme dans le chapitre 7 sur la régression logistique, nous allons créer une variable groupe d'âges et regrouper les modalités de la variable « niveau d'étude ».

```
R> d$grpage <- cut(d$age, c(16, 25, 45, 65, 93), right = FALSE, include.lowest = TRUE)
R> d$etud <- d$nivetud
R> levels(d$etud) <- c("Primaire", "Primaire", "Primaire", "Secondaire", "Secondaire",
+   "Technique/Professionnel", "Technique/Professionnel", "Supérieur")
```

Ensuite, nous allons créer un tableau de données ne contenant que les variables que nous souhaitons prendre en compte pour notre analyse factorielle.

```
R> dt <- d[, c("grpage", "sexe", "etud", "peche.chasse", "cinema", "cuisine", "bricol",
+   "sport", "lecture.bd")]
```

Le calcul de l'ACM se fait tout simplement avec la fonction `dudi.acm`.

```
R> acm <- dudi.acm(dt)
```

1. Pour une présentation plus détaillée, voir <http://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf>.

```
R> screeplot(acm)
Error: object 'acm' not found
```

FIGURE 9.1 – Valeurs propres ou inerties de chaque axe

```
R> s.corcircle(acm$co, 1, 2, clabel = 0.7)
Error: could not find function "s.corcircle"
```

FIGURE 9.2 – Cercle de corrélations des modalités sur les deux premiers axes

Par défaut, la fonction affichera le graphique des valeurs propres de chaque axe (nous y reviendrons) et vous demandera le nombre d'axes que vous souhaitez conserver dans les résultats. Le plus souvent, cinq axes seront largement plus que suffisants. Vous pouvez également éviter cette étape en indiquant directement à `dudi.acm` de vous renvoyer les cinq premiers axes ainsi.

```
R> acm <- dudi.acm(dt, scanff = FALSE, nf = 5)
Error: could not find function "dudi.acm"
```

Le graphique des valeurs propres peut être reproduit avec `screeplot` (voir figure 9.1 de la présente page). Les mêmes valeurs pour les premiers axes s'obtiennent également avec `summary`².

```
R> summary(acm)
Error: object 'acm' not found
```

L'inertie totale est de 1,451 et l'axe 1 en explique 0,1474 soit 17 %. L'inertie projetée cumulée nous indique que les deux premiers axes expliquent à eux seuls 29 % des variations observées dans notre échantillon.

Pour comprendre la signification des différents axes, il importe d'identifier quelles sont les variables/-modalités qui contribuent le plus à chaque axe. Une première représentation graphique est le cercle de corrélation des modalités. Pour cela, on aura recours à `s.corcircle` (voir figure 9.2 de la présente page). On indiquera d'abord `acm$co` si l'on souhaite représenter les modalités ou `acm$li` si l'on souhaite représenter les individus. Les deux chiffres suivant indiquent les deux axes que l'on souhaite afficher (dans le cas présent les deux premiers axes). Enfin, le paramètre `clabel` permet de modifier la taille des étiquettes.

On pourra avoir également recours à `boxplot` pour visualiser comment se répartissent les modalités de chaque variable sur un axe donné³. Voir les figures 9.3 et 9.4 de la présente page.

Le tableau `acm$cr` contient les rapports de corrélation (variant de 0 à 1) entre les variables et les axes choisis au départ de l'ACM. Pour représenter graphiquement ces rapports, utiliser

2. On pourra également avoir recours à la fonction `inertia.dudi` pour l'ensemble des axes.
3. La fonction `score` constituera également une aide à l'interprétation des axes.

```
R> boxplot(acm)
Error: object 'acm' not found
```

FIGURE 9.3 – Répartition des modalités selon le premier axe

```
R> boxplot(acm, 2)
Error: object 'acm' not found
```

FIGURE 9.4 – Répartition des modalités selon le second axe

```
R> par(mfrow = c(2, 2))
R> for (i in 1:4) barplot(acm$cr[, i], names.arg = row.names(acm$cr), las = 2,
+   main = paste("Axe", i))
Error: object 'acm' not found
R> par(mfrow = c(1, 1))
```

FIGURE 9.5 – Rapports de corrélation des variables sur les 4 premiers axes

`barplot(acm$cr[,num],names.arg=row.names(acm$cr),las=2)` où `num` est le numéro de l'axe à représenter (voir figure 9.5 de la présente page). Pour l'interprétation des axes, se concentrer sur les variables les plus structurantes, c'est-à-dire dont le rapport de corrélation est le plus proche de 1.

Pour représenter, les individus ou les modalités dans le plan factoriel, on utilisera la fonction `s.label`. Il est bien sûr possible de préciser les axes à représenter. L'argument `boxes` permet d'indiquer si l'on souhaite tracer une boîte pour chaque modalité.

La figure 9.6 de la présente page représente les modalités sur les deux premiers axes tandis que la figure 9.7 les représente selon les axes 3 et 4. La figure 9.8, quant à elle, représente les individus selon les deux premiers axes. En indiquant `clabel=0` (une taille nulle pour les étiquettes), `s.label` remplace chaque observation par un symbole qui peut être spécifié avec `pch` (pour les différentes valeurs possibles, voir la figure 3.10 page 40).



Lorsque l'on réalise une ACM, il n'est pas rare que plusieurs observations soient identiques, c'est-à-dire correspondent à la même combinaison de modalités. Dès lors, ces observations seront projetées sur le même point dans le plan factoriel. Une représentation classique des observations avec `s.label` ne permettra pas de rendre des effectifs de chaque point. Vous trouverez à cette adresse (<http://joseph.larmarange.net/?article147>) une petite fonction `s.freq` représentant chaque point par un carré proportionnel au nombre d'individus.



Gaston Sanchez propose un graphique amélioré des modalités dans le plan factoriel à cette adresse : <http://rpubs.com/gaston/MCA>.

```
R> s.label(acm$co, clabel = 0.7)
Error: could not find function "s.label"
```

FIGURE 9.6 – Répartition des modalités selon les deux premiers axes

```
R> s.label(acm$co, 3, 4, clabel = 0.7, boxes = FALSE)
Error: could not find function "s.label"
```

FIGURE 9.7 – Répartition des modalités selon les axes 3 et 4

```
R> s.label(acm$li, clabel = 0, pch = 17)
Error: could not find function "s.label"
```

FIGURE 9.8 – Répartition des individus selon les deux premiers axes

La fonction `s.value` permet notamment de représenter un troisième axe factoriel. Sur la figure 9.9 de la présente page, nous projettons les individus selon les deux premiers axes factoriels. La taille et la couleur des carrés dépendent pour leur part de la coordonnée des individus sur le troisième axe factoriel. Le paramètre `csi` permet d'ajuster la taille des carrés.

`s.arrow` permet de représenter les vecteurs variables ou les vecteurs individus sous la forme d'une flèche allant de l'origine du plan factoriel aux coordonnées des variables/individus (voir figure 9.10 page suivante).

`s.hist` permet de représenter des individus (ou des modalités) sur le plan factoriel et d'afficher leur distribution sur chaque axe (figure 9.11 page suivante).

`s.class` et `s.chull` permettent de représenter les différentes observations classées en plusieurs catégories. Cela permet notamment de projeter certaines variables. `s.class` représente les observations par des points, lie chaque observation au barycentre de la modalité à laquelle elle appartient et dessine une ellipse représentant la forme générale du nuage de points (figure 9.12 page suivante). `s.chull` représente les barycentres de chaque catégorie et dessine des lignes de niveaux représentant la distribution des individus de cette catégorie. Les individus ne sont pas directement représentés (figure 9.13 page 133).



Il est préférable de fournir une liste de couleurs (via le paramètre `col`) pour rendre le graphique plus lisible. Si vous avez installé l'extension `RColorBrewer`, vous pouvez utiliser les différentes palettes de couleurs proposées. Pour afficher les palettes disponibles, utilisez `display.brewer.all`. Pour obtenir une palette de couleurs, utilisez la fonction `brewer.pal` avec les arguments `n` (nombre de couleurs demandées) et `pal` (nom de la palette de couleurs désirée, voir figure 9.14 page 133).

```
R> s.value(acm$li, acm$li[, 3], 1, 2, csi = 0.5)
Error: could not find function "s.value"
```

FIGURE 9.9 – Répartition des individus selon les trois premiers axes

```
R> s.arrow(acm$co, clabel = 0.7)
Error: could not find function "s.arrow"
```

FIGURE 9.10 – Vecteurs des modalités selon les deux premiers axes

```
R> s.hist(acm$li, clabel = 0, pch = 15)
Error: could not find function "s.hist"
```

FIGURE 9.11 – Distribution des individus dans le plan factoriel



La variable catégorielle transmise à `s.class` ou `s.chull` n'est pas obligatoirement une des variables retenues pour l'ACM. Il est tout à fait possible d'utiliser une autre variable. Par exemple :

```
R> s.class(acm$li, d$trav.imp, col = brewer.pal(4, "Set1"))
```

Les fonctions `scatter` et `biplot` sont équivalentes : elles appliquent `s.class` à chaque variable utilisée pour l'ACM. Voir la figure 9.15 page ci-contre.

9.3 ACM avec FactoMineR

Comme avec `ade4`, il est nécessaire de préparer les données au préalable (voir section précédente). L'ACM se calcule avec la fonction `MCA`, `ncp` permettant de choisir le nombre d'axes à retenir :

```
R> acm2 <- MCA(dt, ncp = 5, graph = FALSE)
Error: could not find function "MCA"

R> acm2
Error: object 'acm2' not found

R> acm2$eig
Error: object 'acm2' not found

R> sum(acm2$eig$eigenvalue)
Error: object 'acm2' not found
```

```
R> library(RColorBrewer)
R> s.class(acm$li, dt$sex, col = brewer.pal(4, "Set1"))

Error: could not find function "s.class"
```

FIGURE 9.12 – Individus dans le plan factoriel selon le sexe (`s.class`)

```
R> s.chull(acm$li, dt$sex, col = brewer.pal(4, "Set1"))
Error: could not find function "s.chull"
```

FIGURE 9.13 – Individus dans le plan factoriel selon le sexe (`s.chull`)

```
R> library(RColorBrewer)
R> display.brewer.all(8)
```

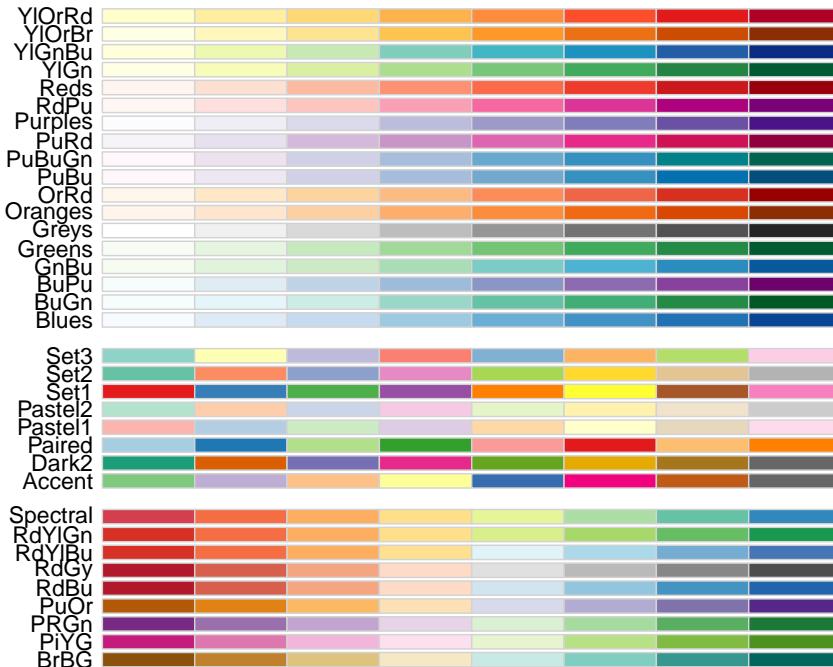


FIGURE 9.14 – Palettes de couleurs disponibles dans RColorBrewer

```
R> scatter(acm, col = brewer.pal(4, "Set1"))
Error: could not find function "scatter"
```

FIGURE 9.15 – La fonction `scatter` appliquée au résultat d'une ACM

```
R> plot(acm2)

Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour
la fonction 'plot' : Error: object 'acm2' not found
```

FIGURE 9.16 – Plan factoriel (deux premiers axes)

```
R> plot(acm2, axes = c(3, 4))

Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour
la fonction 'plot' : Error: object 'acm2' not found
```

FIGURE 9.17 – Plan factoriel (axes 3 et 4)

En premier lieu, il apparaît que l'inertie totale obtenue avec MCA est différente de celle observée avec `dudi.acm`. Cela est dû à un traitement différents des valeurs manquantes. Alors que `dudi.acm` exclut les valeurs manquantes, MCA les considèrent, par défaut, comme une modalité additionnelle. Pour calculer l'ACM uniquement sur les individus n'ayant pas de valeur manquante, on aura recours à `complete.cases` :

```
R> acm2 <- MCA(dt[complete.cases(dt), ], ncp = 5, graph = FALSE)

Error: could not find function "MCA"

R> acm2$eig

Error: object 'acm2' not found

R> sum(acm2$eig$eigenvalue)

Error: object 'acm2' not found
```

Les possibilités graphiques de `FactoMineR` sont différentes de celles de `ade`. Un recours à la fonction `plot` affichera par défaut les individus, les modalités et les variables. La commande `?plot.MCA` affichera toutes les options graphiques. L'argument `choix` permet de spécifier ce que l'on souhaite afficher (« `ind` » pour les individus et les catégories, « `var` » pour les variables). L'argument `invisible` quant à lui permet de spécifier ce que l'on souhaite masquer. Les axes à afficher se précisent avec `axes`. Voir figures 9.16, 9.17, 9.18, 9.19 et 9.20.

La fonction `plotellipses` trace des ellipses de confiance autour des modalités de variables qualitatives. L'objectif est de voir si les modalités d'une variable qualitative sont significativement différentes les unes des autres. Par défaut (`means=TRUE`), les ellipses de confiance sont calculées pour les coordonnées moyennes de chaque catégorie (voir figure 9.21 page suivante). L'option `means=FALSE` calculera les ellipses de confiance pour l'ensemble des coordonnées des observations relevant de chaque catégorie (voir figure

```
R> plot(acm2, choix = "ind")

Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour
la fonction 'plot' : Error: object 'acm2' not found
```

FIGURE 9.18 – Plan factoriel (seulement les individus et les catégories)

```
R> plot(acm2, choix = "ind", invisible = "ind")  
  
Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour  
la fonction 'plot' : Error: object 'acm2' not found
```

FIGURE 9.19 – Plan factoriel (seulement les catégories)

```
R> plot(acm2, choix = "var")  
  
Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour  
la fonction 'plot' : Error: object 'acm2' not found
```

FIGURE 9.20 – Plan factoriel (seulement les variables)

9.22 page suivante).

```
R> plotellipses(acm2)  
  
Error: could not find function "plotellipses"
```

FIGURE 9.21 – Ellipses de confiance (`means=TRUE`) dans le plan factoriel

```
R> plotellipses(acm2, means = FALSE)  
Error: could not find function "plotellipses"
```

FIGURE 9.22 – Ellipses de confiance (means=FALSE) dans le plan factoriel

La fonction `dimdesc` aide à décrire et interpréter les dimensions de l'ACM. Cette fonction est très utile quand le nombre de variables est élevé. Elle permet de voir à quelles variables les axes sont le plus liés : quelles variables et quelles modalités décrivent le mieux chaque axe.

« Pour les variables qualitatives, un modèle d'analyse de variance à un facteur est réalisé pour chaque dimension ; les variables à expliquer sont les coordonnées des individus et la variable explicative est une des variables qualitatives. Un test F permet de voir si la variable a un effet significatif sur la dimension et des tests T sont réalisés modalité par modalité (avec le contraste somme des alpha_i=0). Cela montre si les coordonnées des individus de la sous-population définie par une modalité sont significativement différentes de celles de l'ensemble de la population (i.e. différentes de 0). Les variables et modalités sont triées par probabilité critique et seules celles qui sont significatives sont gardées dans le résultat. »

Source : <http://factominer.free.fr/factosbest/description-des-dimensions.html>

```
R> dimdesc(acm2, axes = 1:2)  
  
Error: could not find function "dimdesc"
```

Partie 10

Classification ascendante hiérarchique (CAH)

Il existe de nombreuses techniques statistiques visant à partitionner une population en différentes classes ou sous-groupes. La *classification ascendante hiérarchique* (CAH) est l'une d'entre elles. On cherche à ce que les individus regroupés au sein d'une même classe (homogénéité intra-classe) soient le plus semblables possibles tandis que les classes soient le plus dissemblables (hétérogénéité inter-classe).

Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une *matrice de distances*, exprimant la distance existante entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. La CAH va ensuite rassembler les individus de manière itérative afin de produire un *dendrogramme* ou *arbre de classification*. La classification est *ascendante* car elle part des observations individuelles ; elle est *hiérarchique* car elle produit des classes ou groupes de plus en plus vastes, incluant des sous-groupes en leur sein. En découplant cet arbre à une certaine hauteur choisie, on produira la partition désirée.



