Fiche TD avec le logiciel R : tdr316

Variables Estudiantines (examen de contrôle continu depuis 2003)

A.B. Dufour & J.R. Lobry

Une initiation de niveau L3 visant à familiariser les étudiants aux problèmes concrets rencontrés lors d'une analyse statistique dans la vraie vie : vous allez analyser vos propres données (Yes!). On part de la mesure et de la saisie des données sur la promotion actuelle, puis, via l'importation des données des *maudites* promotions précédentes, on rédige un rapport **reproductible** sur une problématique donnée. Le choix de la problématique est libre à ce niveau, l'originalité est encouragée (si elle est reproductible).

1 Objectif

L'objectif de cet exercice est de vous faire pratiquer ce que l'on appelle en anglais le « *data pre-processing* », c'est à dire le traitement en amont des données, une étape cruciale puisqu'elle conditionne tout le reste. Ce n'est pas une compétence que l'on peut acquérir par la simple lecture de livres ou de la documentation des fonctions, il faut pratiquer (se tromper), re-pratiquer (se re-tromper) et pratiquer encore (et se tromper encore).

Ce qui est demandé ici c'est de rassembler dans un unique `data.frame` exploitable des données de même nature mais de différentes origines (étudiants australiens, étudiants de diverses promotions lyonnaises). Le format des fichiers et le codage des données peut varier selon les origines, le but est donc d'homogénéiser le tout.

Pour un professionnel de l'analyse des données cet exercice prend, montre en main, de l'ordre d'une heure. Comme vous n'avez pas encore l'habitude il devrait vous prendre de l'ordre de quatre heures. Les compétences que vous allez développer sont les suivantes :

1. Savoir inspecter le résultat d'un `summary()` pour détecter les problèmes lors d'une importation de données avec `read.table()`.
2. Être capable de remédier aux problèmes les plus courants :
   (a) Présence ou absence du nom des colonnes (`header =`).

(b) Choix du séparateur des colonnes (`sep =`).

(c) Choix du séparateur décimal (`dec =`).

(d) Modifier les noms des colonnes (`colnames()`).

(e) Modifier le codage des modalités des variables qualitatives (`levels()`).

(f) Transformer une variable de type chaîne de caractères en variable qualitative (`as.factor()`).

(g) Consolider des connaissances supposées déjà acquises en L3.

3. Faire appel à un ami si vous êtes bloqué. Il n'y a rien de plus frustrant qu'une collation de données qui ne fonctionne pas. Mais pour demander de l'aide de façon efficace à un ami il faut qu'il puisse reproduire ce que vous avez essayé. L'avantage de R est que vous pouvez lui donner le code exact de vos tentatives. Votre ami pourra alors essayer de reproduire exactement ce que vous avez tenté, et donc de vous aider en connaissance de cause.

2 Les données sur les étudiants australiens

Adelaide, Australie

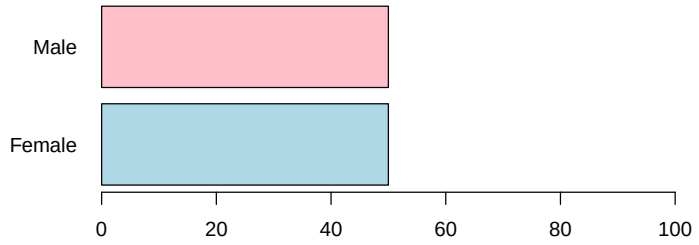
Le jeu de données `survey` du paquet `MASS` [1] contient les réponses à 12 questions de 237 étudiants en statistiques à l'Université d'Adelaide en Australie :

