Analyse statistique de données catégorielles
4303076 – Responsable : Prof. Gilbert Ritschard
Master
Automne 2013

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES

Département des sciences économiques

Université de Genève
Boulevard du Pont d'Arve 40
1211 Genève 4

Livret de travaux pratiques

Support de séminaire

version 1.7.1

Avec solutions

Enseignant Prof. Gilbert Ritschard
Département des sciences économiques
Uni-Mail, bureau 5232
Gilbert.Ritschard@unige.ch

Assistants Emmanuel Rousseaux
Département des sciences économiques
Uni-Mail, bureau 5228
Emmanuel.Rousseaux@unige.ch

Danilo Bolano
Institut d'études démographiques et du parcours de vie
Uni-Mail, bureau 5335
Danilo.Bolano@unige.ch

Remerciements à Matthias Studer pour toutes ses contributions à ce séminaire.

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES

Table des matières

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES

Informations générales

Remarques :

- Durant les TP, de nombreux éclaircissements seront fournis à l'oral en fonction des questions dont vous nous faites part. Pensez à bien prendre notes de ces éléments. Ces notes seront une aide précieuse pour vous préparer à la séance de questions lors de la défense orale de votre travail.
- L'objectif de ce livret de TP est de vous permettre d'obtenir les connaissances pratiques nécessaires pour réaliser les analyses de votre études de cas. Il présuppose que vous maîtrisez déjà le cours avant de vous lancer dans les exercices. Pour vous aider à trouver à quels endroits dans le cours vous trouverez les explications vous permettant de répondre aux questions de l'exercice, des références vers les parties adéquates du cours sont souvent présentées au sein des exercices. Ces références sont indiquées par une étoile dans la marge, comme donné en exemple ici. ★**[CH x.x]**

Objectifs des séminaires

Les séminaires ont pour objectif de vous donner une illustration concrète des outils d'analyse étudiées en cours et de vous faire acquérir un savoir faire pratique sur la mise en œuvre de ces outils.

Modalités d'évaluation

L'évaluation est effectuée sous deux formes :

- Une étude de cas, dans laquelle vous êtes chargé d'apporter d'avantage de compréhension sur une certaine problématique. Le choix de cette problématique est laissé (relativement) libre à l'auditeur. Le travail consiste alors à poser vos hypothèses socio-économiques sur cette problématique, puis de les confronter aux données afin de tenter de les confirmer ou de les infirmer. Appuyé par des tests statistiques, vous discuterez vos hypothèses en toute rigueur et poserez des conclusions. Le travail que vous devez réaliser ici est analogue (bien que simplifié) à celui que l'on vous demandera en tant que sociologue, que ce soit en recherche ou en organisation indépendante (publique ou privée).
- Une interrogation orale portant sur votre étude de cas.

Précisions sur l'évaluation :

- *Problématique :* Nous vous évaluons sur la mise en place, l'utilisation, et l'analyse des outils statistique qui vous sont présentés durant le cours, pas sur l'aspect sociologique du travail. Attention donc à ne pas construire une problématique trop complexe, qui vous empêcherait de consacrer le temps nécessaire pour faire des analyses statistique de qualité, et engendrerait alors une perte de points dans l'évaluation votre travail.
- *Choix des données :* pour répondre à votre problématique vous réaliserez vos analyses sur données réelles. Les données réelles peuvent arriver sous différents formats et différentes qualités. Ainsi, certains jeux de données peuvent être directement prêts à l'emploie, alors que d'autres peuvent nécessiter une longue phase de nettoyage et de clarification avant de pouvoir être utilisées pour une analyse. L'objectif de ce cours n'est pas d'effecter de la préparation de données. Nous vous fournissons pour réaliser vos analyses les données du Panel Suisse des Ménages, qui sont prêtes

à l'emploie. Si pour votre problématique vous souhaitez utiliser un jeu de données différent (données d'un autre pays, récoltées par votre équipe de recherche, etc.) cela est possible mais veillez bien à ce que les données soient déjà suffisamment prêtes pour l'analyse. Le rythme du cours est rapide, et si vous devez passez trop de temps à préparer les données, ceci se ferait au détriment des analyses, et vous perdriez alors des points dans l'évaluation de votre travail. Gardez bien à l'esprit que nous vous évaluons sur la qualité des analyses, pas sur le travail de nettoyage.

- *Une seule variable dépendante :* le rythme du cours est rapide, vous ne pourrez pas réaliser le travail correctement avec plus d'une variable dépendante. Vous devez en choisir une seule.
- *Nombre de variables explicatives :* nous vous conseillons de ne pas vouloir utiliser trop de variables explicatives. En effet, pour chaque variable, vous aurez un pré-traitement à effectuer (gestion des valeurs manquantes, recodage, etc.). Si vous avez beaucoup de variables, vous aurez beaucoup de pré-traitements, et donc moins de temps pour les analyses. Il est préférable d'avoir moins de variables mais de réaliser toutes les analyses demandées avec une bonne qualité, plutôt que de partir avec beaucoup de variables, perdre du temps en recodage, et réaliser des analyses incomplètes ou de mauvaise qualité. Vous perdriez des points dans l'évaluation. Nous conseillons d'utiliser environ 5 variables explicatives. Cas particuliers :
  - Si vous êtes seul(e) pour le travail, vous pouver travailler avec une variable de moins.
  - Si vous êtes trois pour le travail, vous pouvez travailler avec une variable de plus.

Directives pour l'étude de cas

Le travail peut être réalisé seul mais nous conseillons de le réaliser par groupe de deux.

Le corps du texte devrait comprendre entre 8 et 12 pages, 15 au maximum. À cela pourra s'ajouter quelques annexes présentant les *outputs* les plus utiles. Le corps du texte ne doit pas inclure des copies d'*outputs*, mais uniquement des tableaux synthétiques transcrivant uniquement l'information utile au lecteur pour comprendre les analyses.

Nous vous demandons de travailler avec *une seule variable dépendante* et environ *5 variables explicatives*. Parmis les variables explicatives nous vous demandons d'utiliser :

- le sexe
- l'âge
- le niveau d'éducation ou la catégorie socio-professionnelle

Les autres variables explicatives sont en libre choix. Vous choisirez ces variables en fonction de votre problématique et des hypothèses que vous souhaitez tester.

La structure pourrait par exemple être comme suit :

1. Introduction (1 page). Présentation du cadre du travail (travail de Master en X, cours X, etc.), description de la problématique considérée, présentation *claire* des hypothèses que vous souhaitez tester, avec éventuellement des références bibliographiques, et énumération succinte des données et méthodes qui seront utilisées. L'introduction peut aussi annoncer les principaux enseignements qu'apporte votre étude.
2. Données (1-2 pages). Il s'agit de donner toutes les indications utiles sur les données : sources, population concernée, nombre de cas, définition précise des variables retenues, des codages et éventuels recodages . Une personne tierce doit pouvoir reproduire votre analyse.
3. Analyse exploratoire (1-2 pages). Tableau synthétique et commentaire des distributions univariées des variables, avec indication notamment du nombre de valeurs manquantes. Étude des associations bivariées entre la variable dépendante (à expliquer) et les facteurs explicatifs potentiels. Commentaires.

4. Modélisation (2-5 pages). Arbres de classification : détection des interactions les plus pertinentes. Régression logistique : spécification initiale du modèle avec justification. Explication de la démarche de modélisation suivie (essais de plusieurs spécifications, utilisation ou non du *stepwise*,...). Présentation des résultats quantitatifs et évaluation statistique.
5. Interprétation et discussion (2-4 pages) des résultats par rapport à la problématique envisagée.
6. Conclusion (1 page). Récapitulation de la démarche suivie et des principaux résultats trouvés.
7. Bibliographie. Liste des références consultées classées par ordre alphabétique de leur premier auteur, et liste du ou des logiciel(s) que vous avez utilisés pour réaliser vos analyses.

Cette structure est indicative, et il pourra être judicieux dans certains cas de regrouper certains points, comme 4 et 5 par exemple, dans une même section.

N'oubliez pas, pour les tableaux et figures, de mettre des libellés clairs et de donner toutes les permettant leur lecture, sans laisse d'ambiguité d'interprétation au lecteur.

Livrables

- Votre étude de cas, au format PDF, respectant les directives énoncées ci-dessus. En particulier le rapport ne devra en aucun cas excéder 15 pages (sans les annexes).
- Une note indiquant ce que chaque membre du groupe à réalisé dans le rapport.

Dates de rendus et de l'oral

Le rapport devra être rendu pour le vendredi 6 décembre. La défense orale du rapport se déroulera la semaine du 16 au 20 décembre (dernière semaine avant les vacances). Un planning sera mis en place en essayant de tenir compte de vos disponibilités.

Attention : dates à confirmer par Gilbert Ritschard vers le milieu du semestre

Planning des TP

Voici un planning (prévisionnel) des séances de TP.

Séance 1 Présentation de l'organisation des TP et du rendu du rapport. Prendre en main R et Rstudio. Manipuler des données d'enquête avec la librairie Dataset. Comprendre les différents types de variables, introduction aux valeurs manquantes. Graphiques : diagramme en barre, pie.

Séance 2 Mise en place d'une problématique sociologique. Recherche d'un jeu de données pour valider les hypothèses. Chargement d'une base SPSS. Configuration de la pondération. Exclusion des individus non concernés. Opérationnalisation de la problématique. Vérification des cas valides. Vérification des mesures utilisées. Exporter un jeu de données. Présentation succincte des données du Panel Suisse des Ménages.

Séance 3 Consulter les fréquences d'apparition des modalités d'une variable catégorielle. Effectuer un recodage par opération arithmétique, découpage d'une variable continue, et fusion de modalités d'une variable catégorielle. Valider les recodages en consultant les fréquences. Renommer des variables et des valeurs de variable. Présenter un recodage. Apparier plusieurs bases de données. Présentation d'un recodage.

Séance 4 Rédaction de l'interprétation des résultats. Définition des variables de contrôle. Retrait des individus qui sont manquants sur une ou plusieurs variables. Analyses bivariées.

Séance 5 Arbres d'induction.

Séance 6 Régression logistique niveau 1 : régression logistique simple, catégorie de référence, rapports de cotes et leur significativité, statistique du χ^2 du rapport de vraisamblance, régression logistique multiple, profil de référence, calcul de logit, calcul de probabilité prédite, ajout d'interactions, effets transverses, variables de contrôle, modèle complet.

Séance 7 Régression logistique niveau 2 : modèles imbriqués, χ^2 du rapport de vraisemblance entre modèles, χ^2 du rapport de vraisemblance pour une variable, test de l'apport d'une variable, évaluation globale du modèle (χ^2, déviance, AIC, BIC, etc.). Contraste et changement des catégories de référence. Citation de `R` et des librairies `R` utilisées dans un travail.

Séance 8 Coaching sur le dossier personnel.

Séance 9 Coaching sur le dossier personnel, présentation orale de test (entrainement pour la présentation orale).

Supports de cours utilisés pour ces séminaires

Les supports fournis pour suivre le séminaire sont les suivants :
- Ce présent guide des TP
- Les documentations PDF générées par la librairie Dataset pour la vague 2006 du Panel Suisse des ménages.

Travaux pratiques – Séance 1

Objectifs de la séance : présentation de l'organisation des TP et du rendu du rapport. Prendre en main R et Rstudio. Manipuler des données d'enquête avec la librairie Dataset. Comprendre les différents types de variables, introduction aux valeurs manquantes. Graphiques : diagramme en barre, pie.

Remarque :

Les questions des exercices des TP sont faites pour être réaliser dans l'ordre. Vérifiez bien que vous comprenez pleinement la réponse à une question avant de passer à la suivante.

Exercice 1.1 (Préliminaires)**.**

1. Pour le suivi de vos travaux nous aurons besoin de converser avec vous durant le semestre. Veuillez ainsi commencer par vous inscrire au cours sur la plate-forme Chamilo (chamilo.unige.ch).
   Connectez-vous à Chamilo avec votre identifiant étudiant (comme pour le *webmail*), puis faites une recherche du cours à l'aide du code du cours : `4303076` et cliquez sur s'inscrire.
2. Vous devez stocker vos documents dans le disque H :\
   Rendez-vous sur ce disque (passer par l'icône 'Poste de travail') et créez-y un dossier nommé `AnCat`.
3. Des supports de cours vous seront transmis durant le semestre via le site

   mephisto.unige.ch

   Accédez au site web à l'aide du navigateur Firefox, puis rendez-vous dans la partie « supports de cours » du cours `AnCat`. Entrez le mot de passe qui vous sera remis durant la séance. Entrez dans le dossier `tp`, puis 2013. Vous pouvez télécharger le livret de TP et l'enregistrer le dans le dossier `AnCat` précédemment créé.
4. Dans ce cours nous travaillerons avec la librairie R `Dataset`, permettant de gérer des données d'enquête. Vous pourrez avoir durant le semestre des questions sur l'utilisation de cette librairie qui ne seront pas traitées dans ce cours, car spécifique aux besoins de votre problématique. Nous vous conseillsons ainsi de vous inscrire à la liste de diffusion des utilisateurs de la librairie, sur laquelle vous pourrez ensuite poser vos questions et obtenir des réponses de la part des autres utilisateurs de la librairie.
   Suivez l'annexe 2 pour vous abonner à cette liste de diffusion.

Exercice 1.2 (Prise en main de R et Rstudio, graphiques)**.**

1. Lancer Rstudio (Démarrer > Applications SES) et analyser les différents panneaux de l'interface.
2. Créer un fichier de script (File > New > R script ; ou par le raccourci `ctrl+shift+n`).
3. Écrivez dans ce fichier « `# AnCat - Séance 1` » puis sauvegarder le dans votre dossier `AnCat` (File > Save ; ou par le raccourci `ctrl+s`). Nommez le `tp1.R`
   Remarques :

- Une fenêtre s'affiche vous demandant de choisir un encodage. L'encodage est la structure spécifiant comment les éléments textuels sont enregistrés dans un fichier texte. Actuellement, l'encodage le plus universel est l'`UTF-8`. Il est vivement conseillé d'utiliser cet encodage lorsqu'une application vous demande de choisir, sur chacun de vos appareil (PC, Mac, Unix, etc.) afin d'éviter des problèmes d'affichages (signes @, ©, ®, etc. qui apparaissent).
- Le caractère `#` indique la présence d'un commentaire. Tout le texte se situant à droite de ce caractère ne sera pas interprété par R. Les commentaires servent principalement à expliquer les opérations que vous demandez à R de réaliser. Un fichier de script correctement commenté doit pouvoir être compris et utilisé par une tierce personne sans qu'elle ait besoin de vous poser de question.
- Nous ne sommes jamais à l'abri d'une perte de données : coupure de courant, perte du réseau, panne subite de la machine, etc. Un bon réflexe à acquérir est de sauvegarder *à chaque ligne que vous écrivez*. Ceci se fait sans perte de productivité avec `ctrl+s`.

4. Nous commençons par demander à R de nous faire un simple calcul. Tapez `1+1` dans le fichier de script et exécuter le en cliquand sur `Run` ou avec le raccourci `ctrl+enter`.

Solution:

```
1 + 1
```

```
## [1] 2
```

5. Souvent, nous aurons besoin de stocker le résultat d'un calcul pour le réutiliser ailleurs. Pour cela, on stocke le résultat dans un objet. Tapez `a <- 1+1` et exécutez la commande. Maintenant, si j'ai besoin du résultat, je peux le rappeler en utilisant le nom de l'objet.

Solution:

```
a <- 1 + 1
```

```
a
```

```
## [1] 2
```

6. Lorsque nous avons plusieurs valeurs à stocker, nous utilisons un vecteur. En R on crée un vecteur avec la fonction `c()`.
`c()` réfère à *column*, qui est la façon dont on représente un vecteur en mathématique. Tapez `longueur.cheveux <- c(15, 75, 10)`, exécutez le code, et appelez l'objet pour vérifier ce qu'il contient.

Solution:

```
longueur.cheveux <- c(15, 75, 30)
```

```
longueur.cheveux
```

```
## [1] 15 75 30
```

7. Il est conseillé d'écrire des noms d'objets explicites, ainsi vous (et les éventuelles personnes avec qui vous pourrez échanger du code) éviterez des confusions sur ce que vous manipulez exactement dans vos objets, ce qui est source d'erreurs fréquentes. Vous constatez aussi qu'il est possible d'utiliser le symbole `'.'` dans le nom d'un objet, ce qui est pratique pour séparer les mots. On pourrait penser qu'il est fastidieux d'écrire des noms longs, il n'en n'est rien car Rstudio offre un système de completion. Il s'active avec la touche `Tab`.
   Écriver `lon` puis faites `Tab` pour retrouver votre objet. Faîtes encore `Tab` pour exécuter la completion.
8. R propose une batterie de fonctions pour étudier une série statistique. Calculez la moyenne de la longueur des cheveux avec la fonction `mean()`, puis son écart-type avec la fonction `sd()`.

Solution:

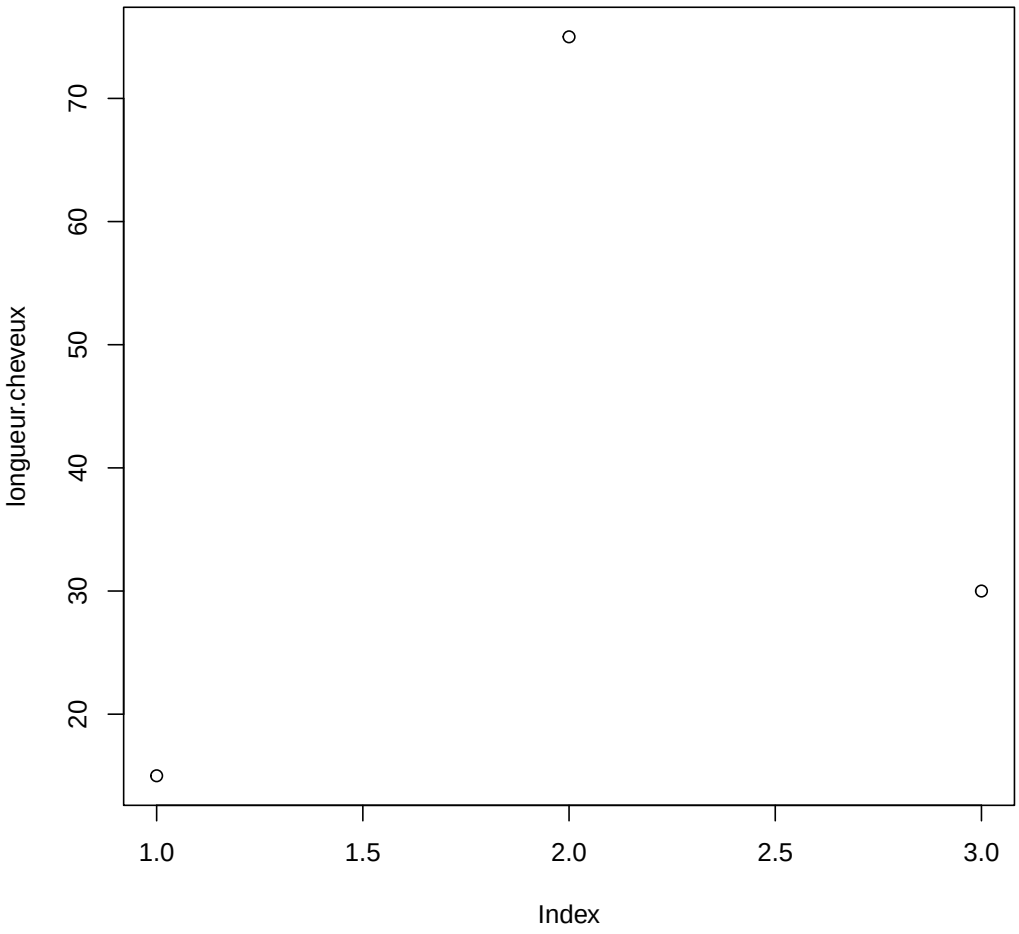
```
mean(longueur.cheveux)
## [1] 40

sd(longueur.cheveux)
## [1] 31.22
```

9. Un point fort de R est le nombre de fonctions graphiques qu'il propose. La méthode standard pour faire un graphique est `plot`.
   Exécutez `plot` sur votre objet `longueur.cheveux`.

Solution:

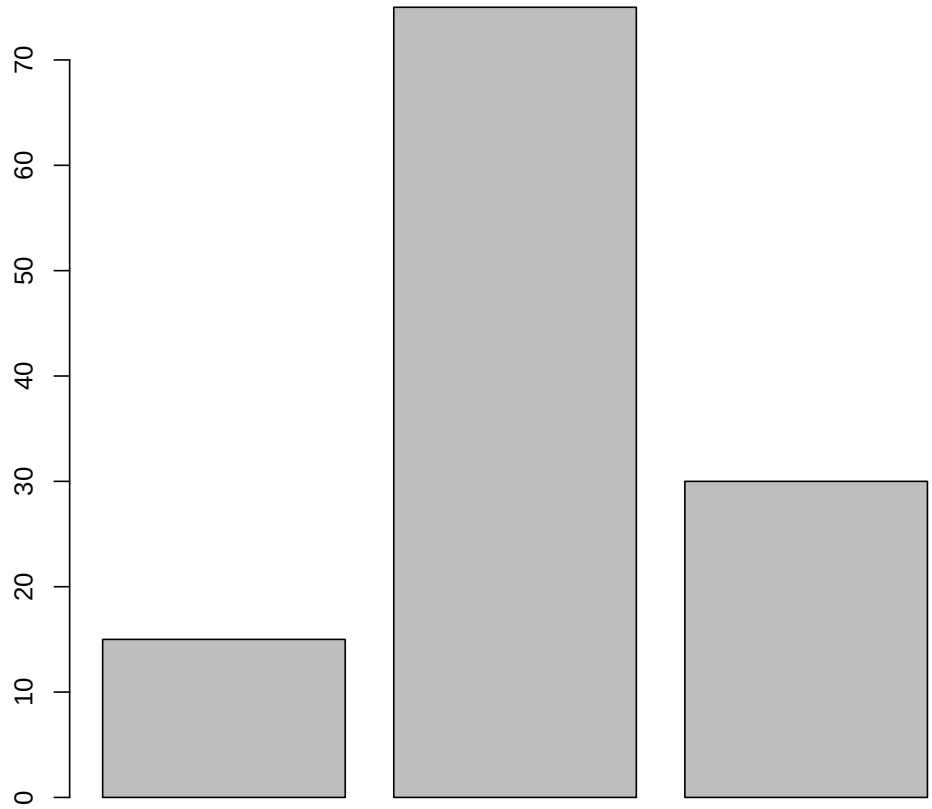
```
plot(longueur.cheveux)
```



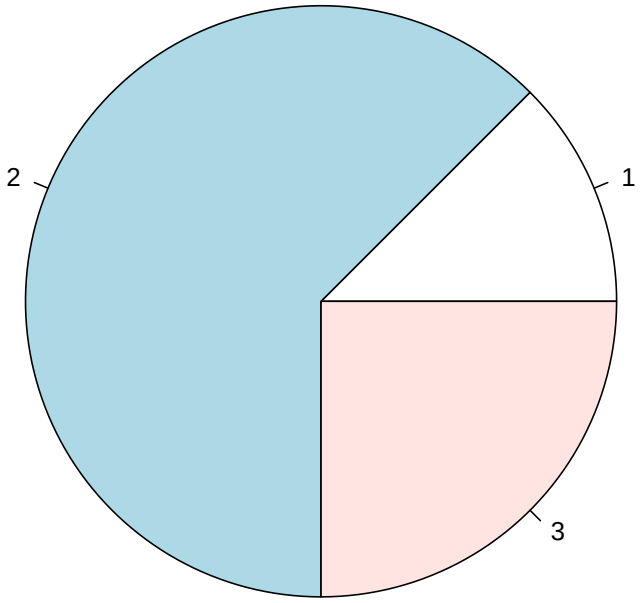
10. Des fonctions plus spécifiquent existent suivant le type de graphique que vous voulez créer.
    Testez successivement les fonctions `barplot`, `pie` et `boxplot` sur votre objet `longueur.cheveux`.

Solution:

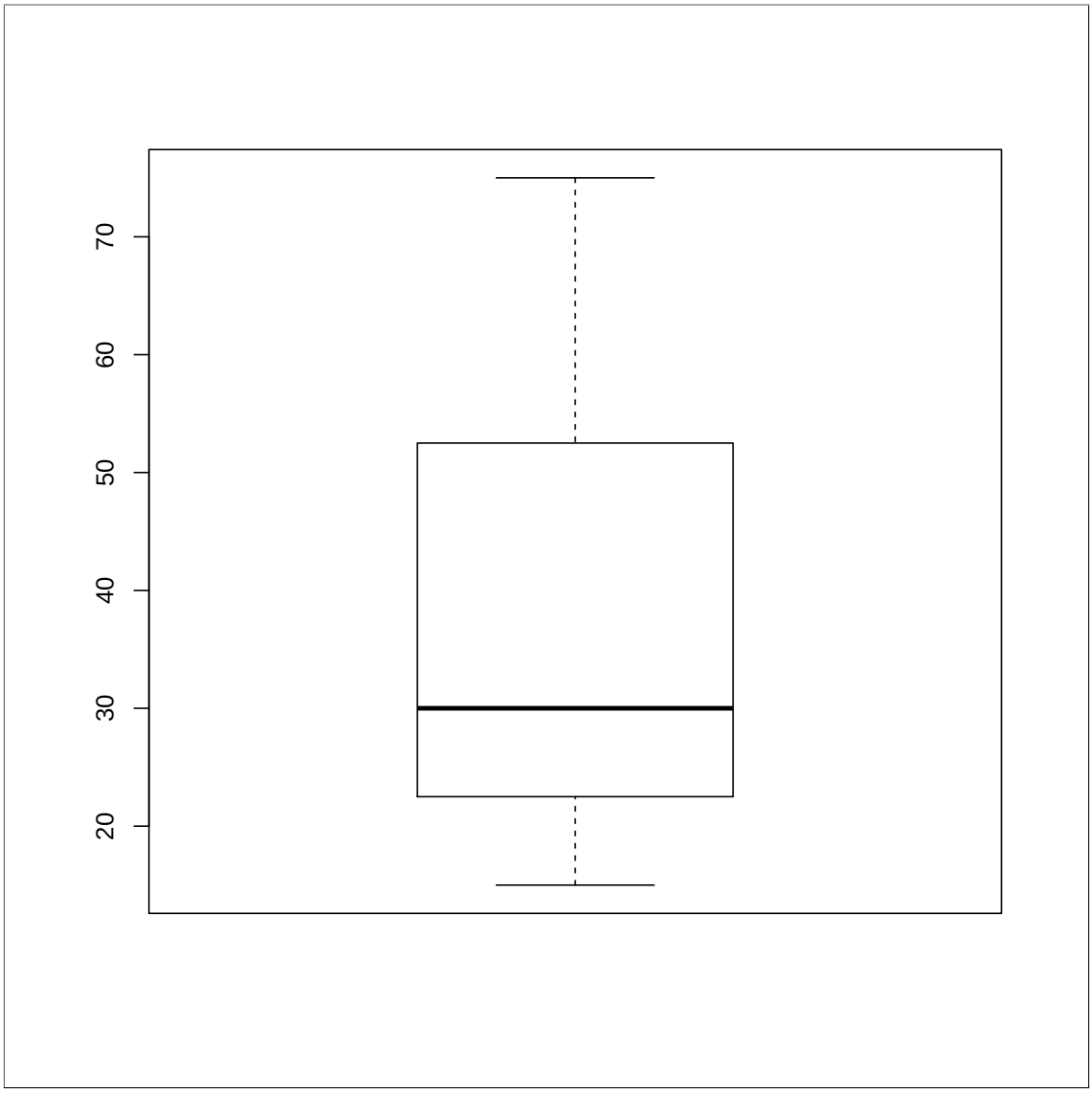
```
barplot(longueur.cheveux)
```




```
pie(longueur.cheveux)
```




```
boxplot(longueur.cheveux)
```



Remarque : une fois les graphiques affichés, vous pouvez naviguer rapidement entre eux grâce aux flèches (gauche/droite) du panneau `Plots`.

11. Un graphique doit contenir en lui-même toute l'information permettant au lecteur de l'interpréter correctement. On ne doit pas avoir besoin de texte annexe pour en expliquer la lecture. Dans un rapport, un texte qui commente un graphique doit contenir uniquement les interprétations que nous pouvons tirer à partir du graphique.
    Nous allons donc ajouter une légende propre à notre graphique. Pour cela nous allons ajouter des options à la fonction réalisant le graphique, par exemple `barplot()`.
    Nous ne pouvons pas deviner nous-même quelles sont les options, nous devons nous référer au manuel utilisateur. Dans R, le manuel d'une fonction se lance avec la commande `?nomfonction`.
    Lancez le manuel de la fonction `barplot`, et étudiez sa structure. Remarquez notamment l'ensemble des exemples fournis en bas de page.

