

Solutions des exercices de la fiche tdr203

A.B. Dufour & M. Royer


Cette fiche comprend les solutions des exercices de la fiche tdr203. Il ne s'agit pas de résultats exhaustifs mais de la manière de présenter quelques résultats afin d'alimenter la réflexion.

Table des matières

Exercice 1	1
Exercice 2	9
Exercice 3	12

Exercice 1

L'étude porte sur les données de l'enquête réalisée au près des étudiants de la filière "Activités Physiques Adaptées" (UFR STAPS, Lyon, 2006). Le fichier de données `L3APA06.txt` contient les réponses des étudiants.

Les données sont importées dans  et constituent un data frame appelé par exemple `l3apa06`. La fonction `names(l3apa06)` affiche le nom de toutes les variables contenues dans le data frame.

```
l3apa06 <- read.table("http://pbil.univ-lyon1.fr/R/donnees/L3APA06.txt",
  h = T)
names(l3apa06)
[1] "groupe"      "identifiant" "sexe"      "poids"      "taille"
[6] "rythmcard"  "age"         "baccalaureat" "mention"    "hmental"
[11] "hmoteur"    "hsensoriel"  "pblesocial" "pratique"   "sport"
[16] "niveau"     "mecriture"   "mfourchette" "pballon"    "oeil"
[21] "rotation"   "pappui"
```

De plus, nous utilisons la commande `attach(l3apa06)` afin de ne pas traîner le nom du data frame dans la suite de l'étude. Enfin, nous introduisons une option sur les résultats des calculs : affichage de 4 chiffres au lieu des 7 proposés par défaut (sur les 22 possibles).

```
options(digits = 4)
```

Le data frame contient des variables qualitatives (nominales et ordinales) et des variables quantitatives.

- a) Variables qualitatives nominales : `groupe`, `sexe`, `baccalaureat`, `hmental`, `hmoteur`, `hsensoriel`, `pblesocial`, `pratique`, `sport`, `mecriture`, `mfourchette`, `pballon`, `oeil`, `rotation`, `pappui`
- b) Variables qualitatives ordinales : `mention`, `niveau`
- c) Variables quantitatives : `poids`, `taille`, `rythmcard`, `age`.

1) Commençons par étudier les variables qualitatives. Nous avons sélectionné quatre variables qualitatives d'intérêt : le sexe, l'intérêt pour le handicap mental, le sport pratiqué et la mention au baccalauréat.

a) La variable "sexe" est une variable qualitative nominale prenant deux modalités : féminin, masculin. Ces deux informations se retrouvent à l'aide des commandes :

```
class(sexe)
[1] "factor"
levels(sexe)
[1] "féminin" "masculin"
```

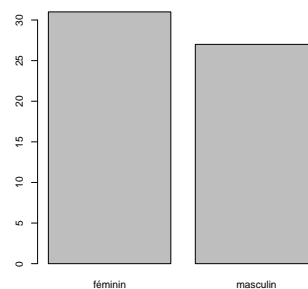
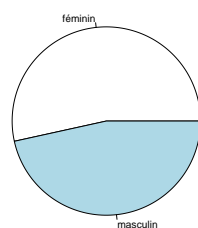
Nous pouvons calculer les fréquences absolues et les fréquences relatives à l'aide du `summary`.

```
summary(sexe)
féminin masculin
    31      27

summary(sexe)/length(sexe)
féminin masculin
0.5345  0.4655
```

Une étude statistique est incomplète si aucun graphe n'est présenté. Les deux représentations essentielles liées aux variables qualitatives nominales sont le camembert et la représentation en bâtons.

```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
pie(summary(sexe))
barplot(summary(sexe))
par(old.par)
```



Commen-

ter les résultats.

b) La variable "handicap mental" est une variable qualitative nominale prenant deux modalités : intérêt (noté 1) ou non intérêt (noté 0).

```
class(hmental)
[1] "integer"
```