On trouvera également de nombreux supports de cours en français sur la CAH sur le site de François Gilles Carpentier : <http://geai.univ-brest.fr/~carpentier/>.

10.1 Calculer une matrice des distances

La notion de *ressemblance* entre observations est évaluée par une distance entre individus. Plusieurs types de distances existent selon les données utilisées.

Il existe de nombreuses distances mathématiques pour les variables quantitatives (euclidiennes, Manhattan...) que nous n'aborderons pas ici¹. La plupart peuvent être calculées avec la fonction `dist`.

Usuellement, pour un ensemble de variables qualitatives, on aura recours à la distance du Φ^2 qui est celle utilisée pour l'analyse des correspondances multiples (voir chapitre 9 page 127). Avec l'extension `ade4`, la distance du Φ^2 s'obtient avec la fonction `dist.dudi`². Le cas particulier de la CAH avec l'extension

1. Pour une présentation de ces différentes distances, on pourra se référer à http://old.biodiversite.wallonie.be/outils/methodo/similarite_distance.htm ou encore à ce support de cours par D. Chessel, J. Thioulouse et A.B. Dufour disponible à <http://pbil.univ-lyon1.fr/R/pdf/stage7.pdf>.

2. Cette même fonction peut aussi être utilisée pour calculer une distance après une analyse en composantes principales ou une analyse mixte de Hill et Smith.

FactoMineR sera abordée dans une section spécifique (section 10.4 page 144). Nous évoquerons également la distance de Gower qui peut s'appliquer à un ensemble de variables à la fois qualitatives et quantitatives et qui se calcule avec la fonction `daisy` de l'extension `cluster`. Enfin, dans le chapitre 11 sur l'analyse de séquences page 148, nous verrons également la fonction `seqdist` (extension `TraMineR`) permettant de calculer une distance entre séquences.

10.1.1 Distance de Gower

En 1971, Gower a proposé un indice de similarité qui porte son nom. L'objectif de cet indice consiste à mesurer dans quelle mesure deux individus sont semblables. L'indice de Gower varie entre 0 et 1. Si l'indice vaut 1, les deux individus sont identiques. À l'opposé, s'il vaut 0, les deux individus considérés n'ont pas de point commun. Si l'on note S_g l'indice de similarité de Gower, la distance de Gower D_g s'obtient simplement de la manière suivante : $D_g = 1 - S_g$. Ainsi, la distance sera nulle entre deux individus identiques et elle sera égale à 1 entre deux individus totalement différents. Cette distance s'obtient sous R avec la fonction `daisy` du package `cluster`.

L'indice de similarité de Gower entre deux individus x_1 et x_2 se calcule de la manière suivante :

$$S_g(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p s_{12j} \quad (10.1)$$

p représente le nombre total de caractères (ou de variables) descriptifs utilisés pour comparer les deux individus. s_{12j} représente la similarité partielle entre les individus 1 et 2 concernant le descripteur j . Cette similarité partielle se calcule différemment s'il s'agit d'une variable qualitative ou quantitative :

- **variable qualitative** : s_{12j} vaut 1 si la variable j prend la même valeur pour les individus 1 et 2, et vaut 0 sinon. Par exemple, si 1 et 2 sont tous les deux « grand », alors s_{12j} vaudra 1. Si 1 est « grand » et 2 « petit », s_{12j} vaudra 0.
- **variable quantitative** : la différence absolue entre les valeurs des deux variables est tout d'abord calculée, soit $|y_{1j} - y_{2j}|$. Puis l'écart maximum observé sur l'ensemble du fichier est déterminé et noté R_j . Dès lors, la similarité partielle vaut $S_{12j} = |y_{1j} - y_{2j}| / R_j$.

Dans le cas où l'on n'a que des variables qualitatives, la valeur de l'indice de Gower correspond à la proportion de caractères en commun. Supposons des individus 1 et 2 décrits ainsi :

1. homme / grand / blond / étudiant / urbain
2. femme / grande / brune / étudiante / rurale

Sur les 5 variables utilisées pour les décrire, 1 et 2 ont deux caractéristiques communes : ils sont grand(e)s et étudiant(e)s. Dès lors, l'indice de similarité de Gower entre 1 et 2 vaut $2/5 = 0,4$ (soit une distance de $1 - 0,4 = 0,6$).

Plusieurs approches peuvent être retenues pour traiter les valeurs manquantes :

- supprimer tout individu n'étant pas renseigné pour toutes les variables de l'analyse ;
- considérer les valeurs manquantes comme une modalité en tant que telle ;
- garder les valeurs manquantes en tant que valeurs manquantes.

Le choix retenu modifiera les distances de Gower calculées. Supposons que l'on ait :

1. homme / grand / blond / étudiant / urbain
2. femme / grande / brune / étudiante / *manquant*

Si l'on supprime individus ayant des valeurs manquantes, 2 est retirée du fichier d'observations et aucune distance n'est calculée. Si l'on traite les valeurs manquantes comme une modalité particulière, 1 et 2 partagent alors 2 caractères sur les 5 analysés, la distance de Gower entre eux est alors de $1 - 2/5 = 1 - 0,4 = 0,6$. Si on garde les valeurs manquantes, l'indice de Gower est dès lors calculé sur les seuls descripteurs renseignés à la fois pour 1 et 2. La distance de Gower sera calculée dans le cas présent uniquement sur les 4 caractères renseignés et vaudra $1 - 2/4 = 0,5$.

10.1.2 Distance du Φ^2

Il s'agit de la distance utilisée dans les analyses de correspondance multiples (ACM). C'est une variante de la distance du χ^2 . Nous considérons ici que nous avons Q questions (soit Q variables initiales de type facteur). À chaque individu est associé un *patron* c'est-à-dire une certaine combinaison de réponses aux Q questions. La distance entre deux individus correspond à la distance entre leurs deux patrons. Si les deux individus présentent le même patron, leur distance sera nulle. La distance du Φ^2 peut s'exprimer ainsi :

$$d_{\Phi^2}^2(L_i, L_j) = \frac{1}{Q} \sum_k \frac{(\delta_{ik} - \delta_{jk})^2}{f_k} \quad (10.2)$$

où L_i et L_j sont deux patrons, Q le nombre total de questions. δ_{ik} vaut 1 si la modalité k est présente dans le patron L_i , 0 sinon. f_k est la fréquence de la modalité k dans l'ensemble de la population.

Exprimé plus simplement, on fait la somme de l'inverse des modalités non communes aux deux patrons, puis on divise par le nombre total de question. Si nous reprenons notre exemple précédent :

1. homme / grand / blond / étudiant / urbain
2. femme / grande / brune / étudiante / rurale

Pour calculer la distance entre 1 et 2, il nous faut connaître la proportion des différentes modalités dans l'ensemble de la population étudiée. En l'occurrence :

- hommes : 52 % / femmes : 48 %
- grand : 30 % / moyen : 45 % / petit : 25 %
- blond : 15 % / châtain : 45 % / brun : 30 % / blancs : 10 %
- étudiant : 20 % / salariés : 65 % / retraités : 15 %
- urbain : 80 % / rural : 20 %

Les modalités non communes entre les profils de 1 et 2 sont : homme, femme, blond, brun, urbain et rural. La distance du χ^2 entre 1 et 2 est donc la suivante :

$$d_{\Phi^2}^2(L_1, L_2) = \frac{1}{5} \left(\frac{1}{0,52} + \frac{1}{0,48} + \frac{1}{0,15} + \frac{1}{0,30} + \frac{1}{0,80} + \frac{1}{0,20} \right) = 4,05 \quad (10.3)$$

Cette distance, bien que moins intuitive que la distance de Gower évoquée précédemment, est la plus employée pour l'analyse d'enquêtes en sciences sociales. Il faut retenir que la distance entre deux profils est dépendante de la distribution globale de chaque modalité dans la population étudiée. Ainsi, si l'on recalcule les distances entre individus à partir d'un sous-échantillon, le résultat obtenu sera différent. De manière générale, les individus présentant des caractéristiques rares dans la population vont se retrouver éloignés des individus présentant des caractéristiques fortement représentées.

10.1.3 Exemple

Nous allons reprendre l'ACM calculée avec `dudi.acm` (`ade4`) au chapitre 9. La matrice des distances s'obtient dès lors avec la fonction `dist.dudi` :

```
R> md <- dist.dudi(acm)
Error: could not find function "dist.dudi"
```

10.2 Calcul du dendrogramme

Il faut ensuite choisir une méthode d'agrégation pour construire le dendrogramme. De nombreuses solutions existent (saut minimum, distance maximum, moyenne, Ward...). Chacune d'elle produira un

```
R> plot(arbre, labels = FALSE, main = "Dendrogramme")
Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour
la fonction 'plot' : Error: object 'arbre' not found
```

FIGURE 10.1 – Dendrogramme obtenu avec `hclust`

dendrogramme différent. Nous ne détaillerons pas ici ces différentes techniques³. Cependant, à l’usage, on privilégiera le plus souvent la *méthode de Ward*. Cette méthode se distingue de toutes les autres en ce sens qu’elle utilise une analyse de la variance approchée afin d’évaluer les distances entre groupes. La méthode de Ward se justifie bien lorsque lorsque l’on utilise le carré de la distance. Choisir de regrouper les deux individus les plus proches revient alors à choisir la paire de points dont l’agrégation entraîne la diminution minimale de l’inertie du nuage. En résumé, cette méthode cherche à minimiser l’inertie intra-classe et à maximiser l’inertie inter-classe afin d’obtenir des classes les plus homogènes possibles.

En raison de la variété des distances possibles et de la variété des techniques d’agrégation, on pourra être amené à réaliser plusieurs dendrogrammes différents sur un même jeu de données jusqu’à obtenir une classification qui fait « sens ».

La fonction de base pour le calcul d’un dendrogramme est `hclust` en précisant le critère d’aggrégation avec `method`. Dans notre cas, nous allons opter pour la méthode de Ward appliquée au carré des distances (ce qu’on indique avec `md^2`) :

```
R> arbre <- hclust(md^2, method = "ward")
Error: object 'md' not found
```



Le temps de calcul d’un dendrogramme peut être particulièrement important sur un gros fichier de données. L’extension `flashClust` permet de réduire significativement le temps de calcul. Il suffit d’installer puis d’appeler cette extension. La fonction `hclust` sera automatiquement remplacée par cette version optimisée.

```
R> library(flashClust)
R> arbre <- hclust(md^2, method = "ward")
```

Le dendrogramme obtenu peut être affiché simplement avec `plot`. Lorsque le nombre d’individus est important, il peut être utile de ne pas afficher les étiquettes des individus avec `labels=FALSE` (voir figure 10.1 de la présente page).

La fonction `agnes` de l’extension `cluster` peut également être utilisée pour calculer le dendrogramme. Cependant, à l’usage, elle semble être un peu plus lente que `hclust`.

```
R> arbre2 <- agnes(md^2, method = "ward")
```

Le résultat obtenu n’est pas au même format que celui de `hclust`. Il est possible de transformer un objet `agnes` au format `hclust` avec `as.hclust`.

3. On pourra consulter le cours de FG Carpentier déjà cité ou bien des ouvrages d’analyse statistique.

```
R> inertie <- sort(arbre$height, decreasing = TRUE)
Error: object 'arbre' not found

R> plot(inertie[1:20], type = "s", xlab = "Nombre de classes", ylab = "Inertie")
Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour
la fonction 'plot' : Error: object 'inertie' not found

R> points(c(2, 5, 8), inertie[c(2, 5, 8)], col = c("green3", "red3", "blue3"),
+           cex = 2, lwd = 3)
Error: object 'inertie' not found
```

FIGURE 10.2 – Sauts d'inertie du dendrogramme



De nombreuses possibilités graphiques sont possibles avec les dendrogrammes. Des exemples documentés sont disponibles à cette adresse : <http://rpubs.com/gaston/dendrograms>.

10.3 Découper le dendrogramme

Pour obtenir une partition de la population, il suffit de découper le dendrogramme obtenu à une certaine hauteur. En premier lieu, une analyse de la forme du dendrogramme pourra nous donner une indication sur le nombre de classes à retenir. Dans notre exemple, deux branches bien distinctes apparaissent sur l'arbre.

Pour nous aider, nous pouvons représenter les sauts d'inertie du dendrogramme selon le nombre de classes retenues (voir figure 10.2 de la présente page). On voit trois sauts assez nets, à 2, 5 et 8 classes, représentés respectivement en vert, en rouge et en bleu.

La fonction `rect.hclust` permet de visualiser les différentes partitions directement sur le dendrogramme (voir figure 10.3 page ci-contre).

L'extension FactoMineR (que nous aborderons dans la section 10.4 page 144) suggère d'utiliser la partition ayant la plus grande perte relative d'inertie. Nous avons développer une fonction `best.cutree` qui permet de calculer cette indicateur à partir de n'importe quel dendrogramme calculé avec `hclust` ou `agnes`. Le code de cette fonction est disponible à <http://joseph.larmarange.net/?article149>. Il suffit de le recopier et de le coller dans R :

```
R> best.cutree <- function(hc, min = 3, max = 20, loss = FALSE, graph = FALSE,
+   ...) {
+   if (class(hc) != "hclust")
+     hc <- as.hclust(hc)
+   max <- min(max, length(hc$height))
+   inert.gain <- rev(hc$height)
+   intra <- rev(cumsum(rev(inert.gain)))
+   relative.loss = intra[min:(max)]/intra[(min - 1):(max - 1)]
+   best = which.min(relative.loss)
+   names(relative.loss) <- min:max
+   if (graph) {
+     temp <- relative.loss
```

```
R> plot(arbre, labels = FALSE, main = "Partition en 2, 5 ou 8 classes", xlab = "",  
+       ylab = "", sub = "", axes = FALSE, hang = -1)  
  
Error: erreur d'évaluation de l'argument 'x' lors de la sélection d'une méthode pour  
la fonction 'plot' : Error: object 'arbre' not found  
  
R> rect.hclust(arbre, 2, border = "green3")  
  
Error: object 'arbre' not found  
  
R> rect.hclust(arbre, 5, border = "red3")  
  
Error: object 'arbre' not found  
  
R> rect.hclust(arbre, 8, border = "blue3")  
  
Error: object 'arbre' not found
```

FIGURE 10.3 – Différentes partitions du dendrogramme

```
R> best.cutree(arbre, min = 2, graph = TRUE, xlab = "Nombre de classes", ylab = "Perte relative d'inertie")  
  
Error: object 'arbre' not found
```

FIGURE 10.4 – Perte relative d'inertie selon le nombre de classes

```
+      temp[best] <- NA  
+      best2 <- which.min(temp)  
+      pch <- rep(1, max - min + 1)  
+      pch[best] <- 16  
+      pch[best2] <- 21  
+      plot(min:max, relative.loss, pch = pch, bg = "grey75", ...)  
+ } else {  
+   if (loss)  
+     relative.loss else best + min - 1  
+ }  
+ }  
R>  
R> best.cutree(arbre)  
  
Error: object 'arbre' not found
```

Par défaut, cette fonction regarde quelle serait la meilleure partition entre 3 et 20 classes, en l'occurrence il s'agirait d'une partition en 5 classes. Il est possible de modifier le minimum et le maximum des partitions recherchées avec `min` et `max`.

```
R> best.cutree(arbre, min = 2)  
  
Error: object 'arbre' not found
```

On peut également représenter le graphique des pertes relatives d'inertie avec `graph=TRUE` (voir figure 10.4 de la présente page). La meilleure partition selon ce critère est représentée par un point noir et la seconde par un point gris. Un découpage en deux classes minimise ce critère. Cependant, si l'on souhaite

```
R> par(mfrow = c(1, 2))
R> s.class(acm$li, as.factor(typo), col = brewer.pal(5, "Set1"), sub = "Axes 1 et 2")
Error: could not find function "s.class"

R> s.class(acm$li, as.factor(typo), 3, 4, col = brewer.pal(5, "Set1"), sub = "Axes 3 et 4")
Error: could not find function "s.class"

R> par(mfrow = c(1, 1))
```

FIGURE 10.5 – Projection de la typologie obtenue par CAH selon les 4 premiers axes

```
R> op = par(bg = "#EFEFEF")
R> A2Rplot(arbre, k = 5, boxes = FALSE, col.up = "gray50", col.down = brewer.pal(5,
+      "Dark2"), show.labels = FALSE)
Error: object 'arbre' not found

R> par(op)
```

FIGURE 10.6 – Un dendrogramme coloré

réaliser une analyse un peu plus fine, un nombre de classes plus élevé serait pertinent. Nous allons donc retenir un découpage en cinq classes. Le découpage s'effectue avec la fonction `cutree`.

```
R> typo <- cutree(arbre, 5)
Error: object 'arbre' not found

R> freq(typo)
Error: object 'typo' not found
```

La typologie obtenue peut être représentée dans le plan factoriel avec `s.class` (voir figure 10.5 de la présente page).

Enfin, Romain François a développé une fonction `A2Rplot` permettant de réaliser facilement un dendrogramme avec les branches colorées. En premier lieu, il faut récupérer le code de cette fonction :

```
R> source("http://addictedor.free.fr/packages/A2R/lastVersion/R/code.R")
```

Puis réaliser le graphique en indiquant le nombre de classes et les couleurs à utiliser pour chaque branche de l'arbre (voir figure 10.6 de la présente page).