Solution:

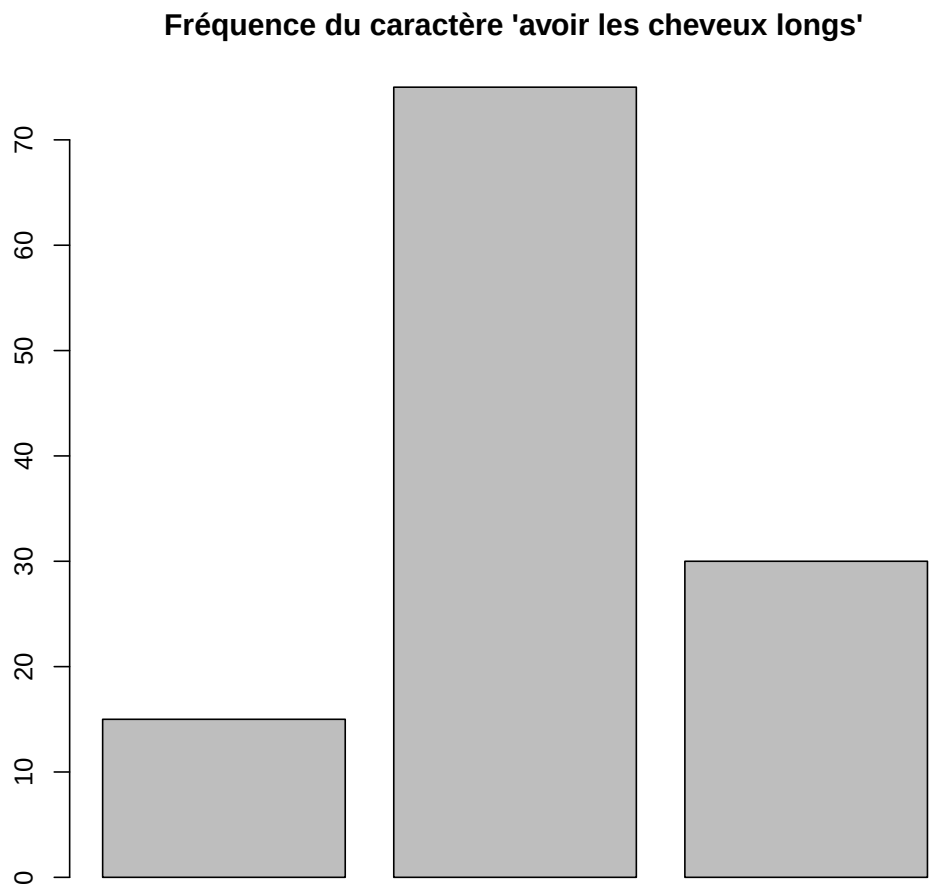
```
?barplot
```

12. Nous connaissons maintenant les options de la fonction `barplot`, remarquons que l'argument `longueur.cheveux` que nous avons précédemment donné correspond à

l'option `height`, et qu'ajouter un titre au graphique se fait avec l'option `main`.
Rstudio permet d'afficher l'ensemble des options d'une fonction, sans avoir besoin d'entrer dans le manuel.
Écrivez `barplot(` puis faite une demande de completion avec `Tab`. Choisissez l'option `height` et donnez l'objet `longueur.cheveux` en argument, séparez ensuite par une virgule, ajoutez l'option `main` et lui donner en argument le titre
`Fréquence du caractère 'avoir les cheveux longs'`
L'argument doit être donné entre des guillements doubles pour indiquer qu'il s'agit d'une chaîne de caractère. Si on ne met pas les guillements, R va croire que l'on écrit des noms d'objets.

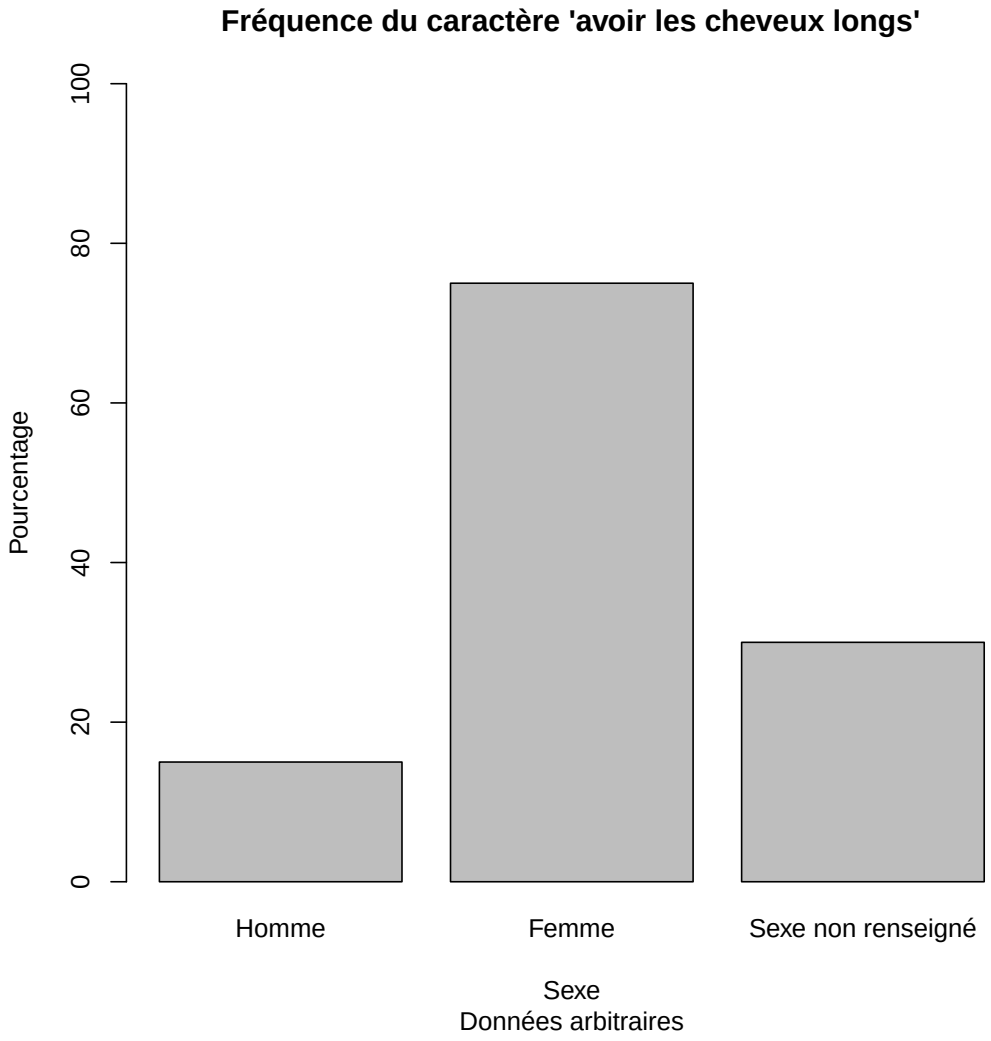
Solution:

```
barplot(
  height=longueur.cheveux,
  main = "Fréquence du caractère 'avoir les cheveux longs'"
)
```



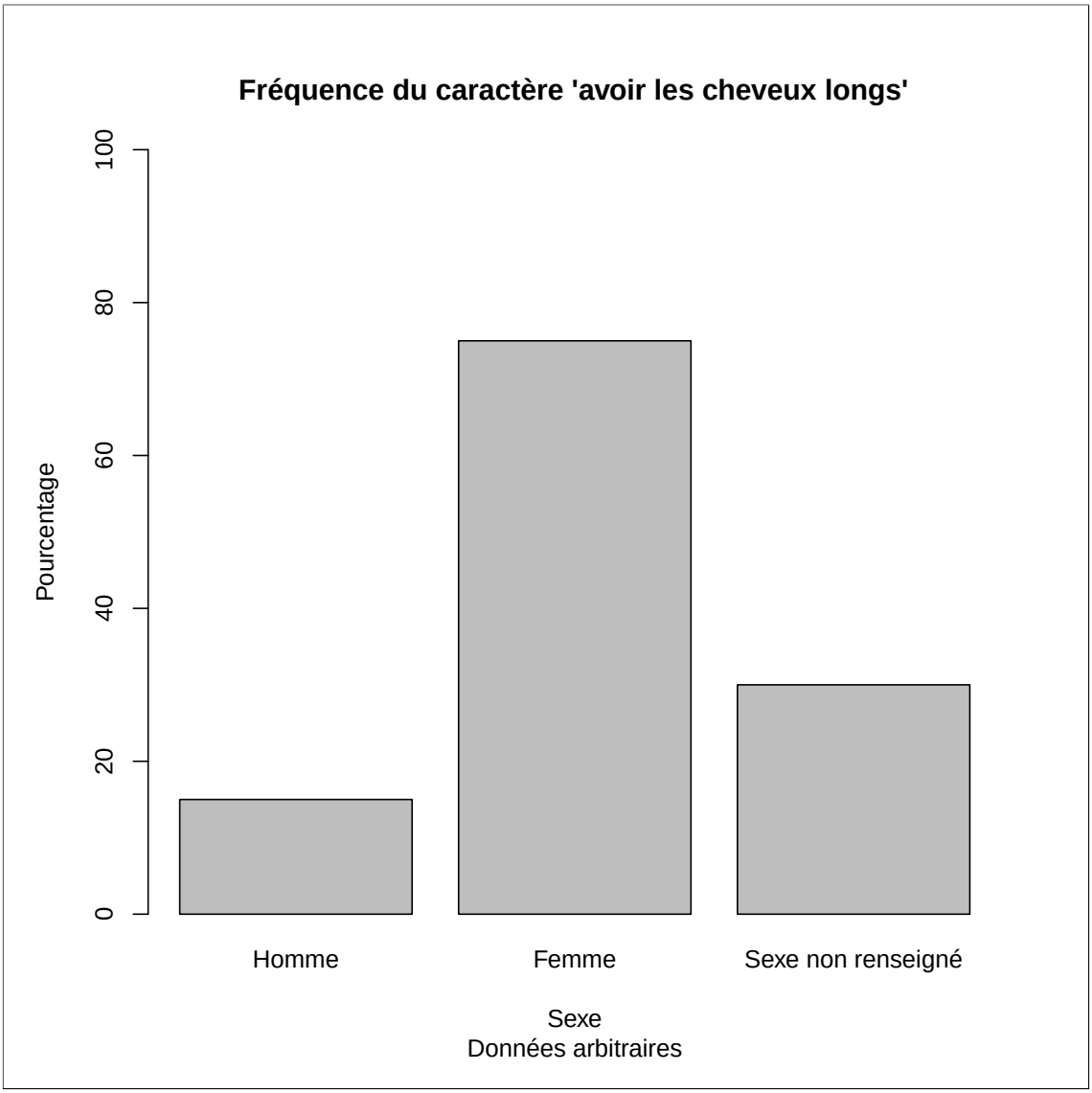
13. Ajoutez les options nécessaires pour vérifier les contraintes suivantes :
    - L'axe des `x` porte le label `Sexe`
    - L'axe des `y` porte le label `Pourcentage`
    - Les barres portent respectivement les noms `Homme`, `Femme`, `Sexe non renseigné`

– L'axe des y varie de 0 à 100
– Ajoutez un titre secondaire `Données arbitraires`



Solution:

```
barplot(
  height=longueur.cheveux,
  main = "Fréquence du caractère 'avoir les cheveux longs'",
  ylab='Pourcentage',
  xlab='Sexe',
  names.arg=c('Homme', 'Femme', 'Sexe non renseigné'),
  ylim=c(0,100),
  sub="Données arbitraires"
)
```



14. Maintenant que notre graphique est terminé, nous allons l'exporter. Nous pourrons ainsi le réutiliser ailleurs, dans une étude par exemple.
    Dans l'onglet `Plots` du panneau sud-est de Rstudio cliquez sur `Export`. Rstudio propose d'exporter en tant qu'image ou en tant que fichier PDF. Il est conseiller d'exporter au format PDF, car dans ce cas, Rstudio crée une image *vectorielle* capable de s'adapter proprement à toute taille de support. Ce n'est pas le cas de l'image *bitmap* qui a une résolution fixe, et qui va se retrouver compressée ou étirée si on modifie sa taille par la suite. Si, pour une raison ou une autre, vous deviez exporter en tant qu'image, le format conseillé est le PNG.
    Exporter l'image en PDF et enregistrez là dans votre dossier `AnCat`. Allez ensuite double-cliquer dessus pour vérifier que l'exportation à bien fonctionnée.

Exercice 1.3 (Introduction aux types de données)**.**
Nous avons vu que dans R nous manipulons des données, que nous stockons dans des objets. Chaque donnée à un type, et le comportement des fonctions que nous allons utiliser dessus peut changer suivant le type de données que l'on manipule. Il est donc très important de toujours avoir en tête quel type de donné nous avons stocké dans chaque objet. Pour regarder cela de plus prêt nous allons charger la base de données `iris`, livrée par défaut avec R.

1. Pour charger une base données connue de R nous utilisons la fonction `data(nomBDD)`. La base de données est alors stockée dans un objet portant le même nom que la base de données.
   Chargez le jeu de données `iris`.

Solution:

```
data(iris)
```

2. La fonction `class` permet de connaître le type d'un objet. Quel est le type de l'objet `iris` ?

Solution:

```
class(iris)
## [1] "data.frame"
```

3. Un `data.frame` permet de stocker des variables. La fonction `names()` permet de récupérer le nom des variables stockées dans un objet `data.frame`.
   Quelles sont les variables stockées dans la base `iris` ?

Solution:

```
names(iris)
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length"
## [4] "Petal.Width"  "Species"
```

4. Pour extraire une variable d'un jeu de données nous utilisons l'opérateur `$` qui s'utilise `nomBDD$nomVariable`.
   Extraire la variable `Species` de `iris` et stockez là dans un objet nommée `reponse`.
   Attention, il faut respecter la casse (les majuscules et minuscules).

Solution:

```
reponse <- iris$Species
```

5. Afficher dans la console R l'objet `reponse`. Quel est le type de cet objet ?

Solution:

```
class(reponse)
## [1] "factor"
```

6. Extraire maintenant la variable `Sepal.Length` et stockez la dans un objet nommée `descripteur1`. Quel est le type de cet objet ?

Solution:

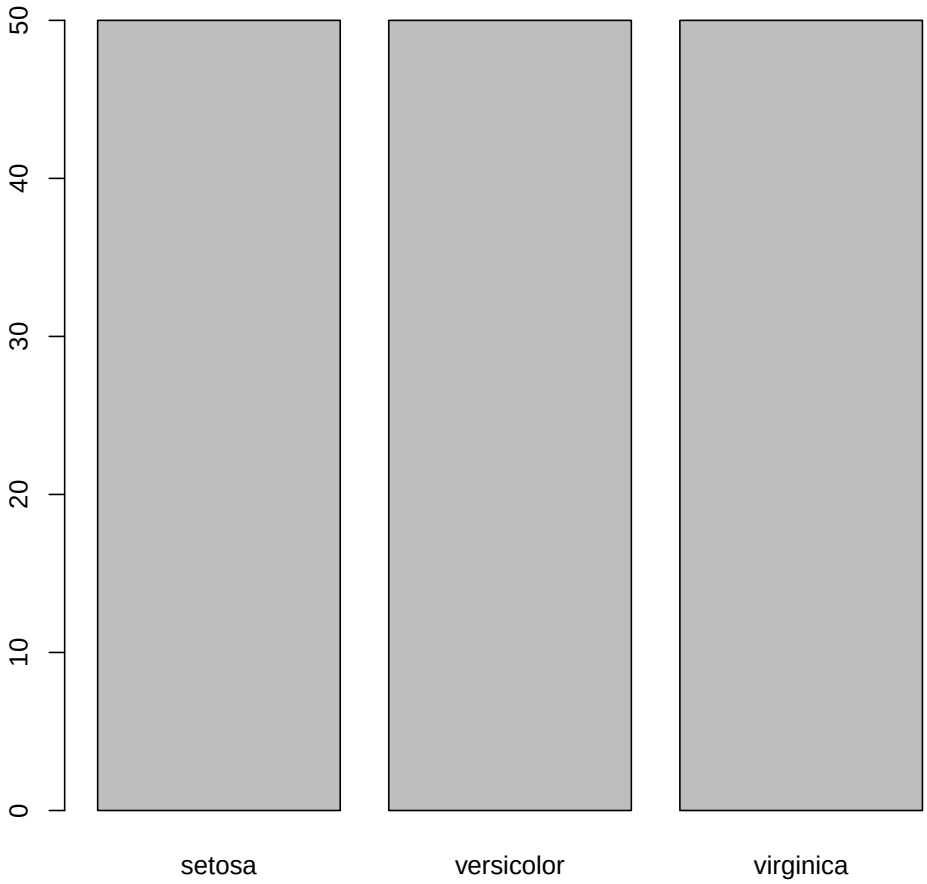
```r
descripteur1 <- iris$Sepal.Length

class(descripteur1)
## [1] "numeric"
```

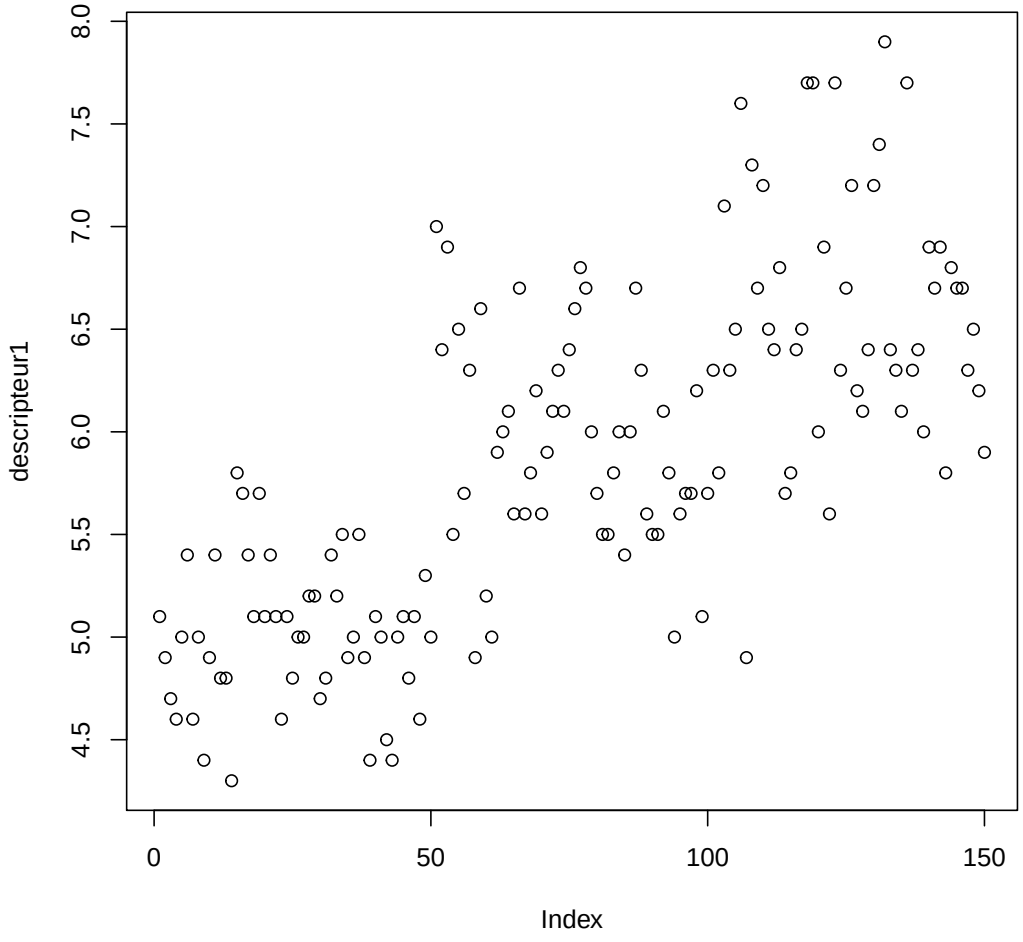
7. Effectuez maintenant un `plot` sur l'objet `reponse`, puis sur l'objet `descripteur1`. Que constatez-vous ?

Solution:

```r
plot(reponse)
```



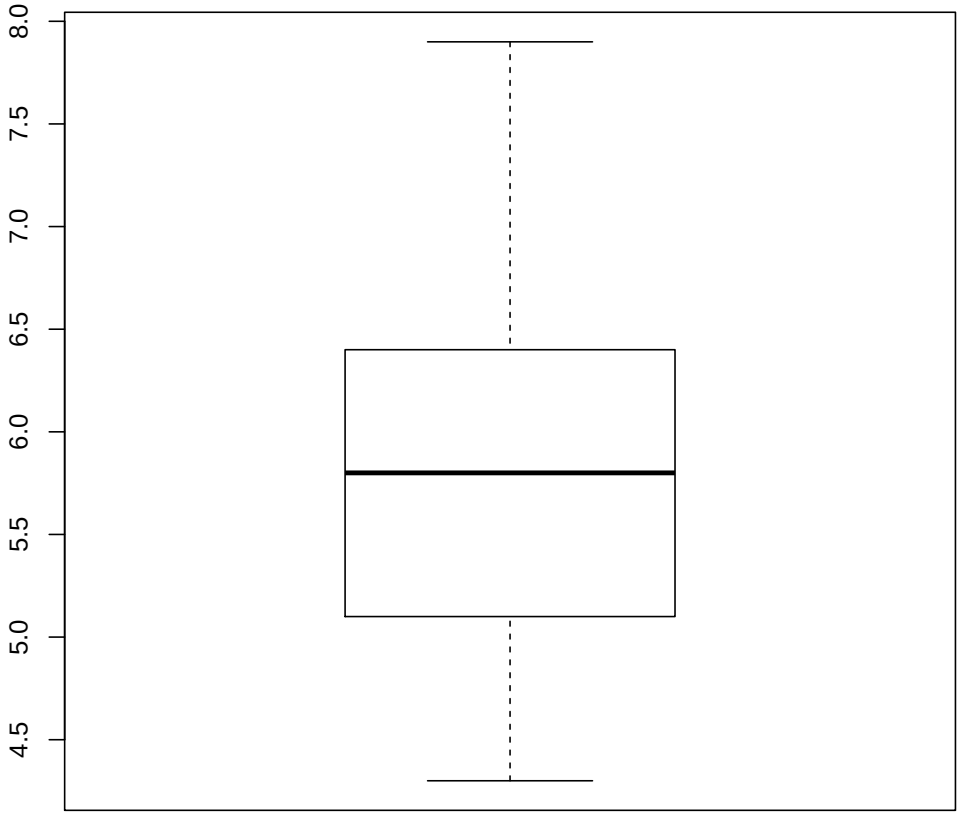

```r
plot(descripteur1)
```



8. Le graphique choisi par la fonction `plot` pour les objets numériques n'est pas très pratique pour résumer de manière intelligible les données. Tracez plutôt un `boxplot`.

Solution:

```
boxplot(descripteur1)
```



Conclusion : nous constatons que R offre un cadre spécifique pour stocker les variables catégorielles, nommé `factor`, et un cadre spécifique pour les variables numériques, nommé `numeric`. Nous avons pu avoir une vue synthétique des données grâce aux fonctions graphiques.

9. Pour un grand nombre de types d'objet de R, nous pouvons demander une vue synthétique des données à l'aide de la fonction `summary`
   Utilisez la fonction `summary` sur les objets `reponse` et `descripteur1` et comparer les résultats.

Solution:

```
summary(reponse)
```

```
##     setosa versicolor  virginica
##         50         50         50
```

```
summary(descripteur1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.30    5.10    5.80    5.84    6.40    7.90
```

Remarque : durant tout ce second exercice, j'espère pour vous que vous n'avez pas écris le nom de chaque fonction et objet à la main, mais avez profité du système de completion !

Exercice 1.4 (Prise en main de la librairie Dataset)**.**

Le cadre de gestion de variables offert par défaut par R ne permet pas de gérer efficacement des données d'enquêtes. En effet, pour des données d'enquêtes, connaître le nom des variables et des valeurs prises par les variables ne suffit pas. Si une variable porte le nom `santé`, de quelle santé parlons nous ? De la santé physique, de la santé mentale, est-ce une mesure de santée auto-rapportée ou évaluée par un médecin, est-ce que l'on parle de la satisfaction que l'individu a avec sa santé ? Un jeu de données d'enquête doit précisemment décrire le sens de chaque variable contenue dans la base de données et ne doit pas laisser d'ambiguïté quand au sens socio-économique de la variable. D'autres critères importants, tel que la gestion des valeurs manquantes et la pondération des individus entre en jeux, nous verrons cela plus en détail par la suite.
R ne fournit pas de cadre *ad hoc* pour gérer les données issues d'enquêtes, mais possède un système de librairies, permettant d'ajouter des modules spécifiques pour telles ou telles tâches.
Pour la gestion des données d'enquêtes nous allons utiliser la librairie `Dataset`. Cette librairie est déjà installée sur vos postes étudiants. Nous verrons au prochain TP comment installer et mettre à jour une librairie R, afin que vous puissiez aussi travailler sur votre ordinateur personnel.

1. Pour charger une librairie R, nous utilisons la fonction `library(nomDeLaLibrairie)`. Chargez la librairie `Dataset`.

Solution:

```
library(Dataset)
```

2. La librairie `Dataset` fournit une base de données exemple nommée `dds` (pour *Dataset Data Sample*). Charger la base de données à l'aide de la fonction `data`.

Solution:

```
data(dds)
```

3. Quel est le type de l'objet `dds` ?

Solution:

```
class(dds)
```

```
## [1] "Dataset"
## attr(,"package")
## [1] "Dataset"
```

4. Un jeu de données contient généralement beaucoup de variables et il est difficile de s'y repérer. Lorsque nous voulons répondre à une problématique, nous avons besoin de sélectionner les variables pertinentes pour notre analyse. La fonction `exportPDF()` permet d'avoir une vue synthétique de l'ensemble des variables contenues dans notre

base de données.
Exécuter la fonction `exportPDF` sur la base `dds`.

Solution:

```
exportPDF(dds)
```

5. Consultez le fichier PDF généré.
6. Utilisez l'option `description.chlenght` avec pour valeur 80 pour réduire la largeur du tableau produit dans le fichier PDF.

Solution:

```
exportPDF(dds, description.chlength = 80)
```

7. Nous allons maintenant regarder comment est stocké une variable dans un objet `Dataset`.
   À l'aide de l'opérateur `$`, extraire la variable `sexe` et la stocker dans un objet nommée `desc.sexe`.

Solution:

```
desc.sexe <- dds$sexe
```

8. Quelle est le type de cette variable ?

Solution:

```
class(desc.sexe)
```

```
## [1] "BinaryVariable"
## attr(,"package")
## [1] "Dataset"
```

9. Le type `BinaryVariable` est une spécialisation du type `Variable`. Les principales commandes permettant de manipuler des objets de type `Variable` sont : `description`, `codes`, `values`, `valids` et `missings`.
   Exécutez ces commandes sur l'objet `desc.sexe` et analysez les résultats.

Solution:

```
description(desc.sexe)
```

```
##                 sexe
## "Sexe du répondant"
```

```
codes(desc.sexe)
```

```
##  [1] 0 0 0 1 1 1 1 0 1 0 1 1 1 0 1
```

```r
values(desc.sexe)
```

```
## N'a pas répondu           Homme           Femme
##              -2               0               1
```

```r
valids(desc.sexe)
```

```
## Homme Femme
##     0     1
```

```r
missings(desc.sexe)
```

```
## N'a pas répondu
##              -2
```

Exercice 1.5. [Sauvegarde]

1. Même votre fichier enregistré, vous pouvez perdre vos données (*crash* du disque dur, vol de la machine, etc.). Il convient donc d'avoir toujours (au moins) 2 exemplaires de ses fichiers à des emplacements géographiquement éloignés. Un réflexe intéressant à acquérir est le suivant : si un de vos fichiers n'est sauvegardé suivant cette contrainte, sauvegardez-le, ou supprimez-le dès maintenant.
   Lorsque le fichier n'est pas trop volumineux, une solution pratique est de s'envoyer le fichier par courriel, il est ainsi stocké dans votre messagerie. Envoyez-vous le fichier `tp1.R` par courriel.

À préparer pour la prochaine séance :

– Finir le TP1. Pour vous aider à terminer le TP, un corrigé sera rendu disponible sur `mephisto.unige.ch`.

– Tester les fonctions `ncol`, `nrow`, et `alldescriptions` sur l'objet `dds`. N'hésitez pas aussi à consulter leurs pages de manuel.

Travaux pratiques – Séance 2

Objectifs de la séance : mise en place d'une problématique sociologique Recherche d'un jeu de données pour valider les hypothèses. Chargement d'une base SPSS. Configuration de la pondération. Exclusion des individus non concernés. Opérationnalisation de la problématique. Vérification des cas valides. Vérification des mesures utilisées. Exporter un jeu de données. Présentation succincte des données du panel.

Exercice 2.6 (Mise en place d'une problématique)**.**

Une problématique est la formulation d'un problème. En sociologie, une problématique va généralement chercher à mettre en compétition des comportements sociaux face à des déterminant sociaux. De manière formelle, lors de l'opérationnalisation d'une problématique, nous nommons les déterminants sociaux « variables explicatives », et les comportements sociaux « variables dépendantes ». De ce point de vue, une problématique met en relation une ou plusieurs variables explicatives, avec une ou plusieurs variables dépendantes.
La problématique se construit autours d'hypothèses. C'est lorsque vous avez clairement posé les hypothèses que vous souhaitez tester que vous pouvez formuler la problématique. Lorsque vous formulez vos hypothèses, veillez à ne pas choisir des déterminants sociaux qui véhiculeraient le même sens que le ou les comportements sociaux étudiés. Par exemple, si vous vous intéressez à l'utilisation ou non d'Internet à la maison, choisir en déterminant social le fait de posséder un ordinateur risque d'être trop proche du comportement ciblé par notre analyse.
Il en résulterait des difficultés à tester nos hypothèses, l'influence des autres determinanx sociaux sélectionnés étant atténuée, masquée par la forte influence du premier. Pour terminer, la problématique doit englober dans sa formulation l'ensemble des hypothèses que vous souhaitez tester, mais ne doit pas faire référence à des hypothèses que vous ne testerez pas.

Durant ces séances de TP, la problématique que nous choisissons est la suivante :

Influence de la formation secondaire sur la réussite au Bachelor chez les étudiants SES de moins de 25 ans

Exercice 2.7 (Chargement d'une base SPSS dans R)**.**

Nous allons utiliser ici les données de résultats aux examens de la Faculté SES de 1998.

1. Chargez ces données dans R au format `Dataset`.
   (a) Les données sont enregistrées au format SPSS dans le fichier `ses_98_500.sav` accessible sur mephisto.unige.ch. Téléchargez les données et enregistrez-les dans votre dossier `AnCat`.
   (b) Lancez Rstudio et chargez la librarie `Dataset`.

Solution:

```
library(Dataset)
```

(c) Nous allons donc charger dans R un fichier présent sur le disque dur. Lorsque vous appelez à partir de R un fichier sans préciser son emplacement exact, R regarde si le fichier se trouve dans ce qu'il appelle le « répertoire de travail » (ou *working directory* en anglais). La commande `getwd`, qui s'utilise `getwd()`, permet de connaître le répertoire de travail actuellement configuré.
Quel est votre répertoire de travail actuel ?