℞ considère cette variable comme un entier. Mais ce n'est pas ce que nous voulons ici. Nous pouvons cependant obtenir le résumé statistique en utilisant `table`.

```
table(hmental)
hmental
0 1
35 23

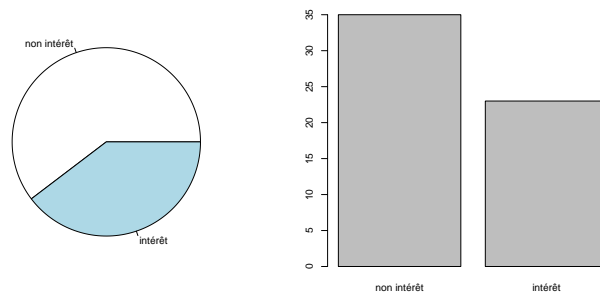
table(hmental)/length(hmental)
hmental
0 1
0.6034 0.3966
```

Mais nous pouvons également rendre la variable "handicap mental" qualitative.

```
hmental2 <- factor(hmental)
class(hmental2)
[1] "factor"
levels(hmental2) <- c("non intérêt", "intérêt")
summary(hmental2)
non intérêt    intérêt
      35         23
```

Dans le cas de l'intérêt ou non des étudiants pour le handicap mental, les deux représentations sont également le camembert et la représentation en bâtons.

```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
pie(summary(hmental2))
barplot(summary(hmental2))
par(old.par)
```



Commen-

ter les résultats.

- c) La variable "sport" est une variable qualitative nominale prenant plusieurs modalités non connues *a priori*.

```
class(sport)
[1] "factor"
levels(sport)
[1] "VTT"           "badminton"    "basket"       "boules"
[5] "danse"        "football"     "gymnastique"  "handball"
[9] "judo"         "natation"     "rugby"        "ski_alpin"
[13] "snowboard"    "tennis"       "tennis_de_table" "volleyball"
[17] "équitation"
```

Malgré le nombre de modalités important, nous pouvons calculer les fréquences absolues de cette variable.

```
summary(sport)
      VTT      badminton      basket      boules      danse
      1          1          4          1          3
football  gymnastique  handball      judo      natation
      8          4          5          1          1
rugby     ski_alpin    snowboard  tennis  tennis_de_table
      2          1          2          2          1
volleyball  équitation      NA's
      1          1          19
```

Nous constatons qu'il y a beaucoup de non réponses (NA) provenant des étudiants ne pratiquant pas un sport plus de deux fois par semaine. C'est un résultat qu'il est intéressant de noter dans un premier temps mais qui ne fait pas partie de l'analyse en soi. Nous pouvons donc obtenir les fréquences de deux manières différentes.

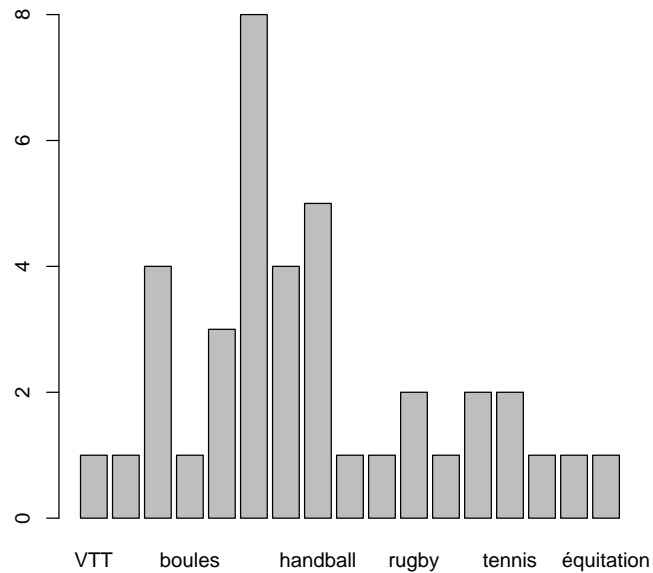
```
summary(na.omit(sport))
      VTT      badminton      basket      boules      danse
      1          1          4          1          3
football  gymnastique  handball      judo      natation
      8          4          5          1          1
rugby     ski_alpin    snowboard  tennis  tennis_de_table
      2          1          2          2          1
volleyball  équitation
      1          1

table(sport)
sport
      VTT      badminton      basket      boules      danse
      1          1          4          1          3
football  gymnastique  handball      judo      natation
      8          4          5          1          1
rugby     ski_alpin    snowboard  tennis  tennis_de_table
      2          1          2          2          1
volleyball  équitation
      1          1
```