10.4 CAH avec l'extension FactoMineR

L'extension `FactoMineR` fournit une fonction `HCPC` permettant de réaliser une classification hiérarchique à partir du résultats d'une analyse factorielle réalisée avec la même extension (voir section 9.3 page 132). `HCPC` réalise à la fois le calcul de la matrice des distances, du dendrogramme et le partitionnement de la population en classes. Par défaut, `HCPC` calcule le dendrogramme à partir du carré des distances et avec

```
R> plot(cah, choice = "tree")
```

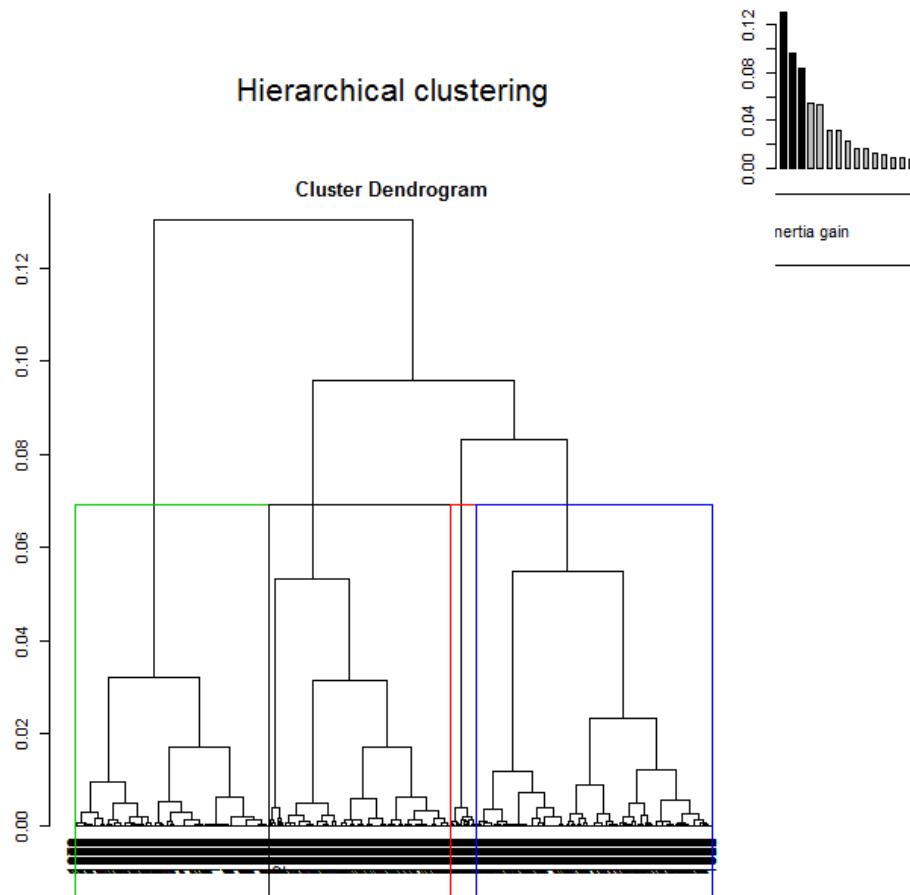


FIGURE 10.7 – Dendrogramme obtenu avec HCPC

la méthode de Ward. L’arbre est affiché à l’écran et vous pouvez indiquer où vous souhaitez le couper à la souris. Cependant, si vous utilisez RStudio, il y a un bug lié à cette fonctionnalité. Vous devrez dès lors appeler HCPC avec `graph=FALSE`. Utilisez l’argument `nb.clust` pour indiquer le nombre de classes désirées. Si vous appelez la fonction avec `nb.clust=-1`, l’arbre sera coupé selon la partition ayant la plus grande perte relative d’inertie (comme avec `best.cutree`).

```
R> cah <- HCPC(acm2, nb.clust = -1, graph = FALSE)
```

```
Error: could not find function "HCPC"
```

On pourra représenter le dendrogramme avec `plot` et l’argument `choice="tree"` (voir figure 10.7 de la présente page).

Il apparait que le dendrogramme obtenu avec HCPC diffère de celui que nous avons calculé précédemment en utilisant la matrice des distances fournies par `dist.dudi`. Cela est dû au fait que HCPC procède différemment pour calculer la matrice des distances en ne prenant en compte que les axes retenus dans le cadre de l’ACM. Pour rappel, nous avions retenu que 5 axes dans le cadre de notre ACM :

```
R> acm2 <- MCA(dt[complete.cases(dt), ], ncp = 5, graph = FALSE)
```

HCPC n'a donc pris en compte que ces 5 premiers axes pour calculer les distances entre les individus, considérant que les autres axes n'apportent que du « bruit » rendant la classification instable. Cependant, comme le montre `summary(acm2)`, nos cinq premiers axes n'expliquent que 54 % de la variance. Il est généralement préférable de garder un plus grande nombre d'axes afin de couvrir au moins 80 à 90 % de la variance⁴. De son côté, `dist.dudi` prends en compte l'ensemble des axes pour calculer la matrice des distances. On peut reproduire cela avec FactoMineR en indiquant `ncp=Inf` lors du calcul de l'ACM.

```
R> acm2 <- MCA(dt[complete.cases(dt), ], ncp = Inf, graph = FALSE)
```

```
Error: could not find function "MCA"
```

```
R> cah <- HCPC(acm2, nb.clust = -1, graph = FALSE)
```

```
Error: could not find function "HCPC"
```

On obtient bien cette fois-ci le même résultat (voir figure 10.8 page suivante). D'autres graphiques sont disponibles, essayez par exemple les commandes suivantes :

```
R> plot(cah, choice = "3D.map")
R> plot(cah, choice = "bar")
R> plot(cah, choice = "map")
```

L'objet renvoyé par HCPC contient de nombreuses informations. La partition peut notamment être récupérée avec `cah$data.clust$clust`. Il y a également diverses statistiques pour décrire les catégories.

```
R> cah

Error: object 'cah' not found

R> freq(cah$data.clust$clust)

Error: object 'cah' not found
```

4. Voir par exemple <http://factominer.free.fr/classical-methods/classification-hierarchique-sur-composantes-principales.html>

```
R> plot(cah, choice = "tree")
```

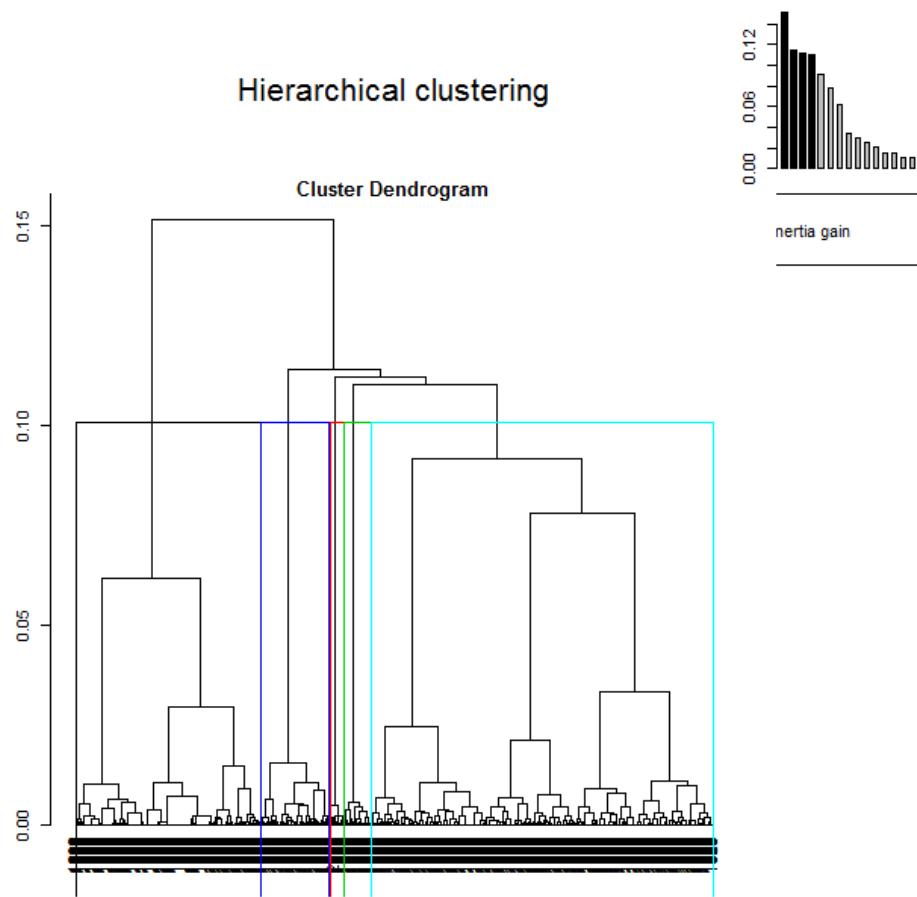


FIGURE 10.8 – Dendrogramme obtenu avec HCPC

Partie 11

Analyse de séquences



Le texte de ce chapitre reprend, avec l'aimable autorisation de son auteur, un article de Nicolas Robette ^a intitulé *L'analyse de séquences : une introduction avec le logiciel R et le package TraMineR* et publié le 24 octobre 2012 sur le blog Quanti ^b.

- a. Maître de conférences à l'Université de Versailles Saint-Quentin-en-Yvelines.
- b. <http://quanti.hypotheses.org/686/>

Depuis les années 1980, l'étude quantitative des trajectoires biographiques (*life course analysis*) a pris une ampleur considérable dans le champ des sciences sociales. Les collectes de données micro-individuelles longitudinales se sont développées, principalement sous la forme de panels ou d'enquêtes rétrospectives. Parallèlement à cette multiplication des données disponibles, la méthodologie statistique a connu de profondes évolutions. L'analyse des biographies (*event history analysis*) — qui ajoute une dimension diachronique aux modèles économétriques *mainstream* — s'est rapidement imposée comme l'approche dominante : il s'agit de modéliser la durée des situations ou le risque d'occurrence des événements.

11.1 L'analyse de séquences

Cependant, ces dernières années ont vu la diffusion d'un large corpus de méthodes descriptives d'analyse de séquences, au sein desquelles l'appariement optimal (*optimal matching*) occupe une place centrale ¹. L'objectif principal de ces méthodes est d'identifier — dans la diversité d'un corpus de séquences constituées de séries d'états successifs — les régularités, les ressemblances, puis le plus souvent de construire des typologies de « séquences-types ». L'analyse de séquences constitue donc un moyen de décrire mais aussi de mieux comprendre le déroulement de divers processus.

La majeure partie des applications de l'analyse de séquences traite de trajectoires biographiques ou de carrières professionnelles. Dans ces cas, chaque trajectoire ou chaque carrière est décrite par une séquence, autrement dit par une suite chronologiquement ordonnée de « moments » élémentaires, chaque moment correspondant à un « état » déterminé de la trajectoire (par exemple, pour les carrières professionnelles : être en emploi, au chômage ou en inactivité). Mais on peut bien sûr imaginer des types de séquences plus originaux : Andrew Abbott ², le sociologue américain qui a introduit l'*optimal matching* dans les sciences

1. Pour une analyse des conditions sociales de la diffusion de l'analyse de séquences dans le champ des sciences sociales, voir Robette, 2012.

2. <http://home.uchicago.edu/~aabott/>

sociales dans les années 1980, s'en est par exemple servi pour étudier la structure rhétorique d'articles scientifiques ou des séquences de pas de danses traditionnelles.

En France, les premiers travaux utilisant l'appariement optimal sont ceux de Claire Lemercier³ sur les carrières des membres des institutions consulaires parisiennes au XIXe siècle (Lemercier, 2005), et de Laurent Lesnard⁴ sur les emplois du temps (Lesnard, 2008). Mais dès les années 1980, les chercheurs du Céreq construisaient des typologies de trajectoires d'insertion à l'aide des méthodes d'analyse des données « à la française » (analyse des correspondances, etc.)⁵. Au final, on dénombre maintenant plus d'une centaine d'articles de sciences sociales contenant ou discutant des techniques empruntées à l'analyse de séquences.

Pour une présentation des différentes méthodes d'analyse de séquences disponibles et de leur mise en œuvre pratique, il existe un petit manuel en français, publié l'année dernière aux éditions du Ceped (collection « Les clefs pour »⁶) et disponible en pdf⁷ (Robette, 2011). De plus, un article récemment publié dans le Bulletin de Méthodologie Sociologique compare de manière systématique les résultats obtenus par les principales méthodes d'analyse de séquences (Robette & Bry, 2012). La conclusion en est qu'avec des données empiriques aussi structurées que celles que l'on utilise en sciences sociales, l'approche est robuste, c'est-à-dire qu'un changement de méthode aura peu d'influence sur les principaux résultats. Cependant, l'article tente aussi de décrire les spécificités de chaque méthode et les différences marginales qu'elles font apparaître, afin de permettre aux chercheurs de mieux adapter leurs choix méthodologiques à leur question de recherche.

Afin d'illustrer la démarche de l'analyse de séquences, nous allons procéder ici à la description « pas à pas » d'un corpus de carrières professionnelles, issues de l'enquête *Biographies et entourage* (INED, 2000)⁸. Et pour ce faire, on va utiliser le logiciel R, qui propose la solution actuellement la plus complète et la plus puissante en matière d'analyse de séquences. Les méthodes d'analyse de séquences par analyses factorielles ou de correspondances ne nécessitent pas de logiciel spécifique : tous les logiciels de statistiques généralistes peuvent être utilisés (SAS, SPSS, Stata, R, etc.). En revanche, il n'existe pas de fonctions pour l'appariement optimal dans SAS ou SPSS. Certains logiciels gratuits implémentent l'appariement optimal (comme Chesa⁹ ou TDA¹⁰) mais il faut alors recourir à d'autres programmes pour dérouler l'ensemble de l'analyse (classification, représentation graphique). Stata propose le module sq¹¹, qui dispose d'un éventail de fonctions intéressantes. Mais c'est R et le package TraMineR¹², développé par des collègues de l'Université de Genève (Gabadinho *et al.*, 2011), qui fournit la solution la plus complète et la plus puissante à ce jour : on y trouve l'appariement optimal mais aussi d'autres algorithmes alternatifs, ainsi que de nombreuses fonctions de description des séquences et de représentation graphique.

11.2 Installer TraMineR et récupérer les données

Tout d'abord, à quoi ressemblent nos données ? On a reconstruit à partir de l'enquête les carrières de 1000 hommes. Pour chacune, on connaît la position professionnelle chaque année, de l'âge de 14 ans jusqu'à 50 ans. Cette position est codée de la manière suivante : les codes 1 à 6 correspondent aux groupes socioprofessionnels de la nomenclature des PCS de l'INSEE¹³ (agriculteurs exploitants ; artisans, commerçants et chefs d'entreprise ; cadres et professions intellectuelles supérieures ; professions intermédiaires ; employés ; ouvriers) ; on y a ajouté « études » (code 7), « inactivité > (code 8) et « service

3. <http://lemmercier.ouvaton.org/document.php?id=62>

4. http://laurent.lesnard.free.fr/article.php3?id_article=22

5. Voir par exemple l'article d'Yvette Grelet (2002).

6. <http://www.ceped.org/?Les-Clefs-pour>

7. http://nicolas.robette.free.fr/Docs/Robette2011_Manuel_TypoTraj.pdf

8. Pour une analyse plus poussée de ces données, avec deux méthodes différentes, voir Robette & Thibault, 2008. Pour une présentation de l'enquête, voir Lelièvre & Vivier, 2001.

9. <http://home.fsw.vu.nl/ch.elzinga/>

10. <http://steinhaus.stat.ruhr-uni-bochum.de/tda.html>

11. <http://www.stata-journal.com/article.html?article=st0111>

12. <http://mephisto.unige.ch/traminer/>

13. <http://www.insee.fr/fr/methodes/default.asp?page=nomenclatures/pcs2003/pcs2003.htm>

militaire » (code 9). Le fichier de données comporte une ligne par individu et une colonne par année : la variable 1 correspond à la position à 14 ans, la variable 2 à la position à 15 ans, etc. Par ailleurs, les enquêtés étant tous nés entre 1930 et 1950, on ajoute à notre base une variable « génération » à trois modalités, prenant les valeurs suivantes : 1=1930-1938 ; 2=1939-1945 ; 3=1946-1950. Au final, la base est constituée de 500 lignes et de 37+1=38 colonnes et se présente sous la forme d'un fichier texte au format csv (téléchargeable à <http://nicolas.robette.free.fr/Docs/trajpro.csv>).