Solution:

```
getwd()
```

(d) Vous pouvez à tout moment modifier ce répertoire de travail à l'aide de la commande `setwd`, qui s'utilise `setwd("cheminVersLeDossier")`.
Définissez comme répertoire de travail le dossier `AnCat`. Pensez à utiliser la completion pour retrouver le chemin vers le dossier !

Solution:

```
setwd("H:/AnCat")
```

(e) La fonction `list.files` permet de lister l'ensemble des fichiers se trouvant dans un dossier donné. Par défaut, elle regarde dans le répertoire de travail.
Listez les fichiers contenus dans le dossier `AnCat` et vérifiez que le jeu de données `ses_98_500.sav` s'y trouve.

(f) La librairie `Dataset` permet de charger un jeu de données d'enquête stockée au format SPSS à l'aide de la fonction `get.spss.file`
Consulter son manuel.

(g) Chargez les données en spécifiant un nom et une description et les stocker dans un objet nommé `ses98`. Le nom pourrait être `ses-98-500` et la description `Résultats aux examens de 1998 pour 500 étudiants de la Faculté SES de l'Université de Genève`.

Solution:

```
ses98 <- get.spss.file(
  file = 'ses_98_500.sav',
  name = 'ses-98-500',
  description = "Résultats aux examens de 1998 pour
    500 étudiants de la Faculté SES de l'Université
    de Genève"
)
```

2. Générez une vue synthétique de la base de données au format PDF. Consultez-la.

Solution:

```
exportPDF(ses98)
```

3. Nous remarquons qu'une variable de poids est présente dans le jeu de données. Cette variable est très importante. En effet, lorsqu'une enquête est réalisée, l'Institut produisant l'enquête va constuire un échantillon d'individus représentant au mieux possible la population étudiée, en vérifiant la représentativité sur un certain nombre de critères démographiques et socio-économiques. Cependant, certains individus ne vont pas répondre au questionnaire et nous perdons alors la représentativité de notre enquête. Il est difficilement acceptable de considérer que ces non-réponses sont distribuées de manière aléatoire dans la population. Les personnes ne participant pas à une enquête ont généralement des profils plus atypiques, ou peuvent rencontrer des difficultées dans leur vie au moment de l'enquête (de santé, au travail, famillliales, etc.). Si nous nous intéressons à la santé par exemple, cette non-réponse pose un problème, car nous risquons de nous retrouver dans notre échantillon avec des individus étant en meilleure santé que la population générale. L'Institut d'enquête calcul alors des poids affectés à chaque individus permettant de compenser ces non-réponses et ramener l'échantillon au plus proche de la population étudiée.
   Dans la librairie `Dataset`, nous spécifions la variable de pondération à utiliser avec la fonction `weighting` qui s'utilise :
   - `weighting(BDD) <- 'nomVariableDePonderation'` pour définir la variable de pondération à utiliser.
   - `weighting(BDD)` pour récupérer le nom de la variable de pondération actuellement définie.

   (a) Une variable de pondération doit vérifier certains critères (pas de valeurs manquantes, pas de poids négatifs). Ces critères sont vérifiés par le type de variable `WeightingVariable`. Utilisez la fonction `wvar` pour convertir la variable `weights` de `ScaleVariable` à `WeightingVariable`.

Solution:

```
ses98$weights <- wvar(ses98$weights)
```

   (b) Définir la variable `weights` comme variable de pondération sur le jeu `ses98`.

Solution:

```
weighting(ses98) <- "weights"
```

   (c) Afficher la variable de pondération utilisée.

Solution:

```
weighting(ses98)
## [1] "weights"
```

   (d) La fonction `weights` permet de récupérer les poids pour chaque individus. Affichez les poids utilisés.

Solution:

```
weights(ses98)
## Description: ponderation des individus
```

   (e) La fonction `nrow` indique le nombre de lignes contenus dans la base. La fonction `nindividual` indique ce même nombre mais en tenant compte de la pondération,

c'est le nombre d'individus de l'échantillon dont la base rend compte. Comparer les résultats de ces deux fonctions sur la base `ses98`.

Solution:

```
nrow(ses98)
## [1] 500
```

```
nindividual(ses98)
## [1] 500
```

(f) Générez de nouveau la vue de synthèse au format PDF.

Solution:

```
exportPDF(ses98)
```

Exercice 2.8 (Exclusion des individus non concernés)**.**
Lorsque votre problématique est posée, vous devez définir clairement la population concernée par votre étude : travaillez-vous sur la population générale (d'un pays, d'une région particulière) ou sur un sous-groupe particulier de la population ciblé par votre problématique ? Par exemple, si vous vous intéressez à la participation aux votations politiques, les mineurs ne sont pas concernés. Et les étrangers le sont-ils ? Si vous vous intéressez à la satisfaction au travail, les personnes ne travaillant pas (enfants, chômeurs, retraités, etc.) ne sont pas concernées.
Ensuite, lorsque vous récupérez un jeu de données, la première question à se poser est « quelle population contient-elle ? ». Définissez alors les éventuels critères d'exclusions d'individus de la base, permettant de ramener celle-ci à la population concernée par votre problématique.

1. Dans notre problématique nous nous intéressons aux étudiants SES de moins de 25 ans. Le jeu de données contient des étudiants SES. Nous devons sélectionner parmis eux ceux qui ont moins de 25 ans. À l'aide de la fonction `subset`, sélectionner ces individus.

   (a) Il est préférable de garder une sauvegarde de la version originale de la BDD. Copiez alors la BDD dans un objet nommé `ses98.original`.

Solution:

```
ses98.original <- ses98
```

   (b) Généralement les bases de données contiennent l'année de naissance des individus. À l'aide de la fonction `contains`, chercher les variables dont la description contient le mot `naissance`.

Solution:

```
naiss <- contains("naissance", ses98)
##                 Description
## moisnais  mois de naissance
## annenais année de naissance
```

(c) Affichez la variable `annenais` pour savoir comment sont codées les années de naissances.

Solution:

```
ses98$annenais
## Description: année de naissance
```

(d) Nous remarquons que les années de naissance sont stockées sur deux chiffres. Nous souhaitons conserver uniquement les individus ayant moins de 25 ans en 98, c'est-à-dire les individus dont l'année de naissance est supérieur à 73 ($98 - 25$).
Utilisez la fonction `subset` pour récupérer les individus concernés, et stockez la nouvelle base dans un objet nommé `ses98.inf25`

Solution:

```
ses98.inf25 <- subset(ses98, annenais >= 73)
```

(e) Combien d'individus avez-vous retirés ?

Solution:

```
nindividual(ses98) - nindividual(ses98.inf25)
## [1] 42
```

2. Générez une vue synthétique de la base de données `ses98.inf25` au format PDF.

Solution:

```
exportPDF(ses98.inf25)
```

Exercice 2.9 (Opérationnalisation de la problématique sur une base de données)**.**
Opérationnaliser la problématique consiste à sélectionner sur une base de données consiste à selectionner les variables dépendantes (dans ce cours nous vous demandons de vous limiter à une) que nous cherchons à expliquer, et les variables indépendantes, aussi appelés co-facteurs, avec lesquels nous allons tenter d'expliquer notre variable dépendante. Il est important dans cette étape de prendre garde au pourcentage de valeurs manquantes que contiennent les variables, car une variable trop vide ne permettra pas de réaliser des analyses.
Vous pouvez consulter la vue synthétique en PDF pour choisir vos variables, qui indique notamment le pourcentage de valeurs manquantes. Pour cette exercice nous retenons les variables suivantes : `bil_99`, `sexe`, `annenais`, `troncom`, `datimmat`, `dompere`, `dommere`, `profpere`, `nationa2`, `lieudi2`, `diplse2`,

1. Consultez la vue synthétique PDF et vérifiez que chacune de ces variables contienne suffisamment de cas valides.
2. Nous allons constituer notre base de données d'étude. Pour cela, nous commençons par conserver uniquement les variables participant à l'analyse.
   (a) Stockez dans un objet `var.etude` un vecteur contenant le nom des variables participant à l'étude.

Solution:

```
var.etude <- c(
  'bil_99',
  'sexe',
  'annenais',
  'troncom',
  'datimmat',
  'dompere',
  'dommere',
  'profpere',
  'nationa2',
  'lieudi2',
  'diplse2'
)
```

(b) L'opérateur de *subscript* permet de sélectionner des individus et des variables d'une base de données. Il s'utilise :
- `nomBDD[i,j]` : pour sélectionner les individus i et variables j. Les indices peuvent être numériques, indiquant alors les numéros de lignes, respectivement de colonnes à sélectionner, ou des chaînes de caractères, indiquant dans ce cas les noms de lignes, respectivement de colones à sélectionner.
- `nomBDD[,j]` : raccourci pour sélectionner les variables j sur toutes les lignes.
- `nomBDD[i,]` : raccourci pour sélectionner les individus i sur toutes les variables.

Sélectionner les variables d'études de la base `ses98.inf25` en prenant toutes les lignes. Enregistrez la nouvelle base dans un objet nommé `ses98.select`

Solution:

```
ses98.select <- ses98.inf25[, var.etude]
```

(c) À l'aide de la fonction `alldescriptions`, afficher les variables contenues dans la bases `ses98.select`.

Solution:

```
alldescriptions(ses98.select)
##                         Description
## weights  ponderation des individus
## troncom               tronc commun
## dompere            domicile du pere
## dommere         domicile de la mère
## profpere               prof.du père
## sexe                           sexe
## annenais         année de naissance
## datimmat    date d'immatriculation
## nationa2       nationalité regoup.
## lieudi2     lieu dipl.sec.regroup.
## diplse2      dipl. second.regroup.
## bil_99                 bilan oct.99
```

Vous pouvez remarquer que la variable `weights` est toujours présente dans la

BDD, alors que nous ne l'avons pas sélectionnée. En effet, comme nous l'avons précédemment définie comme variable de pondération, elle sera automatiquement conservée dans les opérations que nous effectuerons sur la base.

Exercice 2.10 (Vérification des cas valides)**.**

1. Nous devons maintenant vérifier l'affection en tant que « cas valide » ou « valeur manquante » des valeurs de chaque variable.
   Commençons par la variable `dommere`.
   (a) Affichez les cas valides de la variable avec la fonction `valids` et ses valeurs manquantes avec la fonction `missings`.

Solution:

```
valids(ses98.select$dommere)
##          nr          GE     CH rom.   CH al.+TI
##           0           1           2           3
##      France Europe occ.  Europe or.     Afrique
##           4           5           6           7
##   Amér.nord    Amér.sud        Asie
##           8           9          10
```

```
missings(ses98.select$dommere)
## named numeric(0)
```

   (b) La catégorie `nr` correspond à « non renseigné », nous pouvons choisir de la considérer comme une valeur manquante. À l'aide de la fonction `as.missing` changez ce cas valide en valeur manquante.

Solution:

```
ses98.select$dommere <- as.missing(
  'nr',
  ses98.select$dommere
)
## number of missings: 12 ( 2.62 %)
```

2. La fonction `allvalues` permet d'afficher l'ensemble des valeurs pour toutes les variables d'une BDD.
   Vérifiez l'ensemble des variables d'étude, et déterminez celles devant être corrigées.

Solution:

```
allvalues(ses98.select)
```

⇒ Variables à corriger : `dompere` et `profpere`.

3. Corrigez ces variables.

Solution:

```
ses98.select$dompere <- as.missing(
  'nr',
  ses98.select$dompere
)

## number of missings: 23 ( 5.02 %)
```

```
ses98.select$profpere <- as.missing(
  'nr',
  ses98.select$profpere
)

## number of missings: 40 ( 8.73 %)
```

4. Générez de nouveau la vue synthétique en PDF et vérifiez que nos variables d'étude contiennent toujours suffisamment de cas valides.
5. Dans les enquêtes, les non-réponses classiques sont les suivantes : l'individu n'a pas pu être joigné, l'individu n'a pas répondu, l'individu à répondu mais réponse illisible, la question est innaplicable à l'individu (exemple, à la question « Avez-vous voté aux dernières élections fédérales », les mineurs ne sont pas concernés). Pour certaines problématiques, nous pouvons considérer que certaines non-réponses ont un sens socio-économique, et peuvent avoir une valeur prédictive. Le fait par exemple de ne pas avoir répondu à la question peut être le symptôme d'un certain état chez l'individu.
   La fonction `as.valid` permet de permuter une valeur manquante en valeur valide. Nous considérons que le fait pour l'étudiant de ne pas avoir souhaité renseigner la profession de son père à une valeur prédictive. Changez alors sur la variable `profpere` cette valeur manquante en cas valide.

Solution:

```
ses98.select$profpere <- as.valid(
  'nr',
  ses98.select$profpere
)

## number of missings: 0 ( 0 %)
```

Exercice 2.11 (Vérification des mesures de chaque variable)**.**

À chaque variable est associée une mesure qui définit comment les valeurs doivent être interprétées. La première distinction se fait entre les variables *quantitatives*, représentant des données numériques, et les variables *catégorielles*, représentant des données qualitatives. Parmis ces deux types, des distinctions plus fines peuvent être opérées. Nous allons par exemple travailler différemment sur des variables purement numériques (âge, revenu, etc.) et sur des variables représentant des dates, des durées, etc. Du côté des variables catégorielles, nous distinguerons les variables binaires, les variables nominales (plus de 2 catégories, qui ne peuvent pas être ordonnées), et les variables ordinales (plus de 2 catégories, qui peuvent être ordonnées). En exemple de variable nominale nous pouvons citer le statut marital (célibataire, marié, divorcé, veuf), et

en variable ordinale l'évaluation de la santé (mauvaise < satisfaisante < bonne). Les méthodes d'analyses de données traiteront souvent de manière différente ces types de variables, il est donc important de bien les spécifier.
Plusieurs constructeurs sont à disposition pour créer/changer la mesure d'une variable :
- `bvar` : construit une variable binaire.
- `nvar` : construit une variable nominale.
- `cvar` : construit une variable catégorielle. Ce constructeur choisira lui-même entre un type binaire ou nominal.
- `ovar` : construit une variable ordinale.
- `svar` : construit une variable échelle.
- `tvar` : construit une variable temporelle.
- `wvar` : construit une variable de pondération.

Nous allons vérifier les mesures actuellement définies et éventuellement les corriger.

1. Générez la vue synthétique en PDF de la base `ses98.select` et regardez les mesures actuellement définies pour chaque variable, et repérez les variables pour lesquelles les mesures vous semblent inadaptées.

Solution:

```
exportPDF(ses98.select)
```

2. Nous décidons finalement de passer les variables `bil_99` et `nationa2` en ordinales.
   (a) Passez la variable `bil_99` en ordinale avec la méthode `ovar`.

Solution:

```
ses98.select$bil_99 <- ovar(ses98.select$bil_99)
## number of missings: 0 ( 0 %)
```

   (b) Affichez la variable pour vérifier l'ordre des niveaux.

Solution:

```
ses98.select$bil_99[1:10]
## Description: bilan oct.99
##  [1] echec  réussi réussi réussi echec  réussi
##  [7] réussi réussi réussi echec
## Levels: echec < redouble < réussi
```

Remarque : pour éviter de prendre de la place je n'affiche que les 10 premiers individus, je ferai de même pour les questions suivantes.

   (c) Répétez l'opération avec la variable `dompere`.

Solution:

```
ses98.select$dompere <- ovar(ses98.select$dompere)
## number of missings: 23 ( 5.02 %)
```

```
ses98.select$dompere[1:10]
## Description: domicile du pere
##  [1] GE      GE      Afrique GE      Afrique
##  [6] GE      CH rom. CH rom. GE      Asie
## 10 Levels: GE < CH rom. < ... < Asie
```

3. Dans ces exemples les niveaux sont directement dans le bon ordre. Il peut arriver que nous ayons à inverser ou permuter l'ordre des niveaux. Ces deux tâches se réalisent à l'aide des fonctions `valids.reverse` et `valids.permut`.
   (a) Sans enregistrer le résultat, inversez les niveaux de la variable `bil_99`.

Solution:

```
valids.reverse(ses98.select$bil_99)[1:10]
## Description: bilan oct.99
##  [1] echec  réussi réussi réussi echec  réussi
##  [7] réussi réussi réussi echec
## Levels: réussi < redouble < echec
```

   (b) Sans enregistrer le résultat, permutez les niveaux `réussi` et `echec` de la variable `bil_99`.

Solution:

```
valids.permut(ses98.select$bil_99, "redouble", "echec")[1:10]
## Description: bilan oct.99
##  [1] echec  réussi réussi réussi echec  réussi
##  [7] réussi réussi réussi echec
## Levels: redouble < echec < réussi
```

Exercice 2.12 (Exportation d'une base de donnée)**.**

Nous avons terminé de nettoyer notre base de données. Nous allons l'exporter pour en avoir une sauvegarde, avant de passer aux recodages au prochain TP.

1. La fonction `export` permet d'exporter votre BDD en un fichier `.RData` facilement chargeable dans R. C'est l'analogue du format `.sav` chez SPSS.
   Exporter le jeu de données `ses98.select` et lui donnant le nom `ses98.clean`

Solution:

```
export(
  ses98.select,
  name = 'ses98.clean'
)
```

2. Consultez la vue synthétique au PDF pour vérifier que la base de donnée est correcte

3. Dans votre dossier `AnCat` créer un dossier `problématique-tp`. Dans ce dossier créer un nouveau dossier `bdd-nettoyée` et copiez-y votre base de données en `.RData` et sa vue synthétique en PDF.

Exercice 2.13 (Sauvegarde)**.**

1. Faîtes une archive Zip de votre dossier `Ancat` et nommé-là `AnCat-YYYY.MM.DD`, où YYYY représente l'année, MM le mois et DD le jour.
2. Envoyez-vous cette archive par courriel pour conserver une sauvegarde dans votre boîte de courrier.

À préparer pour la prochaine séance :

1. Discutez entre vous pour formez des groupes de 2 ou 3 par centres d'intérêts socio-économiques.
2. Par groupe, formulez la problématique pour votre étude de cas.
3. Formulez ensuite quelques hypothèses que vous voudrez tester pour répondre à votre problématique (au moins 3 hypothèses, maximum 6).
4. En prochaine étape, vous devriez chercher un jeu de données adapté pour répondre à votre problématique. Cependant cette étape est longue et n'est pas l'objet de ce cours. Nous vous fournissons alors directement les données : nous utiliserons les données du *Panel Suisse des Ménages (PSM), vague 2006*. Vous pouvez récupérer sur mephisto.unige.ch les vues synthétiques des données.
   Consultez les vues synthétiques et vérifiez que vos hypothèses peuvent être testées sur ces données. Si ce n'est pas le cas, adaptez vos hypothèses (et éventuellement votre problématique) pour les rendre testables sur ces données.
5. Un membre du groupe envoie par courriel à Danilo :
   - Les noms et prénoms des membres du groupe.
   - La formulation de votre problématique.
   - Les hypothèses que vous voulez tester.
   - Le groupe d'individus concerné par votre problématique. Quels sont les individus que vous devrez retirer de votre jeu d'étude ?
   - La liste des variables du PSM que vous pensez utiliser pour tester ces hypothèses.

Travaux pratiques – Séance 3

Objectifs de la séance : Consulter les fréquences d'apparition des modalités d'une variable catégorielle. Consulter en terme de fréquences une approximation de la densité d'une variable échelle. Effectuer un recodage par opération arithmétique, découpage d'une variable continue, et fusion de modalités d'une variable catégorielle. Valider les recodages en consultant les fréquences. Renommer des variables et des valeurs de variable. Présenter un recodage. Apparier plusieurs bases de données.

Exercice 3.14 (Chargement d'une base de donnée préalablement exportée)**.**
Dans cette séance, nous avons besoin de la base de données que vous avez nettoyée dans le précédent TP. Dans ce dernier TP, vous avez terminé par exporter votre base de données avec la fonction `export`. Charger un objet sous R se fait avec la fonction `load`

1. Chargez la librairie `Dataset`.

Solution:

```
library(Dataset)
```

2. Définissez comme répertoire de travail le dossier contenant votre base `ses98.clean`. Si vous avez suivi nos conventions, cela devrait être le dossier `AnCat`.

Solution:

```
setwd("H:/AnCat/")
```

3. Chargez la base de donnée que vous avez préparée au précédent TP.

Solution:

```
load("ses98.clean.RData")
```

4. Vérifiez que la base est bien présente en mémoire.

Solution:

```
ls()
```

```
##  [1] "a"                 "dds"
##  [3] "descripteur1"      "desc.sexe"
##  [5] "iris"              "longueur.cheveux"
##  [7] "naiss"             "reponse"
##  [9] "ses98"             "ses98.clean"
## [11] "ses98.inf25"       "ses98.original"
## [13] "ses98.select"      "var.etude"
```

Exercice 3.15 (Fréquences des valeurs d'une variable catégorielle)**.**

Regarder les fréquences d'apparition des modalités d'une variable catégorielle est intéressant du point de vue exploratoire. Cela nous permet d'obtenir de premières informations sur notre variable : la modalité la plus représentée, la moins représentée, s'il y a des trous, comment sont ventilées les valeurs manquantes, etc. En particulier, cela nous permettra donc de décider d'un éventuel recodage, et de l'orienter.

1. Afficher les fréquences de la variable `dommere` à l'aide de la fonction `frequencies`, qui s'utilise `frequencies(nomVariable, nomBDD)`

Solution:

```
frequencies("dommere", ses98.clean)
```

```
##    Coding Missing       Label   N N total Percent Percent (all) Percent total
## 1       1                  GE 250           56.05         54.59
## 2       2             CH rom.  78           17.49         17.03
## 3       3           CH al.+TI  53           11.88         11.57
## 4       4              France  24            5.38          5.24
## 5       5         Europe occ.  13            2.91          2.84
## 6       6          Europe or.   7            1.57          1.53
## 7       7             Afrique  10            2.24          2.18
## 8       8           Amér.nord   1            0.22          0.22
## 9       9            Amér.sud   5            1.12          1.09
## 10     10                Asie   5     446    1.12          1.09         97.38
## 11      0       x          nr  12      12  100.00          2.62          2.62
## 12                                    458                                100
```

2. Quelles sont pour vous les modalités qui seront plus difficiles à analyser de part leurs représentativités ?

Solution: `Europe or.`, `Amér.sud`, `Asie`, et `Amér.nord`.

3. On pourrait se demander pourquoi on ne donne pas simplement la variable à la fonction `frequencies` au lieu de donner le nom de la variable et la base de données entière. En réalité, il faut bien comprendre que la variable n'existe pas indépendemment des autres, mais en tant que variable d'une enquête globale. Elle partage ainsi de manière intrinsèque un sens socio-économique avec les autres variables de l'enquête. En particulier, elle partage la pondération des individus ayant répondus à l'enquête.
   Pour l'exercice, exécutez la méthode `frequencies` sur la variable elle-même. Regardez le message de *warning* qui s'affiche, et comparer les résultats.

Solution:

```
frequencies(ses98.clean$dommere)
```

```
## Warning: no weights defined! Equi-weighting will be used.
## Prefer use the method for Dataset objects to take weights into account.
```

```
##    Coding Missing       Label   N N total Percent Percent (all) Percent total
## 1       1                  GE 250           56.05         54.59
## 2       2             CH rom.  78           17.49         17.03
## 3       3           CH al.+TI  53           11.88         11.57
## 4       4              France  24            5.38          5.24
## 5       5         Europe occ.  13            2.91          2.84
## 6       6          Europe or.   7            1.57          1.53
## 7       7             Afrique  10            2.24          2.18
## 8       8           Amér.nord   1            0.22          0.22
## 9       9            Amér.sud   5            1.12          1.09
## 10     10                Asie   5     446    1.12          1.09         97.38
## 11      0       x          nr  12      12  100.00          2.62          2.62
## 12                                    458                                100
```

4. À quoi sont dues ces différences d'effectifs ?

Solution: Pondération de non-réponse des individus.
Remarque : si vous n'observez pas de différence d'effectif, c'est que je n'ai pas encore eu le temps de vous fabriquer des poids pour ce jeu de données, vous utilisez donc une pondération de 1 pour chaque individu (pondération par défaut).

Exercice 3.16 (Fréquences pour une variable échelle : approximation de la densité)**.**

De la même manière que pour les variables catégorielles, il nous sera important de consulter les fréquences d'apparition des valeurs d'une variable échelle, afin d'avoir une idée de la répartition, et nous informer si un recodage s'impose et si oui, de nous orienter sur comment le réaliser au mieux. Cependant, pour une variable numérique, nous avons une infinité de valeurs possibles. Sur un échantillon donné, nous avons bien sûr un nombre fini de valeurs mais on peut facilement imaginer, vu qu'elles sont prises dans un ensemble infini, que ces valeurs vont être très souvent distinctes. Prenons comme exemple le revenu : quelqu'un peut gagner 74026.32 CHF par an, quelqu'un d'autre 74026.31 CHF par an, ce sont des valeurs distinctes. Faire une table de contingence sur ce type de variable amènera un résultat assez pauvre : nous aurons à peu près chaque individu pris séparément dans le tableau.
Une autre raison fait que nous devons considérer différemment les variables échelles : c'est que nous pouvons mesurer la distance entre deux valeurs. Il est important de tenir compte de l'espacement entre les valeurs lorsque nous regardons les fréquences. C'est pourquoi nous allons regarder à travers les fréquences la densité de la variable, approximée en découpant la variable en intervalles de même largeur.
Pour une variable échelle, la méthode `frequencies` va découper l'étendue de la variable (c'est-à-dire l'intervalle $[min(x); max(x)]$) en au maximum 10 sous-intervalles de même largeur. Si la variable possède moins de 10 valeurs distinctes, alors elle créera autant d'intervalles que de valeurs distinctes.

1. Afficher les fréquences de la variable `age98`.

Solution:

```
frequencies("annenais", ses98.clean)

## number of missings: 0 ( 0 %)

##   Coding Missing   Label   N N total Percent Percent (all) Percent total
## 1      1         [73,74]  36           7.86          7.86
## 2      2         (74,75]  30           6.55          6.55
## 3      3         (75,76]  56          12.23         12.23
## 4      4         (76,77]  99          21.62         21.62
## 5      5         (77,78] 135          29.48         29.48
## 6      6         (78,79]  84          18.34         18.34
## 7      7         (79,80]  16           3.49          3.49
## 8      8         (80,81]   2     458    0.44          0.44        100.00
## 9                          458                                      100
```

2. Combien d'intervalles avez-vous ?

Solution: 8.

3. Quelle est la longueur de chaque intervalle ?

Solution: Chaque intervalle a une longueur de 1, sauf le 1er qui a une longueur de 2 car il inclu la première valeur.

4. Quel est le mode de la variable ?

Solution: Le mode est la classe la plus dense, c'est-à-dire la plus représentée vis-à-vis de sa largeur. Pour avoir le mode, nous divisons donc l'effectif de chaque classe par sa largeur. Le mode est l'intervalle $(77, 78]$.

5. Au vue de la densité de la variable, quel découpage proposeriez-vous ?

Solution: Dans notre problématique nous sommes intéressé à la réussite à l'examen. En hypothèse nous pouvons poser que le fait d'être plus âgé que la normale amène un taux d'échec plus important, tandis qu'être plus jeune n'amène pas de différence significative. Nous serions alors intéressé par un découpage en 3 catégories : plus jeune, normal, plus âgé. Maintenant comment déterminer ce qui est plus jeune ou plus âgé en pratique ? Nous pouvons nous baser sur l'âge théorique sans redoublement. Où de manière plus pragmatique, nous pouvons regarder la distribution de la variable, repérer la classe la plus représentée (qui serait alors la classe « normale ») et faire un découpage autour. En procédant de cette manière nous pourrions considérer le découpage $[73; 75], [76 : 78], [79; 81]$.

Exercice 3.17 (Renommer des variables et des valeurs)**.**

La fonction `rename` permet de renommer des noms de variable dans un objet `Dataset`, respectivement de renommer des noms de valeur dans un objet `Variable` ou ses héritant (`BinaryVariable`, `NominalVariable`, `OrdinalVariable`). Elle s'utilise comme suit :

```
rename(
  objetDataset,
  'ancienNom1' = 'nouveauNom1',
  'ancienNom1' = 'nouveauNom1',
  ...
)
```

pour renommer des variables d'un objet `Dataset`, et

```
rename(
  objetVariable,
  'ancienNom1' = 'nouveauNom1',
  'ancienNom1' = 'nouveauNom1',
  ...
)
```

pour renommer les valeurs d'un objet `Variable`.

1. Renommer la variable `dommere` en `domicile.mere`.

Solution:

```
ses98.clean <- rename(
  ses98.clean,
  'dommere' = 'domicile.mere'
)
```

```
names(ses98.clean)
```

```
##  [1] "weights"       "troncom"
##  [3] "dompere"       "domicile.mere"
##  [5] "profpere"      "sexe"
##  [7] "annenais"      "datimmat"
##  [9] "nationa2"      "lieudi2"
## [11] "diplse2"       "bil_99"
```

2. Renommer la catégorie `GE` de la variable `domicile.mere` en `Genève`, et la catégorie `CH rom.` en `CH romande`.

Solution:

```
ses98.clean$domicile.mere <- rename(
  ses98.clean$domicile.mere,
  'GE' = 'Genève',
  'CH rom.' = 'CH romande'
)
```

```
valids(ses98.clean$domicile.mere)
```

```
##      Genève  CH romande   CH al.+TI      France Europe occ.  Europe or.
##           1           2           3           4           5           6
##     Afrique   Amér.nord   Amér.sud        Asie
##           7           8           9          10
```

Exercice 3.18 (Transformer une variable échelle par opérations arithmétique)**.**

Il est possible d'appliquer des opérations arithmétiques directement sur l'ensemble des individus d'une variable échelle. Par exemple, si `v1` est une variable échelle, alors l'opération `v1 + 2` ajoute 2 à la valeur prise par chaque individu de la variable.

1. La variable `ses98.clean$annenais` contient l'année de naissance des individus. Nous trouvons intéressant de travailler avec l'âge qu'avait l'individu en 1998. Ajoutez à la base de données `ses98.clean` une variable nommée `age98` contenant pour chaque individu son âge en 1998.

Solution:

```
ses98.clean$age98 <- 98 - ses98.clean$annenais
```

2. Que se passe-t-il si l'opération arithmétique que vous appliquez crée une collision avec une des valeurs réservées aux valeurs manquantes ?