`summary(na.omit(sport))` permet de réaliser le résumé statistique en enlevant (omettant) les données manquantes.

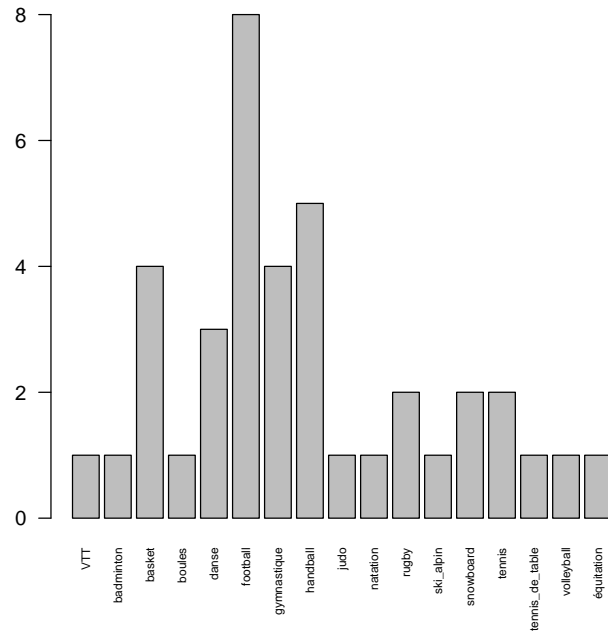
Dans le cas où les modalités de la variable qualitative sont nombreuses, la représentation en camembert n'a aucun sens. Nous ne représentons donc que la représentation en bâtons.

```
barplot(summary(na.omit(sport)))
```

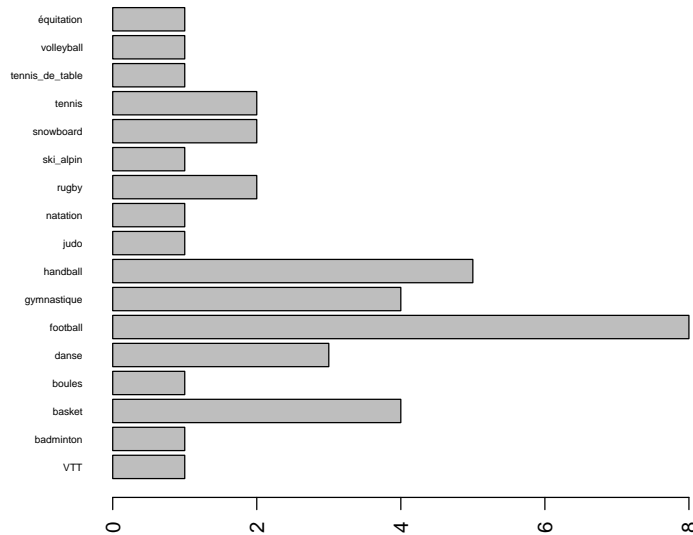


Nous constatons que les légendes des sports ne sont pas bien claires. Nous pouvons réaliser le graphe en mettant les noms des sports en vertical, voire pour plus de clareté la représentation complète en horizontal, ou encore introduire alors le graphe de Cleveland. La lisibilité s'en trouve améliorée.