Une fois R ouvert, on commence par installer les extensions nécessaires à ce programme (opération à ne réaliser que lors de leur première utilisation) et par les charger en mémoire. L'extension TraMineR propose de nombreuses fonctions pour l'analyse de séquences. L'extension cluster comprend un certain nombre de méthodes de classification automatique¹⁴.

```
R> install.packages(c("TraMineR"))
```

```
R> library(TraMineR)
R> library(cluster)
```

On importe ensuite les données, on recode la variable « génération » pour lui donner des labels plus explicites. On jette également un coup d'œil à la structure du tableau de données :

```
R> donnees <- read.csv("http://nicolas.robette.free.fr/Docs/trajpro.csv", header = T)
```

```
R> donnees$generation <- factor(donnees$generation, labels = c("1930-38", "1939-45",
+ "1946-50"))
R> str(donnees)

'data.frame': 1000 obs. of 38 variables:
 $ csp1      : int  1 7 6 7 7 6 7 7 7 6 ...
 $ csp2      : int  1 7 6 7 7 6 7 7 6 6 ...
 $ csp3      : int  1 7 6 6 7 6 7 7 6 6 ...
 $ csp4      : int  1 7 6 6 7 6 7 7 6 6 ...
 $ csp5      : int  1 7 6 6 7 6 7 7 6 6 ...
 $ csp6      : int  1 7 6 6 7 6 9 7 6 6 ...
 $ csp7      : int  6 9 6 6 7 6 9 7 9 6 ...
 $ csp8      : int  6 9 9 6 7 6 9 7 4 6 ...
 $ csp9      : int  6 6 9 6 7 6 9 3 4 9 ...
 $ csp10     : int  6 6 9 6 7 6 4 3 4 9 ...
 $ csp11     : int  6 6 6 6 3 6 4 3 4 6 ...
 $ csp12     : int  6 6 6 6 3 6 4 3 4 6 ...
 $ csp13     : int  6 6 6 6 3 6 4 3 4 6 ...
 $ csp14     : int  6 4 6 6 3 6 4 3 4 6 ...
 $ csp15     : int  6 4 6 6 3 6 4 3 4 6 ...
 $ csp16     : int  6 4 6 6 3 6 6 3 4 6 ...
 $ csp17     : int  6 4 6 6 3 6 6 3 4 6 ...
 $ csp18     : int  6 4 6 6 3 6 6 3 4 6 ...
 $ csp19     : int  6 4 6 6 3 6 6 3 4 6 ...
 $ csp20     : int  6 4 6 6 3 6 6 3 4 6 ...
 $ csp21     : int  6 4 6 6 6 6 6 3 4 6 ...
 $ csp22     : int  6 4 6 6 6 6 6 3 4 4 ...
 $ csp23     : int  6 4 6 6 6 6 6 3 4 4 ...
```

14. Pour une présentation plus détaillée de la classification ascendante hiérarchique, voir le chapitre 10 page 138.

```
$ csp24      : int  6 6 6 6 5 6 6 3 4 4 ...
$ csp25      : int  6 6 6 6 5 6 6 3 4 4 ...
$ csp26      : int  6 6 6 6 5 6 6 3 4 4 ...
$ csp27      : int  6 6 6 6 5 6 6 3 4 4 ...
$ csp28      : int  6 6 6 6 5 6 6 3 4 4 ...
$ csp29      : int  6 6 6 6 5 6 6 3 4 4 ...
$ csp30      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp31      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp32      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp33      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp34      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp35      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp36      : int  4 6 6 6 5 6 6 3 4 4 ...
$ csp37      : int  4 6 6 6 5 6 6 3 4 4 ...
$ generation: Factor w/ 3 levels "1930-38","1939-45",...: 2 1 1 3 2 3 1 1 2 1 ...

```

On a bien 1000 observations et 38 variables. On définit maintenant des labels pour les différents états qui composent les séquences et on crée un objet « séquence » avec `seqdef` :

```
R> labels <- c("agric", "acce", "cadre", "pint", "empl", "ouvr", "etud", "inact",
+       "smil")
R> seq <- seqdef(donnees[, 1:37], states = labels)
```

11.3 Appariement optimal et classification

Ces étapes préalables achevées, on peut comparer les séquences en calculant les dissimilarités entre paires de séquences. On va ici utiliser la méthode la plus répandue, l'appariement optimal (*optimal matching*). Cette méthode consiste, pour chaque paire de séquences, à compter le nombre minimal de modifications (substitutions, suppressions, insertions) qu'il faut faire subir à l'une des séquences pour obtenir l'autre. On peut considérer que chaque modification est équivalente, mais il est aussi possible de prendre en compte le fait que les « distances » entre les différents états n'ont pas toutes la même « valeur » (par exemple, la distance sociale entre emploi à temps plein et chômage est plus grande qu'entre emploi à temps plein et emploi à temps partiel), en assignant aux différentes modifications des « coûts » distincts. Dans notre exemple, on va créer avec `seqsubm` une « matrice des coûts de substitution » dans laquelle tous les coûts sont constants et égaux à 2¹⁵ :

```
R> couts <- seqsubm(seq, method = "CONSTANT", cval = 2)
```

Ensuite, on calcule la matrice de distance entre les séquences (i.e contenant les « dissimilarités » entre les séquences) avec `seqdist`, avec un coût d'insertion/suppression (*indel*) que l'on fixe ici à 1,1 :

```
R> seq.om <- seqdist(seq, method = "OM", indel = 1.1, sm = couts)
```

Cette matrice des distances ou des dissimilarités entre séquences peut ensuite être utilisée pour une classification ascendante hiérarchique (CAH), qui permet de regrouper les séquences en un certain nombre de « classes » en fonction de leur proximité :

¹⁵. Le fonctionnement de l'algorithme d'appariement optimal — et notamment le choix des coûts — est décrit dans le chapitre 3 du manuel de TraMineR.

```
R> plot(as.dendrogram(seq.agnes), leaflab = "none")
```

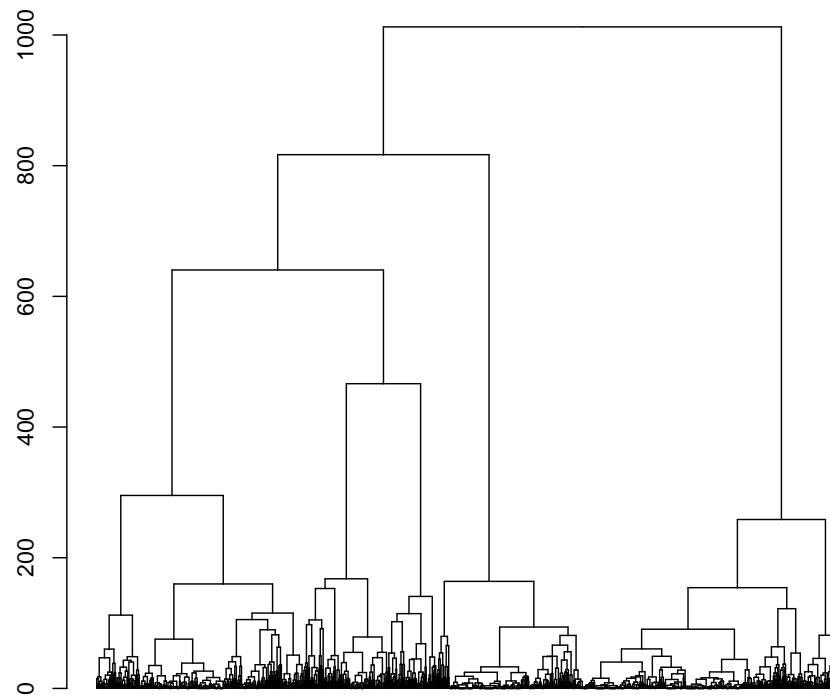


FIGURE 11.1 – Dendrogramme de la classification des séquences

```
R> seq.agnes <- agnes(as.dist(seq.om), method = "ward", keep.diss = FALSE)
```

Avec la fonction `plot`, il est possible de tracer l'arbre de la classification (dendrogramme). L'observation, sur ce dendrogramme (figure 11.1 de la présente page) ou sur la courbe des sauts d'inertie (figure 11.2 page ci-contre), des sauts d'inertie des dernières étapes de la classification peut servir de guide pour déterminer le nombre de classes que l'on va retenir pour la suite des analyses. Une première inflexion dans la courbe des sauts d'inertie apparaît au niveau d'une partition en 5 classes. On voit aussi une seconde inflexion assez nette à 7 classes. Mais il faut garder en tête le fait que ces outils ne sont que des guides, le choix devant avant tout se faire après différents essais, en fonction de l'intérêt des résultats par rapport à la question de recherche et en arbitrant entre exhaustivité et parcimonie.

On fait ici le choix d'une partition en 5 classes :

```
R> nbcl <- 5
R> seq.part <- cutree(seq.agnes, nbcl)
R> seq.part <- factor(seq.part, labels = paste("classe", 1:nbcl, sep = "."))
```

```
R> plot(sort(seq.agnes$height, decreasing = TRUE)[1:20], type = "s", xlab = "nb de classes",
+       ylab = "inertie")
```

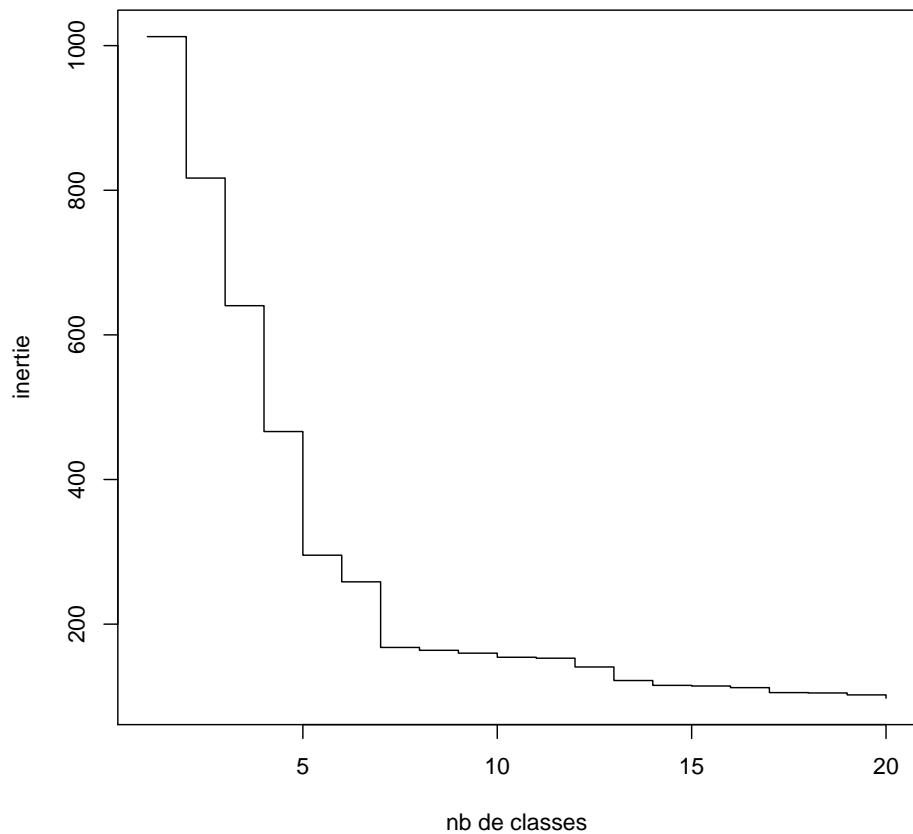


FIGURE 11.2 – Sauts d’inertie de la classification des séquences

```
R> seqdplot(seq, group = seq.part, xlab = 14:50, border = NA, withlegend = T)
```

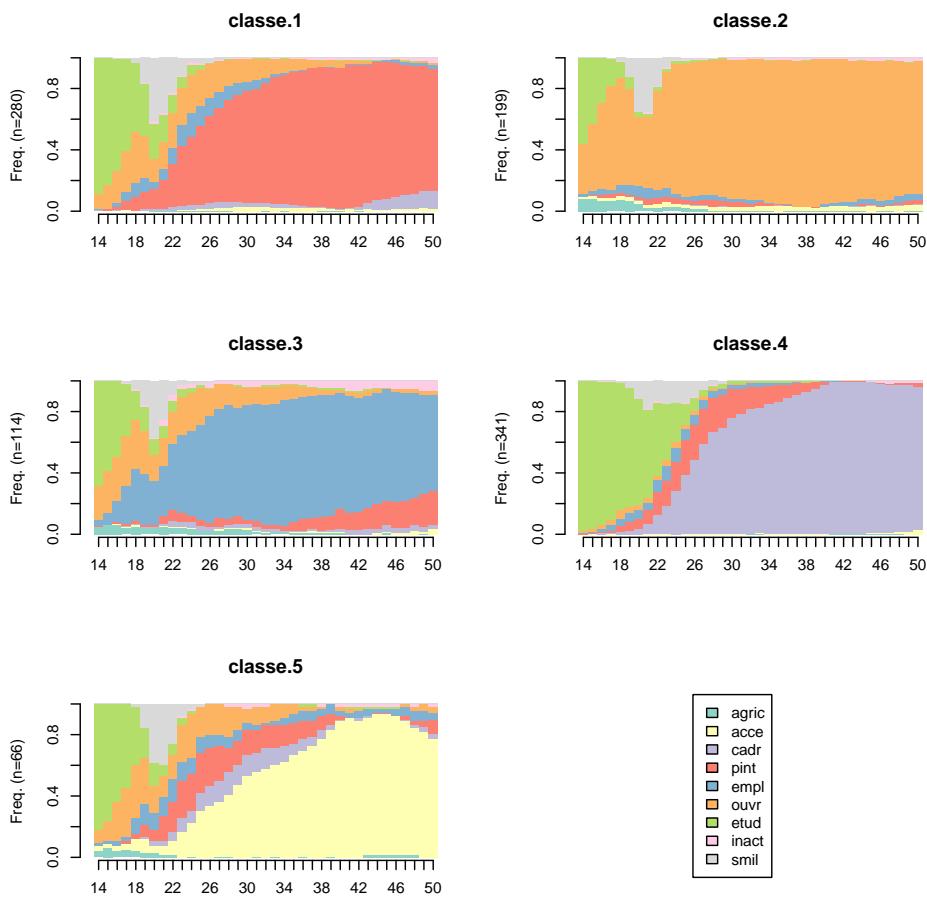


FIGURE 11.3 – Chronogrammes

11.4 Représentations graphiques

Pour se faire une première idée de la nature des classes de la typologie, il existe un certain nombre de représentations graphiques. Les chronogrammes (*state distribution plots*) présentent une série de coupes transversales : pour chaque âge, on a les proportions d'individus de la classe dans les différentes situations (agriculteur, étudiant, etc.). Ce graphique s'obtient avec `seqdplot` (voir figure 11.3 154)

Chacune des classes semble caractérisée par un groupe professionnel principal : professionnel intermédiaire pour la classe 1, ouvrier pour la 2, employé pour la 3, cadre pour la 4 et indépendant pour la 5. Cependant, on aperçoit aussi des « couches » d'autres couleurs, indiquant que l'ensemble des carrières ne sont probablement pas stables.

Les « tapis > (*index plots*) , obtenus avec `seqiplot`, permettent de mieux visualiser la dimension individuelle des séquences. Chaque segment horizontal représente une séquence, découpée en sous-segments correspondant aux différents états successifs qui composent la séquence (voir figure 11.4 page suivante).