Exercice 3.19 (Transformer une variable échelle en variable catégorielle par découpage)**.**

Selon notre problématique, certaines mesures numériques auront davantage de pertinence dans un regroupement par classe. C'est typiquement le cas de l'âge, où généralement ce n'est pas l'âge exact d'un individu qui nous intéresse mais son appartenance à une certaine tranche d'âge. Le découpage sera orienté par notre problématique : si par exemple nous nous intéressons aux chômage, un découpage « jeune actif », « actif »et « fin de carrière »pourra être pertinent. Si nous nous intéressons à la santé tout au long de la vie, un découpage suivant chaque grande étape de la vie (enfance, adolescence, vie adulte, retraite) pourra être pertinent.
Dans ces exemples, nous avons d'abord proposer un découpage en 3 classes, puis un découpage en 4 classes. Le nombre de classes ne se détermine pas par une règle définie, plusieurs découpages peuvent être pertinents pour une même problématique, suivant les nuances que l'on souhaite étudier. Il est cependant important que le découpage soit cohérent avec notre problématique et les hypothèses que nous souhaitons tester, tout en garantissant des modalités suffisamment représentées pour participer participer à l'analyses. En effet, pour une modalité ayant trop peu de représentants il sera difficile d'obtenir des résultats significatifs.
Ainsi, à chaque découpage, une vérification des fréquences sera effectuée.

1. Le découpage d'une variable echelle s'effectue à l'aide de la fonction `cut`. Consultez son manuel d'utilisation.
2. Créez une variable `age98.3` par découpage de la variable `age` suivant les points de coupures suivants : `<= 19`, `20-23`, et `>= 24`.
   Enregistrez-là dans votre base de données `ses98.clean`.

Solution:

```
ses98.clean$age98.3 <- cut(
  ses98.clean$age98,
  breaks = c(19,23)
)

## number of missings: 0 ( 0 %)
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.

##
##      [17,19] (19,23] (23,25]
##   17       2       0       0
##   18      16       0       0
##   19      84       0       0
##   20       0     135       0
##   21       0      99       0
##   22       0      56       0
##   23       0      30       0
##   24       0       0      20
##   25       0       0      16
```

3. La fonction `cut` affiche le tableau croisée de la répartition des individus de l'ancien codage dans le nouveau. Ceci vous permet de vérifier que le recodage effectué était bien celui que vous attendiez.
4. Vérifiez les fréquences de chacune des modalités de la nouvelle variable. Est-ce acceptable ?

Solution:

```
frequencies("age98.3", data = ses98.clean)
```

```
##   Coding Missing   Label   N N total Percent Percent (all) Percent total
## 1      1         [17,19] 102           22.27         22.27
## 2      2         (19,23] 320           69.87         69.87
## 3      3         (23,25]  36     458    7.86          7.86        100.00
## 4                                458                                100
```

Exercice 3.20 (Fusion des modalités d'une variable catégorielle)**.**

Dans la même logique, une variable catégorielle peut contenir des modalités qui s'intègrent mal dans notre problématique, où des modalités trop peu représentées pour être analysables. Nous pouvons alors envisager de grouper ces modalités.
Une fusion de modalités s'effectue avec la fonction `recode`.

1. Consulter le manuel de la fonction `recode`.
2. Affichez les cas valides de la variable `domicile.mere`.

Solution:

```
valids(ses98.clean$domicile.mere)
```

```
##      Genève  CH romande   CH al.+TI      France Europe occ.  Europe or.
##           1           2           3           4           5           6
##     Afrique   Amér.nord    Amér.sud        Asie
##           7           8           9          10
```

3. Pour notre problématique, nous posons en hypothèse que le domicile de la mère, située soit à Genève, soit en Suisse hors Genève, soit à l'étranger, aura une influence sur la réussite à l'examen. Effectuer le recodage permettant de tester cette hypothèse.

Solution:

```
ses98.clean$domicile.mere.3 <- recode(
  ses98.clean$domicile.mere,
  'GE' = 1,
  'Suisse hors GE' = 2:3,
  'Étranger' = 4:10
)
```

```
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.
```

```
##
##                GE Suisse hors GE Étranger
##   Genève      250              0        0
##   CH romande    0             78        0
##   CH al.+TI     0             53        0
##   France        0              0       24
##   Europe occ.   0              0       13
##   Europe or.    0              0        7
##   Afrique       0              0       10
##   Amér.nord     0              0        1
##   Amér.sud      0              0        5
##   Asie          0              0        5
```

4. Vérifier le bon déroulement du recodage avec le tableau croisé.
5. Vérifiez les fréquences de chacune des nouvelles modalités. Cela vous semble-t-il acceptable ?

Solution:

```
frequencies("domicile.mere.3", data = ses98.clean)
```

```
##   Coding Missing          Label   N N total Percent Percent (all) Percent total
## 1      1                     GE 250           56.05         54.59
## 2      2         Suisse hors GE 131           29.37         28.60
## 3      4               Étranger  65     446   14.57         14.19         97.38
## 4      0       x             nr  12      12  100.00          2.62          2.62
## 5                                  458                                     100
```

Exercice 3.21 (Entraînement aux recodages)**.**

Nom de variable	Description	Codage et étiquettes
`datimmat.3`	date immatriculation	1 avant 97 2 97 3 98
`nationa.3`	nationalité	1 Genève 2 Suisse hors GE 3 étranger
`lieudi.3`	lieu obt. dipl. secondaire	1 Genève 2 Suisse hors GE 3 étranger
`diplse.3`	type de dipl. secondaire	1 classique, latine 2 moderne, scientifique, économique 3 autre
`diplse.3b`	type de dipl. secondaire	1 classique, latine, scientifique 2 moderne, économique 3 autre
`profpere.4`	profession du père en 4 catégories	0 non renseigné 1 ouvrier, employé 2 artisan, commerçant, cadre moyen 3 cadre supérieur, prof. libérale 4 sans profession + chômeur

TABLE 1 – Recodages supplémentaires à effectuer

1. Réalisez l'ensemble des recodages présentés dans le tableau 1.
   (a) Variable `datimmat.3`.

Solution:

```
class(ses98.clean$datimmat)
## [1] "ScaleVariable"
## attr(,"package")
## [1] "Dataset"
```

C'est une variable échelle, on la découpe alors suivant les modalités qui font sens pour les hypothèses à tester.

```
ses98.clean$datimmat.3 <- cut(
  ses98.clean$datimmat,
  breaks = c(96,97,98)
)
## number of missings: 1 ( 0.22 %)
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.
##
##       [92,96] (96,97] (97,98]
##    92       2       0       0
##    93       4       0       0
##    94       4       0       0
##    95       7       0       0
##    96      23       0       0
##    97       0      49       0
##    98       0       0     368
```

```
# Nous vérifions ensuite les fréquences
frequencies("datimmat.3", ses98.clean)
##   Coding Missing              Label   N N total Percent Percent (all)
## 1      1         [92,96]  40            8.75          8.73
## 2      2         (96,97]  49           10.72         10.70
## 3      3         (97,98] 368     457   80.53         80.35
## 4     -1       x Unspecified missing   1       1  100.00          0.22
## 5                                       458
##   Percent total
## 1
## 2
## 3         99.78
## 4          0.22
## 5           100
```

(b) Variable `nationa.3`.

Solution:

```
class(ses98.clean$nationa2)
## [1] "NominalVariable"
## attr(,"package")
## [1] "Dataset"
```

```
valids(ses98.clean$nationa2)
##         Genève Suisse Romande  ch-al.+Tessin         Europe    hors Europe
##              1              2              3              4              5
```

C'est une variable nominale, des modalités n'ont pas assez de représentants, nous utilisons donc `recode` pour les fusionner.

```r
ses98.clean$nationa.3 <- recode(
  ses98.clean$nationa2,
  'GE' = 1,
  'Suisse hors GE' = 2:3,
  'Étranger' = 4:5
)
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.
##
##                    GE Suisse hors GE Étranger
##   Genève          133               0        0
##   Suisse Romande    0              93        0
##   ch-al.+Tessin     0             112        0
##   Europe            0               0       92
##   hors Europe       0               0       28
```

```r
# Nous vérifions ensuite les fréquences
frequencies("nationa.3", ses98.clean)
##   Coding Missing          Label   N N total Percent Percent (all) Percent total
## 1      1                     GE 133           29.04         29.04
## 2      2         Suisse hors GE 205           44.76         44.76
## 3      4               Étranger 120     458   26.20         26.20        100.00
## 4                                       458                                100
```

(c) Variable `lieudi.3`.

Solution:

```r
class(ses98.clean$lieudi2)
```

```r
valids(ses98.clean$lieudi2)
```

```r
ses98.clean$lieudi.3 <- recode(
  ses98.clean$lieudi2,
  'GE' = 1,
  'Suisse hors GE' = 2:3,
  'Étranger' = 4:6
)
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.
```

```r
# Nous vérifions ensuite les fréquences
frequencies('lieudi.3', ses98.clean)
```

(d) Variable `diplse.3`.

Solution:

```r
class(ses98.clean$diplse2)
```

```r
valids(ses98.clean$diplse2)
```

```r
ses98.clean$diplse.3 <- recode(
  ses98.clean$diplse2,
  'classique, latine' = 1,
  'moderne, scientifique, économique' = 2:4,
  'autre' = 5:8
)
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.

# Nous vérifions ensuite les fréquences
frequencies('diplse.3', ses98.clean)
```

(e) Variable `diplse.3b`.

Solution:

```r
class(ses98.clean$diplse2)

valids(ses98.clean$diplse2)

ses98.clean$diplse.3b<- recode(
  ses98.clean$diplse2,
  'classique, latine, scientifique' = c(1,3),
  'moderne, économique' = c(2,4),
  'autre' = 6:8
)
## Error: invalid class "NominalVariable" object: FALSE

# Nous vérifions ensuite les fréquences
frequencies('diplse.3b', ses98.clean)
## Error: unable to find an inherited method for function 'frequencies'
for signature '"NULL", "missing"'
```

(f) Variable `profper4`.

Solution:

```r
class(ses98.clean$profpere)

valids(ses98.clean$profpere)

# Nous souhaitons garder 'nr' comme un cas valide
ses98.clean$profpere.4 <- recode(
  ses98.clean$profpere,
  'ouvrier, employé' = 1:2,
  'artisan, commerçant, cadre moyen' = 3:4,
  'cadre supérieur, prof.libérale' = 5,
  'sans profession + chômeur' = 6:8,
  'non renseigné' = 0
)
## Operation completed successfully.
## Here is the allocation of the rows in the different classes.
```

```r
# Nous vérifions ensuite les fréquences
frequencies("profpere.4", ses98.clean)
```

Remarque : 5 modalités est beaucoup pour l'analyse, il serait sans doute plus pertinent de continuer de fusionner des modalités, de manière cohérente avec notre problématique bien sûr.

Exercice 3.22 (Export et sauvegarde)**.**

Nous avons terminé de construire notre jeu d'étude pour notre problématique. Nous allons l'exporter et la sauvegarder. Ainsi, si un de mes lecteurs est intéressé par mon travail et souhaite réaliser les analyses lui-même, je pourrais lui transmettre la base prête à l'emploie, et la vue synthétique en PDF pour lui faciliter la prise en main de la base.

1. Exportez la base avec la fonction `export` en nommant l'objet `ses.study`.

Solution:

```r
export(
  ses98.clean,
  name = 'ses98.study'
)
```

2. Envoyez-vous la base et sa vue synthétique en PDF par courriel pour garder une sauvegarde dans votre boîte de courrier électronique.

Exercice 3.23 (Présentation d'un recodage)**.**

À venir...

Exercice 3.24 (Appariement de bases de données)**.**
Les grandes enquêtes peuvent être livrées en plusieurs fichiers de données, généralement classés par thème, ou par années (données longitudinales).
Pour notre problématique nous pouvons avoir besoin de variables présentes dans différentes bases. Nous devons alors extraire ces variables de leur base puis réaliser un appariement pour lier chaque individu à sa valeur de la première variable et sa valeur sur la seconde variable.
Pour cela il est nécessaire d'avoir à disposition un identifiant unique pour chaque individus dans chacune des bases sur lesquelles nous voulons réaliser l'appariemment.
Nous allons ici travailler sur les données du PSM, que vous pouvez récupérer à l'arborescence `V:\SSS3\AnCat-SHP-release2012`.

1. Récupérez les fichiers `SHP_MP.sav` et `SHP06_P_USER.sav` et enregistrez-les dans un dossier nommé `shp` dans votre répertoire `AnCat`.
2. Le fichier `SHP_MP.sav` contient les données sur les individus communes à toutes les vagues, tandis que le fichier `SHP06_P_USER.sav` contient les données récoltées durant l'enquête de 2006.
   Consultez les vues synthétiques PDF et chercher un identifiant unique permettant d'identifier les individus dans chacune des bases.

Solution: `idpers`.

3. Charger la base `SHP_MP.sav` dans un objet nommé `shp.all` et la base `SHP06_P_USER.sav` dans un objet nommé `shp.w2006`.

Solution:

```
setwd("H:/AnCat")
shp.all <- get.spss.file(
  file = 'shp/SHP_MP.sav',
  name = 'SHP all MP'
)
shp.w2006 <- get.spss.file(
  file = 'shp/SHP06_P_USER.sav',
  name = 'SHP wave 2006'
)
```

4. Combien d'individus avez vous dans chacune de ces bases ?

Solution:

```
nrow(shp.all)
## [1] 22976
```

```
nrow(shp.w2006)
## [1] 10863
```

5. La fonction `merge` permet de réaliser un appariement. Elle s'utilise
`merge(BDD1, BDD2, by = ID)`
où `ID` est l'identifiant sur lequel vous réalisez l'appariement. Appariez ces deux bases sur l'identifiant `idpers`

Solution:

```
shp.merged <- merge(shp.all, shp.w2006, by = "idpers")
```

6. Vérifiez que vous avez bien l'ensemble dans votre base `shp.merged` l'ensemble des variables de chacunes des deux autres bases.

Solution:

```
ncol(shp.all) + ncol(shp.w2006)
## [1] 505
```

```
ncol(shp.merged)
## [1] 504
```

7. Pourquoi a-t-on perdu une colonne ?

Solution: La colonne `idpers` était présente dans chacune des deux bases, mais présente une seule fois dans la base appariée.

8. Combien d'individus avez-vous dans votre nouvelle base ?

Solution:

```
nrow(shp.merged)
```

```
## [1] 10863
```

Nous avons donc tous les individus de la vague 2006, pour lesquels nous avons maintenant leurs valeurs sur les variables de la base `shp.all`.

À préparer pour la prochaine séance :

La semaine dernière vous avez défini votre problématique, et opérationnalisée celle-ci sur le PSM vague 2006, ou vos propres données. Cette semaine, vous devez créer votre jeu d'étude, tout comme nous l'avons fait ensemble en séance 2 sur les données `ses98`, mais cette fois-ci sur vos propres données, avec vos variables d'analyse retenues.

1. Les données d'enquêtes sont généralement distribuées sous une licence d'utilisation. Les données du PSM s'acquierent à titre gratuit sous signature d'un contrat de licence. Rendez vous sur le site du PSM et consulter la procédure d'acquisition des données (www.swisspanel.ch > Données PSM > Acquérir les données).
   - Télécharger ensuite le contrat de licence (fichier « contrat de données »), remplissez-le, et
   - Scannez-le et envoyez-le par courriel aux responsables du PSM (swisspanel@fors.unil.ch), en mettant Danilo et Emmanuel en copie.

   Attention, cette étape fait partie de l'évaluation.
   Remarques :
   - si vous ne travaillez pas sur les données du PSM, envoyez-nous la copie de licence d'utilisation des données que vous utilisez. Si vos données ne sont pas sous licence, faites l'exercice pour une licence du PSM.
   - le contrat est individuel, il ne suffit donc pas de remplir un seul contrat par groupe. Tout utilisateur doit signer le contrat pour accéder aux données. *Tous les membres du groupe* doivent donc signer et envoyer le contrat aux responsables du PSM.
2. Créez votre jeu d'étude :
   - Fusionnez les différents fichiers du Panel pour récupérer vos variables d'étude.
   - Configurez la pondération.
   - Exclure les individus non concernés par votre problématique
   - Vérifiez pour chaque variable la configuration des cas valides et valeurs manquantes.
   - Vérifiez pour chaque variable la mesure utilisée.
   - Exportez le jeu de données et faites-en une sauvegarde.
3. Envoyez à Danilo ou à Emmanuel la vue synthétique au format PDF de votre jeu d'étude.
4. Notez au brouillon dans votre rapport les principales étapes que vous avez suivies, afin de ne pas avoir à vous les remémorer, au risque d'en oublier, au moment où vous écrirez votre rapport.

Travaux pratiques – Séance 4

Objectifs de la séance : Rédaction de l'interprétation des résultats. Définition des variables de contrôle. Retrait des individus qui sont manquants sur une ou plusieurs variables. Analyses bivariées.

Exercice 4.25 (Interprétation des résultats)**.**
Présenter les résultats de vos analyses et les conclusions que vous portez grâce à celles-ci n'est pas simple. Voici quelques points clés à toujours avoir en tête :

- À qui je m'adresse ? Quel est son niveau de compétence ? Quel type de document je rédige ? Un rapport interne, un article scientifique, un article pour les journaux quotidiens, ...
- Présentez les tenants et aboutissants de votre interprétation : introduisez, concluez.
- Présentez les « preuves » (statistiques) de vos affirmations.
- Interprétez les nombres qui doivent l'être.
- Présentez uniquement ce qui est pertinent pour votre propos : *soyez le plus concis possible !* Si une phrase n'apporte pas d'information nouvelle, il faut la retirer.

Pour chaque hypothèse que vous testez, vous devez rappeler au lecteur le contexte de l'analyse. Commencez par introduire précisément ce que vous testez :

1. **la question de recherche :** pourquoi ce lien est-il intéressant ?
2. **Théorie(s) :** quelle(s) « théorie(s) » permet d'expliquer ce lien ?
3. **Hypothèse testée :** selon cette « théorie », quelle forme le lien devrait-il prendre ?
4. **Hypothèse(s) alternative(s) :** à quels autres formes de lien pourrait-on s'attendre ?

Par exemple, si nous souhaitons tester la relation entre le type d'étude suivi et l'ambiance dans la filière (ceci est la question de recherche), les points précédents pourraient se formuler :

- **Théorie :** Les enseignements de sciences économiques renforcent l'esprit de compétition des étudiants.
- **Autre théorie :** sélection *a priori* (les plus compétitifs vont en sciences économiques).
- **Hypothèse testée :** les sciences économiques génèrent un comportement plus compétitif que les sciences sociales.
- **Hypothèse alternative :** il peut aussi ne pas y avoir de relation.

En seconde étape vous lancez les analyses et obtenez donc des résultats. Vous devez alors les interpréter. Voici quelques points à suivre lors de l'interprétation :

1. **Détaillez la relation :** il est nécessaire de décrire l'association, dire uniquement que la relation est significative ne suffit pas. Y'a-t-il une catégorie sur-représentée ?
2. **Interprétez les nombres :** est-ce grand, petit ? Est-ce que cela implique une direction dans la relation ?
3. **Lien avec l'hypothèse :** est-ce que cela confirme ou infirme la théorie ? A-t-on des hypothèses sur l'origine (ou l'absence) de lien à la lumière des résultats ?
4. **Conclusion :** qu'avez vous apporté ?

1. Que pensez-vous de l'interprétation suivante ?
   « Afin de vérifier si l'étude des sciences économiques engendre un comportement plus compétitif que celle des sciences sociales, nous avons mesuré l'association entre la

section et l'ambiance compétitive dans la faculté des SES. L'association est statistiquement significative selon le test du chi-carré (Chi2 = 78.764 ; dl = 4 ; p < 0.001) et le V de Cramer est élevé (0.535 ; p < 0.001). Le D de Somers (0.599 ; p < 0.001) est positif, ce qui nous indique que lorsqu'on "monte" dans les sections (on passe de sciences sociales à économiques), l'ambiance devient plus compétitive. »

Solution: Réponse relativement bien formulée : l'hypothèse est présentée, puis uniquement les résultats importants sont rapportés, et justifiés par des tests statistiques. Les tests statistiques effectués sont interprétés. Manque une conclusion, notamment sur le fait que l'on n'arrive pas à décider avec le travail effectué si la compétitivité est la cause d'une formation en économie, ou provient d'une sélection a priori des individus.

2. Que pensez-vous de l'interprétation suivante ?
« Une lecture attentive du tableau croisé nous montre que 32% des étudiants en sciences sociales ne ressentent pas d'esprit de compétition contre 5.7% des étudiants en sciences économiques. Cette tendance se vérifie si l'on regarde la ligne "assez mal" puisqu'ils sont 40.8% en sciences sociales contre 17.9% en sciences économiques. À partir de "plus ou moins", la tendance s'inverse puisque les pourcentages sont plus élevés pour les sciences économiques que pour les étudiants en sciences sociales avec 29.2% contre 18.3%. Assez bien est la catégorie modale en sciences économiques avec 30.2% pour seulement 8.3% en sciences sociales. Finalement, ils sont 17% à juger qu'un "esprit de compétition" correspond "tout à fait" à l'ambiance de sciences économiques pour seulement 0.6% en sciences sociales. Ces différences de pourcentages sont suffisamment importantes pour que l'association soit statistiquement significative au seuil de 5%. »

Solution: Catastrophique. Ceci est un commentaire du tableau croisé, pas une analyse sociologique. On ne sait même pas ce que l'on teste, il n'y a pas de présentation de l'hypothèse testée. Aucun test de statistique pour justifier que les écarts entre les pourcentages sont significativement différents. Aucune interprétation des résultats (d'un côté interpréter n'aurait pas de sens puisqu'il n'y a pas d'hypothèse). Pas de conclusion sur ce que l'analyse nous a apprise. Texte beaucoup trop long.

3. Que pensez-vous de l'interprétation suivante ?
« L'association entre une ambiance compétitive et la section est statistiquement significative selon le test du chi-carré (Chi2 = 78.764 ; dl = 4 ; p < 0.001). Cette relation est confirmée par l'ensemble des mesures puisque le V de Cramer (0.535 ; p < 0.001) et le D de Somers (0.599 ; p < 0.001) sont élevés. »

Solution: Travail insuffisant. L'hypothèse testée n'est pas présentée. Il est dit que la relation est confirmée par le D de Somers, sans préciser ce que le résultat obtenu cette mesure nous permet de conclure. Aucune conclusion, on nous dit qu'il y a une association significative, mais on ne sait pas laquelle, est-ce en sciences économiques qu'il y a plus de compétitivité ou en sciences sociales ?

Exercice 4.26 (Chargement d'une base de donnée préalablement exportée)**.**
Dans cette séance, nous avons besoin de la base de données que vous avez préparée pour les analyses au précédent TP, et que vous avez exporté sous le nom `ses98.study`.

1. Chargez la librairie `Dataset`.

Solution:

```
library(Dataset)
```

2. Définissez comme répertoire de travail le dossier contenant votre base `ses98.study`. Si vous avez suivi nos conventions, cela devrait être le dossier `AnCat`.

Solution:

```
setwd("H:/AnCat/")
```

3. Chargez la base de donnée que vous avez préparée au précédent TP.

Solution:

```
load("ses98.study.RData")
```

4. Vérifiez que la base est bien présente en mémoire.

Solution:

```
"ses98.study" %in% ls()
## [1] TRUE
```

Exercice 4.27 (Variables de contrôle et retrait d'individus par non-réponse)**.**
Dans notre jeu d'analyse, nous avons des valeurs manquantes sur nos variables, c'est-à-dire des individus qui n'ont pas répondus à certaines questions. Avant chaque analyse nous devons réfléchir à comment vont être traitées ces valeurs manquantes dans la méthode que nous allons utiliser. La plupart du temps, la façon de traiter les valeurs manquantes est de retirer des données chaque individu qui possèderait une valeur manquante sur au moins une des variables utilisées dans l'analyse à réaliser. Cette façon de procéder est légitime : nous n'avons pas l'information suffisante sur l'individu en question, donc nous le retirons. Cependant cette manipulation peut entraîner un biais de non-réponse, qu'il est important de considérer.
La classe `Dataset` permet de définir des variables de contrôles, à l'aide de la fonction `checkvars(BDD) <- VecteurNomDesVariables`
Ces variables de contrôle seront utilisées lors du retrait d'individus de la base de données. Pour chaque variable de contrôle, les marges des modalités de cette variable seront calculées avant le retrait et après le retrait, puis un test du Chi2 sera effectuer pour évaluer si il y a eu une perturbation significatives des marges de la variable. Nous pourrons alors prendre en compte cette information lors de l'interprétation des résultats : par exemple, si le retrait des individus a amené à une sur-représentation des hommes par rapport aux femmes, il sera intéressant de prendre en compte cette information lors de la discussion des résultats.

1. Définissez les variables `sexe` et `age98.3` comme variables de contrôle pour la base `ses98.study`.

Solution:

```
checkvars(ses98.study) <- c("sexe", "age98.3")
```

2. Affichez les variables de contrôle que vous venez de définir.

Solution:

```
checkvars(ses98.study)
## [1] "sexe"    "age98.3"
```

3. À l'aide de la fonction `subset`, créez un jeu ne contenant que les hommes.

Solution:

```
ses98.study.homme <- subset(ses98.study, sexe == "homme")
## => control on sexe: warning, p-value < 0.05
## homme are overrepresented
## femme are underrepresented
## => control on age98.3: ok
```

   (a) Que se passe-t-il au niveau de la variable de contrôle `sexe` ?

Solution: Nous sommes prévenu que les hommes sont significativement sur-représentés dans le jeu de données extrait, et que réciproquement les femmes sont sous-représentées.

   (b) Que se passe-t-il au niveau de la variable de contrôle `age98.3` ?

Solution: La distribution de notre sous-échantillon ne contenant que les hommes n'est pas significativement différente de l'échantillon principal en terme de classes d'âge.

4. La fonction `only.complete`, permet de retirer les individus contenant des non-réponses sur une ou plusieurs variables. Elle s'utilise :
   `only.complete(vecteurVariables, BDD)`
   (a) Créez un jeu de données `dommere.ok` en retirant les individus n'ayant pas donné de réponse au domicile de leur mère (utilisez la variable recodée, car nous avons peut-être déjà fait un traitement des valeurs manquantes auparavant).

Solution:

```
dommere.ok <- only.complete("domicile.mere", ses98.study)
## => control on sexe: ok
## => control on age98.3: ok
```

   (b) Avez-vous eu une perturbation signficative de la représentativité ?

Solution: Non, les tests effectués sur chacune des variables de contrôle n'ont pas décelé d'écarts significatifs.

(c) Combien d'individus avez-vous retirés ?

Solution:

```
nindividual(ses98.study) - nindividual(dommere.ok)
## [1] 12
```

(d) Créez un jeu de données `dompere.ok` en retirant les individus n'ayant pas donné de réponse au domicile de leur père (utilisez pour la même raison que précedemment la variable recodée).

Solution:

```
dompere.ok <- only.complete("dompere", ses98.study)
## => control on sexe: ok
## => control on age98.3: ok
```

(e) Avez-vous eu une perturbation signficative de la représentativité ?

Solution: Non, les tests effectués sur chacune des variables de contrôle n'ont pas décelé d'écarts significatifs.

(f) Combien d'individus avez-vous retirés ?

Solution:

```
nindividual(ses98.study) - nindividual(dompere.ok)
## [1] 23
```

(g) Créez un jeu de données `dom.mere.pere.ok` en retirant les individus n'ayant pas donné de réponse au domicile de leur mère ou au domicile de leur père.

Solution:

```
dom.mere.pere.ok <- only.complete(
  c("domicile.mere", "dompere"),
  ses98.study
)
## => control on sexe: ok
## => control on age98.3: ok
```

(h) Avez-vous eu une perturbation signficative de la représentativité ?

Solution: Non, les tests effectués sur chacune des variables de contrôle n'ont pas décelé d'écarts significatifs.

(i) Combien d'individus avez-vous retirés ?

Solution:

```
nindividual(ses98.study) - nindividual(dom.mere.pere.ok)
## [1] 23
```

(j) Que pouvons nous en conclure ?

Solution: Si nous testons l'association entre le domicile de la mère et le bilan en 99, et respectivement le domicile du père et le bilan en 99, nous devons avoir en tête que les analyses n'auront pas été effectuées sur les mêmes individus : 23 individus seront retirés pour le domicile du père, tandis que 12 seront retirés pour le domicile de la mère.
Par ailleurs, nous pouvons constater que les individus n'ayant pas renseigné le domicile de la mère n'ont pas renseigné non plus le domicile du père.

Exercice 4.28 (traitement des valeurs manquantes sur la variable dépendante)**.**

Dans votre étude de cas, vous souhaitez expliquer les comportements sociaux modélisés par une variable dépendante. Dans ces séances de séminaires, notre variable dépendante est le bilan en 99.

1. Y'a-t-il des valeurs manquantes sur notre variable dépendante ?

Solution:

```
frequencies("bil_99", ses98.study)
```

```
##   Coding Missing    Label   N N total Percent Percent (all) Percent total
## 1      1            echec 115           25.11         25.11
## 2      2         redouble  79           17.25         17.25
## 3      3           réussi 264     458   57.64         57.64        100.00
## 4                                458                                 100
```

Non, aucune valeur manquante.

2. À l'issue des analyses, que pourront nous dire sur ces valeurs manquantes ? Est-ce une hypothèse que vous voulez tester ?

Solution: Lors des analyses nous allons essayer de comprendre à l'aide des variables explicatives comment les individus sont associés à chacune des modalités de la variable dépendante. Si nous avons des valeurs manquantes sur la variable dépendante, une partie de notre analyse signifie donc « comprendre pourquoi les individus n'ont pas répondus ».
Dans notre cas, ce n'est pas une hypothèse que nous souhaitons tester.

3. Quelle stratégie pensez-vous la plus intéressante à adoper face aux valeurs manquantes sur la variable dépendante ?

Solution: Si vous avez une hypothèse sur un certain type de valeur manquante, faîte entrer cette valeur manquante dans les cas valides.
Pour toutes les valeurs manquantes sur lequel votre étude ne porte pas, il est préférable de les retirer du jeu de données. En particulier, il est plus intéressant de connaître la répartition des individus parmis les cas valides de la variable, non parmis l'ensemble des valeurs possibles.