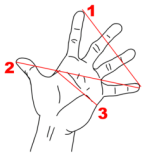
```
library(MASS)
data(survey)
summary(survey)
```

```
    Sex          Wr.Hnd           NW.Hnd          W.Hnd          Fold     
Female:118   Min.   :13.00   Min.   :12.50   Left : 18   L on R : 99  
Male  :118   1st Qu.:17.50   1st Qu.:17.50   Right:218   Neither: 18  
NA's  :  1   Median :18.50   Median :18.50   NA's :  1   R on L :120  
             Mean   :18.67   Mean   :18.58                            
             3rd Qu.:19.80   3rd Qu.:19.73                            
             Max.   :23.20   Max.   :23.50                            
             NA's   :1       NA's   :1                                
     Pulse             Clap       Exer       Smoke        Height      
Min.   : 35.00   Left   : 39   Freq:115   Heavy: 11   Min.   :150.0  
1st Qu.: 66.00   Neither: 50   None: 24   Never:189   1st Qu.:165.0  
Median : 72.50   Right  :147   Some: 98   Occas: 19   Median :171.0  
Mean   : 74.15   NA's   :  1              Regul: 17   Mean   :172.4  
3rd Qu.: 80.00                            NA's :  1   3rd Qu.:180.0  
Max.   :104.00                                        Max.   :200.0  
NA's   :45                                            NA's   :28     
       M.I           Age        
Imperial: 68   Min.   :16.75  
Metric  :141   1st Qu.:17.67  
NA's    : 28   Median :18.58  
              Mean   :20.37  
              3rd Qu.:20.17  
              Max.   :73.00  
```

`Sex` Le sexe des étudiants. Variable qualitative nominale à deux modalités `Female` et `Male`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(Sex)/sum(table(Sex)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```

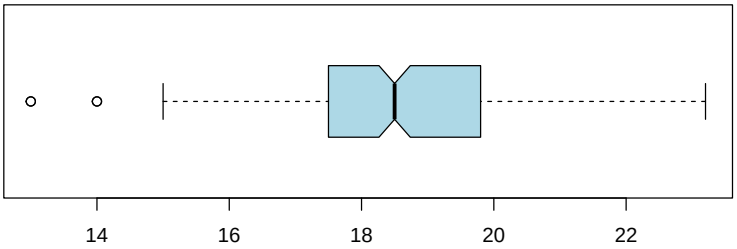




Les mesures de la main : 1) Palme 2) Empan 3) Paume (source : Wikipedia)

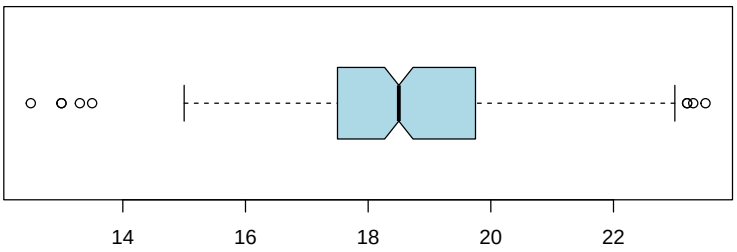
`Wr.Hnd` L'empan (distance entre l'extrémité du pouce et l'extrémité de l'auriculaire doigts écartés au maximum, le poignet bien à plat sur le plan de mesure) de la main dominante, en centimètres.

```
par(mar = c(2, 0.1, 0.1, 0.1))
with(survey, boxplot(Wr.Hnd, horizontal = TRUE, notch = TRUE, col = "lightblue"))
```



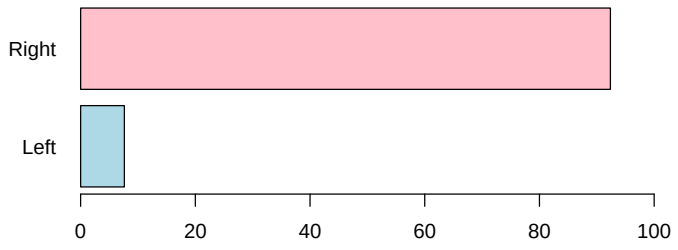
`NW.Hnd` Empan de la main non-dominante.

```
par(mar = c(2, 0.1, 0.1, 0.1))
with(survey, boxplot(NW.Hnd, horizontal = TRUE, notch = TRUE, col = "lightblue"))
```