Il est possible de trier les séquences pour rendre les tapis plus lisibles (on trie ici par *multidimensional scaling*, voir figure 11.5 page 156).

```
R> seqiplot(seq, group = seq.part, xlab = 14:50, tlim = 0, space = 0, border = NA,
+           withlegend = T, yaxis = FALSE)
```

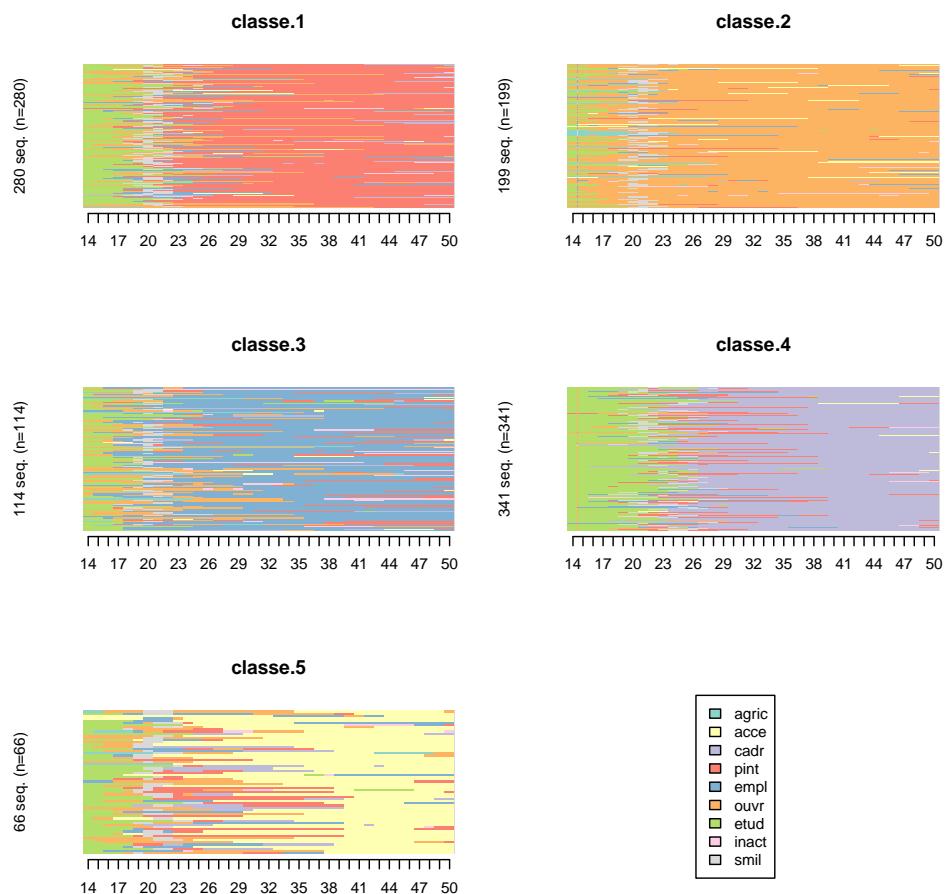


FIGURE 11.4 – Tapis des séquences

```
R> ordre <- cmdscale(as.dist(seq.om), k = 1)
R> seqiplot(seq, group = seq.part, sortv = ordre, xlab = 14:50, tlim = 0, space = 0,
+ border = NA, withlegend = T, yaxis = FALSE)
```

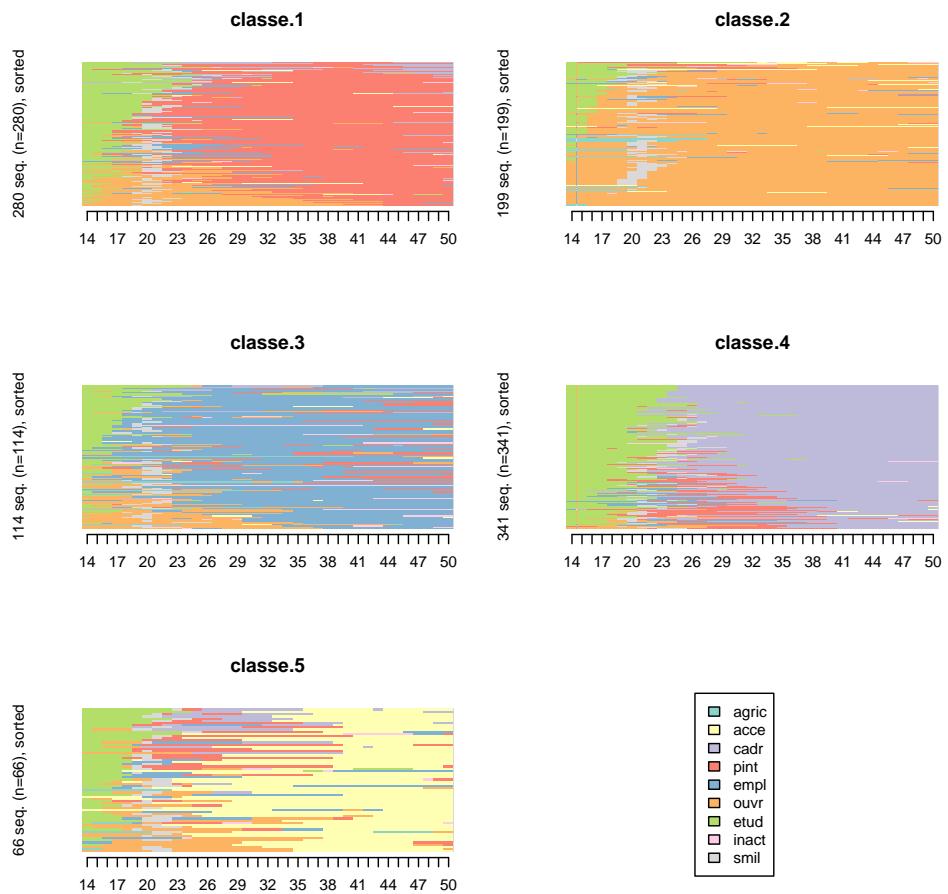


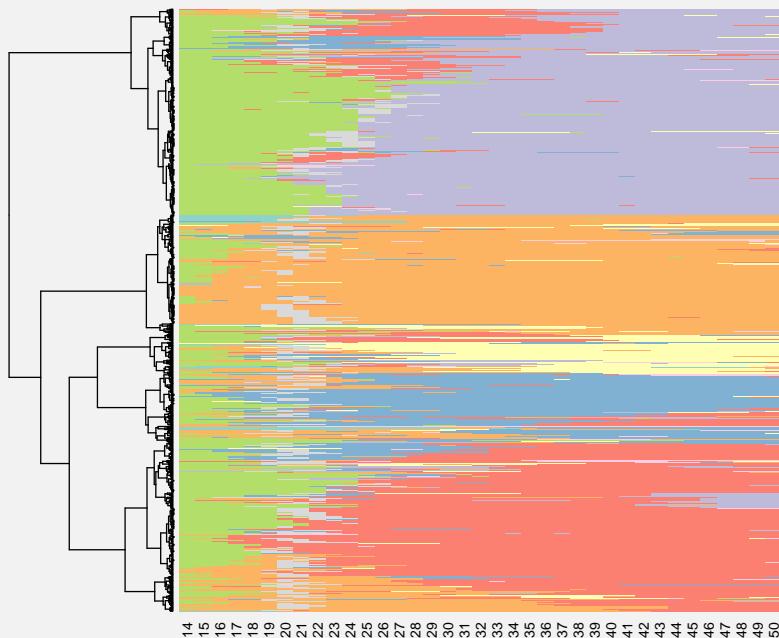
FIGURE 11.5 – Tapis des séquences triés par multidimensional scaling

On voit mieux apparaître ainsi l'hétérogénéité de certaines classes. Les classes 1, 3 et 4, par exemple, semblent regrouper des carrières relativement stables (respectivement de professions intermédiaires, d'employés et de cadres) et des carrières plus « mobiles » commencées comme ouvrier (classes 1 et 3, en orange) ou comme profession intermédiaire (classe 4, en rouge). De même, la majorité des membres de la dernière classe commencent leur carrière dans un groupe professionnel distinct de celui qu'ils occuperont par la suite (indépendants). Ces distinctions apparaissent d'ailleurs si on relance le programme avec un nombre plus élevé de classes (en remplaçant le 5 de la ligne `nbcl <- 5` par 7, seconde inflexion de la courbe des sauts d'inertie, et en exécutant de nouveau le programme à partir de cette ligne) : les stables et les mobiles se trouvent alors dans des classes distinctes.



Une astuce publiée sur <http://joseph.larmarange.net/?article137> permet de représenter le tapis de l'ensemble des séquences selon l'ordre du dendrogramme. On commencera par recopier dans R le code de la fonction `seq.heatmap` qu'il suffit ensuite d'appeler.

```
R> seq.heatmap(seq, seq.agnes, labCol = 14:50)
```



```
R> seqfplot(seq, group = seq.part, withlegend = T)
```

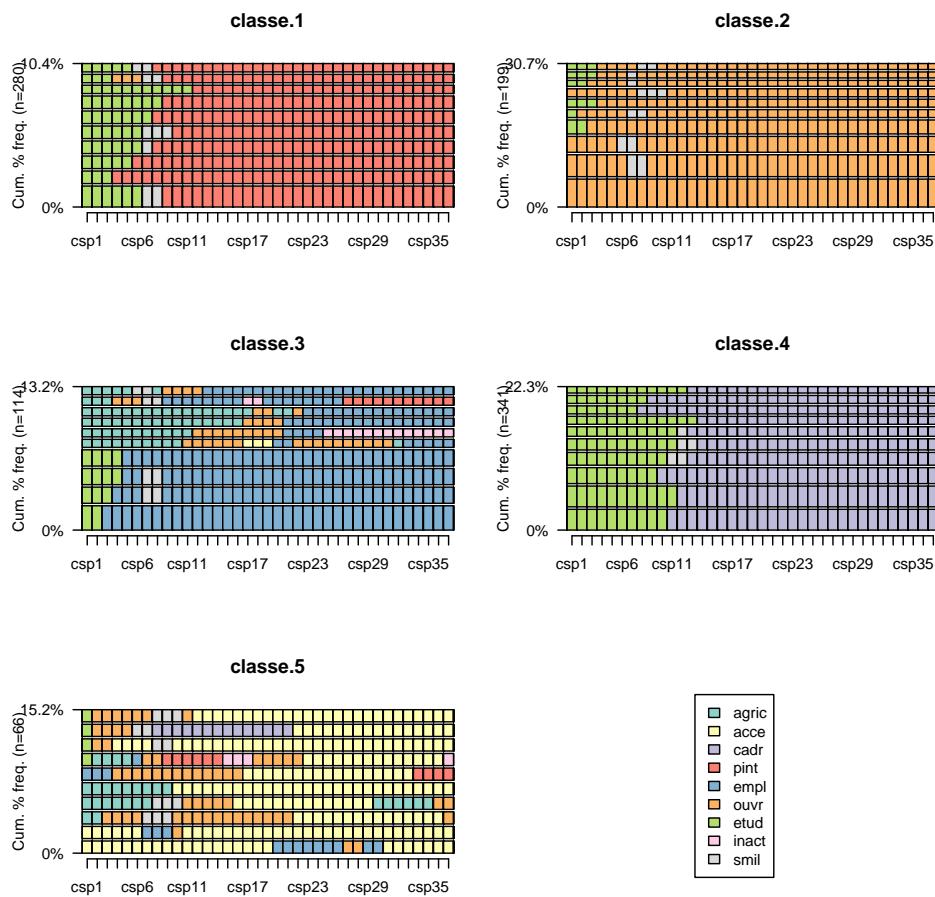


FIGURE 11.6 – Séquences les plus fréquentes de chaque classe

La distance moyenne des séquences d'une classe au centre de cette classe, obtenue avec `disscenter`, permet de mesurer plus précisément l'homogénéité des classes :

```
R> round(aggregate(disscenter(as.dist(seq.om)), group = seq.part, list(seq.part),
+      mean)[, -1], 1)
[1] 13.1  8.9 15.6  9.7 16.5
```

Cela nous confirme que les classes 1, 3 et 5 sont nettement plus hétérogènes que les autres, alors que la classe 2 est la plus homogène.

D'autres représentations graphiques existent pour poursuivre l'examen de la typologie. On peut visualiser les 10 séquences les plus fréquentes de chaque classe avec `seqfplot` (figure 11.6 de la présente page).

On peut aussi visualiser avec `seqmsplot` l'état modal (celui qui correspond au plus grand nombre de séquences de la classe) à chaque âge (figure 11.7 page ci-contre).

On peut également représenter avec `seqmplot` les durées moyennes passées dans les différents états (figure 11.8 page 160).

```
R> seqmsplot(seq, group = seq.part, xlab = 14:50, withlegend = T, title = "classe")
```

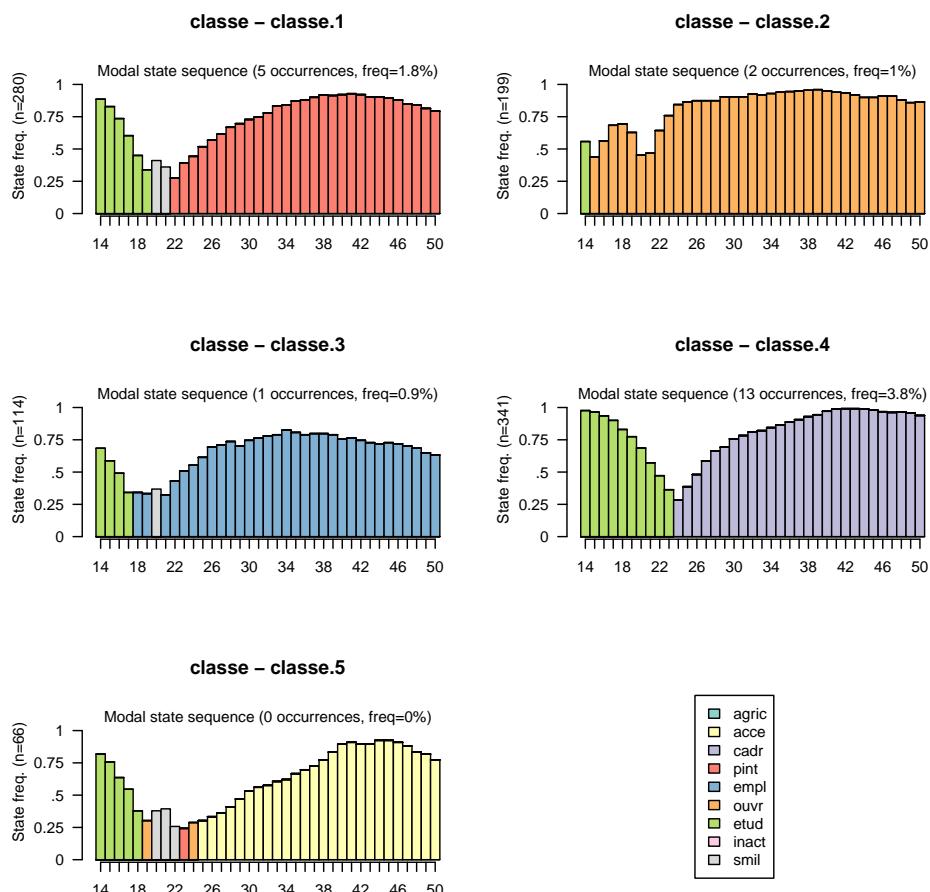


FIGURE 11.7 – Statut modal à chaque âge

```
R> seqmtpplot(seq, group = seq.part, withlegend = T)
```

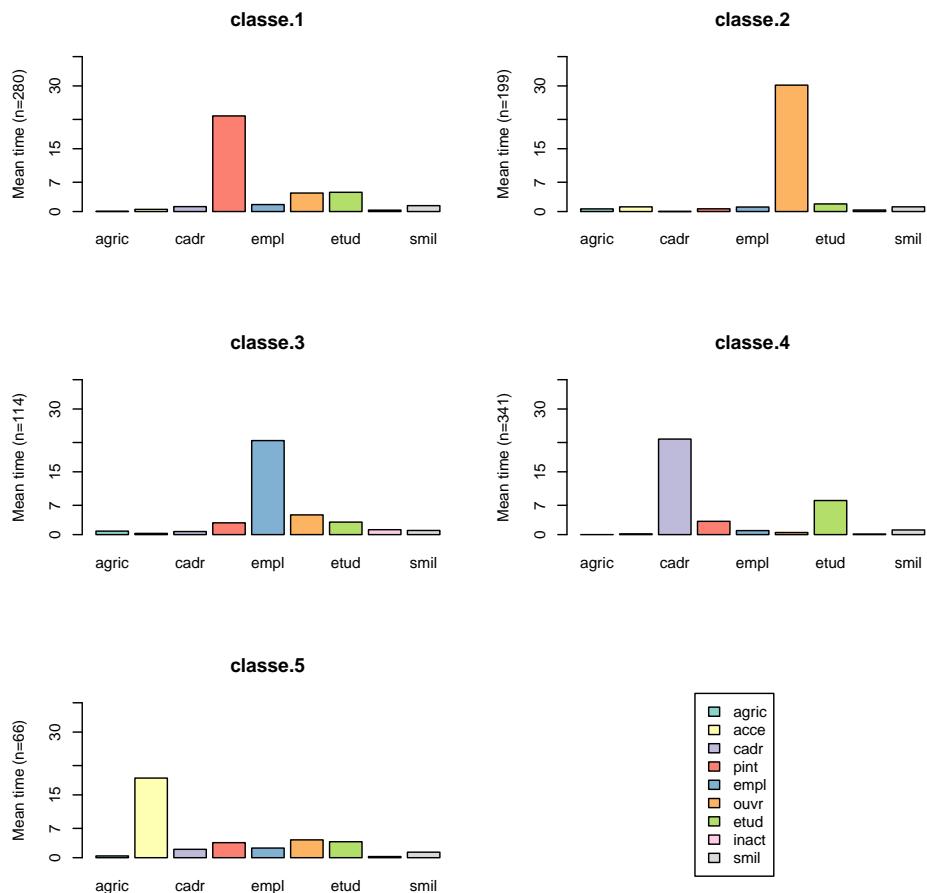


FIGURE 11.8 – Durée moyenne dans chaque statut

```
R> seqHtplot(seq, group = seq.part, xlab = 14:50, withlegend = T)
```

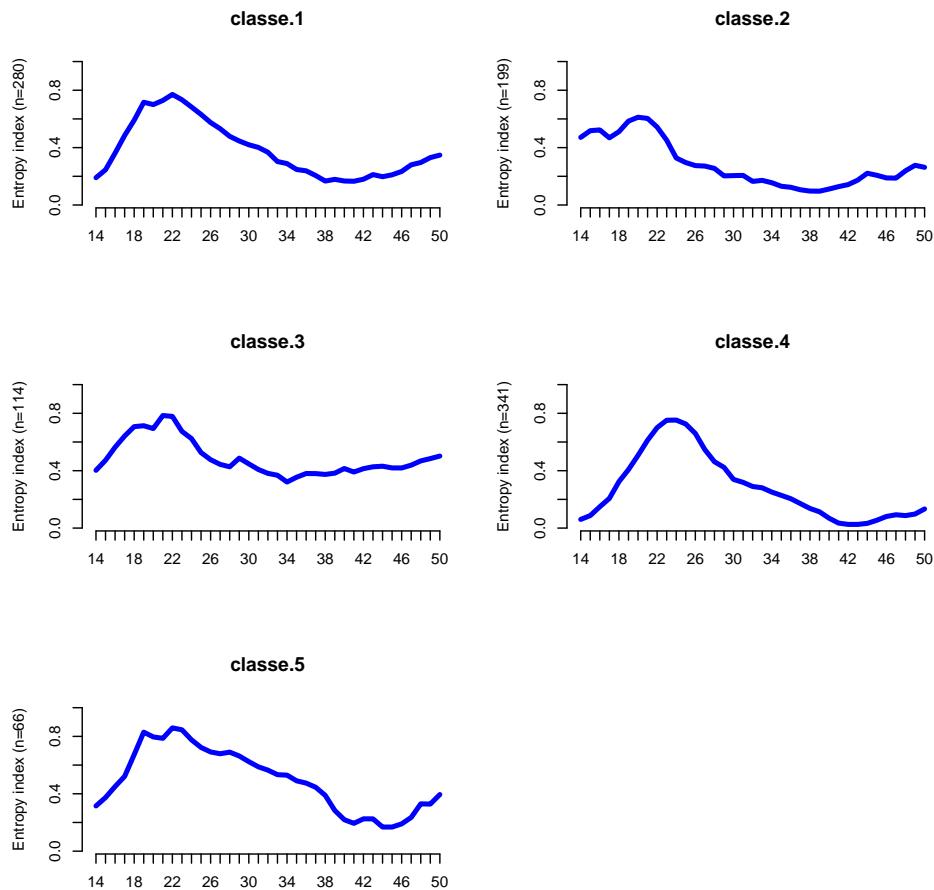


FIGURE 11.9 – Entropie transversale

Enfin, l'entropie transversale décrit l'évolution de l'homogénéité de la classe. Pour un âge donné, une entropie proche de 0 signifie que tous les individus de la classe (ou presque) sont dans la même situation. À l'inverse, l'entropie est de 1 si les individus sont dispersés dans toutes les situations. Ce type de graphique produit par `seqHtplot` peut être pratique pour localiser les moments de transition, l'insertion professionnelle ou une mobilité sociale ascendante (voir par exemple la figure 11.9 de la présente page).