4. Bien qu'ici nous n'ayons pas de valeur manquante sur la variable dépendante, écrivez tout de même le code permettant de retirez ces individus.

Solution:

```
bil99.ok <- only.complete("bil_99", ses98.study)

## => control on sexe: ok
## => control on age98.3: ok
```

Exercice 4.29 (Prise en main de la fonction bivan)**.**
La fonction `bivan` permet de réaliser des tests d'association bivariée, c'est-à-dire, de tester l'association pouvant exister entre deux variables. La fonction fournit une grande variété de tests qui permettent, lorsqu'il y a association, d'en cerner les différentes nuances.

1. Consultez le manuel de la fonction `bivan`. Quelles sont les mesures s'appliquant à des données catégorielles, et celles s'appliquant à des données numériques ?
2. Parmis les mesures pour données catégorielles, quelles sont : **★[CH 2.2]**
   (a) les mesures symétriques ?

Solution: Le Chi carré de Pearson, le Phi, le t de Tschuprow, le v de Cramer, le coefficient de contingence de Pearson, et le Chi carré du rapport de vraisemblance.

   (b) les mesures directionnelles ?

Solution: Le *lambda* de Goodman et Kruskal, le *tau* de Goodman et Kruskal et le u de Theil.

   (c) les mesures pour variables ordinales ?

Solution: Les *tau a* et *b* de Kendall, le *tau c* de Stuart, le *gamma* de Goodman et Kruskal, le d de Somers et le e de Wilson.

3. Exécutez la fonction bivan sur les variables `bil_99` et `age98.3`, avec `bil_99` en variable dépendante.

Solution:

```
biv1 <- bivan(bil_99 ~ age98.3, ses98.study)

## Global association measures
##              chi2  cramer.v gk.tau.sqrt   somers.d
## age98.3 8.64 +    0.10 +      0.10 *    -0.12 **
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

   (a) Quelle est la valeur du Chi2 de Pearson ?

Solution: La statistique du Chi2 de Pearson entre l'âge et le bilan en 99 est : 8,64.

   (b) Combien de degrés de liberté avez-vous pour ce test ?

Solution: Nous avons 3 modalités pour le bilan, et 3 pour l'âge. Le nombre de degrés de liberté est donc $(3-1)*(3-1)=4$.

(c) Au seuil de 10%, est-elle significative ? ★[CH 2.3]

Solution: Oui, à un seuil de 10%, l'association est significative.

(d) Au seuil de 5%, est-elle significative ?

Solution: Non, à un seuil de 5%, l'association n'est pas suffisante pour être significative.

(e) La fonction `global` appliquées à un objet de type `Bivan` permet de récupérer les valeurs numériques des mesures globales. Donnez une valeur approchée en pourcentage à 1 décimale de la p-valeur du test. Que représente cette valeur du point de vue statistique ?

Solution:

```
as.numeric(round(global(biv1)[2], 3)) * 100
## [1] 7.1
```

Nous considérons donc qu'il y a 7.1 pourcents de probabilité que l'association que l'on observe ici soit du au hasard de l'échantillonnage.

(f) À votre avis, est-il raisonnable dans ce cas de porter des conclusions sur la population générale en utilisant le V de Cramer ?

Solution: Non, nous considérons trop précipité de conclure à une association entre l'âge et le bilan à la fin de l'année en se basant sur le test du Chi2, et allons plutôt regarder les autres mesures d'association.

Exercice 4.30 (Analyse des résidus)**.**
L'analyse des résidus du Chi2 est un outil important, il permet de détecter rapidement des sur-représentation ou sous-représentation au croisement de certaines modalités.
La fonction `std.res` permet de récupérer les valeurs des résidus standardisés. Les résultats sont enregistrés dans un objet de type `Statdf`, permettant de stocker des données statistiques avec leurs p-valeurs. Vous pouvez utiliser la fonction ★[CH 2.1.3]
`summary(x, merge = 'left')`
pour faire afficher les étoiles de significativité.

1. Récupérez dans un objet les résidus standardisés de l'association entre l'âge et le bilan.

Solution:

```
residus <- std.res(biv1)
```

2. Affichez-les, en valeur exacte, puis avec les étoiles de significativité.

Solution:

```
residus
```

```
##                   echec echec signif. redouble redouble signif. réussi réussi signif.
## age98.3 / [17,19] -2.748      0.005992   0.7152           0.4745  1.865        0.06219
## age98.3 / (19,23]  1.797      0.072376  -0.8616           0.3889 -0.918        0.35862
## age98.3 / (23,25]  1.185      0.235833   0.3632           0.7164 -1.318        0.18747
```

```
summary(residus, merge = "left")
```

```
## Standardized Residuals table
```

```
##                           echec   redouble     réussi
## age98.3 / [17,19] -2.75 **    0.72       1.86 +
## age98.3 / (19,23]  1.80 +    -0.86      -0.92
## age98.3 / (23,25]  1.19       0.36      -1.32
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

3. Quels sont les individus sous-représentés et sur-représentés ?

Solution: Les individus âgés entre 17 et 19 ans et ayant échoué en 1999 sont sous-représentés. Les individus âgés entre 17 et 19 ans et ayant réussi en 1999 ainsi que les individus âgés entre 19 et 23 ans et ayant échoué en 1999 semblent sur-représentés. Cependant ces deux dernières associations ne sont pas significatives à moins de 5 %, il est peut-être plus raisonnable de ne pas porter de conclusion dessus.

4. Que pouvez vous dire à propos de la représentation des individus de 24 à 25 ans qui ont échoués ?

Solution: Rien, car la statistique n'est pas significative.

Exercice 4.31 (Mesures directionnelles)**.**
Les mesures directionnelles visent à déterminer l'influence de la variable explicative sur les modalités prises par la variable dépendante. **★[CH 2.4.2]**

1. Lancez la fonction bivan avec les mesures suivantes `gk.tau.sqrt` et `theil.u.sqrt`.

Solution:

```
biv2 <- bivan(
  bil_99 ~ age98.3,
  ses98.study,
  chi2=F,
  cramer.v=F,
  gk.tau.sqrt=T,
  theil.u.sqrt=T,
  somers.d=F
)
```

```
## Global association measures
##         gk.tau.sqrt theil.u.sqrt
## age98.3   0.10 *       0.10 +
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

2. Interprétez la valeur de la racine du tau.

Solution: La valeur du *tau* de Goodman et Kruskal est significative au seuil de 5%, ce qui nous permet de conclure que l'âge est un relativement bon prédicteur du succès à l'examen.
Remarques :
- il faut comprendre par bon prédicteur que si on me donne l'âge d'un individu, je pourrais affirmer avec davantage de certitude s'il a échoué ou réussi que si on ne me donne pas son âge.
- en pratique, la valeur de la réduction de l'incertitude ne nous intéresse pas tellement, on va surtout comparer les valeurs de réduction de l'incertitude *entre variables* pour déterminer laquelle est la « meilleure prédictrice ».

3. Les valeurs de la racine du tau et de la racine du u sont-elles exactement les mêmes ?

Solution:

```
global(biv2)[c("gk.tau.sqrt", "theil.u.sqrt")]
##         gk.tau.sqrt theil.u.sqrt
## age98.3      0.1022       0.1023
```

Non, les valeurs sont très proches, mais ne sont pas exactement les mêmes.

4. Pourquoi regarde-t-on la racine du *tau* de Goodman & Kruskal et non la valeur de la statistique en elle-même ?

Solution: Le *tau* de Goodman & Kruskal mesure la réduction de l'entropie quadratique. Ceci signifie que dans le calcul, chaque entropie a été élevée au carrée. Cela engendre que le coefficient ne varie pas linéairement avec l'association mais quadratiquement. Nous devons donc compenser par l'application de la racine carrée sur ce coefficient (de la même manière que nous le faisons avec la variance, pour obtenir une mesure de dispersion variant linéairement avec les écarts à la moyenne, que nous appelons l'écart-type).

5. La racine du *tau* de Goodman & Kruskal vaut 10%, peux-t-on en conclure que connaissant l'âge l'incertitude sur le succès à l'examen est réduite d'environ 10% ?

Solution: Non, la réduction de l'incertitude est mesurée par le *tau* de Goodman & Kruskal, et non sa racine. Pour discuter de la valeur de la réduction de l'incertitude, nous devons élever au carré la racine : $0.10^2 = 0.01$. On peux alors conclure que connaissant l'âge l'incertitude sur le succès à l'examen est réduite d'environ 1%.
Remarque : avec l'argument `gk.tau=TRUE` vous pouvez obtenir directement la valeur du *tau* de Goodman & Kruskal.

Exercice 4.32 (Mesures pour variables ordinales)**.**

Les mesures pour variables ordinales permettent d'évaluer si une association entre l'ordonnancement des niveaux de chacune des variables existe. Par exemple, lorsque je suis plus âgé (i.e. je monte dans les catégorie d'âge), est ce que je vais tendre vers un meilleur bilan, ou un bilan plus mitigé ?

1. Avant de démarrer l'analyse, nous devons avoir clairement en tête les niveaux utilisés et leur ordonnancement.
   (a) La variable bilan en 99 est-elle ordinale ? Consultez ses niveaux

Solution:

```
class(ses98.study$bil_99)
## [1] "OrdinalVariable"
## attr(,"package")
## [1] "Dataset"
```

```
valids(ses98.study$bil_99)
##    echec redouble   réussi
##        1        2        3
```

L'ordre des niveaux est donnée de gauche à droite. Ce n'est pas l'ordre du codage qui compte mais bien l'ordre de gauche à droite des noms de modalités.

   (b) La variable `age98.3` est-elle ordinale ? Consultez ses niveaux.

Solution:

```
class(ses98.study$age98.3)
## [1] "OrdinalVariable"
## attr(,"package")
## [1] "Dataset"
```

```
valids(ses98.study$age98.3)
## [17,19] (19,23] (23,25]
##       1       2       3
```

L'ordre des niveaux est donnée de gauche à droite. Ce n'est pas l'ordre du codage qui compte mais bien l'ordre de gauche à droite des noms de modalités.

2. Lancez la fonction `bivan` avec les mesures `kendall.tau.b`, `somers.d`, et `wilson.e`.

Solution:

```
biv3 <- bivan(
  bil_99 ~ age98.3,
  ses98.study,
  chi2=F,
  cramer.v=F,
  gk.tau.sqrt=F,
  kendall.tau.b=T,
  somers.d=T,
  wilson.e = T
)
```

```
## Global association measures

##         kendall.tau.b   somers.d   wilson.e
## age98.3     -0.11 *     -0.12 **   -0.07 **

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

3. Que pouvez-vous conclure quand à l'effet d'âge sur le bilan en 99 ? ★**[CH 2.4.3]**

Solution: Le d de Somers est significatif à moins de 1% et à une valeur négative. Ceci signifie que plus on était âgé, moins on avait de chances de réussir en 1999.
Rappel : la valeur est négative, ceci signifie que lorsque l'on monte dans les niveaux de la variable prédictrice, on descent dans les niveaux de la variable dépendante.

4. À votre avis, quel serait la valeur du D de Somers si on inverse les niveaux :
   (a) De la variable `bil_99` ?

Solution: Le d de Somers aurait même valeur en valeur absolue, mais serait de signe positif.

   (b) De la variable `age98.3` ?

Solution: Même chose.

   (c) Des deux ?

Solution: Le d de Somers aurait même valeur en valeur absolue et serait toujours de signe négatif.

5. Et qu'en sera-t-il de la significativité pour chacun de ces changements ?

Solution: La significativité sera la même pour chacun de ces changement : il faut comprendre que la force de l'association est la même, c'est juste la direction qui change. Remarque : par contre si nous mettons l'âge en variable dépendante et le bilan en prédicteur, nous n'obtiendront a priori pas les mêmes résultats, à la fois en terme de valeur et de significativté.

6. Vérifiez vos réponse en inversant les niveaux de la variable `bil_99` et en relançant `bivan`.

Solution:

```r
ses98.study$bil_99.reverse <- valids.reverse(ses98.study$bil_99)
biv4 <- bivan(
  bil_99.reverse ~ age98.3,
  ses98.study,
  chi2=F,
  cramer.v=F,
  gk.tau.sqrt=F,
  kendall.tau.b=T,
  somers.d=T,
  wilson.e = T
)
```

```
## Global association measures

##         kendall.tau.b  somers.d  wilson.e
## age98.3      0.11 *    0.12 **   0.07 **

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

Exercice 4.33 (Analyses bivariées complètes)**.**
Nous avons jusqu'ici regardé l'association entre l'âge en 98 et le bilan en 99. Dans vos analyses vous lancerez directement les tests sur l'ensemble des variables permettant de tester vos hypothèses. Pour ajouter d'autres variables explicative, utilisez le symbole `+` entre chaque variable.

1. Lancez les analyses bivariées, en prenant soin de choisir au moins une mesure basée sur le Chi2, une mesure directionnelle, et une mesure ordinale, sur les variables : `sexe`, `age98.3`, `troncom`, `domicile.mere.3`, `nationa.3`, `profpere.4`, et `lieudi.3`.

Solution:

```r
biv4 <- bivan(
  bil_99 ~ sexe + age98.3 + troncom +
    domicile.mere.3 + nationa.3 + profpere.4 + lieudi.3,
  ses98.study
)
```

```
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
```

```
## Global association measures

##                      chi2  cramer.v gk.tau.sqrt   somers.d
## sexe              1.18      0.05        0.03       0.03
## age98.3           8.64 +    0.10 +      0.10 *    -0.12 **
## troncom           7.49 *    0.13 *      0.08 *     0.05
## domicile.mere.3   9.23 +    0.10 +      0.11 *    -0.04
## nationa.3        11.68 *    0.11 *      0.12 *    -0.08 *
## profpere.4       10.02      0.10        0.11      -0.06 +
## lieudi.3         11.90 *    0.11 *      0.12 **   -0.03

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

2. La fonction affiche des *warnings*, que signifient-t-ils ?

Solution: Le message *Chi-squared approximation may be incorrect* signifie qu'aux croisements de certaines modalités, il y a trop peu d'individu pour que l'approximation du Chi2 de Pearson avec la loi du Chi2 théorique soit suffisamment robuste pour être affirmée comme correcte. Il est communément admis qu'il est nécessaire d'avoir au moins 5 individus par case dans le tableau des effectifs théorique pour que l'approximation par une loi du Chi soit acceptable.

3. Cherchez pour quelles variables ces *warnings* sont affichés.

Solution: Il faut tester les variables une par une, en les mettant à tour de rôle en seul descripteur. Nous constatons que les *warnings* apparaissent uniquement pour la variable `profpere4`.

```
biv5 <- bivan(
  bil_99 ~ profpere.4,
  ses98.study
)

## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
```

Dans le tableau des effectifs attendus nous constatons en effet que le profil `sans profession + chômeur` et `redouble` est attendu avec un effectif inférieur à 5.

```
expected(biv5)

## $profpere.4
##                                    echec redouble réussi
## ouvrier, employé                  27.369    18.80  62.83 109
## artisan, commerçant, cadre moyen  25.611    17.59  58.79 102
## cadre supérieur, prof.libérale    44.945    30.88 103.18 179
## sans profession + chômeur          7.031     4.83  16.14  28
## non renseigné                     10.044     6.90  23.06  40
##                                  115.000    79.00 264.00 458
```

4. Il est généralement intéressant de repérer quelle est la variable la plus associée à notre variable réponse. Ici, est-ce évident ?

Solution: Non. Les variables `lieudi3`, `nationa3`, `age98.3`, `troncom` et `domicile.mere3` sont fortement associées, en particulier en terme d'incertitude. Les variables `age98.3` et `nationa3` présentent de plus une association ordinale significative. Mais il ne semble pas y avoir une variable clairement plus associée que les autres.

5. En terme de réduction de l'incertitude, quelle semble-être la variable la plus associée ?

Solution: Le lieu du diplôme.

6. Pouvez-vous affirmer que cet écart avec les autres variables est significatif ?

Solution: Non, il faudrait pour cela tester la significativité de l'écart.

7. À votre avis, les analyses pour chaque variable explicative ont-elle été réalisées sur exactement les mêmes individus ?

Solution: Pour chaque analyse bivariée, les individus ayant une valeur manquante sur le prédicteur ou la variable cible sont retirés de l'analyse. Si le jeu de données ne contient aucune valeur manquante, ou si les valeurs manquantes sont distribuées soit sur toutes les variables soit sur aucune, alors oui les mêmes individus ont été utilisés pour les analyses. Ce n'est pas le cas ici : le sexe n'a aucune valeur manquante, on va donc faire l'analyse sur tous les individus, alors que le domicile de la mère en a, on retirera ces individus.
Il est conseillé de toujours vérifié sur combien d'individus ont été réalisées chacune des analyses.

À préparer pour la prochaine séance :

Partie I

Par groupe, interprêtez l'association entre le bilan en 99 et la nationalité recodée en 3 catégories, avec le bilan en variable dépendante. Utilisez dans votre analyse :

- Les résidus standardisés ;
- Une mesure basée sur le Chi2 de Pearson ;
- Une mesure directionnelle ;
- Une mesure ordinale (précisez les niveaux et leur ordre).

Envoyez-nous par courriel la vue synthétique en PDF de l'analyse, et vos interprétations.

Partie II

La semaine dernière vous avez préparé votre jeu d'étude à partir du PSM vague 2006, ou de vos propres données. Cette semaine, vous devez préparer votre jeu d'étude pour l'analyse, tout comme nous l'avons fait ensemble en séance 3 sur les données `ses98`.

1. Effectuez les corrections éventuelles suite aux retours de Danilo et Emmanuel sur votre vue synthétique envoyée la semaine dernière.
2. Sauvegarder votre jeu d'étude avec la fonction `export`.
3. Créez votre jeu d'analyse :
   - Renommez éventuellement les variables de manière à ce que chacune ait un nom clair.
   - Renommez éventuellement les valeurs des variables de manière ce que chacune ait un nom clair et court. Pensez à regarder à la fois les cas valides et les valeurs manquantes.
   - Effectuez l'ensemble des recodages nécessaires pour tester les hypothèses que vous souhaitez tester. Nous conseillons de faire intervenir dans votre rapport les 3 types de recodages présentés au séminaire.

     Remarques :
     - dans le cours AnCat, nous effectuerons les analyses uniquement entre variables catégorielles. Toutes les variables échelles doivent donc être recodées en une variable catégorielle ordinale.
     - bien sûr vous pouvez effectuer plusieurs recodages différents pour une même variable, suivant les hypothèses que vous voulez tester (c'est ce que nous avons fait en séance 3 avec la variable `diplse`).
     - il est préférable de créer une nouvelle variable pour chaque recodage : évitez de remplacer la variable d'origine par la variable recodée. (regardez comment nous avons fait en séance 3)
   - Exportez le jeu de données et faites-en une sauvegarde.
4. Envoyez à Danilo ou à Emmanuel la vue synthétique au format PDF de votre jeu d'analyse.
5. Notez au brouillon dans votre rapport les principales étapes que vous avez suivies, afin de ne pas avoir à vous les remémorer, au risque d'en oublier, au moment où vous écrirez votre rapport.

Travaux pratiques – Séance 5

Objectifs de la séance : Arbres de classification

Exercice 5.34 (Chargement d'une base de donnée préalablement exportée)**.**
Dans cette séance, nous avons besoin de la base de données que vous avez préparée pour les analyses, et que vous avez exporté sous le nom `ses98.study`.

1. Chargez la librairie `Dataset`.

Solution:

```
library(Dataset)
```

2. Définissez comme répertoire de travail le dossier contenant votre base `ses98.study`. Si vous avez suivi nos conventions, cela devrait être le dossier `AnCat`.

Solution:

```
setwd("H:/AnCat/")
```

3. Chargez la base de donnée.

Solution:

```
load("ses98.study.RData")
```

4. Vérifiez que la base est bien présente en mémoire.

Solution:

```
"ses98.study" %in% ls()
## [1] TRUE
```

Exercice 5.35 (Vérification des mesures utilisées)**.**
Les mesures utilisées pour chaque variable vont avoir un impact sur la construction de l'arbre. Par exemple, considérons une variable catégorielle ayant les modalités : `faible`, `moyen`, et `fort`. Si cette variable est définie comme nominale, il sera autorisé de regrouper les catégories `faible` et `fort` ensemble. Ce regroupement, qui sera peut-être efficace au niveau statistique, ne fera pas nécessairement de sens au niveau socio-économique, car il n'est pas naturel que des individus ayant des valeurs extrêmes sur une variable partagent un point commun permettant une meilleure compréhension de notre variable dépendante. Cela peut dans certains cas avoir un sens, à nous de voir si nous pouvons mettre en lien le résultat trouvé avec les hypothèses que nous testons. Dans le cas général, lorsqu'une variable catégorielle possède un « ordre naturel », nous allons plutôt la définir comme ordinale.

Ceci aura pour effet d'autoriser uniquement des regroupements de modalités adjacentes. Par exemple, les regroupements [`faible`, `moyen`] et [`moyen`, `fort`] seront acceptés, alors que le regroupement [`faible`, `fort`] sera refusé.

1. Nous considérons les variables `profpere.4` et `diplse.3` comme nominales. Vérifiez pour chacune que leur mesure est bien réglée.

Solution:

```
class(ses98.study$profpere.4)
```

```
## [1] "NominalVariable"
## attr(,"package")
## [1] "Dataset"
```

```
class(ses98.study$diplse.3)
```

```
## [1] "NominalVariable"
## attr(,"package")
## [1] "Dataset"
```

2. Nous considérons les variables `age98.3`, `domicile.mere.3`, et `lieudi.3` comme ordinales. Vérifiez que leur mesure sont bien réglées.

Solution:

```
class(ses98.study$age98.3)
```

```
## [1] "OrdinalVariable"
## attr(,"package")
## [1] "Dataset"
```

```
class(ses98.study$domicile.mere.3)
```

```
## [1] "NominalVariable"
## attr(,"package")
## [1] "Dataset"
```

La variable après recodage n'est pas ordinale, on la transforme donc en ordinale :

```
ses98.study$domicile.mere.3 <- ovar(ses98.study$domicile.mere.3)
```

```
## number of missings: 12 ( 2.62 %)
```

```
class(ses98.study$domicile.mere.3)
```

```
## [1] "OrdinalVariable"
## attr(,"package")
## [1] "Dataset"
```

```
class(ses98.study$lieudi.3)
```

```
## [1] "NominalVariable"
## attr(,"package")
## [1] "Dataset"
```

La variable après recodage n'est pas ordinale, on la transforme donc en ordinale :

```
ses98.study$lieudi.3 <- ovar(ses98.study$lieudi.3)
## number of missings: 0 ( 0 %)
class(ses98.study$lieudi.3)
## [1] "OrdinalVariable"
## attr(,"package")
## [1] "Dataset"
```

3. Vérifiez, pour chacune des variables ordinales, l'ordre des niveaux.

Solution:

```
valids.ordering(ses98.study$age98.3)
## [1] "[17,19] < (19,23] < (23,25]"
```

Des plus jeunes aux plus âgés.

```
valids.ordering(ses98.study$domicile.mere.3)
## [1] "GE < Suisse hors GE < Étranger"
```

Du plus proche de Genève au plus éloigné.

```
valids.ordering(ses98.study$lieudi.3)
## [1] "GE < Suisse hors GE < Étranger"
```

Du plus proche de Genève au plus éloigné aussi.

4. Nous considérons les variables `sexe` et `troncom` comme binaires. Vérifiez que ces variables sont bien définies comme binaires.

Solution:

```
class(ses98.study$sexe)
## [1] "BinaryVariable"
## attr(,"package")
## [1] "Dataset"
```

```
class(ses98.study$troncom)
## [1] "BinaryVariable"
## attr(,"package")
## [1] "Dataset"
```

Exercice 5.36 (Construction de l'objet de paramétrage d'un arbre)**.**
La construction d'un arbre commence par la définition des paramètres que nous allons utiliser pour contrôler son développement.
Dans la librairie `Dataset`, l'objet stockant les paramètres de développement d'un arbre est le `Treecontrol`. Vous pouvez créer un objet `Treecontrol` avec la fonction `tree.control()`.

Cette fonction possède des valeurs par défaut pour tous ses paramètres, vous pouvez donc la lancez sans arguments.

1. Construisez un `Treecontrol` avec les paramètres par défaut, stockez-le dans un objet nommé `treectrl`, puis affichez-le.

Solution:

```
treectrl <- tree.control()
```

```
treectrl
```

```
##                                  CHAID CART
## min.for.splitting          20.000     x    x
## min.terminal.node.n         7.000     x    x
## min.terminal.node.percent   1.000     x
## min.complexity.reduction    0.001          x
## max.height                  3.000     x    x
## max.pvalue.merge            0.050     x
## alpha3                     -1.000     x
## max.pvalue.split            0.050     x
## stump                       0.000     x
```

2. L'objet `Treecontrol` se manipule comme un objet `list`.
Récupérez la valeur de l'option `min.terminal.node.n`, qui spécifie le nombre minimal d'individus que doit posséder un nœud pour exister.

Solution:

```
treectrl$min.terminal.node.n
```

```
## [1] 7
```

3. Pour la construction des arbres dans ce TP nous allons commencer avec les paramètres suivants : au moins 25 individus par nœud, au moins 50 individus pour pouvoir éclater un nœud, une hauteur maximale de 5 étages et, pour la méthode CHAID, une significativité minimal de 5% pour qu'il y ait éclatement.
Spécifiez ces paramètres dans l'objet `treectrl`.

Solution:

```
treectrl$min.terminal.node.n <- 25
treectrl$min.for.splitting <- 50
treectrl$max.height <- 5
treectrl$max.pvalue.split <- 0.05
```

```
treectrl
```

```
##                                  CHAID CART
## min.for.splitting          50.000     x    x
## min.terminal.node.n        25.000     x    x
## min.terminal.node.percent   1.000     x
## min.complexity.reduction    0.001          x
## max.height                  5.000     x    x
## max.pvalue.merge            0.050     x
## alpha3                     -1.000     x
## max.pvalue.split            0.050     x
## stump                       0.000     x
```

Exercice 5.37 (Arbre d'induction par méthode CHAID)**.**
La construction d'un arbre par la méthode CHAID se fait sous la librairie `Dataset` à l'aide de la fonction `tree.learn.chaid` qui s'utilise :
`tree.learn.chaid(formula, data, control)`

1. Construire un arbre CHAID avec le bilan en 99 comme variable dépendante et les descripteurs suivants : `sexe`, `troncom`, `profpere.4`, `diplse.3`, `age98.3`, `domicile.mere.3`, `lieudi.3`.

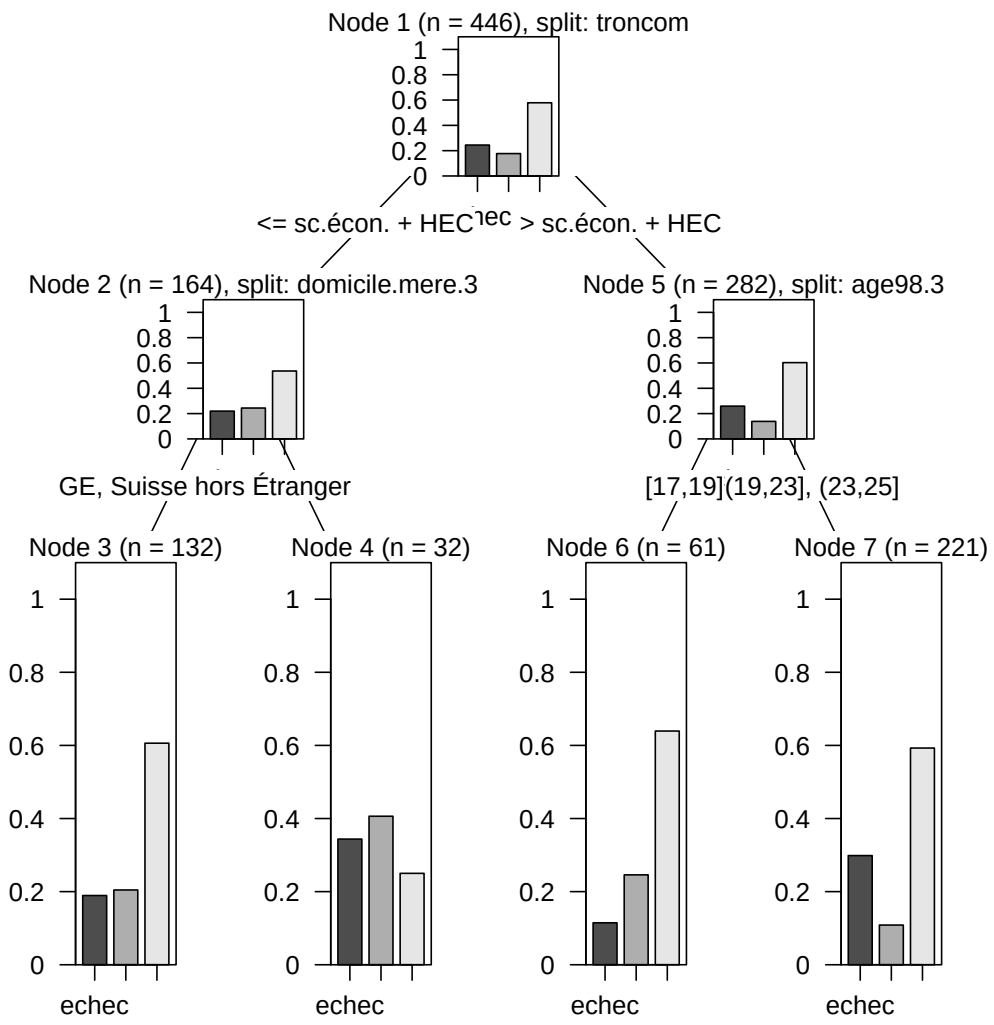
Solution:

```
tree1 <- tree.learn.chaid(
  bil_99 ~ sexe + troncom + profpere.4 +
    diplse.3 + age98.3 + domicile.mere.3 + lieudi.3,
  ses98.study,
  treectrl
)
```

2. À l'aide de la méthode `plot`, tracez l'arbre obtenu.