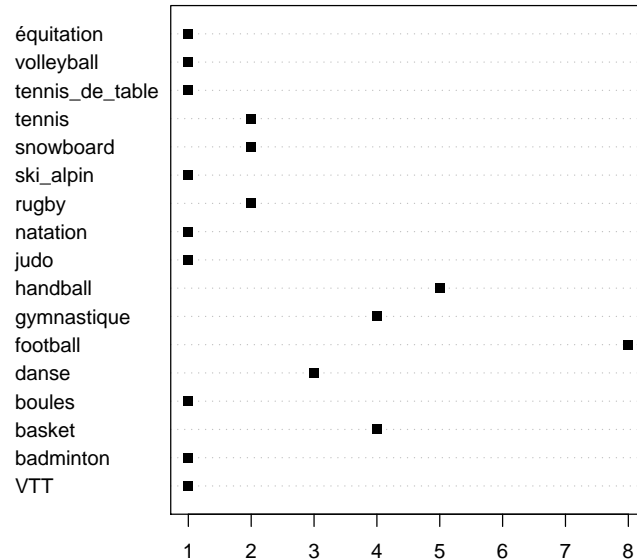
```
barplot(summary(na.omit(sport)), las = 2, cex.names = 0.6)
```



```
barplot(summary(na.omit(sport)), las = 2, cex.names = 0.5, horiz = TRUE)
```



```
dotchart(summary(na.omit(sport)), pch = 15)
```



Commen-

ter les résultats.

- d) La variable "mention" au baccalauréat est une variable qualitative ordinaire prenant quatre modalités : passable, assez-bien, bien et très-bien.

```
class(mention)
[1] "factor"
levels(mention)
[1] "AB" "B" "P"
```

Nous pouvons émettre deux constats. La modalité très-bien n'est pas présente. Les modalités sont rangées par ordre alphabétique et non ordonnées.

```
summary(mention)
  AB  B  P NA's
11  1 40  6
```

Nous constatons qu'aucun étudiant n'a eu la mention très-bien, ce qui explique son absence dans les modalités de la variable sous \mathcal{R} . Remarquons que là aussi il y a des données manquantes : les étudiants ignorent que le fait d'avoir le baccalauréat avec une moyenne comprise entre 10 et 12 correspond à la mention passable.

```
mention2 <- mention
levels(mention2) <- list(P = "P", AB = "AB", B = "B", TB = "TB")
levels(mention2)
[1] "P" "AB" "B" "TB"
summary(na.omit(mention2))
  P AB  B TB
40 11  1  0
```

Attention aux calculs des fréquences relatives dans le cas des données manquantes.

```
summary(na.omit(mention2))/length(mention2)
```

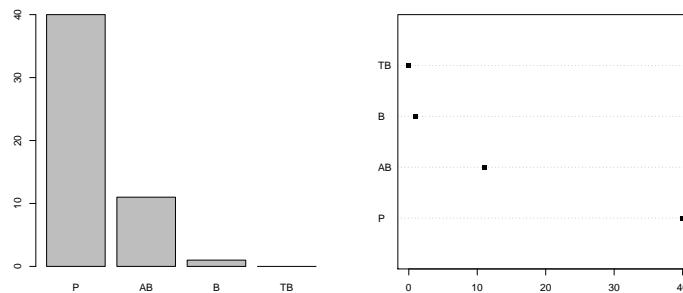
```

      P      AB      B      TB
0.68966 0.18966 0.01724 0.00000
sum(summary(na.omit(mention2))/length(mention2))
[1] 0.8966
summary(na.omit(mention2))/length(na.omit(mention2))
      P      AB      B      TB
0.76923 0.21154 0.01923 0.00000
sum(summary(na.omit(mention2))/length(na.omit(mention2)))
[1] 1
    
```

Dans le cas d'une variable qualitative à modalités ordonnées, seules les deux représentations suivantes ont un sens : la représentation en bâtons et le graphe de Cleveland.

```

old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
barplot(summary(na.omit(mention2)))
dotchart(summary(na.omit(mention2)), pch = 15)
par(mfrow = c(1, 1))
par(old.par)
    
```



Commen-

ter les résultats.