On souhaite maintenant connaître la distribution de la typologie (en effectifs et en pourcentages) :

```
R> freq(seq.part)
```

	n	%
classe.1	280	28.0
classe.2	199	19.9
classe.3	114	11.4
classe.4	341	34.1
classe.5	66	6.6
NA	0	0.0

On poursuit ensuite la description des classes en croisant la typologie avec la variable `generation` :

```
R> cprop(table(seq.part, donnees$generation))

seq.part    1930-38 1939-45 1946-50 Ensemble
  classe.1   25.6     27.1     31.0     28.0
  classe.2   20.9     20.0     18.9     19.9
  classe.3    7.9     13.6     12.9     11.4
  classe.4   38.2     31.9     32.1     34.1
  classe.5    7.4      7.5      5.2      6.6
  Total      100.0    100.0    100.0    100.0

R> chisq.test(table(seq.part, donnees$generation))

Pearson's Chi-squared test

data: table(seq.part, donnees$generation)
X-squared = 12.03, df = 8, p-value = 0.1498
```

Le lien entre le fait d'avoir un certain type de carrières et la cohorte de naissance est significatif à un seuil de 15 %. On constate par exemple l'augmentation continue de la proportion de carrières de type « professions intermédiaires » (classe 1) et, entre les deux cohortes les plus anciennes, l'augmentation de la part des carrières de type « employés » (classe 3) et la baisse de la part des carrières de type « cadres » (classe 4).

Bien d'autres analyses sont envisageables : croiser la typologie avec d'autres variables (origine sociale, etc.), construire l'espace des carrières possibles, étudier les interactions entre trajectoires familiales et professionnelles, analyser la variance des dissimilarités entre séquences en fonction de plusieurs variables « explicatives »¹⁶... Mais l'exemple proposé est sans doute bien suffisant pour une première introduction !

11.5 Bibliographie

- Abbott A., 2001, *Time matters. On theory and method*, The University of Chicago Press.
- Abbott A., Hrycak A., 1990, « Measuring ressemblance in sequence data : an optimal matching analysis of musicians' careers », *American journal of sociology*, (96), p.144-185. <http://www.jstor.org/stable/10.2307/2780695>
- Abbott A., Tsay A., 2000, « Sequence analysis and optimal matching methods in sociology : Review and prospect », *Sociological methods & research*, 29(1), p.3-33. <http://smr.sagepub.com/content/29/1/3.short>
- Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M., 2011, « Analyzing and visualizing state sequences in R with TraMineR », *Journal of Statistical Software*, 40(4), p.1-37. <http://archive-ouverte.unige.ch/downloader/vital/pdf/4hff8pe6uhukqiavvgaluqmjq2/out.pdf>
- Grelet Y., 2002, « Des typologies de parcours. Méthodes et usages », *Document Génération* 92, (20), 47 p. http://www.cmh.greco.ens.fr/programs/Grelet_tropolparc.pdf

¹⁶. L'articulation entre méthodes « descriptives » et méthodes « explicatives » est un prolongement possible de l'analyse de séquences. Cependant, l'analyse de séquences était envisagée par Abbott comme une alternative à la sociologie quantitative *mainstream*, i.e le « paradigme des variables » et ses hypothèses implicites souvent difficilement tenables (Abbott, 2001). Une bonne description solidement fondée théoriquement vaut bien des « modèles explicatifs » (Savage, 2009).

- Lelièvre É., Vivier G., 2001, « Évaluation d'une collecte à la croisée du quantitatif et du qualitatif : l'enquête Biographies et entourage », *Population*, (6), p.1043-1073. http://www.persee.fr/web/revues/home/prescript/article/pop_0032-4663_2001_num_56_6_7217
- Lemercier C., 2005, « Les carrières des membres des institutions consulaires parisiennes au XIXe siècle », *Histoire et mesure*, XX (1-2), p.59-95. <http://histoiremesure.revues.org/786>
- Lesnard L., 2008, « Off-Scheduling within Dual-Earner Couples : An Unequal and Negative Externality for Family Time », *American Journal of Sociology*, 114(2), p.447-490. http://laurent.lesnard.free.fr/IMG/pdf/lesnard_2008_off-scheduling_within_dual-earner_couples-2.pdf
- Lesnard L., Saint Pol T. (de), 2006, « Introduction aux Méthodes d'Appariement Optimal (Optimal Matching Analysis) », *Bulletin de Méthodologie Sociologique*, 90, p.5-25. <http://bms.revues.org/index638.html>
- Robette N., 2011, *Explorer et décrire les parcours de vie : les typologies de trajectoires*, CEPED (Les Clefs pour), 86 p. http://nicolas.robette.free.fr/Docs/Robette2011_Manuel_TypoTraj.pdf
- Robette N., 2012, « Du prosélytisme à la sécularisation. Le processus de diffusion de l'Optimal Matching Analysis », document de travail. http://nicolas.robette.free.fr/Docs/Proselytisme_secularisation_NRobette.pdf
- Robette N., Bry X., 2012, « Harpoon or bait ? A comparison of various metrics to fish for life course patterns », *Bulletin de Méthodologie Sociologique*, 116, p.5-24. http://nicolas.robette.free.fr/Docs/Harpoon_maggot_RobetteBry.pdf
- Robette N., Thibault N., 2008, « L'analyse exploratoire de trajectoires professionnelles : analyse harmonique qualitative ou appariement optimal ? », *Population*, 64(3), p.621-646. <http://www.cairn.info/revue-population-2008-4-p-621.htm>
- Savage M., 2009, « Contemporary Sociology and the Challenge of Descriptive Assemblage », *European Journal of Social Theory*, 12(1), p.155-174. <http://est.sagepub.com/content/12/1/155.short>

Partie 12

Analyse de survie

L'analyse de survie sous R s'effectue principalement avec l'extension `survival`. Nous n'aborderons pas l'analyse de survie ici. Mais plusieurs ressources sont disponibles en ligne :

- <http://www-irma.u-strasbg.fr/~geffray/cours/cours-nantes/M2polyR2008-2009.pdf>
- http://www.mastergbf.fr/_media/members/slemler/courssurvie.pdf
- http://www.mastergbf.fr/_media/members/slemler/survie5.pdf
- http://ljk.imag.fr/membres/Anatoli.Iouditski/cours/M1MAI/dm_ch17.pdf

Partie 13

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.

13.1 Export manuel de tableaux

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 `copy` 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

13.1.1 Copier/coller vers Excel et Word via le presse-papier

La première possibilité est d'utiliser les options par défaut de `copy`. 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 Windows².

```
R> copy(tab)
```

Loading required package: R2HTML

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

13.1.2 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 `copy` l'option `file=TRUE` qui enregistre le contenu de l'objet dans un fichier plutôt que de le placer dans le presse-papier :

```
R> copy(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> copy(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

13.2 Export de graphiques

13.2.1 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.

13.2.2 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()
R>
```

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).

13.3 Génération automatique de documents avec OpenOffice ou LibreOffice

Les méthodes précédentes permettent d'exporter tableaux et graphiques, mais cette opération reste manuelle, un peu laborieuse et répétitive, et surtout elle ne permet pas de mise à jour facile des documents externes en cas de modification des données analysées ou du code.

R et son extension `odfWeave` permettent de résoudre en partie ce problème. Le principe de base est d'inclure du code R dans un document de type traitement de texte, et de procéder ensuite au remplacement automatique du code par le résultat sous forme de texte, de tableau ou de figure.

À noter qu'`odfWeave` n'est pas la seule extension proposant ce type de fonctionnalités, on citera notamment `knitr`, présentée section 13.4 page 172, plus utilisée et plus versatile. `odfWeave` a l'avantage de fournir directement en sortie un document au format OpenDocument, mais présente l'inconvénient de devoir saisir le code R dans LibreOffice, sans les facilités d'édition d'un outil spécifique à R, et sans pouvoir exécuter ce code de manière interactive.

13.3.1 Prérequis

`odfWeave` ne fonctionne qu'avec des documents au format OpenDocument (extension `.odt`), donc en particulier avec OpenOffice ou LibreOffice mais pas avec Word. L'utilisation d'OpenOffice est cependant très proche de celle de Word, et les documents générés peuvent être ensuite ouverts sous Word pour édition.

L'installation de l'extension se fait de manière tout à fait classique :

```
R> install.packages("odfWeave", dep = TRUE)
```

Un autre prérequis est de disposer d'applications permettant de compresser et décompresser des fichiers au format `zip`. Or ceci n'est pas le cas par défaut sous Windows. Pour les récupérer, téléchargez l'archive à l'adresse suivante :

<http://alea.fr.eu.org/public/files/zip.zip>

Décompressez-là et placez les deux fichiers qu'elle contient (`zip.exe` et `unzip.exe`) dans votre répertoire système, c'est à dire en général soit `c:\windows`, soit `c:\winnt`.



FIGURE 13.1 – Exemple de fichier odfWeave



FIGURE 13.2 – Résultat de l'exemple de la figure 13.1

13.3.2 Exemple

Prenons tout de suite un petit exemple. Soit le fichier OpenOffice représenté figure 13.1 de la présente page.

On voit qu'il contient à la fois du texte mis en forme (sous forme de titre notamment) mais aussi des passages plus ésotériques qui ressemblent plutôt à du code R.

Ce code est séparé du reste du texte par les caractères «»=, en haut, et @, en bas.

Créons maintenant un nouveau fichier R dans le même répertoire que notre fichier OpenOffice, et mettons-y le contenu suivant :

```
R> library(odfWeave)
R> odfWeave("odfWeave_exemple1.odt", "odfWeave_exemple1_out.odt")
```

Puis exécutons le tout... Nous devrions alors avoir un nouveau fichier nommé odfWeave_exemple1_out.odt dans notre répertoire de travail. Si on l'ouvre avec OpenOffice, on obtient le résultat indiqué figure 13.2 de la présente page.

Que constate-t-on ? Le passage contenant du code R a été remplacé par le code R en question, de couleur bleue, et par son résultat, en rouge.

Tout ceci est bien sympathique mais un peu limité. La figure 13.3 page suivante, montre un exemple plus complexe, dont le résultat est indiqué figure 13.4, page 171.

Le premier bloc de code R contient des options entre les séparateurs « et »=. L'option echo=FALSE supprime l'affichage du code R (en bleu) dans le document résultat. L'option results=hide supprime

```

<<echo=FALSE, .results=hide>>=
library(questionr)
@|
|
Tableau-croisé¶
|
<<echo=FALSE, .results=xml>>=
data(iris)|
tab<-table(cut(iris$Sepal.Length, 4),iris$Species)|
odfTable.matrix(tab)|
@|
|
L'effectif total du tableau vaut \Sexpr{sum(tab)}.

Graphique¶
|
<<echo=FALSE, .fig=TRUE>>=
boxplot(Sepal.Length ~ Species, .data=iris)|
@|

```

FIGURE 13.3 – Un fichier odfWeave un peu plus compliqué

l'affichage du résultat du code (en rouge). Au final, le code `library(questionr)` est exécuté, mais caché dans le document final.

Dans le deuxième bloc, l'option `results=xml` indique que le résultat du code ne sera pas du simple texte mais un objet déjà au format OpenOffice (en l'occurrence un tableau). Le code lui-même est ensuite assez classique, sauf la dernière instruction `odfTable.matrix`, qui, appliquée à un objet de type `table`, produit le tableau mis en forme dans le document résultat.

Plus loin, on a dans le cours du texte une chaîne `\Sexpr{sum(tab)}` qui a été remplacée par le résultat du code qu'elle contient.

Enfin, dans le dernier bloc, l'option `fig=TRUE` indique que le résultat sera cette fois une image. Et le bloc est bien remplacé par la figure correspondante dans le document final.

13.3.3 Utilisation

Le principe est donc le suivant : un document OpenOffice classique, avec du texte mis en forme, stylé et structuré de manière tout à fait libre, à l'intérieur duquel se trouve du code R. Ce code est délimité par les caractères «»= (avant le code) et @ (après le code). On peut indiquer des options concernant le bloc de code R entre les caractères « et » de la chaîne ouvrante. Parmi les options possibles les plus importantes sont :

eval si TRUE (par défaut), le bloc de code est exécuté. Sinon il est seulement affiché et ne produit pas de résultat.

echo si TRUE (par défaut), le code R du bloc est affiché dans le document résultat (par défaut en bleu). Si FALSE, le code est masqué.

results indique le type de résultat renvoyé par le bloc. Si l'option vaut `verbatim` (par défaut), le résultat de la commande est affiché tel quel (par défaut en rouge). Si elle vaut `xml`, le résultat attendu est un objet OpenOffice : c'est l'option qu'on utilisera lorsqu'on fait appel à la fonction `odfTable`. Si l'option vaut `hide`, le résultat est masqué.

fig si TRUE, indique que le résultat du code est une image.

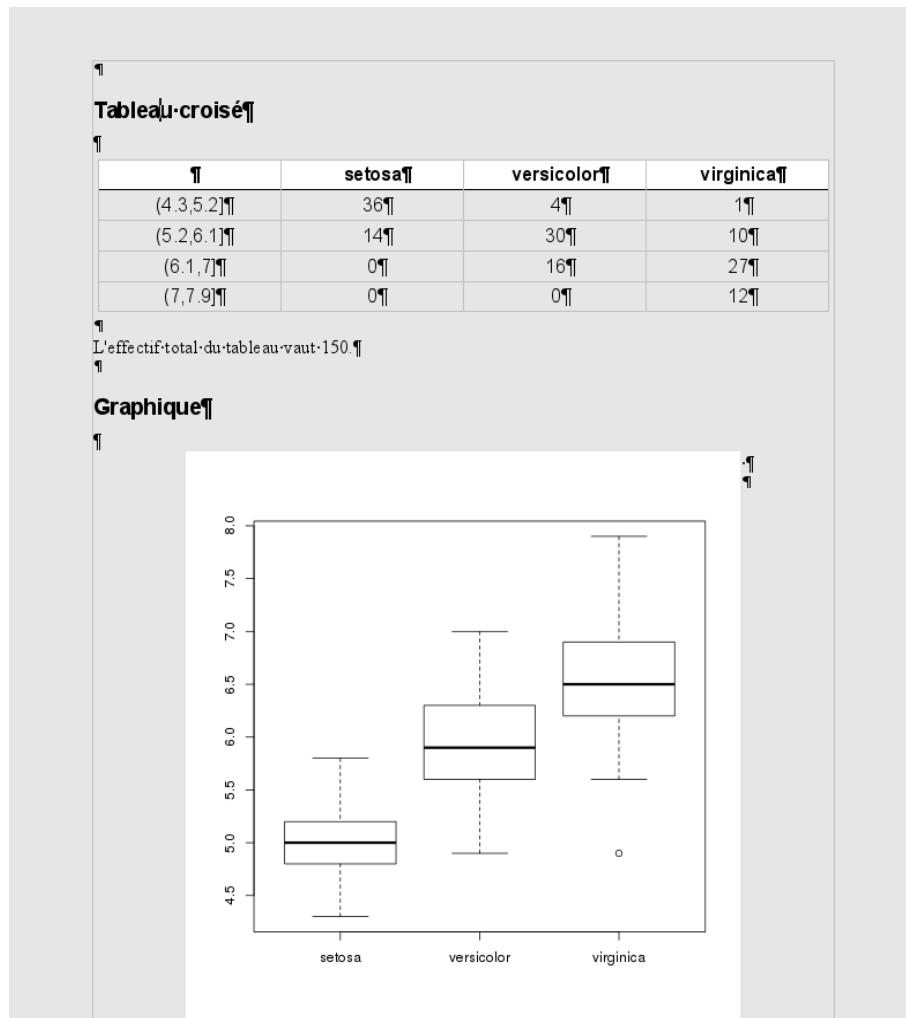


FIGURE 13.4 – Résultat de l'exemple de la figure 13.3

En résumé, si on souhaite utiliser un bloc pour charger des extensions sans que des traces apparaissent dans le document final, on utilise «`echo=FALSE,results='hide'`». Si on veut afficher un tableau généré par `odfTable`, on utilise «`echo=FALSE,results='xml'`». Si on souhaite insérer un graphique, on utilise «`echo=FALSE,fig=TRUE`». Si on souhaite afficher du code R et son résultat « tel quel », on utilise simplement «`>>`».