Solution:

```
plot(tree1)
```



3. Sur cet arbre, quelle est la variable la plus associée au bilan en 99 ?

Solution: La variable la plus associée au sens du χ^2 est la variable sur laquelle l'éclatement est réalisé en premier. Ici, nous constatons que c'est le tronc commun qui conditionne en premier lieu la réussite en 99.

4. Nous parlons d'interaction entre deux variables (ou plus) lorsque l'effet d'une variable est modulé par la valeur prise par l'autre.
   Repérez un effet d'interaction.

Solution: Nous observons ici une interaction entre le tronc commun et le domicile de la mère : chez les étudiants en sciences économiques nous observons un éclatement suivant le domicile de la mère (donc effet significatif), cependant nous ne retrouvons pas cet éclatement chez les étudiants en sciences sociales (c'est suivant l'âge que l'éclatement a eu lieu, le domicile de la mère n'a pas été retenu). Le domicile de la mère ne semble donc influencer les chances de réussite *que dans le cas des sciences économiques*. C'est là qu'est l'interaction.

Nous observons aussi une interaction entre le tronc commun et l'âge : chez les étudiants en sciences sociales nous observons que l'âge a un effet significatif (puisqu'il y a éclatement) tandis que chez les étudiants en sciences économiques celui-ci ne semble pas avoir d'effet (l'éclatement a eu lieu suivant le domicile de la mère et non suivant l'âge).

5. Ici nous avons observé des interactions à chaque éclatement. Dans quel cas nous n'observerions pas d'interaction ?

Solution: Reprenons l'interaction entre le tronc commun et le domicile de la mère : si nous avions observé aussi l'éclatement chez les étudiants en sciences sociales, cela voudrait dire que le domicile de la mère à un effet significatif peu importe le tronc commun choisi. Pour conclure sur l'existence d'une interaction, il faudrait alors regarder la nature de l'effet pour chacun des troncs communs, puis comparer.
Par exemple, dans le cas des sciences économiques, nous constatons que le fait que le domicile de la mère soit situé à l'étranger amène majoritairement un résultat d'échec ou de redoublement (nœud 4), tandis que pour les étudiants dont le domicile de la mère est situé en Suisse les résultats sont majoritairement la réussite à l'examen (nœud 3). Si dans le cas des sciences sociales nous observons le même effet alors il n'y a pas d'interaction. Il y aurait un effet significatif du domicile de la mère pour chacun des cursus, mais cet effet serait le même. Si par contre nous observions un effet différent (par exemple les étudiants en Suisse échouent davantage et les étudiants à l'étranger réussissent d'avantage) alors il y aurait effet d'interaction.

Remarque : nous avons observé une interaction, mais il est difficile de conclure sur la force de cette interaction. En effet, nous ne pouvons pas vraiment conclure que le domicile de la mère n'a pas d'effet chez les étudiants en sciences sociales, il en a peut-être un (et même sûrement), mais qui serait moins fort que celui de l'âge et donc masqué par celui-ci. Dans un second temps, nous pourrons mesurer de manière effective l'impact des interactions que nous venons de découvrir, à l'aide de méthodes de régression.
Mais déjà maintenant, nous aurions deux façons de faire pour essayer de quantifier cet effet :

- Retirer la variable d'âge, et regarder si elle est remplacée par le domicile de la mère, et regarder alors la nature de l'effet. Si nous observons par exemple le même effet du domicile pour chacun des troncs, c'est que l'interaction est faible.
- Essayer de développer plus profondemment l'arbre, en changeant les paramètres de contrôle, et regarder si le domicile de la mère apparaît plus loin. Si par exemple il n'apparaît jamais, c'est que l'interaction est forte, car l'effet du domicile de la mère serait réservé aux sciences économiques.

autre remarque : on comprend donc qu'un arbre ne permet généralement pas de trouver tous les effets d'interactions, mais de découvrir seulement les plus marquant. Si on veut tester spécifiquement si une interaction existe entre deux variables (si c'est l'une de nos hypothèses de notre problématique par exemple) il faudrait construire un arbre avec uniquement ces deux variables là.

6. Nous souhaitons maintenant découvrir des interactions sur 3 niveaux. Pour cela nous allons autoriser un développement plus profond de l'arbre en réduisant le nombre d'individus nécessaire pour éclater un nœud.
   Passez les paramètres `min.terminal.node.n` à 5 et `min.for.splitting` à 15, puis

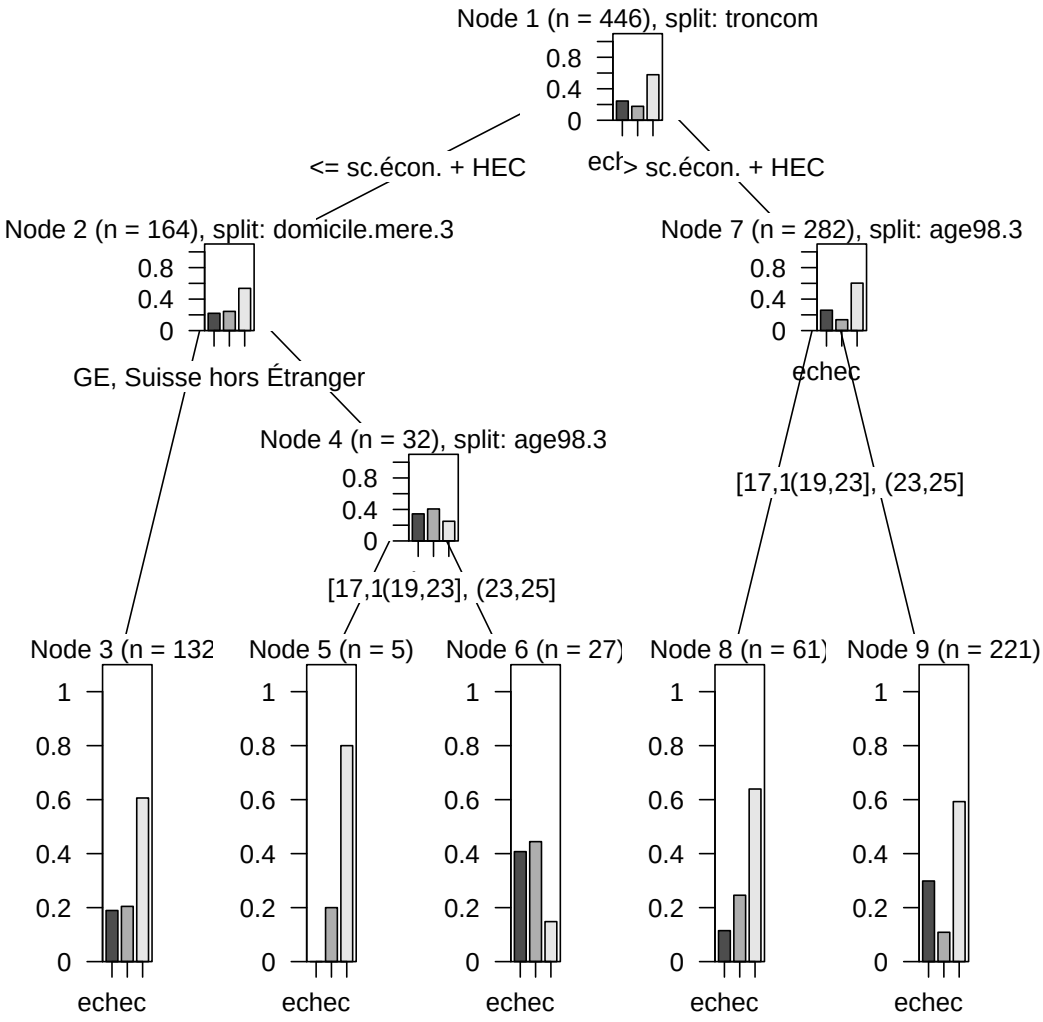
relancez la construction de l'arbre, que vous nommerez `tree2`.

Solution:

```
treectrl$min.terminal.node.n <- 5
treectrl$min.for.splitting <- 15
```

```
tree2 <- tree.learn.chaid(
  bil_99 ~ sexe + troncom + profpere.4 +
    diplse.3 + age98.3 + domicile.mere.3 + lieudi.3,
  ses98.study,
  treectrl
)
```

```
plot(tree2)
```



7. Observez-vous une interaction d'ordre 3 ?

Solution: Oui, entre les variables tronc commun, domicile de la mère et l'âge en 98.

8. Décrivez la différence de l'effet de l'âge dans cette interaction.

Solution: En regardant la distribution des étudiants en sciences sociales (nœud 7), nous constatons que l'effet de l'âge est principalement de discriminer les individus qui vont plutôt redoubler (nœud 8) des individus qui vont plutôt échouer (n9) alors que la proportion des individus ayant réussi reste à peu près la même (environ 0.6) dans les trois nœuds.
Si nous regardons maintenant la distribution des étudiants de sciences économiques dont le domicile de la mère est situé à l'étranger, nous constatons que l'effet est principalement porté sur la réussite : 20% de réussite dans le nœud 4 pour 80% de réussite chez les individus jeunes (nœud 5) et moins de 20% chez les individus plus âgés (nœud 6).

Remarque : pour étudier ces interactions d'ordre supérieur ou égal à 3, il est important de comparer les distributions par rapport au nœud parent de chaque éclatement. En effet si dans les nœuds parents les distributions sont différentes, il est évident que les distributions dans les nœuds enfants seront différentes. Ce qui nous intéresse pour avoir une interaction c'est l'effet de la variable et non de la distribution. Donc pour évaluer l'effet de la variable, il faut tenir compte de la distribution que l'on avait déjà dans le nœud parent.

9. Est-il dangereux de conclure immédiatement ?

Solution: Oui, car dans le nœud 5 il y a seulement 5 individus. On peut alors se demander si l'interaction trouvée est une spécificité de l'échantillon récolté ou est suffisamment prononcée pour être généralisable à la population entière.
Les méthodes de régression nous permettront de quantifier l'impact de ces interactions, et alors de conclure (confirmer, infirmer, ou impossibilité de décider) sur ce résultat.

10. L'arbre stocke l'indice du nœud terminal dans lequel se trouve chaque individu dans l'attribut `where` que vous pouvez accéder via l'opérateur `$`. Récupérez cet objet et stockez le dans une variable nommée `chaid1.nodes` dans le jeu `ses98.study`, puis calculez les fréquences d'apparition de chaque catégorie.

Solution:

```
ses98.study$chaid.nodes <- tree.where(tree2)
```

```
frequencies("chaid.nodes", ses98.study)
```

```
##   Coding Missing Label   N N total Percent Percent (all) Percent total
## 1      1             3 138           30.13         30.13
## 2      2             5   5            1.09          1.09
## 3      3             6  27            5.90          5.90
## 4      4             8  61           13.32         13.32
## 5      5             9 227     458   49.56         49.56        100.00
## 6                              458                                 100
```

Remarque : on calcule les fréquences de la variable au sein du jeu de données, et non pas indépendemment du jeu de données, ceci pour que la pondération soit prise en compte. Dans le cas contraire, vous verrez un message d'alerte s'afficher.

11. Nous voulons mesurer la qualité statistique de l'arbre comme prédicteur du bilan en 99. Pour cela nous allons regarder l'association entre la variable contenant l'indice des nœuds terminaux et la variable bilan en 99. Quel type de mesure d'analyse bivariée proposez vous pour tester cette association ?

Solution: Une mesure directionnelle.

12. Réalisez les tests d'association qui vous semblent pertinents pour comparer l'apport de l'arbre en terme de force d'association comparé à chaque variable prise indépendemment.

Solution:

```
chaid.bivan <- bivan(
  bil_99 ~ sexe + troncom + profpere.4 + diplse.3
    + age98.3 + domicile.mere.3 + lieudi.3 + chaid.nodes,
  ses98.study,
  chi2 = F,
  cramer.v = F,
  somers.d = F,
  gk.tau = T,
  theil.u.sqrt = T,
  theil.u = T
)
```

```
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
```

```
## Global association measures

##                    gk.tau gk.tau.sqrt   theil.u theil.u.sqrt
## sexe             0.00         0.03      0.00         0.04
## troncom          0.01 *       0.08 *    0.01 *       0.09 *
## profpere.4       0.01         0.11      0.01         0.11
## diplse.3         0.02 ***     0.14 ***  0.02 **      0.13 **
## age98.3          0.01 *       0.10 *    0.01 +       0.10 +
## domicile.mere.3 0.01 *       0.11 *    0.01 +       0.10 +
## lieudi.3         0.02 **      0.12 **   0.01 *       0.12 *
## chaid.nodes      0.05 ***     0.21 ***  0.05 ***     0.22 ***

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

Conclusion : le meilleur descripteur obtient 2% de réduction d'incertitude d'après le u de Theil, notre arbre obtient sur ce même indicateur une réduction de 5% et sa significativité est plus importante. Notre arbre a donc apporter de l'information supplémentaire.
Remarque : l'arbre est aussi un modèle plus complexe, et plus il y aura de nœud, plus notre arbre sera bon statistiquement, mais aussi plus il sera complexe. Il faudra trouver un compromis entre la force prédictive et la complexité.

13. Générez la vue synthétique au format PDF de votre analyse par arbre CHAID.

Solution:

```
exportPDF(tree2)
```

Exercice 5.38 (Paramétrage intéracif des arbres)**.**

Un point important à retenir sur les méthodes d'arbres, c'est que ce sont des outils d'exploration des données, qui vont nous permettre de découvrir certaines finesses au sein de nos hypothèses et ainsi d'aller plus loin dans l'analyse. Pour explorer nos données, nous avons un certain nombre de paramètres que nous pouvons faire varier. Il est *impératif* de faire varier ces paramètres pour deux raisons principales :

– Explorer une grande partie des données. Avec 1 seul paramétrage vous n'aurez qu'une seule « vue » de vos données, et vous pouvez passer à côter d'éléments pertinents.

– Contrôler la robustesse de la vue que vous avez découverte. Si en changeant très faiblement le paramétrage l'éclatement que vous avez observé n'existe plus, alors c'est qu'il n'est pas si fort, et donc il vaut mieux s'abstenir de le discuter.

Pour faire varier facilement les paramètres, les méthodes d'arbres intégrés à la librairie Dataset possèdent une vue interactive, que vous pouvez lancer en ajoutant `interactive = TRUE` à la fonction d'apprentissage de l'arbre :

```
tree1 <- tree.learn.chaid(
  bil_99 ~ sexe + troncom + profpere.4 +
    diplse.3 + age98.3 + domicile.mere.3 + lieudi.3,
  ses98.study,
  treectrl,
  interactive = TRUE
)
```

1. Contrôlez la sensibilité de l'éclatement précédemment découvert.

Exercice 5.39 (Arbre d'induction par méthode CART)**.**
La construction d'un arbre par la méthode CART se fait sous la librairie `Dataset` à l'aide de la fonction `tree.learn.cart` qui s'utilise de la même manière que `tree.learn.chaid` : `tree.learn.cart(formula, data, control)`

1. Construire un arbre CART avec le bilan en 99 comme variable dépendante en utilisant les mêmes descripteurs que pour l'arbre CHAID précédemment calculé.

Solution:

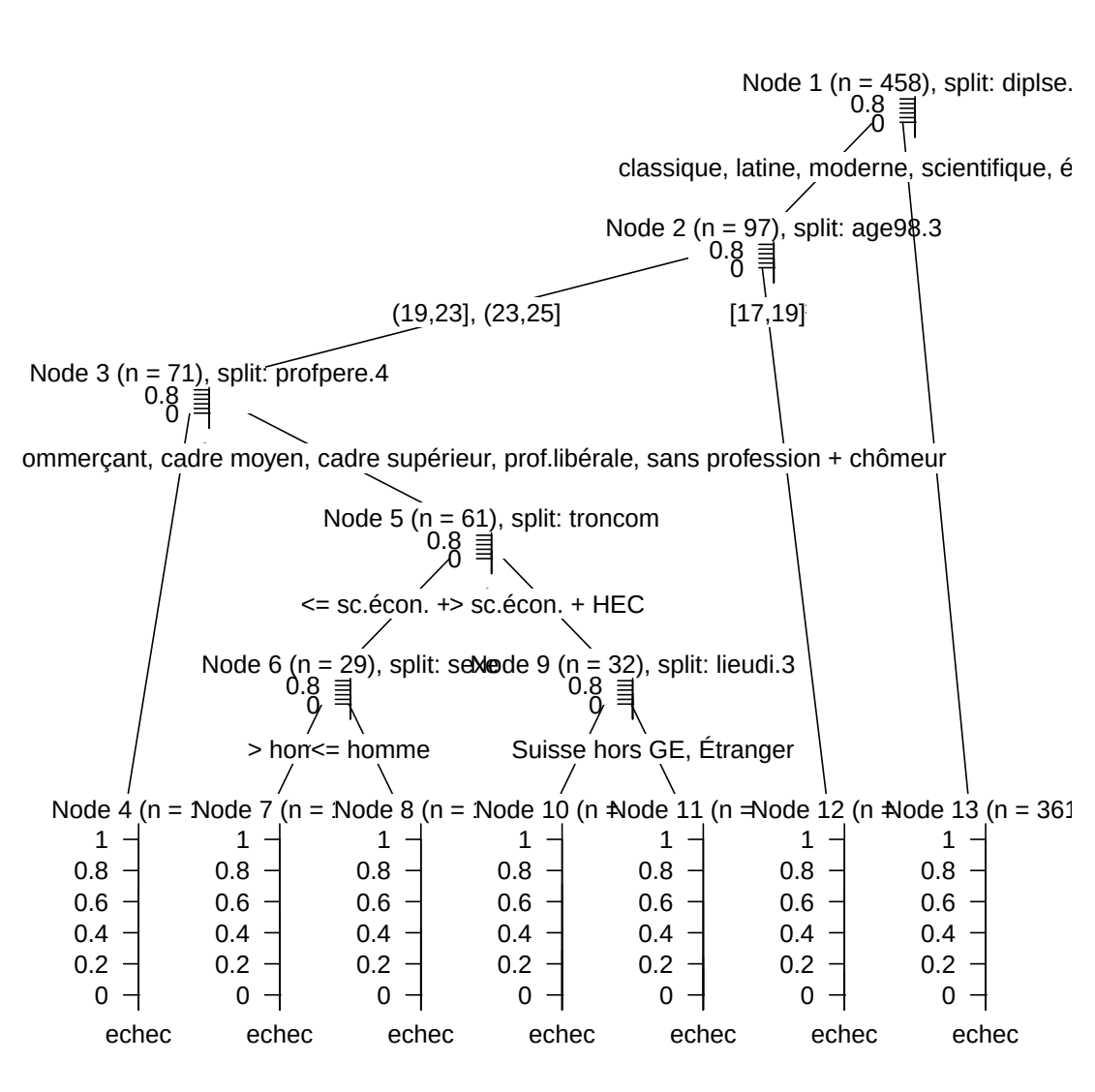
```
tree3 <- tree.learn.cart(
  bil_99 ~ sexe + troncom + profpere.4 +
    diplse.3 + age98.3 + domicile.mere.3 + lieudi.3,
  ses98.study
)

## Loading required package: rpart
## number of missings: 0 ( 0 %)
```

2. À l'aide de la méthode `plot`, tracez l'arbre obtenu.

Solution:

```
plot(tree3)
```



3. Sur cet arbre, quelle est la variable la plus associée au bilan en 99 ?

Solution: Le diplôme de la scolarité du secondaire.

4. Récupérez l'objet `where` de cet arbre et stockez le dans une variable nommée `cart1.nodes` dans le jeu `ses98.study`, puis calculez les fréquences d'apparition de chaque catégorie.

Solution:

```
ses98.study$cart.nodes <- tree.where(tree3)
```

```
frequencies("cart.nodes", ses98.study)
```

```
##   Coding Missing Label   N N total Percent Percent (all) Percent total
## 1      1             4  10           2.18          2.18
## 2      2             7  15           3.28          3.28
## 3      3             8  14           3.06          3.06
## 4      4            10   7           1.53          1.53
## 5      5            11  25           5.46          5.46
## 6      6            12  26           5.68          5.68
## 7      7            13 361     458   78.82         78.82        100.00
## 8                         458                                     100
```

5. Réalisez les tests d'association qui vous semblent pertinents pour comparer l'apport de l'arbre en terme de force d'association comparé à chaque variable prise indépendemment.

Solution:

```
cart.bivan <- bivan(
  bil_99 ~ sexe + troncom + profpere.4 + diplse.3
    + age98.3 + domicile.mere.3 + lieudi.3 + cart.nodes,
  ses98.study,
  chi2 = F,
  cramer.v = F,
  somers.d = F,
  gk.tau = T,
  theil.u.sqrt = T,
  theil.u = T
)
```

```
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
```

```
## Global association measures

##                    gk.tau gk.tau.sqrt   theil.u theil.u.sqrt
## sexe             0.00         0.03      0.00         0.04
## troncom          0.01 *       0.08 *    0.01 *       0.09 *
## profpere.4       0.01         0.11      0.01         0.11
## diplse.3         0.02 ***     0.14 ***  0.02 **      0.13 **
## age98.3          0.01 *       0.10 *    0.01 +       0.10 +
## domicile.mere.3 0.01 *       0.11 *    0.01 +       0.10 +
## lieudi.3         0.02 **      0.12 **   0.01 *       0.12 *
## cart.nodes       0.06 ***     0.24 ***  0.05 ***     0.23 ***

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

Conclusion : le meilleur descripteur obtient 2% de réduction d'incertitude d'après le u de Theil, notre arbre obtient sur ce même indicateur une réduction de 5%.

6. Générez la vue synthétique au format PDF de votre analyse par arbre CART.

Solution:

```
exportPDF(tree3)
```

Exercice 5.40 (Comparaison des modèles d'arbre)**.**
Nous avons construit plusieurs arbres. Ces arbres, générés avec des méthodes différentes ou des paramétrages différents, vont pouvoir nous apporter des informations complémentaires qui pourront nous donner des pistes (comprendre de nouvelles hypothèses à tester) pour expliquer les mécanismes socio-économiques sous-jacents dans notre problématique.

Lorsque nous construisons plusieurs modèle, il est important d'avoir le réflexe de les comparer, en terme de performance et de complexité. Nous donnons ici quelques éléments basiques de comparaison de modèles d'arbre.

1. Statistiquement, quel est l'arbre le plus associé à la variable réponse ?

Solution:

```
biv.trees <- bivan(
  bil_99 ~ chaid.nodes + cart.nodes,
  ses98.study
)
```

```
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
## Warning: Chi-squared approximation may be incorrect
```

```
## Global association measures
```

```
##                   chi2  cramer.v gk.tau.sqrt   somers.d
## chaid.nodes 40.90 ***  0.21 ***    0.21 ***  -0.01
## cart.nodes  45.93 ***  0.22 ***    0.24 ***   0.20 ***
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
```

2. Est-ce étonnant ?

Solution: Non. Nous avons utilisé comme mesures directionnelles des mesures basées sur la réduction de l'incertitude, or l'arbre CART est développé en visant une certaine amélioration de la réduction de l'incertitude (plus précisemment, il utilise un indice de réduction de la complexité).

3. Quelle est la profondeur de chaque arbre ? Quel est alors l'arbre le plus profond ?

Solution: Arbre CHAID : profondeur 3.
Arbre CART : profondeur 5.

4. Combien de feuilles possède chaque arbre ? Quel est alors l'arbre le plus feuillu ?

Solution: Arbre CHAID : nombre de feuilles 5.
Arbre CART : nombre de feuilles 7.
L'arbre le plus feuillu est l'arbre CART.

5. Combien de nœuds possède chaque arbre ? Quel est alors l'arbre le plus complexe ?

Solution: Arbre CHAID : nombre de nœuds 9.
Arbre CART : nombre de nœuds 13.
L'arbre le plus complexe est l'arbre CART.

6. Si un arbre apparaissait comme trop complexe, peu lisible, quels moyens auriez-vous pour faire construire un arbre plus simple ?

Solution: Adapter les paramètres du `tree.control`.

À préparer pour la prochaine séance :

Partie I – Acceptation des modalités d'évaluation en vue de passer l'examen

Les modalités d'évaluation de ce cours étant un peu différentes des modalités habituelles (examen sous forme d'un rapport écris, rendu et défendu oralement avant la session d'examens officielle), nous vous demandons de bien vouloir manifester votre acceptation des conditions d'évaluation en envoyant un courriel à Danilo et Emmanuel contenant (mot pour mot) le texte suivant :

Je soussigné(e), [prénom] [NOM], accepte les conditions d'examen du cours et du séminaire AnCat telles qu'elles ont été spécifiées sur le livret de TP du séminaire AnCat, section « Modalités d'évaluation ».

Remarque : Ce courriel doit *impérativement* être envoyé depuis votre adresse étudiant de l'Université de Genève.

Partie II – Pratique de l'analyse par arbres de classification

Par groupe, effectuez une analyse par arbre d'induction de la variable `troncom` en utilisant les descripteurs suivants `sexe`, `diplse.3`, `profpere.4`, et `lieudi.3` :

- Commencez par un arbre avec la méthode CHAID.
- Repérez au moins une interaction, en précisant précisemment les variables entrant en jeu.
- Réalisez un autre arbre avec la méthode CART.
- Du point de vue statistique, lequel est le meilleur ?
- Quel arbre vous semble le plus intéressant ?

Envoyez-nous par courriel les réponses à ces question, en joignant les vues synthétiques en PDF de chacun des arbres, et des associations bivariées réalisées.

Partie III – Avancement sur votre étude de cas

La semaine dernière vous avez terminé de préparer votre jeu d'étude, en recodant de manière adéquate les variables en fonction des hypothèses que vous allez tester pour répondre à votre problématique. Vous avez aussi renommé les variables et les modalités des variables de manière à ce que chaque intitulé soit facile à lire et sans ambiguïté pour la personne qui vous lira. Vous avez de plus vérifié les mesures de chaque variables, ainsi que l'ordre des niveaux pour les variables ordinales.

Vous pouvez maintenant démarrer pleinement les analyses. Nous ne vous demanderons à partir de maintenant plus aucun suivi sur votre étude de cas, c'est à vous de nous le demander, par exemple lorsque vous faites face à une difficulté ou avez des questions vis-à-vis d'un résultat obtenu qui vous semble surprenant.

Néanmoins nous vous conseillons pour la semaine à venir les objectifs suivants :

1. Rédiger les parties 1 et 2 de votre étude de cas (voir la section « Directives pour l'étude de cas »). Il est en effet préférable de mettre sur papier dès maintenant les recodages que vous avez effectués : cela vous permet de vérifier que vos recodages sont bien en accord avec vos hypothèses à tester, et d'écrire plus facilement les choses car vous avez encore les idées fraîches sur comment vous avez réalisé les recodages.
2. Écrire le code R réalisant l'ensemble des analyses bivariées que vous devez effectuer pour tester vos hypothèses.
3. Noter au brouillon vos premières interprétations sur les analyses bivariées.

Travaux pratiques – Séance 6

Objectifs de la séance : Régression logistique niveau 1 : régression logistique simple, catégorie de référence, rapports de cotes et leur significativité, statistique du χ^2 du rapport de vraisamblance, régression logistique multiple, profil de référence, calcul de logit, calcul de probabilité prédite, ajout d'interactions, effets transverses, variables de contrôle, modèle complet.

Exercice 6.41 (Chargement d'une base de donnée préalablement exportée)**.**
Dans cette séance, nous avons besoin de la base de données que vous avez préparée pour les analyses, et que vous avez exporté sous le nom `ses98.study`.

1. Chargez la librairie `Dataset`.

Solution:

```
library(Dataset)
```

2. Définissez comme répertoire de travail le dossier contenant votre base `ses98.study`. Si vous avez suivi nos conventions, cela devrait être le dossier `AnCat`.

Solution:

```
setwd("H:/AnCat/")
```

3. Chargez la base de donnée.

Solution:

```
load("ses98.study.RData")
```

4. Vérifiez que la base est bien présente en mémoire.

Solution:

```
"ses98.study" %in% ls()
## [1] TRUE
```

Exercice 6.42 (Régression logistique binaire simple)**.**
La régression logistique binaire simple, dite aussi régression logistique simple ou régression logistique bivariée, fait intervenir une variable à prédire binaire que l'on va expliquer à l'aide d'une autre variable. L'avantage de cet outil est de permettre de quantifier en terme de probabilités l'impact de la variable explicative si elle est quantitiative ou l'impact de chacune des modalités si elle est catégorielle, sur une valeur spécifique de la variable cible. La régression logistique binaire fonctionne comme son nom l'indique de manière binaire,

c'est-à-dire que nous allons opposer une modalité spécifique face à toutes les autres.
Par exemple, nous nous intéressons dans ce TP à la réussite des étudiants de 1998. La variable `bil_99` possède trois modalités : `echec`, `redouble` et `réussi`. En spécifiant la modalité `réussi` dans la régression, nous allons étudier les effets des facteurs explicatifs sur les chances de réussir, plutôt que de ne pas réussir, c'est-à-dire soit d'échouer ou de redoubler.
Dans la librairie `Dataset`, la regression logistique se lance avec la commande `reglog`, qui s'utilise comme suit :

```
reglog(
  formula = MODELE,
  target = CATÉGORIE CIBLÉE,
  data = BDD
)
```

où `formula` est une formulation d'un modèle (comme dans la fonction `bivan`).