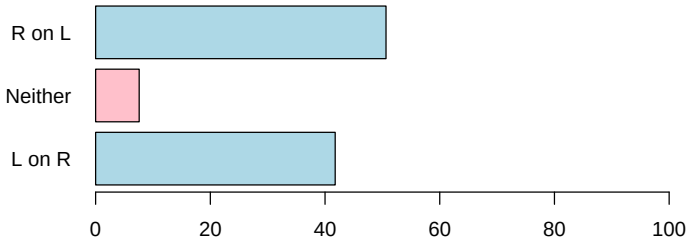
`W.Hnd` Quelle est la main dominante. Variable qualitative nominale à deux modalités `Left` et `Right`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(W.Hnd)/sum(table(W.Hnd)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```



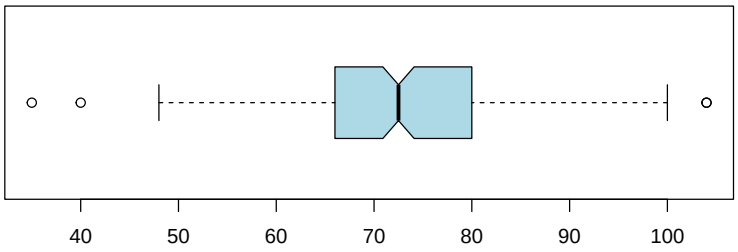
`Fold` Croisez vos bras ! Lequel est au-dessus ? Variable qualitative à trois modalités `R on L`, `L on R`, `Neither`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(Fold)/sum(table(Fold)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```



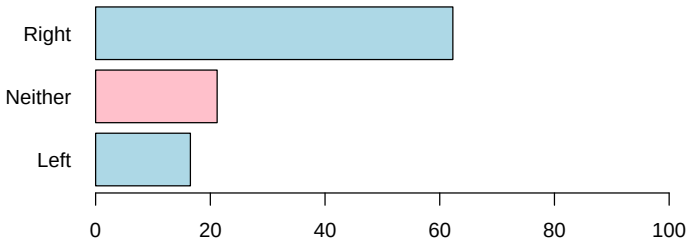
`Pulse` Rythme cardiaque au repos en pulsations par minute.

```
par(mar = c(2, 0.1, 0.1, 0.1))
with(survey, boxplot(Pulse, horizontal = TRUE, notch = TRUE, col = "lightblue"))
```



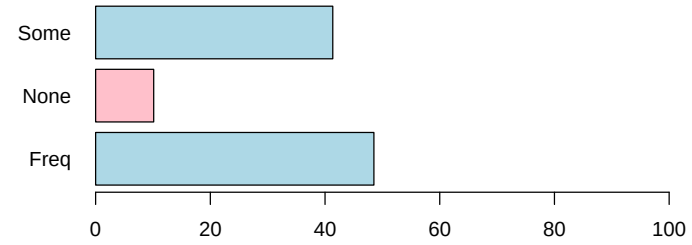
`Clap` Applaudissez ! Quelle main est au-dessus ? Variable qualitative nominale à trois modalités `Right`, `Left`, `Neither`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(Clap)/sum(table(Clap)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```



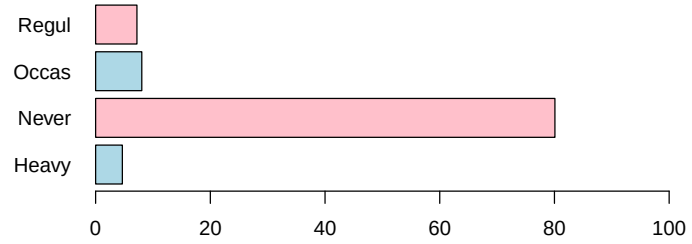
`Exer` Fréquence des activités physiques et sportives. Variable qualitative ordonnée à trois modalités : `Freq` (fréquemment), `Some`, `None`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(Exer)/sum(table(Exer)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```