- 2) Etudions les variables quantitatives. Nous avons sélectionné deux variables quantitatives : le poids et le rythme cardiaque. Notons ici que l'étude est réalisée hommes et femmes confondus. Ceci limite fortement l'intérêt de l'interprétation dans le cas du poids. Une étude plus fine est proposée dans l'exercice 3.

- a) La variable poids est exprimée en kg.

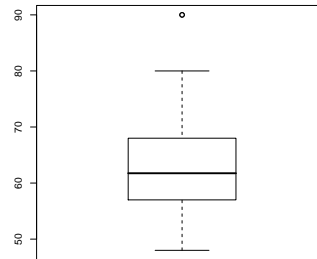
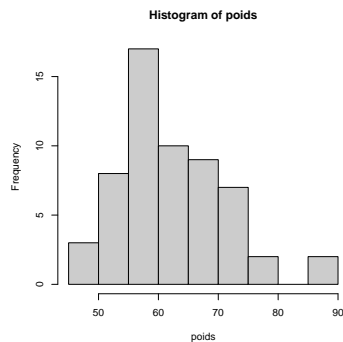
```

class(poids)
[1] "numeric"
summary(poids)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.0   57.0   61.8   63.2   68.0   90.0
    
```

Deux représentations classiques sont utilisées : l'histogramme et la boîte à moustaches.

```

old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
hist(poids, col = grey(0.8))
boxplot(poids)
par(old.par)
    
```



Commen-

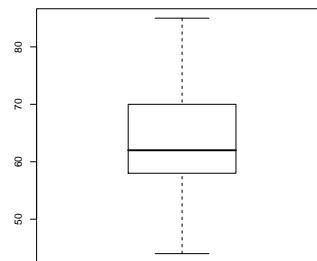
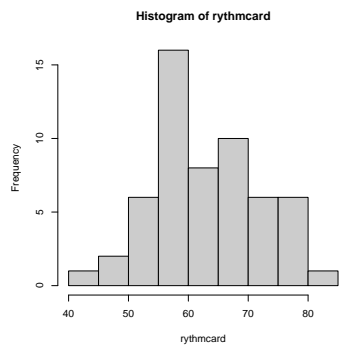
ter les résultats.

- b) La variable rythme cardiaque est exprimée en nombre de pulsations par minute.

```
class(rythmcard)
[1] "integer"
summary(rythmcard)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
  44.0   58.0   62.0   64.3   70.0   85.0    2.0
summary(na.omit(rythmcard))
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  44.0   58.0   62.0   64.3   70.0   85.0
```

Deux représentations classiques sont utilisées : l'histogramme et la boîte à moustaches.

```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
hist(rythmcard, col = grey(0.8))
boxplot(rythmcard)
par(mfrow = c(1, 1))
par(old.par)
```



Commen-

ter les résultats.

Exercice 2

Pour représenter une variable qualitative, on peut utiliser trois types de représentation :

- un diagramme en secteur (`pie`),
- une représentation en bâtons (`barplot`)
- un graphe de Cleveland (`dotchart`).

L'objectif de cet exercice est de montrer l'intérêt du graphe de Cleveland sur les deux autres représentations.

Considérons le jeu de données portant sur 592 étudiants (extrait de Snee, R. D. (1974) Graphical display of two-way contingency tables. The American Statistician, 28 :9-12). Pour chaque étudiant on a observé 3 variables qualitatives : la couleur des cheveux, la couleur des yeux et le sexe. Les données se trouvent dans le fichier "qualitatif.txt", que vous pouvez télécharger à partir du site <http://pbil.univ-lyon1.fr/R/donnees/>.