1. Calculez le modèle de régression en ciblant la modalité `réussi` de la variable `bil_99` sur la variable `lieudi3`.

Solution:

```
reg1 <- reglog(
  formula = bil_99 ~ lieudi.3,
  target = 'réussi',
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```
print(reg1)

## Table 1:
## Estimated coefficients (odds ratios)

##                            Model 1
## lieudi.3Suisse hors GE 1.493 +
## lieudi.3Étranger       0.569 *
## (Intercept)            1.339 *

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures

##                   Model 1
## Deviance       613.57
## Deviance H0    624.18
## Model Chi2      10.61 **
## Model DF         2.00
## Block Chi2      10.61 **
## Block DF         2.00
## R2 Cox-Snell     0.02
## R2 Nagelkerke    0.03
## N parameters     3.00
## AIC            619.57
## BIC            631.95
## N              458.00

## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

2. Générez la vue synthéthique au format PDF du modèle.

Solution:

```
exportPDF(reg1)
```

3. Est-il raisonnable d'utiliser ce modèle pour l'interprétation ?

Solution: Nous devons avant toute chose vérifier la significativité du χ^2 du rapport de vraisemblance du modèle.

4. Que vaut le χ^2 du rapport de vraisemblance du modèle ? Est-il significatif ?

Solution: Le χ^2 vaut environ 10.61, il est significatif au hauteur de 1%.
Il est donc raisonnable d'interpréter ce modèle.

5. Pour quelle catégorie le modèle estime-t-il le logit ?

Solution: Le modèle estime le logit de la catégorie que nous avons donnée en argument `target`, c'est-à-dire ici la catégorie `réussi`.

6. Dans la colonne `Model 1`, avez-vous les contributions au logit ou les rapports de cotes ?

Solution: Nous avons les rapports de cotes.
Cette information est indiquée dans le titre de la table affichée.

7. Quelle est la catégorie de référence ?

Solution: La catégorie de référence est la catégorie pour laquelle aucun coefficient n'a été estimé, et qui n'est donc pas affichée. Dans ce cas présent, c'est la catégorie `GE`.

8. Pouvez-vous interpréter le coefficient estimé pour les Suisses n'ayant pas eu leur diplôme secondaire à Genève ?

Solution: Avec prudence, car le coefficient est faiblement significatif (signficativité entre 0.05 et 0.1).

9. Le coefficient pour les étudiants ayant passé leur examen secondaire à l'étranger est significatif au seuil de 5%. Interprétez ce coefficient.

Solution: Les étudiants ayant passé leur diplôme secondaire en à l'étranger ont deux fois moins de chances de réussir en 1999 que ceux qui ont passé leur diplôme à Genève.

10. On nomme « individu de référence » l'individu dont le profil est défini par la catégorie de référence pour chaque prédicteur impliqué dans la régression.
Ici nous avons un seul prédicteur. Quel est le profil de l'individu de référence ?

Solution: L'individu de référence est l'individu ayant passé son diplôme secondaire dans le canton de Genève.

11. Il est possible d'afficher dans la console ou dans la vue synthétique en PDF les contributions au logit au lieu des rapports de cotes. Il suffit pour cela de passer l'argument `odds.ratios = FALSE` respectivement à la fonction `print` et `exportPDF`.
Afficher la vue synthétique avec les contributions au logit.

Solution:

```r
print(reg1, odds.ratios = FALSE)
```

```
## Table 1:
## Estimated coefficients
```

```
##                             Model 1
## lieudi.3Suisse hors GE  0.401 +
## lieudi.3Étranger       -0.564 *
## (Intercept)             0.292 *
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                    Model 1
## Deviance        613.57
## Deviance H0     624.18
## Model Chi2       10.61 **
## Model DF          2.00
## Block Chi2       10.61 **
## Block DF          2.00
## R2 Cox-Snell      0.02
## R2 Nagelkerke     0.03
## N parameters      3.00
## AIC             619.57
## BIC             631.95
## N               458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

```r
exportPDF(reg1, odds.ratios = FALSE)
```

12. Quel est le logit calculé par le modèle pour l'individu de référence ?

Solution: Le contraste utilisé est *indicateur*, le logit pour l'individu de référence est donc la constante : 0.29

13. Quel est le logit pour un étudiant ayant obtenu son diplôme secondaire à l'étranger ?

Solution: Le contraste utilisé est *indicateur*, le logit pour cet individu est donc la constante + le coefficient estimé correspondant à sa catégorie : 0.292 + (-0.564) = -0.271

14. En déduire la probabilité de réussir prédite par le modèle pour un étudiant ayant obtenu son diplôme secondaire à l'étranger.

Solution:

```r
logit <- -0.271
prob <- exp(logit)/(1 + exp(logit))
prob
```

```
## [1] 0.4327
```

Vous pouvez aussi formatter le résultat le récupérer comme une chaîne de caractères avec 2 chiffres après la virgule :

```r
formatC(prob, digits = 2, format = "f")
```

```
## [1] "0.43"
```

Exercice 6.43 (Régression logistique binaire multiple)**.**
La régression logistique binaire multiple, dite aussi régression logistique multiple, fait intervenir quant à elle plusieurs predicteurs. De la même manière, cet outil permet de quantifier en terme de probabilités l'impact de chacune des modalités de chaque variable sur la modalité cible de la variable cible.
Important : pour chaque variable explicative, le coefficient estimé pour une de ses modalités, représente l'impact de la modalité en contrôlant pour l'effet des autres variables, c'est-à-dire l'impact dans l'hypothèse où les autres variables resteraient fixes. On dit ainsi que le coefficient s'interprête « toutes choses égales par ailleurs ».

1. Dans la fonction de régression logistique, nous pouvons ajouter d'autres variables à un modèle en utilisant le symbole `+` dans la formule.
   Calculez le modèle de régression en ajoutant la variable `diplse3`.

Solution:

```r
reg2 <- reglog(
  formula = bil_99 ~ lieudi.3 + diplse.3,
  target = 'réussi',
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```
print(reg2)
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                              Model 1
## lieudi.3Suisse hors GE                     1.428
## lieudi.3Étranger                           1.071
## diplse.3moderne, scientifique, économique 0.602 +
## diplse.3autre                              0.335 *
## (Intercept)                                2.121 **
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                    Model 1
## Deviance        607.57
## Deviance H0     624.18
## Model Chi2       16.61 **
## Model DF          4.00
## Block Chi2       16.61 **
## Block DF          4.00
## R2 Cox-Snell      0.04
## R2 Nagelkerke     0.05
## N parameters      5.00
## AIC             617.57
## BIC             638.21
## N               458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

2. Générez la vue synthéthique au format PDF du modèle.

Solution:

```
exportPDF(reg2)
```

3. Que devient le χ^2 du modèle ?

Solution: La valeur du χ^2 a augmenté pour passer à 16.61.

4. Est-il raisonnable d'interpréter ce modèle ?

Solution: Oui, le seuil de significativité n'a pas changé, le modèle apporte toujours une information significative au seuil de 1%.

5. Quel est maintenant le profil de l'individu de référence ?

Solution: L'individu de référence est un étudiant ayant passé son diplôme secondaire dans le canton de Genève et dont le dipôme est de nature classique ou latine.

6. Pouvez-vous interpréter le coefficient estimé pour un étudiant ayant un diplôme secondaire `autre` ?

Solution: Oui, le coefficient est significatif au seuil de 5%.

7. Comment interprétez vous ce coefficient ?

Solution: *Toutes choses égales par ailleurs*, un étudiant ayant passé son examen secondaire d'une nature `autre` a 3 fois moins de chances de réussir qu'un étudiant ayant passé un examen de nature classique ou latine.

8. Comment interprétez vous ce coefficient pour un étudiant ayant passé son diplôme dans le canton de Genève ?

Solution: L'interprétation reste la même. L'étudiant ayant passé son examen secondaire d'une nature `autre` a 3 fois moins de chances de réussir qu'un autre étudiant *ayant aussi passé son diplôme dans le canton de Genève* et ayant passé un examen de nature classique ou latine.
Le « *ayant aussi passé son diplôme dans le canton de Genève* » est une instanciation du « toutes choses égales par ailleurs », appliqué à la seule autre variable prédictrice du modèle : le lieu du diplôme.

9. Comment interprétez vous ce coefficient lorsque l'étudiant a passé son diplôme à l'étranger ?

Solution: Encore une fois, l'interprétation reste la même, l'étudiant ayant passé son examen secondaire d'une nature `autre` a 3 fois moins de chances de réussir qu'un autre étudiant *ayant aussi passé son diplôme dans à l'étranger* et ayant passé un examen de nature classique ou latine.

10. Pouvez-vous dire qu'un étudiant ayant passé son examen secondaire d'une nature `autre` a 3 fois moins de chances de réussir qu'un étudiant ayant passé un examen de nature classique ou latine ?

Solution: Non, cela est faux dans le cas général. Par exemple, un étudiant ayant passé son examen secondaire d'une nature `autre` en Suisse hors Genève n'a pas nécessairement 3 fois moins de chances de réussir qu'un étudiant ayant passé un examen de nature classique ou latine à l'étranger.

11. Nous allons vérifier cela sur notre exemple.
    (a) En affichant les contributions au logit dans la console ou dans la vue synthétique en PDF, déterminez le logit pour l'individu de référence.

Solution:

```
exportPDF(reg2, odds.ratios = F)
```

```
print(reg2, odds.ratios = F)
## Table 1:
## Estimated coefficients
##                                                Model 1
## lieudi.3Suisse hors GE                          0.356
## lieudi.3Étranger                                0.069
## diplse.3moderne, scientifique, économique      -0.507 +
## diplse.3autre                                  -1.093 *
## (Intercept)                                     0.752 **
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
##                    Model 1
## Deviance        607.57
## Deviance H0     624.18
## Model Chi2       16.61 **
## Model DF          4.00
## Block Chi2       16.61 **
## Block DF          4.00
## R2 Cox-Snell      0.04
## R2 Nagelkerke     0.05
## N parameters      5.00
## AIC             617.57
## BIC             638.21
## N               458.00
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

Nous utilisons un contraste indicateur, le logit estimé pour l'individu de référence est donc la constante : 0.75.

(b) Quel est le logit pour un étudiant ayant obtenu son diplôme secondaire en Suisse hors Genève, et dont la spécialisation est `autre` ?
Quelle est alors sa probabilité de réussir en 1999 ?

Solution: Nous utilisons un contraste indicateur. Nous additionnons donc la constante et la contribution au logit des valeurs prises par chacune des variables. Nous obtenons :

```
logit <- 0.752 + 0.356 + (-1.093)
logit
## [1] 0.015
```

```
prob <- exp(logit)/(1 + exp(logit))
formatC(prob, digits = 2, format = "f")
## [1] "0.50"
```

(c) Quel est le logit pour un étudiant ayant obtenu son diplôme secondaire en à l'étranger, et dont la spécialisation est classique ou latine ?
Quelle est alors sa probabilité de réussir en 1999 ?

Solution: Nous utilisons un contraste indicateur. Nous additionnons donc la constante et la contribution au logit des valeurs prises par chacune des variables.
La modalité `classique ou latine` étant la modalité de référence, son coefficient est nul par hypothèse. Nous avons donc :

```
logit <- 0.752 + 0.069 + 0
logit
## [1] 0.821
```

```
prob <- exp(logit)/(1 + exp(logit))
formatC(prob, digits = 2, format = "f")
## [1] "0.69"
```

(d) Que constatez-vous ?

Solution: L'étudiant ayant passé une spécialisation `autre` (en Suisse hors Genève) a une probabilité de réussite de 0.50, tandis que l'étudiant ayant passé une spécialité classique ou latine (à l'étranger) a une probabilité de réussite de 0.69, ce qui n'est pas trois fois plus que la probabilité de réussite avec une spécialité autre.

Remarque : dans ces deux derniers calculs nous avons utilisé pour calculer le logit des coefficients qui n'étaient pas significatifs, nous ne pouvons donc rien conclure sur la population général. Le résultat que nous venons d'effectuer ne s'applique qu'à notre échantillon (mais est bien valide dans notre échantillon : nous l'avons observé).

Exercice 6.44 (Prise en compte des effets d'interaction)**.**
Dans le TP précédent nous avons découvert, grâce à l'analyse par arbres, des effets d'interaction qui pouvaient se révéler pertinents pour comprendre les comportements sociaux que nous souhaitons expliquer dans notre problématique. Nous allons maintenant pouvoir quantifier l'impact de ces interactions grâce à l'analyse par regression.

Nous nous proposons de tester les interactions suivantes :
– entre le tronc commun et le lieu du diplôme ;
– entre le tronc commun et l'âge.

Remarque : nous n'allons pas pouvoir ici tester exactement les interactions découvertes dans l'arbre, il faudra pour cela changer les catégories de références et faire un recodage plus sophistiqué, ce qui sera abordé lors de la prochaine séance seulement.

Sous R, et dans la librairie `Dataset`, une interaction entre deux variables se note : `variable1:variable2` et s'ajoute dans la formule à la liste des variables que nous souhaitons tester.

1. Calculez le modèle de regression en utilisant les variables `troncom`, `lieudi.3` et la variable d'interaction entre ces deux variables.

Solution:

```r
reg3 <- reglog(
  formula = bil_99 ~ troncom + lieudi.3 + troncom:lieudi.3,
  target = 'réussi',
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```r
print(reg3)
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                              Model 1
## troncomsc.sociales                         0.954
## lieudi.3Suisse hors GE                     1.455
## lieudi.3Étranger                           0.291 **
## troncomsc.sociales:lieudi.3Suisse hors GE 1.048
## troncomsc.sociales:lieudi.3Étranger       3.392 *
## (Intercept)                                1.375 +
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                   Model 1
## Deviance       607.61
## Deviance H0    624.18
## Model Chi2      16.58 **
## Model DF         5.00
## Block Chi2      16.58 **
## Block DF         5.00
## R2 Cox-Snell     0.04
## R2 Nagelkerke    0.05
## N parameters     6.00
## AIC            619.61
## BIC            644.37
## N              458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

2. Vérifiez la significativité du modèle, puis du coefficient de la modalité croisée `sc. sociales : étranger`.

Solution: Le modèle apporte une information significative à hauteur de 1%. La constante est aussi significative à hauteur de 5%.

3. Précédemment, lorsque nous avons voulu calculer la valeur du coefficient pour un profil en particulier, nous avons additionner des contributions au logit. Ici, nous allons directement travailler avec les rapports de cotes. Dans la regression logistique, un rapport de cotes est donné en prenant l'exponentiel de sa contribution au logit estimée. De plus, comme nous avons la propriété suivante :

$$\exp(a+b) = \exp(a) * \exp(b)$$

nous devrons multiplier les rapports de cotes entre eux pour calculer celui d'un profil particulier.
Calculez le rapport de cotes estimé pour la modalité croisée. Interprétez.
Remarques :
– la catégorie de référence pour une interaction est le croisement des catégories de références des variables entrant en jeu dans l'interaction ;
– Nous multiplions le coefficient de l'interaction avec celui des effets propres, soit ici

Solution: Nous multiplions le coefficient de l'interaction avec celui des effets propres, soit ici :

```
3.39 * 0.95 * 0.29
```

```
## [1] 0.9339
```

Attention : ici nous ne prenons pas en compte la constante : nous voulons simplement le coefficient estimé pour la modalité croisée. Si nous multiplions aussi par la constante, nous obtenons le coefficent estimé pour un individu ayant le profil donné par la modalité croisée, mais ce n'est pas ici la question.

4. Que concluez-vous ?

Solution: Le rapport de cotes est très proches de 1, nous comprenons donc que le fait que le lieu du diplôme secondaire soit situé à l'étranger n'a pas un effet spécifique pour les étudiants en sciences sociales, par rapport à l'effet déjà existant sur l'ensemble des étudiants.

5. Calculez le coefficient estimé pour la modalité croisée `sciences eco + HEC : Étranger`. Effectuez pour cela un changement de catégorie de référence (attendre pour cela la correction de la prochaine séance)

Solution:

```
reg2.soc <- reglog(
  bil_99 ~ troncom + lieudi.3 + troncom:lieudi.3,
  target = 'réussi',
  reference = list("troncom" = "sc.sociales"),
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```
reg2.soc
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                                Model 1
## troncomsc.écon. + HEC                        1.048
## lieudi.3Suisse hors GE                       1.525
## lieudi.3Étranger                             0.987
## troncomsc.écon. + HEC:lieudi.3Suisse hors GE 0.954
## troncomsc.écon. + HEC:lieudi.3Étranger       0.295 *
## (Intercept)                                  1.311
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                   Model 1
## Deviance       607.61
## Deviance H0    624.18
## Model Chi2      16.58 **
## Model DF         5.00
## Block Chi2      16.58 **
## Block DF         5.00
## R2 Cox-Snell     0.04
## R2 Nagelkerke    0.05
## N parameters     6.00
## AIC            619.61
## BIC            644.37
## N              458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

Le coefficient de la modalité croisée `sciences eco + HEC : Étranger` est donné par :

```
0.295 * 1.48 * 0.987
```

```
## [1] 0.4309
```

Cette fois le coefficient est très différent de 1 (significativité à évaluer...), il semble donc y avoir un effet spécifique du fait que le lieu du diplôme soit situé à l'étranger pour les étudiants en sciences économiques.

6. Calculez le modèle de regression pour tester la seconde interaction. Cette fois, mettez uniquement la variable d'interaction dans le modèle.

Solution:

```
reg4 <- reglog(
  formula = bil_99 ~ troncom*age98.3,
  target = 'réussi',
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```
print(reg4)
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                      Model 1
## troncomsc.sociales                 0.823
## age98.3(19,23]                     0.490 +
## age98.3(23,25]                     0.155 **
## troncomsc.sociales:age98.3(19,23] 1.654
## troncomsc.sociales:age98.3(23,25] 6.769 *
## (Intercept)                        2.154 *
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                    Model 1
## Deviance        611.76
## Deviance H0     624.18
## Model Chi2       12.42 *
## Model DF          5.00
## Block Chi2       12.42 *
## Block DF          5.00
## R2 Cox-Snell      0.03
## R2 Nagelkerke     0.04
## N parameters      6.00
## AIC             623.76
## BIC             648.52
## N               458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

7. Regardez la liste des variables utilisées dans le modèle. Que remarquez-vous ?

Solution: Nous remarquons que les variables `troncom` et `age98.3` ont d'office été intégrées au modèle. Ceci est effectivement indispensable, nous devons tester les variables seules pour pouvoir quantifier l'effet porté par l'interaction en elle-même.

8. Vérifiez la significativité du modèle, puis du coefficient de la modalité croisée `sc. sociales : (23,25]`. Quelle est la valeur de ce coefficient ? Interprétez.

Solution:

```
6.769 * 0.155 * 0.823
```

```
## [1] 0.8635
```

L'effet est présent mais pas très prononcé (coefficient relativement proche de 1).
Ceci signifie que l'effet d'âge (les moindres chances de réussite de ceux ayant entre 23 et 25 ans) semble affecter principalement les individus de sciences économiques.

9. Auriez-vous pensez de vous même à ces hypothèses ? Concluez alors sur l'intérêt de l'analyse par arbres.

Solution: Nous pouvons comprendre a posteriori l'existence de ces effets et mettre en place des justifications socio-économiques, mais nous n'aurions pas forcément pensé de nous-même à poser ces hypothèses.
L'analyse par arbre est ainsi très efficace pour trouver de nouveaux déterminants sociaux, de nouvelles nuances, pour mieux expliquer les comportements sociaux observés. Grâce à l'analyse par arbre combinée à la regression, nous avons ici abouti à une analyse plus fine des déterminants sociaux.
En pratique, on pourra donner à analyser à l'arbre un grand nombre de variables (entre une vingtaine et une centaine), et le laisser détecter les plus associées associés, et les principaux effets d'interaction.

Exercice 6.45 (Effets transverses, variables de contrôle, modèle complet)**.**

1. Reprenez les deux premiers modèles calculés : `formula = bil_99 ~ lieudi3` et `formula = bil_99 ~ lieudi3 + diplse3`. Que remarque-t-on sur la significativité des coefficients de la variable `lieudi3` ?

Solution: Nous remarquons que les coefficients estimés pour la variable du lieu du diplôme sont significatifs lorsque l'on teste la variable seule, mais ne le sont plus lorsque l'on ajoute la variable du type du diplôme.

2. Comment expliquez vous cela ?

Solution: Nous pouvons comprendre que finalement, ce n'est pas tant le lieu du diplôme qui impacte sur la réussite, mais le type de diplôme secondaire que l'on a soutenu. Comme la variable du lieu du diplôme était significative seule, et ne l'est plus en ajoutant le lieu du diplôme, cela signifie qu'elle porte en elle-même une partie de l'information du type de diplôme : le lieu où l'on réalise son diplôme secondaire influe sur le type de diplôme que nous allons soutenir.
Le fait de capter à l'intérieur d'une variable l'effet d'une autre variable s'appelle un *effet transverse*.

3. En déduire pourquoi nous avons intérêt à faire une régression multiple même si nous nous intéressons uniquement à tester une hypothèse bivariée.

Solution: Nous avons en pratique une forte probabilité de se retrouver face à des effets transverses, car les différents aspects de la vie d'un individu sont souvent interconnectés.
En ajoutant d'autres de variables au modèle, nous allons pouvoir limiter les effets transverses et mesurer l'impact réel des différentes modalités de la variable explicative sur la modalité à prédire, au risque cependant d'introduire de la multicolinéarité.

4. Les variables utilisées pour contrôler les effets transverses sont appelées *variables de contrôle*. Dans le modèle précédent, si nous testons une hypothèse sur le lieu du diplôme, alors la variable du type de diplôme est une variable de contrôle, et si nous testons une hypothèse sur le type de diplôme, alors c'est la variable du lieu diplôme qui est une variable de contrôle.

5. Il est alors tentant d'ajouter beaucoup de variables de contrôle pour limiter au maximum les effets transverses, mais pouvons nous ajouter autant de variables de contrôle que nous le voulons ?

Solution: Non, il y aurait trop de coefficients à estimer et nous nous retrouverions avec un modèle qui ne serait plus significatif, par manque d'individus.

6. On appelle *modèle complet* le modèle contenant toutes les variables explicatives liées à notre problématiques, ainsi que les variables de contrôles supplémentaires qui ne feraient pas parties des hypothèses à tester, mais sur lesquelles nous voulons contrôler les effets transverses. Par exemple, peut-être nous n'avons aucune hypothèse à tester sur l'impact du genre, mais il semble tout de même raisonnable d'ajouter la variable `sexe` à notre modèle pour éviter de capter des effets absorbés par le genre. **C'est sur ce modèle complet qu'il faut effectuer les interprétations.**
Estimez le modèle complet pour notre problématique : nous testons des hypothèses sur le tronc commun suivi, le lieu du diplôme, le type de diplôme secondaire, le lieu de résidence de la mère, la profession du père, et nous contrôlons en plus par le sexe et l'âge. Nous testons aussi les interactions découvertes dans notre analyse par arbres.

Solution:

```
reg5 <- reglog(
  formula = bil_99 ~
    troncom +
    lieudi.3 +
    diplse.3 +
    domicile.mere.3 +
    profpere.4 +
    sexe +
    age98.3,
  target = 'réussi',
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```
reg5
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                              Model 1
## troncomsc.sociales                          1.134
## lieudi.3Suisse hors GE                      4.196 +
## lieudi.3Étranger                            2.069
## diplse.3moderne, scientifique, économique   0.577 +
## diplse.3autre                               0.302 *
## domicile.mere.3Suisse hors GE               0.339
## domicile.mere.3Étranger                     0.574
## profpere.4artisan, commerçant, cadre moyen  0.712
## profpere.4cadre supérieur, prof.libérale    0.781
## profpere.4sans profession + chômeur         0.933
## profpere.4non renseigné                     0.434 +
## sexefemme                                   0.935
## age98.3(19,23]                              0.593 *
## age98.3(23,25]                              0.476 +
## (Intercept)                                 4.058 ***
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                   Model 1
## Deviance       579.40
## Deviance H0    607.26
## Model Chi2      27.86 *
## Model DF        14.00
## Block Chi2      27.86 *
## Block DF        14.00
## R2 Cox-Snell     0.06
## R2 Nagelkerke    0.08
## N parameters    15.00
## AIC            609.40
## BIC            670.90
## N              446.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

7. Que remarquez-vous sur la significativité des coefficients ?

Solution: La plupart des coefficients ne sont pas significatifs. Ceci peut signifiquer que :
- nous avons intégré trop de variables au modèle ;
- certaines de nos variables ne sont pas pertinentes ;
- certains de nos recodages ne sont pas pertinents ;
- nous manquons d'individus pour réaliser une analyse aussi fine ;
- nous avons des effets de multicolinéarité entre certaines variables.

Remarque : en ce sens, à la prochaine séance, nous aborderons l'imbrication des

modèles permettant de tester l'apport de chaque variable ou de groupes de variables.

À préparer pour la prochaine séance :

Partie I – Pratique de l'analyse par régression logistique
Par groupe :
– Posez deux hypothèses sur le choix du tronc commun (variable `troncom`), l'une liée au type de diplôme secondaire (variable `diplse3`), l'autre liée à la profession du père (variable `profpere4`).
– Si possible, définissez aussi une hypothèse alternative pour chacune des hypothèses. (Si vous manquez de temps pour votre étude de cas, passez cette question)
– Étudiez ensuite ces hypothèses à l'aide d'un modèle de régression logistique multiple, en contrôlant de plus par le sexe et l'âge.

Remarque : veillez bien sûr à vérifier la significativité de votre modèle et de chaque coefficient que vous allez interpréter.
Envoyez-nous par courriel les réponses à ces questions, en joignant la vue synthétique en PDF de votre modèle de régression.

Partie II – Avancement sur votre étude de cas
Pour cette semaine, nous vous conseillons de travailler de la manière suivante (en supposant que le travail suggéré la semaine dernière a été réalisé) :

1. Rédiger la partie 3 de votre étude de cas (voir la section « Directives pour l'étude de cas ») en vous aidant des notes prises au brouillon pendant la réalisation des analyses bivariées et commencer la rédaction de la partie 5.
2. Écrire le code R réalisant l'ensemble des analyses par arbres que vous devez effectuer pour tester vos hypothèses.
3. Lister l'ensemble des interactions que vous aurez détectées pertinentes pour répondre à votre problématique (il se peut aussi qu'il n'y en ait pas).
4. Noter au brouillon vos premières interprétations sur les analyses par arbre effectuées.

Travaux pratiques – Séance 7

Objectifs de la séance : Régression logistique niveau 2 : modèles imbriqués, χ^2 du rapport de vraisemblance entre modèles, χ^2 du rapport de vraisemblance pour une variable, test de l'apport d'une variable, évaluation globale du modèle (χ^2, déviance, AIC, BIC, etc.). Contraste et changement des catégories de référence. Citation de `R` et des librairies `R` utilisées dans un travail.

Exercice 7.46 (Chargement d'une base de donnée préalablement exportée)**.**
Dans cette séance, nous avons besoin de la base de données que vous avez préparée pour les analyses, et que vous avez exporté sous le nom `ses98.study`.

1. Chargez la librairie `Dataset`.

Solution:

```
library(Dataset)
```

2. Définissez comme répertoire de travail le dossier contenant votre base `ses98.study`. Si vous avez suivi nos conventions, cela devrait être le dossier `AnCat`.

Solution:

```
setwd("H:/AnCat/")
```

3. Chargez la base de donnée.

Solution:

```
load("ses98.study.RData")
```

4. Vérifiez que la base est bien présente en mémoire.

Solution:

```
"ses98.study" %in% ls()
## [1] TRUE
```

Exercice 7.47 (Modèles imbriqués)**.**
Dans la séance précédente nous avons vu que les conclusions que nous allons donner aux hypothèses que nous testons doivent s'appuyer sur le modèle complet. Nous abordons ici la question du choix des variables que nous allons retenir pour la construction du modèle complet. En effet, lorsque nous avons plusieurs variables, comment pouvons nous choisir celles qui sont réellement pertinentes pour l'analyse ? Même si le modèle est significatif, il peut y avoir des variables qui n'apportent que peu d'information : si d'autres variables sont très significatives, elles feront conserver la significativité du modèle complet même en présence des autres. Grâce aux analyses bivariées, nous avons pu évaluer l'association de chaque variable explicative avec la variable dépendante, et peut-être en avons nous déjà éliminée. Avec l'analyse des résidus, nous avons aussi pu découvrir des modalités ayant un effet particulièrement fort sur une ou plusieurs modalités de la variable à expliquer. L'imbrication de modèles va ici se révéler être un outil complémentaire et très efficace pour nous permettre de tester l'apport d'une variable ou d'un groupe de variables, par rapport à ce que nous avons déjà réussi à expliquer avec un modèle plus simple.
L'imbrication de modèles consiste à construire plusieurs modèles à partir d'un modèle initial, en ajoutant au fur et à mesure de nouvelles variables explicatives. Chaque ajout de variable est appelé un *bloc*. Pour chacun des modèles calculés nous regarderons les critères de qualité des modèles, pour évaluer la pertinence des variables que nous avons ajoutées, et en comparant un modèle au modèle précédent, nous pouvons évaluer si le bloc ajouté apporte suffisamment d'information par rapport à la complexité supplémentaire qu'il génère.

Les modèles imbriqués ont aussi un autre rôle, qui va être guidé par nos hypothèses socio-économiques : tester l'importance de différents aspects du profil des individus sur le comportement social que nous souhaitons expliquer. Par exemple, supposons que nous nous intéressons à l'opinion politique (gauche/droite) des individus, et que nous avons posé des hypothèses sur la situation professionnelle de l'individu ainsi que ses pratiques religieuses, pour expliquer cet opinion. Nous pourrions mettre toutes les variables explicatives directement dans le même modèle. Cependant, il serait plus intéressant de regrouper ces variables suivant aspect qu'elles concernent, et d'imbriquer les blocs pour évaluer l'importance de chacun de ces aspects dans l'opinion de l'individu. Par exemple et nous pouvons imaginer un premier bloc contenant les variables démographiques (âge, sexe, statut marital, etc.), un second bloc décrivant la situation professionnelle (revenu, CSP, etc.), et un dernier bloc contenant les variables liées aux pratiques religieuses (fréquence de culte, type de religion, etc.).
Dans la librairie `Dataset`, l'imbrication de modèle se réalise de la manière suivante :

```
reglog(
  formula = MODELE INITIAL,
  nested = list(
    . ~ BLOC 1,
    . ~ BLOC 2,
    etc.
  ),
  target = CATÉGORIE CIBLÉE,
  data = BDD
)
```

Le '.' signifiant que nous réutilisons la même variable cible pour chacun des modèles.

1. Pour expliquer la réussite des étudiants, réalisez une imbrication de modèles, avec en premier bloc les variables démographiques et du cursus suivi (`sexe`, `age98.4` et `troncom`), un second bloc traitant du diplôme secondaire obtenu (`diplse3` et

`lieudi3`), et un dernière bloc sur l'effet du statut parental (modélisé uniquement par `profper4`).