Pour générer le document résultat, on doit lancer une session R utilisant comme répertoire de travail celui où se trouve le document OpenOffice source, et exécuter les deux commandes suivantes :

```
R> library(odfWeave)
R> odfWeave("fichier_source.odt", "fichier_resultat.odt")
```

En pratique, on répartit en général son travail entre différents fichiers R qu'on appelle ensuite dans le document OpenOffice à l'aide de la fonction `source` histoire de limiter le code R dans le document au strict minimum. Par exemple, si on a regroupé le chargement des données et les recodages dans un fichier nommé `recodages.R`, on pourra utiliser le code suivant en début de document :

```
source("recodages.R")
```

Et se contenter dans la suite de générer les tableaux et graphiques souhaités.



Il existe un conflit entre les extensions `R2HTML` et `odfWeave` qui peut empêcher la seconde de fonctionner correctement si la première est chargée en mémoire. En cas de problème on pourra enlever l'extension `R2HTML` avec la commande `detach(package:R2HTML)`.

Enfin, différentes options sont disponibles pour personnaliser le résultat obtenu, et des commandes permettent de modifier le style d'affichage des tableaux et autres éléments générés. Pour plus d'informations, on se référera à la documentation de l'extension :

<http://cran.r-project.org/web/packages/odfWeave/index.html>

et notamment au document d'introduction en anglais :

<http://cran.r-project.org/web/packages/odfWeave/vignettes/odfWeave.pdf>

13.4 Génération automatique de documents avec knitr

`knitr` est une extension R, développée par Yihui Xie, qui permet de mélanger du code R dans des documents de différents formats 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://yihui.name/knitr/>

`knitr` est extrêmement versatile, et permet d'inclure du code R dans des documents suivant différents formats. On pourra ainsi l'utiliser avec du `LATEX`, du `Markdown` ou du `HTML`.

RStudio⁵ propose une interface pratique à `knitr`⁶. On peut ainsi facilement créer un fichier *R Markdown*, *R HTML* ou *R Sweave* et, d'un clic, générer des fichiers HTML pour les deux premiers formats, ou PDF pour le dernier.

5. Pour plus d'informations sur RStudio, voir section A.5 page 183.

6. `knitr` peut aussi parfaitement s'utiliser en ligne de commande sans passer par RStudio

13.4.1 Exemple

Voyons tout de suite un exemple. Dans RStudio, choisissez le menu *File*, puis *New*, puis *R Markdown*. Un nouveau document s'ouvre. Effacez son contenu et remplacez le par quelque chose comme :

```
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 `xtable`, on peut aussi afficher des tableaux avec le code suivant :

```
```{r exemplatab, results='asis'}
library(xtable)
tab <- table(iris$Species, cut(iris$Sepal.Width, breaks=3))
print(xtable(tab), type="html")
```

```

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 13.5 page suivante.

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 appliquer `knitr` à votre fichier *R Markdown* en lançant R dans le même répertoire et en utilisant le code suivant :

```
R> library(knitr)
R> knit2html("test.Rmd")
```

Le résultat se trouvera dans le fichier `test.html` du même répertoire.

13.4.2 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`.

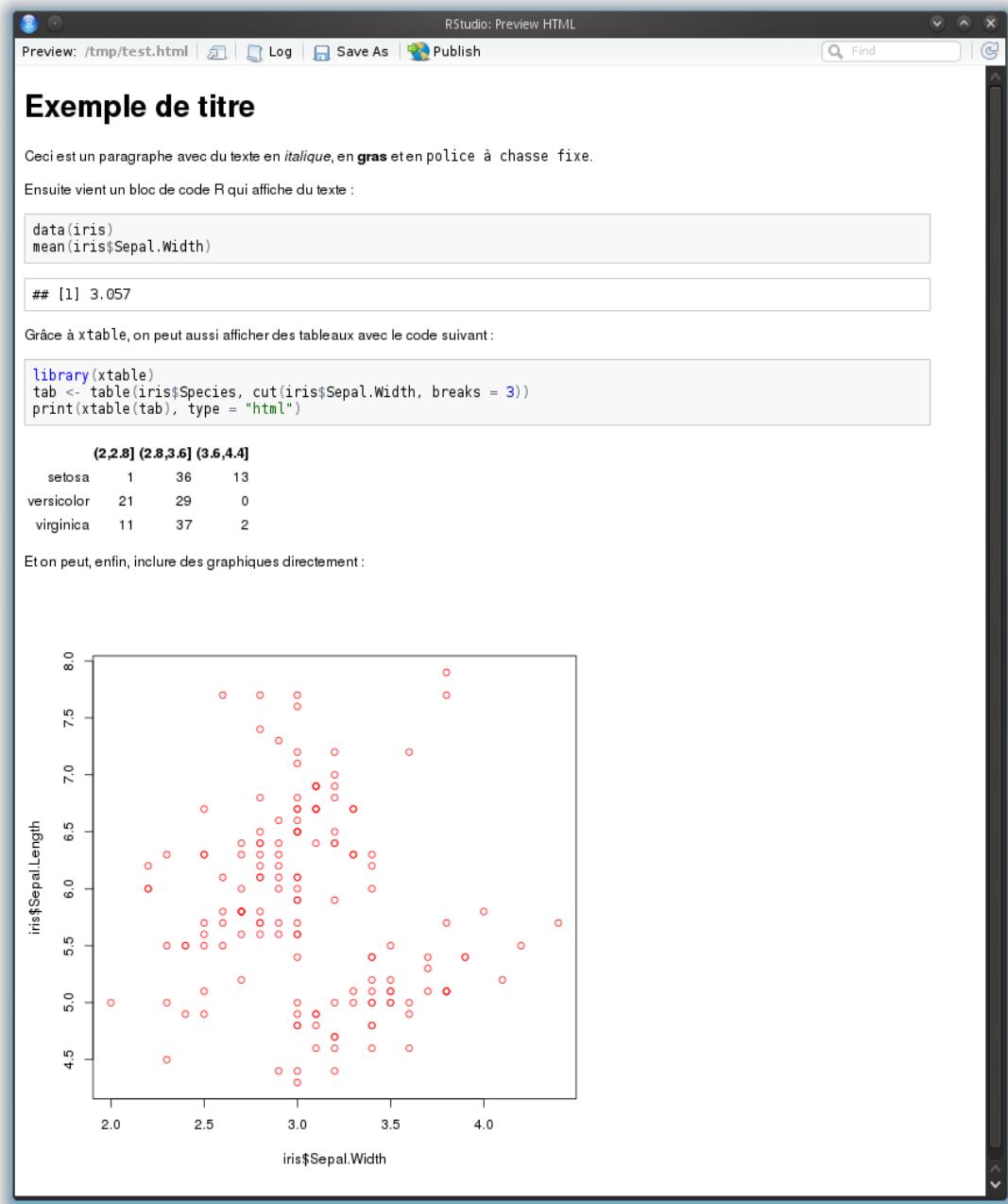


FIGURE 13.5 – Résultat de la génération d'un document HTML par knitr

Markdown est un langage de balisage permettant de mettre en forme du texte en désignant des niveaux de titre, du gras, des listes à puces, etc. Ainsi, du texte placé entre deux astérisques sera mis en italique, une ligne soulignée par des caractères = sera transformée en titre de niveau 1, etc. Dans RStudio, choisissez le 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 suivante⁷ :

```
```{r nom_du_bloc, options}
```

Le bloc commence et se termine par trois quotes inverses suivie, 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` modifera 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>

### 13.4.3 Aller plus loin

L'objectif ici était de présenter un aperçu de l'intérêt et des possibilités de knitr. 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.

`knitr` est très versatile, permet de générer des documents dans de nombreux formats, et évolue rapidement. L'utilisation de programmes auxiliaires comme `pandoc` permettent même de générer des documents au format traitement de texte, par exemple.

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")
```

<sup>7</sup>. Ces délimiteurs seront différents pour d'autres formats de documents, comme *Sweave* ou *Rhtml*.

# Partie 14

## Où trouver de l'aide

### 14.1 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.

#### 14.1.1 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
```

### 14.1.2 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.

## 14.2 Ressources sur le Web

De nombreuses ressources existent en ligne, mais la plupart sont en anglais.

### 14.2.1 Moteur de recherche

Le fait que le logiciel s'appelle R ne facilite malheureusement pas les recherches sur le Web... La solution à ce problème a été trouvée grâce à la constitution d'un moteur de recherche *ad hoc* à partir de Google, nommé Rseek :

<http://www.rseek.org/>

Les requêtes saisies dans Rseek sont exécutées dans des corpus prédéfinis liés à R, notamment les documents et manuels, les listes de discussion ou le code source du programme.

Les requêtes devront cependant être formulées en anglais.

### 14.2.2 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/>

### 14.2.3 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.

**Le Wiki** est un site dont les pages sont éditées par les utilisateurs, à la manière de *Wikipédia*. N'importe quel visiteur du site peut ainsi rajouter ou modifier des informations sur tel aspect de l'utilisation du logiciel :

<http://wiki.r-project.org/>

**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>

Pour les utilisateurs déjà habitués à SAS ou SPSS, le livre *R for SAS and SPSS Users* et le document gratuit qui en est tiré peuvent être de bonnes ressources, tout comme le site Web *Quick-R* :

<http://rforsasandspssusers.com/>

<http://www.statmethods.net/>

#### 14.2.4 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/>

#### 14.2.5 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](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](http://zoonek2.free.fr/UNIX/48_R_2004/all.html)

## 14.3 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.

### 14.3.1 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://listes.cru.fr/sympa/subscribe/r-soc>

Grâce aux services offerts par le site [gmane.org](http://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>

### 14.3.2 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.

### 14.3.3 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.

#### 14.3.4 Canaux 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](irc://irc.freenode.net)) puis de rejoindre l'un des canaux 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.

#### 14.3.5 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 réservé 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<sup>1</sup> 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<sup>2</sup> :

1. Désinstaller R. Pour cela on pourra utiliser l'entrée *Uninstall R* présente dans le groupe R du menu *Démarrer*.

---

1. Voir 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](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 10. 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. Une liste des projets en cours est disponible sur la page suivante :

[http://www.sciviews.org/\\_rgui/](http://www.sciviews.org/_rgui/) (en anglais)

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/>

Par ailleurs, 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 sur 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.

### 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("questionr", "juba")
```

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 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](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
<code>id</code>	Identifiant (numéro de ligne)
<code>poids</code>	Variable de pondération <sup>1</sup>
<code>age</code>	Âge
<code>sexe</code>	Sexe
<code>nivetud</code>	Niveau d'études atteint
<code>occup</code>	Occupation actuelle
<code>qualif</code>	Qualification de l'emploi actuel
<code>freres.soeurs</code>	Nombre total de frères, sœurs, demi-frères et demi-sœurs
<code>cuso</code>	Sentiment d'appartenance à une classe sociale
<code>relig</code>	Pratique et croyance religieuse
<code>trav.imp</code>	Importance accordée au travail
<code>trav.satisf</code>	Satisfaction ou insatisfaction au travail
<code>hard.rock</code>	Ecoute du Hard rock ou assimilés
<code>lecture.bd</code>	Lecture de bandes dessinées
<code>peche.chasse</code>	Pêche ou chasse pour le plaisir au cours des 12 derniers mois
<code>cuisine</code>	Cuisine pour le plaisir au cours des 12 derniers mois
<code>bricol</code>	Bricolage ou mécanique pour le plaisir au cours des 12 derniers mois
<code>cinema</code>	Cinéma au cours des 12 derniers mois
<code>sport</code>	Sport ou activité physique pour le plaisir au cours des 12 derniers mois
<code>heures.tv</code>	Nombre moyen d'heures passées à regarder la télévision par jour

### B.3.4 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
<b>nom</b>	nom de la commune
<b>code</b>	Code de la commune
<b>pop.act</b>	Population active
<b>pop.tot</b>	Population totale
<b>pop15</b>	Population des 15 ans et plus
<b>nb.rp</b>	Nombre de résidences principales
<b>agric</b>	Part des agriculteurs dans la population active
<b>artis</b>	Part des artisans, commerçants et chefs d'entreprises
<b>cadres</b>	Part des cadres
<b>interm</b>	Part des professions intermédiaires
<b>empl</b>	Part des employés
<b>ouvr</b>	Part des ouvriers
<b>retr</b>	Part des retraités
<b>tx.chom</b>	Part des chômeurs
<b>etud</b>	Part des étudiants
<b>dipl.sup</b>	Part des diplômés du supérieur
<b>dipl.aucun</b>	Part des personnes sans diplôme
<b>proprio</b>	Part des propriétaires parmi les résidences principales
<b>hlm</b>	Part des logements HLM parmi les résidences principales
<b>locataire</b>	Part des locataires parmi les résidences principales
<b>maison</b>	Part des maisons parmi les résidences principales

---

1. 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 18

```
R> c(12, 13, 14, 15, 16)
[1] 12 13 14 15 16
```

### Exercice 2.2, page 18

```
R> c(1, 2, 3, 4)
[1] 1 2 3 4

R> 1:4
[1] 1 2 3 4

R> c(1, 2, 3, 4, 8, 9, 10, 11)
[1] 1 2 3 4 8 9 10 11

R> c(1:4, 8:11)
[1] 1 2 3 4 8 9 10 11

R> c(2, 4, 6, 8)
[1] 2 4 6 8

R> 1:4 * 2
[1] 2 4 6 8
```

### Exercice 2.3, page 18

```
R> chef <- c(1200, 1180, 1750, 2100)
R> conjoint <- c(1450, 1870, 1690, 0)
R> nb.personnes <- c(4, 2, 3, 2)
R> (chef + conjoint)/nb.personnes

[1] 662.5 1525.0 1146.7 1050.0
```

### Exercice 2.4, page 18

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

[1] 1180

R> max(chef)

[1] 2100

R> chef.na <- c(1200, 1180, 1750, NA)
R> min(chef.na)

[1] NA

R> max(chef.na)

[1] NA

R> min(chef.na, na.rm = TRUE)

[1] 1180

R> max(chef.na, na.rm = TRUE)

[1] 1750
```

### Exercice 3.5, page 42

```
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 ...
$ cuso : 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 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 42

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 42

```
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")
R> t.test(df$age)
```

### Exercice 3.8, page 42

```
R> table(df$trav.imp)
R> summary(df$trav.imp)
R> freq(df$trav.imp)
R> dotchart(table(df$trav.imp))
R>
```

### Exercice 4.9, page 49

Utilisez la fonction `read.table` ou l'un de ses dérivés, en fonction du tableau 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 49**

Utilisez la fonction `read.dbf` de l'extension `foreign`.

**Exercice 5.11, page 82**

```
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 82**

```
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
[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] 1.764

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

[1] 2.258

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

 Non Oui
2.258 1.764
```

### Exercice 5.16, page 83

```
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 83

```
R> d$fs1 <- cut(d$freres.soeurs, 5)
R> table(d$fs1)

(-0.022,4.39] (4.39,8.8] (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 83

```

R> d$trav.imp2cl[d$trav.imp == "Le plus important" | d$trav.imp == "Aussi important que le reste"] <- "Le plus ou aussi important"
R> d$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
 288 moins ou peu important
 760

R> table(d$trav.imp, d$trav.imp2cl)

 Le plus ou aussi 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(dvar, dsexe)
```

	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(dvar, dbricol)
```

	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(dvar, dlecture.bd)
```

	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

```

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(dvar, dage > 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(dvar, dage > 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