- a) Importez le fichier "qualitatif.txt" dans R à l'aide de la commande `read.table`. Appelez le data frame créé `qualnom`.

```
qualnom <- read.table("http://pbil.univ-lyon1.fr/R/donnees/qualitatif.txt",  
  h = T)
```

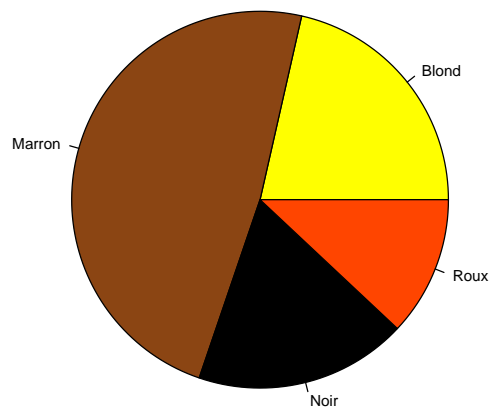
```
summary(qualnom)
```

```
  cheveux      yeux      sexe  
Blond :127   Bleu  :215   Femelle:328  
Marron:286   Marron :220   Male  :264  
Noir  :108   Noisette: 93  
Roux  : 71   Vert   : 64
```

- b) Représentez les données de la couleur des cheveux sous la forme d'un diagramme en secteurs en tapant la commande suivante :

```
pie(table(qualnom$cheveux), col = c("yellow", "chocolate4", "black",  
  "orangered"), main = "Couleur des cheveux de 592 etudiants",  
  cex.main = 1, cex = 0.75)
```

Couleur des cheveux de 592 etudiants



Quelle est la couleur de cheveux dominante ? Dans quel ordre peut-on classer les couleurs ?

Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir ?

- c) La représentation en secteurs n'est pas la représentation optimale ; il faut s'en méfier ! Consultez en effet la documentation de la fonction `pie()` à l'aide de la commande `help(pie)`.

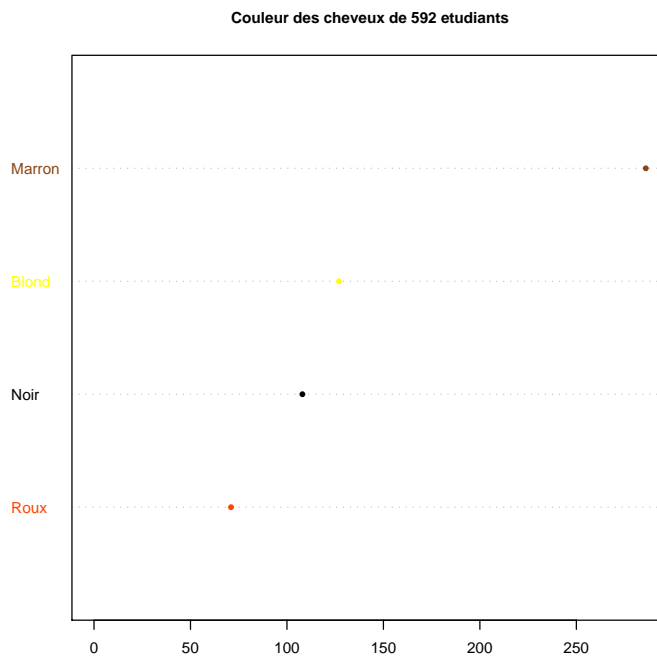
Note:

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

Cleveland (1985), page 264: "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

- d) La commande `dotchart` permet d'obtenir un graphique plus lisible :

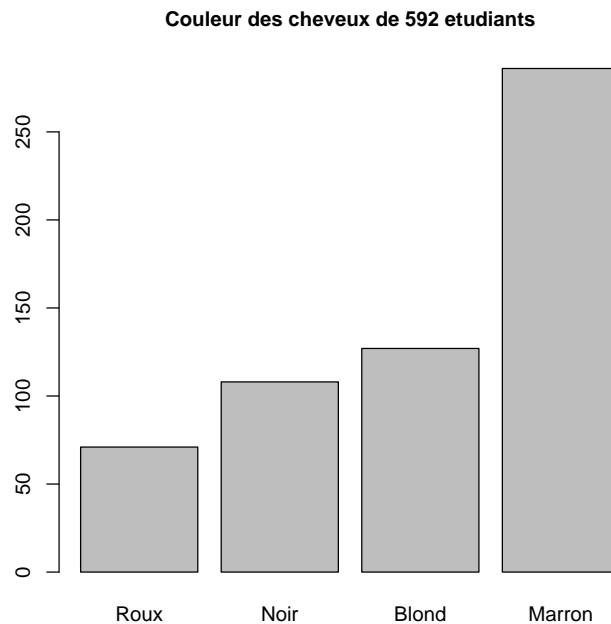
```
dotchart(sort(table(qualnom$cheveux)), xlim = c(0, max(table(qualnom$cheveux))),
  pch = 20, cex = 0.75, color = c("orangered", "black", "yellow",
  "chocolate4"), main = "Couleur des cheveux de 592 étudiants",
  cex.main = 1)
```