Solution:

```
nested1 <- reglog(
  bil_99 ~ sexe + age98.3 + troncom,
  nested = list(
    . ~ diplse.3 + lieudi.3,
    . ~ profpere.4
  ),
  target = 'réussi',
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```
nested1
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                                Model 1    Model 2    Model 3
## sexefemme                                     0.948      0.939      0.941
## age98.3(19,23]                                0.656 +    0.609 *    0.628 +
## age98.3(23,25]                                0.467 +    0.552      0.566
## troncomsc.sociales                            1.382      1.163      1.175
## diplse.3moderne, scientifique, économique                0.641      0.613 +
## diplse.3autre                                            0.347 *    0.348 *
## lieudi.3Suisse hors GE                                   1.434      1.456
## lieudi.3Étranger                                         1.088      1.100
## profpere.4artisan, commerçant, cadre moyen                          0.712
## profpere.4cadre supérieur, prof.libérale                            0.773
## profpere.4sans profession + chômeur                                 0.924
## profpere.4non renseigné                                             0.511 +
## (Intercept)                                   1.630 +    2.808 **   3.542 **
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
##
## Table 2:
## Quality measures
```

```
##                    Model 1     Model 2     Model 3
## Deviance        617.00      602.60      599.06
## Deviance H0     624.18      624.18      624.18
## Model Chi2        7.18       21.58 **    25.12 *
## Model DF          4.00        8.00       12.00
## Block Chi2        7.18       14.40 **     3.54
## Block DF          4.00        4.00        4.00
## R2 Cox-Snell      0.02        0.05        0.05
## R2 Nagelkerke     0.02        0.06        0.07
## N parameters      5.00        9.00       13.00
## AIC             627.00      620.60      625.06
## BIC             647.64      657.74      678.71
## N               458.00      458.00      458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

2. Générez la vue PDF de cette analyse.

Solution:

```
exportPDF(nested1)
```

3. Parmi les trois modèles calculés, lesquels sont significatifs au sens du $\chi 2$ du rapport de vraisemblance ?

Solution: Les modèles 2 et 3 apportent une information significative au seuil de 5%. Le premier modèle n'apporte pas d'information significative.

4. Quel est la significativité de chacun des deux blocs ajoutés ?

> **Solution:** Le premier bloc, qui concerne le diplôme secondaire, amène un χ^2 d'environ 14.40, pour 4 degrés de libertés, ce qui est significatif au seuil de 5%.
> Le second bloc, qui concerne la profession du père, amène un χ^2 d'environ 3.54, pour 4 degrés de libertés, ce qui n'est pas significatif.

5. Que concluez-vous sur l'intérêt du troisème bloc ?

> **Solution:** Le troisième bloc n'apporte pas d'information significative, il n'est donc pas intéressant de prendre la variable `profpere4` dans le modèle complet.

6. La fonction a retourné les valeurs des pseudo-R^2 de Cox-Snell et de Nagelkerke, lequel préférez-vous interpréter ? (indice : voir dans le cours)

> **Solution:** Le pseudo-R^2 de Cox-Snell n'atteint pas nécessairement son maximum en 1, il peut l'atteindre pour une valeur inférieure. Le pseudo-R^2 de Nagelkerke quant à lui a été normalisé pour atteindre son maximum en 1. Nous le préférerons donc pour comparer les modèles.

7. Quel est le modèle ayant la valeur de pseudo-R^2 la plus haute ?

> **Solution:** C'est le troisième modèle, avec une valeur de 0.07, ce qui est tout de même relativement faible.

> **Solution:** C'est le troisième modèle, avec une valeur de 0.07, ce qui est tout de même relativement faible.

8. Pouvez-vous suivre la valeur du pseudo-R^2 et choisir pour l'interprétation des résultats le modèle avec le pseudo-R^2 systématiquement le plus grand ?

> **Solution:** Non, nous venons de voir que l'information apportée par le troisième modèle n'est pas significative. Il serait donc dangereux d'utiliser le bloc supplémentaire dans l'analyse des résultats.
> Nous avons ajouté une variable explicative, il est donc évident que cela va améliorer le R^2, puisque nous utilisons plus d'information pour expliquer la modalité cible. La question est justement « est-ce que l'on ne paye pas trop cher cette information ? », et la non significativité du bloc nous fait comprendre que cela semble être le cas : nous complexifions trop le modèle pour le peu d'information apportée.

9. Le critère AIC permet d'évaluer l'information apportée par un modèle pénalisée par sa complexité. La valeur du AIC ne s'interprête pas, elle doit seulement être utilisée pour comparer les modèles entre eux. En particulier, elle permet, contrairement à la déviance, de comparer des modèles non imbriqués.
   Nous allons ici comparer les AIC des différents modèles.
   (a) Quel est l'écart entre les deux premier modèles ? Est-ce un écart important ?
   (b) Quel est l'écart entre le deuxième et le troisième modèle ? Est-ce un écart important ?

(c) En suivant le critère AIC, quel modèle vous semble le plus pertinent ?

(d) Est-ce cohérent avec ce que nous avions conclu en regardant le χ^2 du rapport de vraisemblance ?

Exercice 7.48 (Changement des catégories de références)**.**
Pour chaque variable, le choix de la catégorie de référence doit se faire en tenant compte ces deux points :

- La population de référence dans l'hypothèse que vous voulez tester. Par exemple, si vous vous intéressez à l'effet d'avoir une pratique religieuse sur la variable cible, il pourrait être intéressant de prendre en catégorie de référence « ne pas avoir de pratique religieuse », ainsi vous testez l'apport d'avoir une pratique par rapport à ne pas en avoir une.
- Le nombre d'individus concernés par la modalité. En effet, les individus de la catégorie de référence seront utilisés pour l'estimation de chaque coefficient estimé pour les autres catégories. Alors, si vous prenez comme catégorie de référence une catégorie assez peu d'individus, vous risquez de ne pas obtenir de coefficients significatifs, par manque d'individus pour valider les résultats.

Souvent, et sauf si nos hypothèses nous suggèrent une autre catégorie, nous prenons en catégorie de référence la catégorie la plus fréquente. Ainsi, nous limitons au mieux les problèmes d'estimation, et socio-économiquement cela fait sens : si la catégorie est la plus fréquente c'est a priori que c'est un comportement « de référence » dans la population étudiée.
Dans la librairie `Dataset`, si vous ne spécifiez rien, la première catégorie sera utilisée comme catégorie de référence. Pour connaître l'ordre des catégories, utilisez la fonction `valids`. Si ce choix ne vous convient pas, vous pouvez changer la catégorie de référence en fournissant l'argument `reference` comme suit :

```
reglog(
  formula = MODELE,
  target = CATÉGORIE CIBLÉE,
  reference = list(
    "VARIABLE A" = "MODALITE REFERENCE POUR A",
    "VARIABLE B" = "MODALITE REFERENCE POUR B",
    etc.
  ),
  data = ses98.study
)
```

1. Calculez le modèle de regression multiple avec les variables `lieudi.3` et `diplse.3`, comme nous l'avions calculé au précédent TP, mais en utilisant cette fois la modalité `sc.sociales` comme modalité de référence pour la variable `troncom`.

Solution:

```r
reg2.lieudi.3 <- reglog(
  formula = bil_99 ~ lieudi.3 + diplse.3,
  target = 'réussi',
  reference = list("lieudi.3" = "Suisse hors GE"),
  data = ses98.study
)
```

```
## Logistic regression model (currently not-weighted)
##
## number of missings: 0 ( 0 %)
## Here is the allocation of the rows in the different classes.
```

```
##
##              1   0
##   echec      0 115
##   redouble   0  79
##   réussi   264   0
```

```r
reg2.lieudi.3
```

```
## Table 1:
## Estimated coefficients (odds ratios)
```

```
##                                              Model 1
## lieudi.3GE                                 0.700
## lieudi.3Étranger                           0.750
## diplse.3moderne, scientifique, économique 0.602 +
## diplse.3autre                              0.335 *
## (Intercept)                                3.029 ***
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1
##
## Table 2:
## Quality measures
```

```
##                   Model 1
## Deviance       607.57
## Deviance H0    624.18
## Model Chi2      16.61 **
## Model DF         4.00
## Block Chi2      16.61 **
## Block DF         4.00
## R2 Cox-Snell     0.04
## R2 Nagelkerke    0.05
## N parameters     5.00
## AIC            617.57
## BIC            638.21
## N              458.00
```

```
## *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, " = NA
```

2. Quel est le rapport de cotes pour un étudiant ayant obtenu son diplôme secondaire en Suisse hors Genève et dont la spécialisation est autre ?

Solution: La catégorie `Suisse hors GE` étant la catégorie de référence, son rapport de cotes est 1. Nous multiplions les rapports de cotes et obtenons :

```
rc <- 3.029 * 1 * 0.335
rc
```

```
## [1] 1.015
```

Nous pouvons aussi calculer la probabilité de réussir pour un individu de ce profil :

```
prob <- rc/(1 + rc)
formatC(prob, digits = 2, format = "f")
```

```
## [1] "0.50"
```

3. Comparez avec le résultat obtenu lors de la séance précédente, et vérifiez que vous obtenez la même valeur.

Solution: Nous avions calculé le logit et avions trouvé :

```
logit <- 0.752 + 0.356 + (-1.093)
```

Alors le rapport de cotes est :

```
rc <- exp(logit)
rc
```

```
## [1] 1.015
```

Nous obtenons donc bien le même résultat.

Exercice 7.49 (Citations du matériel utilisé)**.**
Citer le matériel que vous utilisez est impératif pour trois principales raisons :

1. *Communiquer la littérature.* C'est-à-dire permettre à votre lecteur, qui ne connaît peut-être pas votre littérature ou les logiciels utilisez, de les retrouver pour lui même pouvoir parfaire ses connaissances, et reproduire vos analyses. Vous pourriez dire « il suffit de chercher dans Google », certes oui en cherchant dans Google on peut retrouver des informations, mais on peut en trouver beaucoup sur le même sujet, un auteur peut par exemple avoir écris plusieurs articles sur un même sujet, qui parfois se remettent en cause l'un l'autre. Il faut être précis avec le lecteur, ne pas lui faire perdre son temps, et ne pas le laisser croire que vous avez mal compris le travail de quelqu'un, alors qu'en réalité, c'est simplement que le lecteur n'a pas retrouvé le bon document auquel vous vouliez faire référence.
2. *Donner du crédit à votre travail.* Vous devez en effet obtenir le réflexe que chaque affirmation que vous faites ne doit pas être gratuite. Chaque affirmation doit : soit être appuyée par un article publié (cela veut dire qu'il a été relu, vérifié, et accepté par d'autres chercheurs du domaine, on peut donc faire confiance aux affirmations qu'il contient), soit être démontré vous même par une analyse rigoureuse (ce que vous avez appris à faire durant ce séminaire `AnCat`). Vous montré ainsi au lecteur que vous ne donnez pas des conclusions au hasard, mais en connaissance de cause.
3. *Favoriser la recherche de qualité.* Si vous utilisez le travail de quelqu'un au sein du votre, c'est que vous le trouvez pertinent. Ainsi, en citant un papier, vous donnez une certaine reconnaissance au travail de l'auteur. Si un travail est de nombreuses fois cité, c'est que la

reconnaissance est assez unanime et donc que l'auteur a fait un travail de qualité. Cette « preuve de qualité par les citations »permettra alors à l'auteur de faire plus facilement accepter de nouveaux projets de recherche, et comme il produit du travail de qualité, cela favorisera l'apparition de nouveaux résultats de qualité.

Remarque : le troisième point est finalement une instanciation du principe de sélection naturelle, permettant d'orienter de manière naturelle la recherche vers les travaux de qualité. Nous retrouvons aussi ce système dans l'économie de marché : supposons que vous aimez bien un film et que vous l'achetiez, cet achat a deux effets ; D'une part, il vous permet d'acquérir le droit de profiter du bien, alors que vous ne l'avez pas fabriqué vous-même et ne le possédez pas, et d'autre part cet achat permet de rémunérer l'auteur et de lui fournir les moyens de poursuivre son travail et de créer par la suite d'autres films que vous aimerez sans doute bien. Par contre, les films que vous n'avez pas aimé et alors pas acheté, n'ont pas profité d'un financement, et l'auteur aura alors peut-être du se ré-orienter et ne plus faire de film. Ainsi, en achetant les films que vous aimez bien, vous favorisez par sélection naturelle la création de films que vous aimez bien.

À travers ces trois points, vous comprenons que citer n'est pas un acte anodin.

Dans un travail socio-économique, on peut distinguer au moins trois types de matériel à citer :

- La littérature liée à votre problématique et sur laquelle vous appuyez vos raisonnements ;
- Les données que vous avez utilisées ;
- Les logiciels et méthodes que vous avez utilisés pour les analyses.

Dans votre étude de cas, et vos futurs travaux, il faudra bien penser à citer l'ensemble des éléments qui doivent l'être.

Dans le cas de `R`, un système spécifique permet de récupérer les citations, du logiciel en lui-même et des librairies externes. Nous allons voir ici comment l'utiliser.

1. Quels sont les logiciels que vous avez utilisé pour conduire vos analyses durant ces séances de séminaire ?

Solution:
- `R` comme logiciel d'analyses de données.
- La librairie `Dataset` pour la gestion de données d'enquête.
- Les librairies `rpart`, `party` et `CHAID`, pour l'implémentation des arbres (pour lesquelles `Dataset` fournit une interface homogénéisant leur utilisation).
- La librairie `stats`, pour la regression logistique.

2. Nous allons récupérer les citations pour chacun des logiciels utilisés durant ces séances de TP, afin de pouvoir les indiquer à notre rapport. Sous R, la fonction `citation` permet d'obtenir les citations de `R` et des packages installés.
   Remarque : lorsque vous écrivez un nouveau rapport, il faut éviter de reprendre les citations d'un précédent rapport. En effet, les citations peuvent changer au fil du temps, suivant l'avancée des travaux des chercheurs et développeurs, il faut donc reconsulter les citations à chaque nouveau travail.
   Lisez le manuel de la fonction `citation` pour connaître son fonctionnement.

```
?citation
```

3. Nous allons maintenant récupérer les citations pour les différents logiciels utilisés.
   (a) Récupérer la citation à utiliser pour citer `R` dans un travail.

Solution:

```
citation()
```

(b) Récupérer la citation à utiliser pour citer `Dataset`.

Solution:

```
citation("Dataset")
```

(c) Récupérer les citations à utiliser pour citer `rpart` et `CHAID`.

Solution:

```
citation("rpart")
citation("CHAID")
```

(d) Récupérer la citation à utiliser pour citer `stats`.

Solution:

```
citation("stats")
```

En haut du message affiché à l'écran, il est indiqué que le package `stats` fait partie intégrante de `R`, nous n'auront donc pas besoin de le citer spécifiquement, citer `R` suffit.

À préparer pour la prochaine séance :

Partie I – Pratique de l'analyse par régression logistique

Par groupe :

– Construire un modèle imbriqué pour évaluer les chances d'être étudiant en sciences sociales plutôt que d'être étudiant en sciences économiques (variable `troncom` en `target`) avec en premier bloc le sexe et l'âge (variables démographiques et de contrôle), en deuxième bloc le lieu du diplôme secondaire et le type de diplôme et en troisième bloc la réussite en 1999. Utilisez la modalité `GE` comme modalité de référence pour les variables `lieudi3` et `domicile.mere3`, et la modalité la plus fréquente pour chacune des autres variables.

– À la vue des critères de qualités des modèles, lequel vous semble le plus pertinent ? (justifier)

– Quel est la variable impactant le plus le choix du tronc commun ?

– Nous souhaitons tester l'hypothèse suivante en italique :

« Nous faisons l'hypothèse que les étudiants étranger effectuant leurs études dans le canton de Genève choisissent principalement l'Université de Genève pour réaliser leurs études (note : c'est sans doute faux). En raison de la forte notoriété de Genève comme pôle de compétence au niveau économique et financier, nous faisons l'hypothèse que *les étudiants étranger choisissent davantage un tronc commun économique plutôt que sciences sociales.* »

À l'aide de votre modèle, quelle conclusion apportez-vous ?

– Quel bloc avez-vous utilisé pour faire cette conclusion ?

Envoyez à Danilo et Emmanuel la vue synthétique en PDF du modèle de régression généré, ainsi que les réponses aux questions.

Partie II – Avancement sur votre étude de cas

Pour cette semaine, nous vous conseillons de travailler de la manière suivante (en supposant que le travail suggéré la semaine dernière a été réalisé) :

1. Rédiger la partie 4 de votre étude de cas (voir la section « Directives pour l'étude de cas ») concernant les analyses par arbre.
   Remarque : concernant les interactions, il ne faut bien sur analyser que celles qui vous permettent de tester ou d'avoir des résultats plus fins sur vos hypothèses. Si vous avez des interactions qui sortent de vos questions de recherche, mais que vous pensez tout de même pertinentes pour aborder la problématique sous un autre angle, il ne faut pas en parler maintenant, mais plutôt imaginer une section « Perspectives » à votre rapport, dans laquelle vous appuirez le bien-fondé des questions de recherche que vous proposez à l'aide de ces interactions découvertes.
2. Écrire le code R réalisant l'ensemble des analyses par arbres que vous devez effectuer pour tester vos hypothèses.
3. Lister l'ensemble des interactions que vous aurez détectées pertinentes pour répondre à votre problématique (il se peut aussi qu'il n'y en ait pas).
4. Noter au brouillon vos premières interprétations sur vos analyses de régression.

Partie III (Optionnelle) – Préparer une courte présentation orale du travail

Afin de vous entraîner à la présentation orale d'un travail, nous vous proposons de préparer quelques transparents (4 ou 5 par exemple) que vous nous présenterez en quelques minutes (5 à 10 minutes).

Ceci sera une présentation totalement informelle, sans aucun stress, et il n'y a pas besoin

d'avoir fini le travail pour le présenter, vous pouvez présenter simplement les résultats que vous avez déjà.
Nous pourrons vous formuler des retours sur le travail que vous aurez déjà effectué, ce qui pourra se révéler constructif pour terminer votre travail.
Remarque : Il ne *faut pas* préparer de transparents pour la défense orale. Lors de la défense orale, nous aurons lu votre travail, nous n'aurons donc pas besoin qu'il nous soit présenté, nous démarrerons directement avec les questions qui auront été soulevées durant notre lecture.

Annexe 1

Installation de l'environnement logiciel sur votre ordinateur personnel
Windows

Installation de R

Le logiciel de statistique réalisant les calculs est R. Vous pouvez le télécharger à l'adresse http://www.r-project.org/.

- Rendez-vous dans la section `CRAN`
- Choisissez le CRAN pour la Suisse : http://stat.ethz.ch/CRAN/
- Choisissez `Download R for Windows`
- Puis `base`
- Sélectionnez enfin `Download R 3.0.1 for Windows`, ce qui lancera le téléchargement de l'installateur.

Installez R en double cliquant sur l'installateur et en suivant la procédure d'installation proposée par défaut.

Installation de Rstudio

L'interface utilisateur que nous utilisons pour travailler avec R est Rstudio. Rendez-vous sur http://rstudio.org/download/, télécharger la version `Desktop`, et suivez la procédure d'installation par défaut.

Installation de la librairie Dataset

La librairie `Dataset` est située sur le dépôt `r-forge`. Les librairies sur ce dépôt sont disponibles uniquement pour la version de R courante. Avant d'installer la librairie, vérifier bien que la dernière version de R est installée sur votre machine. À la date de septembre 2013, la version courante de R est la `3.0.x`

Pour installer la librairie `Dataset`, lancez Rstudio et executez la commande suivante :

```
install.packages("Dataset", repos = "http://r-forge.r-project.org")
```

Remarque : le dépôts `r-forge` ne compile plus les librairies pour MacOS, vous devez donc installer le package à partir des fichiers sources :

```
install.packages("Dataset", repos = "http://r-forge.r-project.org",
    type = "source")
```

Installation de LaTeX pour les export PDF de Dataset

Les vues synthétiques des bases de données et résultats d'analyse produites par la librairie `Dataset` sont générés à l'aide de l'éditeur de document LaTeX. Pour profiter de ces fonctionnalités, vous devez donc installer une distribution LaTeX sur votre machine. Sous Windows, nous conseillons d'utiliser la distribution `Miktex`.

1. Rendez-vous sur www.miktex.org et accédez à la section *download*.
2. Télécharger `Miktex 2.9` et installez-le en suivant la procédure d'installation par défaut.

Annexe 2 - Librairie Dataset : nouvelles versions et aide utilisateur

Recevoir une alerte lorsqu'une nouvelle version est disponible

La librairie `Dataset` est régulièrement mise à jour, de nouvelles fonctionnalités apparaissent et les bogues découverts sont corrigés.

Vous pouvez être tenu au courant de la disponibilité d'une nouvelle version de la librairie en vous abonnant à la liste de diffusion `Dataset-updates`. L'inscription se fait sur la page suivante :

http://dataset-updates.eryx.org

Lorsqu'une nouvelle version est disponible, un message est envoyé à tous les abonnés de cette liste de diffusion. Le message contient en particulier la liste des nouvelles fonctionnalités et la liste des corrections apportées.

Recevoir de l'aide sur l'utilisation de la librairie

Comme la plupart des librairie `R`, la librairie `Dataset` fournit un manuel détaillé pour (presque) chaque fonctionnalité, et présente des exemples d'utilisation prêts à l'emploie. Cependant, pour certaines opérations plus spécifiques que vous pouvez être amené à réaliser, ce manuel ne suffit pas et une aide extérieur peut être requise. À cette fin, une liste de diffusion nommée `Dataset-users` est utilisée comme support de discussion entre les utilisateurs de la librairie. Vous pouvez y poser vos questions, et obtenir des réponses de la part des autres utilisateurs. L'inscription à cette liste se fait sur la page suivante :

http://dataset-users.eryx.org

Après vous êtes inscris, vous pouvez poser vos questions en envoyant un message à l'adresse

dataset-users@lists.r-forge.r-project.org

Remarque : sur cette liste vous êtes à la fois poseur de questions et donneur de réponses ! N'hésitez donc pas à aider les autres utilisateurs lorsque vous avez la réponse à leur question.

Faire une demande de fonctionnalité ou de correction de bogue

Vous souhaitez qu'une fonctionnalité soit ajoutée à la librairie, ou vous avez découvert un comportement suspect et vous vous demandez si ce ne serait pas un bogue, alors n'hésitez pas à en faire part à l'équipe de développement en envoyant un message à l'adresse

dataset-requests@lists.r-forge.r-project.org

Annexe 3 – Mise à jour d'une librairie R

Toutes plate-formes

R ainsi que les librairies développées dessus, sont en perpétuelle évolution, leurs développeurs ajoutant des fonctionnalités, améliorant des anciennes, et corrigeant les bugs découverts. Vous pouvez donc être amené à devoir mettre à jour une librairie que vous avez déjà d'installée sur votre machine, pour pouvoir profiter des fonctionnalités offertes par la nouvelle version.

Typiquement, mettre à jour une librairie consiste simplement à la réinstaller. Toutefois il est nécessaire de sortir la librairie de notre espace de travail avant de réaliser la réinstallation. En effet, lorsque l'on charge une librairie dans l'espace de travail, avec la fonction `library()`, R réserve les noms de fonctions définies dans la librairie. Si nous procédons à la réinstallation avec la librairie chargée, tout semblera se dérouler correctement, cependant rien n'aura été effectué, car les noms de fonctions étant déjà occupés, il n'aura pas pu effectuer les remplacements.

Détacher la librairie à mettre à jour (si attachée)

La première étape est donc de détacher la librairie en question de l'espace de travail. Ceci ce fait avec la fonction `detach`, qui s'utilise :

```
detach("package:NomDuPackage", unload = TRUE)
```

Cette étape n'est bien sûr inutile si la librairie n'est pas chargée dans votre espace de travail. Si tel est le cas, la fonction renverra une erreur, mais qui sera sans incidence pour la suite.

Relancer la commande d'installation de la librairie

La seconde étape est d'exécuter la commande d'installation de la librairie, comme pour une première installation.

Si la librairie est située sur le CRAN, vous pouvez directement écrire

```
install.packages("NomDuPackage")
```

Si la librairie est située sur R-forge, la commande s'écrit :

```
install.packages("NomDuPackage", repos = "http://r-forge.r-project.org")
```

Exemple avec la librairie `Dataset`

Voici un exemple complet pour la mise à jour de la librairie `Dataset` :

```
detach("package:Dataset", unload = TRUE)
install.packages("Dataset", repos = "http://r-forge.r-project.org")
```

Annexe 4 – Conseil pour la rédaction de l'étude de cas

Typographie

Avant d'utiliser un logiciel de traitement de texte (Word, Page, LaTeX, etc.), il est important de commercer par vérifier sa configuration typographique. Si le texte est écris en français, nous le configurons pour la typographie française, si le texte est écris en anglais, nous le configurons pour la typographie anglaises, etc.
Pour reconnaître dans quelle typographie nous sommes nous pouvons remarquer quer :

– [Listes d'énumération :] (FR) utilisation de tirets / (EN) utilisation de *bullets*.
– [Alinéas de paragraphe :] (FR) à chaque paragraphe / (EN) à chaque paragraphe sauf au premier.

Tableaux, graphiques, ...

Faîtes bien attention à ce que chaque table, graphique, ou autre structure de présentation de données :

– soit numéroté avec un identifiant unique.
– soit référencé dans le texte.
– possède une légende claire : on doit être capable d'interpréter l'élément sans avoir recours à des informations contenues dans le texte.

Page de garde

La page de présentation doit être faite rigoureusement. On doit avoir la date, le lieu, dans quel cadre le projet s'inscrit, les noms, prénoms des personnes ayant réalisées le dossier, les noms, prénoms des personnes ayant encadrés, et plus généralement les noms, prénoms de toutes les personnes ayant participées à la réalisation du dossier.

Introduction

En introduction, il est important de présenter le cadre du travail, par exemple à l'aide d'une phrase du style « Ce dossier a été réalisé dans le cadre du cours d'Analyse statistique de données catégorielles, du Master en ... ». En effet, le lecteur doit savoir dans quel contexte le document qu'il lit a été constitué.

Présentation des variables utilisées dans les analyses

Lorsque vous présentez les variables que vous allez utiliser pour vos analyses, pensez bien que :

– il est préférable de commencer par retirer les valeurs manquantes sur la variable réponse avant de calculer les répartitions des individus sur les modalités des variables explicatives, ainsi on a les répartitions des individus pour les personnes qui ont répondues à la variable réponse. Il n'est pas très intéressant de donner la distribution des modalités pour l'ensemble des individus, si il n'y en a que 60% concernés par la variable réponse. La distribution risquant en effet d'être significativement différente.
– il est intéressant d'indiquer le pourcentage de cas valides sur chacune des modalités prises par une variable, mais il est important d'écrire le nombre total d'individus valides (on pourrait créer des pourcentages même avec seulement 10 individus valides...).

Annexe 5 – Liste d'erreurs fréquentes

– Fonction `subset` : Sélection d'un sous-groupe de la population avec `subset`, pour faire un test d'égalité il faut utiliser `==` et non `=` qui lui signifie l'affectation (comme utilisé affecté les paramètres aux fonctions)
– Fonction `cut` : il ne faut pas donner les bornes min et max avec la méthode `cut` de `Dataset`, celles-ci sont calculées automatiquement. Il faut juste donner les points de coupure.

Annexe 6 – Importer les résultats des analyses dans MS Word

Toutes plate-formes

R est un logiciel libre, conçu à la base pour être utiliser dans un environnement de logiciels libres. Microsoft Word est un logiciel privé, et R n'offre pas de méthodes natives pour travailler avec, notamment pour la question des imports/exports. Mais R étant libre, d'autres utilisateurs ont développés des packages qui vont nous permettre tout de même de faire discuter R et Word.

La solution que nous allons utiliser ici est d'exporter nos résultats d'analyses sous forme de tables HTML : plusieurs packages de R permettent d'exporter vers du HTML et Word sait importer le HTML. Le package que nous allons utiliser est `R2HTML`.

```
install.packages("R2HTML")
```

```
library(R2HTML)
```

Export d'un data.frame

```
data(iris)
ex.table <- iris
HTML(ex.table, file = "test.iris.html")
```

Puis depuis word faire "Fichier > Importer" et sélectionner le fichier.

Export d'un objet Statdf

Sous Dataset, les tables d'analyses sont stockées dans des objets `Statdf` (comprendre `data.frame` pour des tables de statistiques), qui permettent de stocker des statistiques et leurs p-valeurs associées, et de paramétrer le formattage avec les étoiles.

Depuis Dataset, vous devez donc commencer par extraire des outputs des analyses les tables qui vous intéressent, au format `Statdf`, puis les formatter, et enfin les exporter en `data.frame` pour les exporter en HTML.

Par exemple pour bivan :

```
library(Dataset)
data(iris)
iris$Sepal.Length <- cut(iris$Sepal.Length,
    breaks = 4)
iris$Sepal.Width <- cut(iris$Sepal.Width,
    breaks = 4)
iris$Petal.Length <- cut(iris$Petal.Length,
    breaks = 4)
iris$Petal.Width <- cut(iris$Petal.Width,
```

```
    breaks = 4)
ir <- dataset(iris)
biv1 <- bivan(Species ~ Sepal.Length + Petal.Length,
    data = ir)
biv1
biv1.global <- global(biv1)
class(biv1.global)
biv1.global.formatted <- summary(biv1.global,
    merge = "left")
biv1.global.formatted.df <- v(biv1.global.formatted)  # conversion en data.frame
HTML(biv1.global.formatted.df, file = "test.iris2.html")
```

Vous pouvez aussi récupérer les tables suivantes :

Extraire les outputs de bivan

```
global(biv1)  # statistiques globales
std.res(biv1)  # résidus standardisés
observed(biv1)  # effectifs observés
expected(biv1)  # effectifs attendus sous hypothèse d'indépendance
```

Extraire les outputs de tree.*

Extraire les outputs de reglog

```
reg1 <- reglog(
  formula = Species ~ Sepal.Length + Petal.Length,
  target = 'setosa',
  data = ir
)
}
```

```
exportTAB(reg1) # export all tables contained in the object
exportTAB(reg1)$logit.contributions # contributions to logit
exportTAB(reg1)$odds.ratios # odds ratios
exportTAB(reg1)$quality.measures # fitting adjusment quality measures
}
```