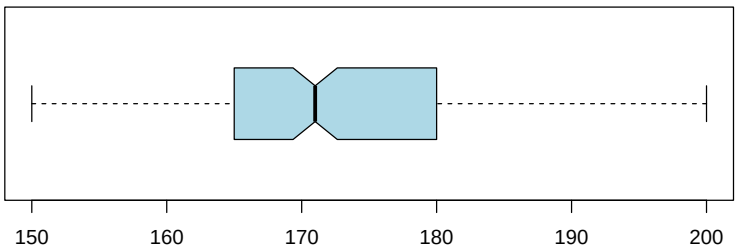
`Smoke` Intensité de l'addiction au tabac. Variable qualitative ordonnée à quatre modalités : `Heavy`, `Regul`, `Occas`, `Never`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(Smoke)/sum(table(Smoke)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```



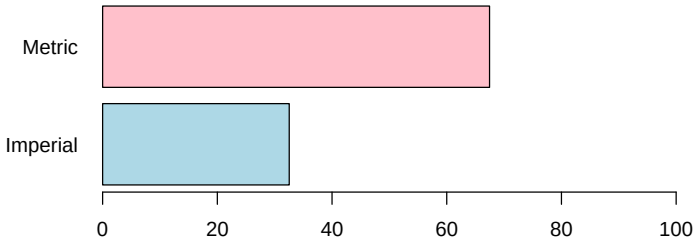
`Height` Taille en centimètres.

```
par(mar = c(2, 0.1, 0.1, 0.1))
with(survey, boxplot(Height, horizontal = TRUE, notch = TRUE, col = "lightblue"))
```



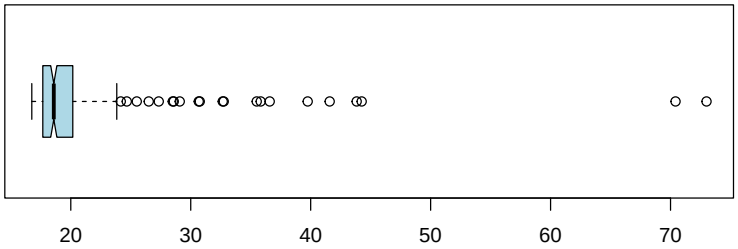
`M.I` Indique si l'étudiant exprime sa taille en unités vernaculaires (pieds/pouces) ou universelles (mètres). Variable qualitative nominale à deux modalités : `Metric` et `Imperial`.

```
par(mar = c(2, 4.1, 0.1, 0.1))
with(survey,
  barplot(100*table(M.I)/sum(table(M.I)), xlim=c(0, 110), las = 1, horiz = TRUE,
  col = c("lightblue", "pink")))
```



`Age` Âge des étudiants en années.

```
par(mar = c(2, 0.1, 0.1, 0.1))
with(survey, boxplot(Age, horizontal = TRUE, notch = TRUE, col = "lightblue"))
```



3 Acquisition de données

Complétez le jeu de données précédent en incluant les réponses de tous les étudiants de votre promotion. Attention, la version électronique de ce jeu de données ne doit en aucun cas, et en aucune façon, contenir des informations nominatives (il doit être impossible d'identifier les individus). Une version au format texte de ce jeu de données sera disponible sur le site du pbil : `https://pbil.univ-lyon1.fr/R/donnees/`. Vous disposez également des données des promotions des années précédentes à partir de 2003 :

```
read.table("https://pbil.univ-lyon1.fr/R/donnees/survey2003.txt", header = TRUE,
  dec = ",") -> s2003
```

```
head(s2003)
```

```
  sexe WrHnd NWHnd WHnd Fold Pulse    Clap Exer Smoke Height MI    Age
1    F  20.0  20.5    R RonL    72       R Some Never    170  M 24.100
2    M  23.0  22.0    R RonL    86 neither None Heavy    182  M 20.900
3    M  23.0  22.0    R LonR    74 neither None Heavy    180  M 24.990
4    F  20.2  20.2    R LonR    68       R Some Never    172  M 21.006
5    M  23.0  23.0    R RonL    76 neither Freq Never    194  M 21.000
6    F  19.5  19.3    R LonR    80 neither Some Never    173  M 20.500
```

Vous devrez donc créer un data frame comportant les données de `survey` et celles de Lyon, avec en plus une variable qualitative indiquant l'origine australienne ou française de l'étudiant.

References

[1] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.