Peut-on maintenant répondre à la question : "Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir?"

e) Utilisez la commande `barplot` pour représenter la variable `qualnom$cheveux`.

```
barplot(sort(table(qualnom$cheveux)), cex.main = 1, main = "Couleur des cheveux de 592 etudiants")
```



Exercice 3

Pour représenter une variable quantitative, on peut utiliser en première approche, soit un histogramme (`hist`), soit une boîte à moustaches (`boxplot`).

1) Voici le poids (en ordre croissant) de 10 marathoniens : 61 62 67 67 68 69 76 77 78 79.

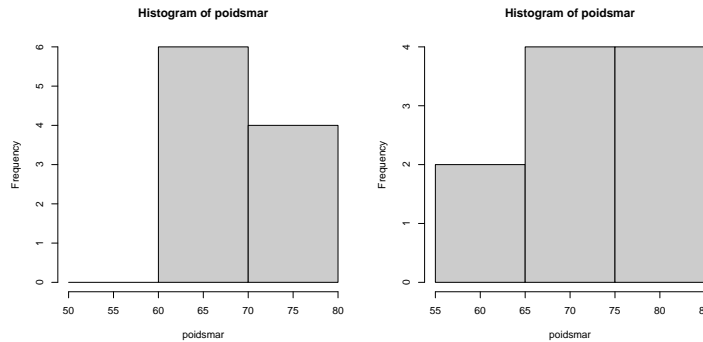
Donnez une représentation sous forme d'histogramme de ces données, en choisissant successivement les séries d'intervalles suivants :

a) `[50, 60]` `]60, 70]` et `]70, 80]` ;

b) `[55, 65]` `]65, 75]` et `]75, 85]`.

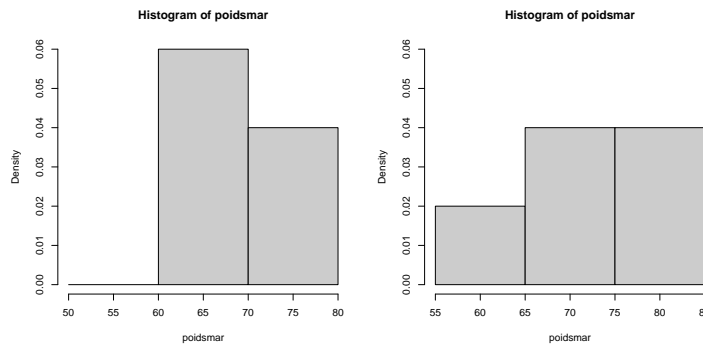
```
poidsmar <- c(61, 62, 67, 67, 68, 69, 76, 77, 78, 79)
poidsmar
[1] 61 62 67 67 68 69 76 77 78 79

old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
hist(poidsmar, col = grey(0.8), breaks = c(50, 60, 70, 80))
hist(poidsmar, col = grey(0.8), breaks = c(55, 65, 75, 85))
par(mfrow = c(1, 1))
par(old.par)
```



On ne peut pas bien comparer les aires car l'échelle n'est pas la même.
On va donc imposer l'échelle.

```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
hist(poidsmar, freq = F, col = grey(0.8), ylim = c(0, 0.06), breaks = seq(50,
80, by = 10))
hist(poidsmar, freq = F, col = grey(0.8), ylim = c(0, 0.06), breaks = seq(55,
85, by = 10))
par(mfrow = c(1, 1))
par(old.par)
```



c) Que peut-on dire sur l'allure de ces 2 figures ?