```

### Exercice 7.21, page 117

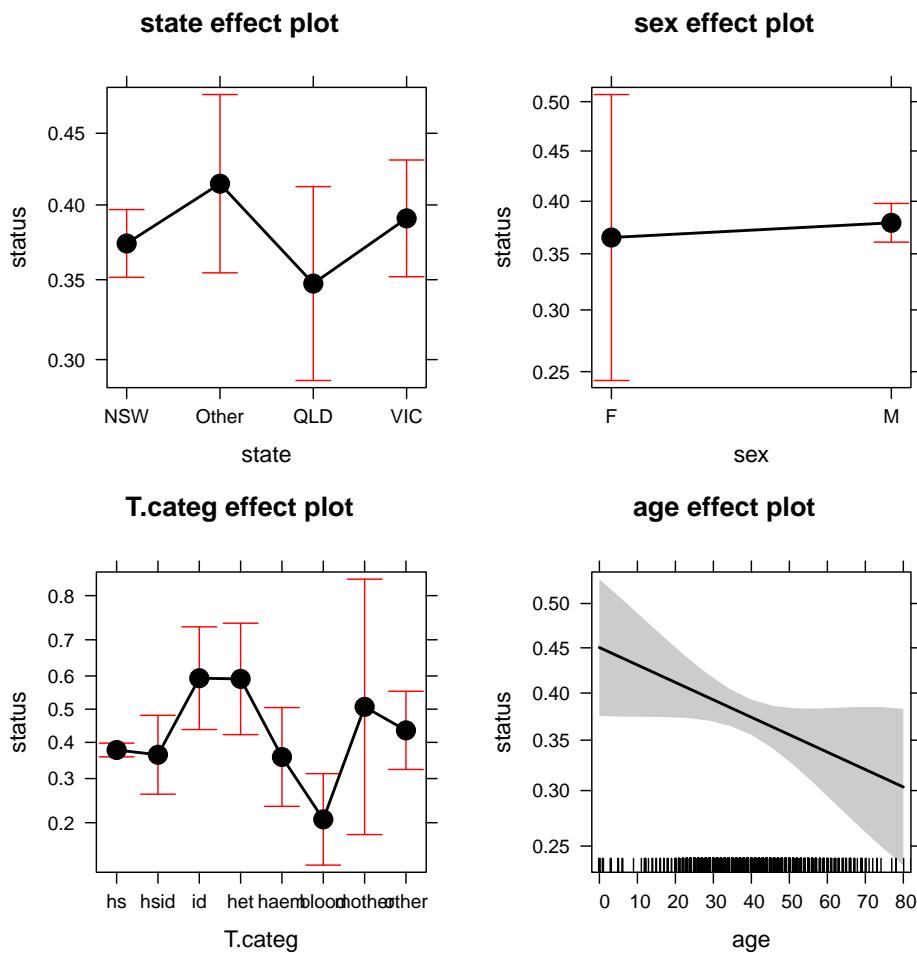
```

R> library(MASS)
R> data(Aids2)
R> freq(Aids2$status) # Pour visualier les modalités

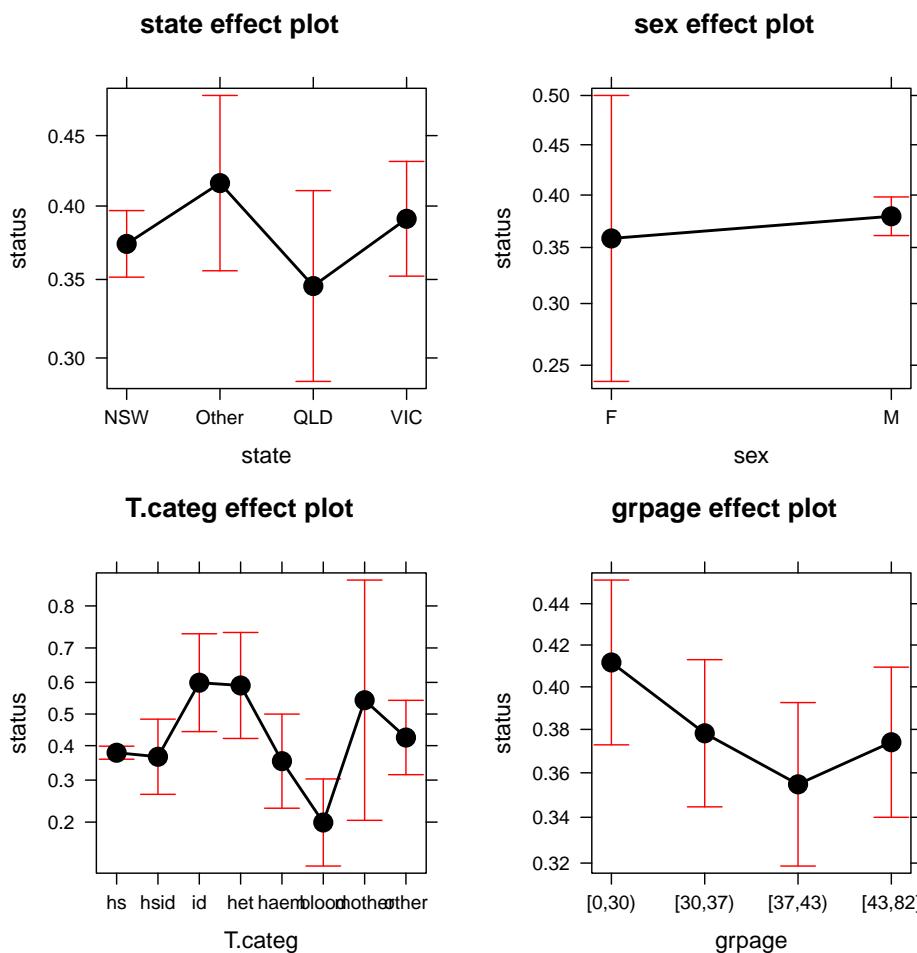
 n %
A 1082 38.1
D 1761 61.9
NA 0 0.0

```

```
R> # On souhaite la probabilité d'être en vie La modalité de référence est donc
R> # le décès
R> Aids2$status <- relevel(Aids2$status, "D")
R> # Premier modèle
R> reg <- glm(status ~ state + sex + T.categ + age, data = Aids2, family = binomial(logit))
R> # Graphique des effets
R> library(effects)
R> plot(allEffects(reg))
```



```
R> # Groupes d'âges
R> Aids2$grpage <- quant.cut(Aids2$age, 4)
R> # Régression
R> reg2 <- glm(status ~ state + sex + T.categ + grpage, data = Aids2, family = binomial(logit))
R> # Graphique des effets
R> plot(allEffects(reg2))
```



```
R> # Calcul des Odds Ratio
R> exp(coef(reg2))

(Intercept) stateOther stateQLD stateVIC sexM
0.6269 1.1933 0.8847 1.0757 1.0945
T.categhsid T.categid T.categhet T.categhaem T.categblood
0.9494 2.4549 2.3688 0.8973 0.4086
T.categmother T.categother grpage[30,37] grpage[37,43] grpage[43,82]
1.9610 1.2151 0.8694 0.7860 0.8541
```

```
R> odds.ratio(reg2) # Alternative
```

Waiting for profiling to be done...

	OR	2.5 %	97.5 %	p
(Intercept)	0.627	0.337	1.16	0.1376
stateOther	1.193	0.907	1.57	0.2041
stateQLD	0.885	0.657	1.18	0.4138
stateVIC	1.076	0.887	1.30	0.4578
sexM	1.094	0.605	2.00	0.7664
T.categhsid	0.949	0.576	1.54	0.8352
T.categid	2.455	1.304	4.73	0.0059 **

```

T.categhet 2.369 1.192 4.81 0.0146 *
T.categhaem 0.897 0.479 1.63 0.7275
T.categblood 0.409 0.224 0.71 0.0022 **
T.categmother 1.961 0.413 10.39 0.3936
T.categother 1.215 0.742 1.97 0.4317
grpage[30,37) 0.869 0.699 1.08 0.2085
grpage[37,43) 0.786 0.624 0.99 0.0407 *
grpage[43,82] 0.854 0.684 1.07 0.1632

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> # Recherche d'un meilleur modèle
R> best.reg <- step(reg2)

Start: AIC=3766
status ~ state + sex + T.categ + grpage

 Df Deviance AIC
- state 3 3739 3763
- sex 1 3736 3764
- grpage 3 3740 3764
<none> 3736 3766
- T.categ 7 3767 3783

Step: AIC=3763
status ~ sex + T.categ + grpage

 Df Deviance AIC
- sex 1 3739 3761
- grpage 3 3743 3761
<none> 3739 3763
- T.categ 7 3772 3782

Step: AIC=3761
status ~ T.categ + grpage

 Df Deviance AIC
- grpage 3 3743 3759
<none> 3739 3761
- T.categ 7 3772 3780

Step: AIC=3759
status ~ T.categ

 Df Deviance AIC
<none> 3743 3759
- T.categ 7 3777 3779

```

# Table des figures

2.1	L'interface de R sous Windows au démarrage . . . . .	10
2.2	L'interface de RStudio sous Windows au démarrage . . . . .	10
3.1	Exemple d'histogramme . . . . .	28
3.2	Un autre exemple d'histogramme . . . . .	29
3.3	Encore un autre exemple d'histogramme . . . . .	30
3.4	Exemple de boîte à moustaches . . . . .	31
3.5	Interprétation d'une boîte à moustaches . . . . .	32
3.6	Boîte à moustaches avec représentation des valeurs . . . . .	33
3.7	Exemple de diagramme en bâtons . . . . .	37
3.8	Exemple de diagramme de Cleveland . . . . .	38
3.9	Exemple de diagramme de Cleveland ordonné . . . . .	39
3.10	Différentes valeurs possibles pour l'argument <code>pch</code> . . . . .	40
6.1	Nombre d'heures de télévision selon l'âge . . . . .	85
6.2	Nombre d'heures de télévision selon l'âge avec semi-transparence . . . . .	86
6.3	Représentation de l'estimation de densité locale . . . . .	87
6.4	Proportion de cadres et proportion de diplômés du supérieur . . . . .	88
6.5	Régression de la proportion de cadres par celle de diplômés du supérieur . . . . .	90
6.6	<i>Boxplot</i> de la répartition des âges (sous-populations) . . . . .	91
6.7	<i>Boxplot</i> de la répartition des âges (formule) . . . . .	92
6.8	Distribution des âges pour appréciation de la normalité . . . . .	94
6.9	Exemple de graphe en mosaïque . . . . .	100
6.10	Exemple de barres cumulées . . . . .	101
7.1	Représentation graphique de l'effet de chaque variable du modèle logistique . . . . .	111
7.2	Représentation graphique de l'effet de chaque variable du modèle logistique . . . . .	117
8.1	Fonctions graphiques de l'extension <code>survey</code> . . . . .	123
9.1	Valeurs propres ou inerties de chaque axe . . . . .	129
9.2	Cercle de corrélations des modalités sur les deux premiers axes . . . . .	129

9.3 Répartition des modalités selon le premier axe . . . . .	129
9.4 Répartition des modalités selon le second axe . . . . .	130
9.5 Rapports de corrélation des variables sur les 4 premiers axes . . . . .	130
9.6 Répartition des modalités selon les deux premiers axes . . . . .	130
9.7 Répartition des modalités selon les axes 3 et 4 . . . . .	131
9.8 Répartition des individus selon les deux premiers axes . . . . .	131
9.9 Répartition des individus selon les trois premiers axes . . . . .	131
9.10 Vecteurs des modalités selon les deux premiers axes . . . . .	132
9.11 Distribution des individus dans le plan factoriel . . . . .	132
9.12 Individus dans le plan factoriel selon le sexe ( <code>s.class</code> ) . . . . .	132
9.13 Individus dans le plan factoriel selon le sexe ( <code>s.chull</code> ) . . . . .	133
9.14 Palettes de couleurs disponibles dans <code>RColorBrewer</code> . . . . .	133
9.15 La fonction <code>scatter</code> appliquée au résultat d'une ACM . . . . .	133
9.16 Plan factoriel (deux premiers axes) . . . . .	134
9.17 Plan factoriel (axes 3 et 4) . . . . .	134
9.18 Plan factoriel (seulement les individus et les catégories) . . . . .	134
9.19 Plan factoriel (seulement les catégories) . . . . .	135
9.20 Plan factoriel (seulement les variables) . . . . .	135
9.21 Ellipses de confiance ( <code>means=TRUE</code> ) dans le plan factoriel . . . . .	135
9.22 Ellipses de confiance ( <code>means=FALSE</code> ) dans le plan factoriel . . . . .	136
 10.1 Dendrogramme obtenu avec <code>hclust</code> . . . . .	141
10.2 Sauts d'inertie du dendrogramme . . . . .	142
10.3 Différentes partitions du dendrogramme . . . . .	143
10.4 Perte relative d'inertie selon le nombre de classes . . . . .	143
10.5 Projection de la typologie obtenue par CAH selon les 4 premiers axes . . . . .	144
10.6 Un dendrogramme coloré . . . . .	144
10.7 Dendrogramme obtenu avec <code>HCPC</code> . . . . .	145
10.8 Dendrogramme obtenu avec <code>HCPC</code> . . . . .	147
 11.1 Dendrogramme de la classification des séquences . . . . .	152
11.2 Sauts d'inertie de la classification des séquences . . . . .	153
11.3 Chronogrammes . . . . .	154
11.4 Tapis des séquences . . . . .	155
11.5 Tapis des séquences triés par multidimensional scaling . . . . .	156
11.6 Séquences les plus fréquentes de chaque classe . . . . .	158
11.7 Statut modal à chaque âge . . . . .	159
11.8 Durée moyenne dans chaque statut . . . . .	160
11.9 Entropie transversale . . . . .	161
 13.1 Exemple de fichier <code>odfWeave</code> . . . . .	169

13.2 Résultat de l'exemple de la figure 13.1 . . . . .	169
13.3 Un fichier odfWeave un peu plus compliqué . . . . .	170
13.4 Résultat de l'exemple de la figure 13.3 . . . . .	171
13.5 Résultat de la génération d'un document HTML par knitr . . . . .	174

# Index des fonctions

!, 59  
!=, 58  
\*, 17  
+, 17  
-, 17  
/, 17  
::, 17  
<, 58  
<-, 7, 12  
<=, 58  
==, 58  
>, 58  
>=, 58  
??, 177  
\$, 25, 50  
%in%, 61, 72  
&, 59  
^, 17  
A2Rplot, 144  
abline, 89  
addNA, 54, 106  
ade, 134

hclust, 141, 142  
 HCPC, 144–146  
 head, 25, 59  
 help.search, 177  
 help.start, 177  
 help.start(), 177, 178  
 hist, 27  
 ifelse, 74  
 image, 84  
 inertia.dudi, 129  
 install.packages, 185, 186  
 install\_github, 186  
 is.na, 63, 72  
 jpeg, 167  
 kde2d, 84  
 length, 16, 17  
 levels, 52, 104  
 library, 186, 187  
 lm, 88, 89, 118  
 load, 48  
 lprop, 97, 119, 122  
 max, 17  
 MCA, 127, 132, 134  
 mca, 127  
 mean, 7, 16, 17, 67, 118  
 median, 27  
 merge, 48, 79, 80  
 min, 17  
 mosaicplot, 99  
 multinom, 113  
 names, 23, 51  
 ncol, 23  
 nrow, 23  
 odds.ratio, 109, 123  
 odfTable, 170, 172  
 odfTable.matrix, 170  
 order, 76  
 PCA, 127  
 pdf, 167  
 pie, 36  
 plot, 37, 111, 134, 141, 145, 152  
 plotellipses, 134  
 png, 167  
 postscript, 167  
 predict, 111, 123  
 princomp, 127  
 print, 98  
 prop.table, 36  
 prop.test, 40, 41  
 quant.cut, 71  
 rbind, 78  
 read.csv, 46  
 read.csv2, 45  
 read.dbf, 48  
 read.delim2, 45  
 read.dta, 47  
 read.spss, 47  
 read.ss, 47  
 read.table, 43, 46, 49  
 read.xport, 47  
 rect.hclust, 142  
 relevel, 41, 104  
 remove.packages, 186  
 rename.variable, 52  
 row.names, 58  
 rug, 33  
 s.arrow, 131  
 s.chull, 131, 132  
 s.class, 131, 132, 144  
 s.corcicle, 129  
 s.freq, 130  
 s.hist, 131  
 s.label, 130  
 s.value, 131  
 sas.get, 47  
 save, 48  
 save.image, 49  
 scatter, 132  
 score, 129  
 screeplot, 129  
 sd, 17  
 seq.heatmap, 157  
 seqdef, 151  
 seqdist, 139, 151  
 seqdplot, 154  
 seqfplot, 158  
 seqHtplot, 161  
 seqiplot, 154  
 seqmsplot, 158  
 seqmtpplot, 158  
 seqsubm, 151  
 setwd, 44, 80  
 shapiro.test, 93  
 sort, 35, 37, 75  
 source, 80, 81, 172  
 step, 112  
 str, 23, 24, 26, 51  
 subset, 66, 126

summary, 7, 27, 35, 61, 75, 107, 109, 129  
svg, 167  
svyboxplot, 120  
svyby, 120, 122  
svychisq, 120, 122  
svyciprop, 120  
svydesign, 120, 124  
svyglm, 120, 123, 124  
svyhist, 120  
svymean, 120  
svyplot, 120  
svyquantile, 120  
svytable, 120–122  
svytotal, 120  
svyvar, 120  
  
t, 37  
t.test, 34, 93  
table, 34, 35, 37, 42, 53, 61, 75, 96–99, 118  
tail, 25  
tapply, 67, 68, 89, 122  
tiff, 167  
  
update.packages, 183  
  
var, 16, 17, 118  
var.test, 95  
  
win.metafile, 167  
write.dbf, 49  
write.foreign, 49  
write.table, 49  
wtd.mean, 118  
wtd.table, 118, 119  
wtd.var, 118