Le choix du découpage en intervalles est un problème délicat qui risque de biaiser fortement notre perception des données.

2) Reprenons l'exemple de l'exercice 1 sur le poids `poids` des étudiants. Le sexe est une variable qu'on ne saurait exclure de l'étude. D'une manière générale, toute étude portant sur la morphologie doit tenir compte du dimorphisme sexuel.

a) Construire les data frames `sexeM` et `sexeF` qui contiennent respectivement les réponses des élèves de chacun des groupes.

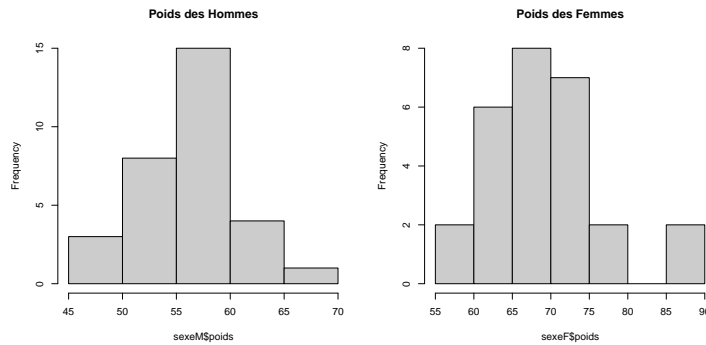
```
l3apa06 <- read.table("http://pbil.univ-lyon1.fr/R/donnees/L3APA06.txt",
h = T)
attach(l3apa06)

sexeM <- l3apa06[sexe == "féminin", ]
sexeF <- l3apa06[sexe == "masculin", ]
```

b) Construire l'histogramme du poids des élèves de chaque groupe.

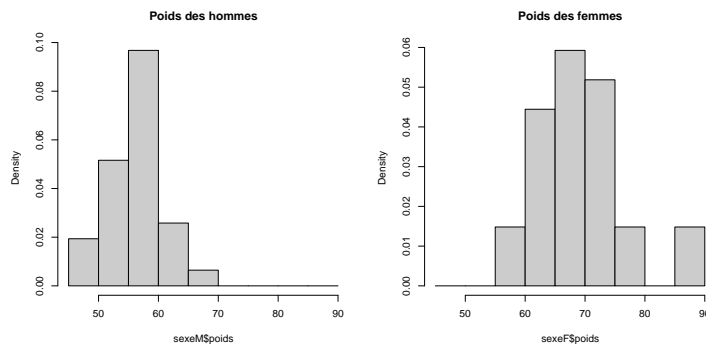
```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
hist(sexeM$poids, col = grey(0.8), main = "Poids des Hommes")
```

```
hist(sexeF$poids, col = grey(0.8), main = "Poids des Femmes")
par(mfrow = c(1, 1))
par(old.par)
```



- c) Afin de pouvoir comparer le poids selon les 2 sexes, imposez les mêmes classes grâce à la commande `breaks`, et travaillez avec les fréquences relatives.

```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
hist(sexeM$poids, freq = F, col = grey(0.8), breaks = seq(45, 90,
  by = 5), main = "Poids des hommes")
hist(sexeF$poids, freq = F, col = grey(0.8), breaks = seq(45, 90,
  by = 5), main = "Poids des femmes")
par(mfrow = c(1, 1))
par(old.par)
```



- d) Utilisez la commande `boxplot(poids ~ sexe)` pour avoir une représentation sous la forme d'une boîte à moustache.

```
boxplot(l3apa06$poids ~ l3apa06$sexe, cex.main = 1, main = "Poids des étudiants en fonction du groupe")
```

Poids des étudiants en fonction du groupe

