

Fiche TD avec le logiciel : tdr204

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.

Exercice 1

Prenons le tableau de données concernant l'ensemble des étudiants de la filière IGAPA depuis 2006 : `enqL3APA.txt` ayant répondu à une enquête au premier cours de statistique.

```
enqL3APA <- read.table("http://pbil.univ-lyon1.fr/R/donnees/enqL3APA.txt",
                        h = T)
names(enqL3APA)
[1] "annee"          "groupe"        "identifiant"   "sexe"         "poids"
[6] "taille"         "rythmcard"     "age"           "baccaureat"   "mention"
[11] "hmental"        "hmeuteur"      "hsensoriel"    "pblesocial"   "pratique"
[16] "sport"          "niveau"        "mecriture"     "mfourchette"  "pballon"
[21] "oeil"           "rotation"      "pappui"
```

- 1) Donner les proportions d'hommes et de femmes en filière APA depuis 2006 ; faire une représentation graphique adéquate. Commenter.

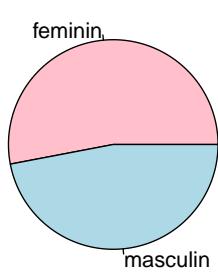
```
summary(enqL3APA$sexe)/length(enqL3APA$sexe)
feminin masculin
0.53      0.47
```

La proportion des hommes en filière L3 APA est 0.47 et la proportion des femmes est 0.53.

En ce qui concerne la représentation graphique liée à la répartition des sexes, le camembert et la représentation en bâtons sont toutes les deux acceptables. En effet, nous avons un seul échantillon à représenter, une variable avec peu de modalités qui sont de nature nominale.

```
par(mfrow = c(1, 2))
pie(summary(enqL3APA$sexe), col = c("pink", "lightblue"), main = "Répartition des sexes en L3 APA")
barplot(summary(enqL3APA$sexe)/length(enqL3APA$sexe), main = "Répartition des sexes en L3 APA")
```

Répartition des sexes en L3 APA



Répartition des sexes en L3 APA



La proportion de femmes dans cette filière est légèrement supérieure à celle des hommes.

- 2) Donner les pourcentages d'étudiants intéressés par les handicaps moteur, sensoriel, mental et par les problèmes sociaux. Commenter.

```
class(enqL3APA$hmoteur)
```

```
[1] "integer"
```

Pour répondre à cette question, nous avons deux solutions. Une solution est celle du cours où la variable 'integer' est transformée en variable qualitative à deux modalités : 1 pour oui et 0 pour non.

```
hmental <- factor(enqL3APA$hmental)
100 * summary(hmental)/length(hmental)
0 1
55 45

hmoteur <- factor(enqL3APA$hmoteur)
100 * summary(hmoteur)/length(hmoteur)
0 1
39 61

hsensoriel <- factor(enqL3APA$hsensoriel)
100 * summary(hsensoriel)/length(hsensoriel)
0 1
66 34

pblesocial <- factor(enqL3APA$pblesocial)
100 * summary(pblesocial)/length(pblesocial)
0 1
68 32
```

Nous pouvons noter également qu'une réponse favorable est donnée par 1 et défavorable par 0. Pour calculer les pourcentages d'étudiants intéressés, il suffit donc de compter le nombre de 1 c'est-à-dire faire la somme des éléments de la colonne.

```
100 * sum(enqL3APA$hmental)/length(enqL3APA$hmental)
```

```
[1] 45
```

```
100 * sum(enqL3APA$hmoteur)/length(enqL3APA$hmoteur)
```

```
[1] 61
```

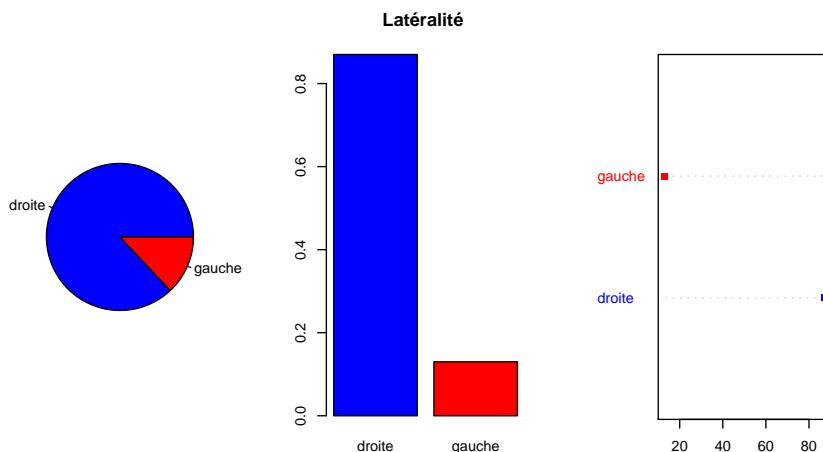
```
100 * sum(enqL3APA$hsensoriel)/length(enqL3APA$hsensoriel)
```

```
[1] 34
100 * sum(enqL3APA$pblesocial)/length(enqL3APA$pblesocial)
[1] 32
```

Parmi tous les thèmes proposés, le handicap moteur est celui qui intéresse le plus les étudiants. Notons toutefois, que la filière propose une spécialisation en direction des personnes âgées et que ce thème n'a pas été proposé dans le questionnaire.

- 3) Faire une représentation graphique liée à la variable 'main d'écriture'.

```
par(mfrow = c(1, 3))
pie(summary(enqL3APA$mecriture), col = c("blue", "red"))
barplot(summary(enqL3APA$mecriture)/length(enqL3APA$mecriture),
       col = c("blue", "red"), main = "Latéralité ")
dotchart(summary(enqL3APA$mecriture), col = c("blue", "red"), pch = 15)
```



- 4) Etudier la variable 'mention au baccalauréat'. Qu'observe-t-on ?

La variable 'mention au baccalauréat' est une variable qualitative.

```
class(enqL3APA$mention)
[1] "factor"
summary(enqL3APA$mention)
AB      B      P NA's
20      2      65    13
```

Nous constatons que certains étudiants n'ont pas répondu à la question.

```
summary(enqL3APA$mention)[4]/length(enqL3APA$mention)
NA's
0.13
```

Cela correspond à 13% des étudiants. Deux explications sont possibles : (i) certains ignorent qu'avoir son baccalauréat avec une moyenne générale comprise entre 10 et 12 exclue signifie avoir une mention passable ; (ii) certains étudiants ont eu une mention mais veulent rester anonymes. Devant ces deux possibilités, nous ne pouvons prendre de décision. L'étude sera donc réalisée sur l'ensemble des étudiants ayant effectivement répondu à la question.

```
table(enqL3APA$mention)
```

```

AB   B   P
20   2   65

table(enqL3APA$mention)/sum(table(enqL3APA$mention))
          AB      B      P
0.22988506 0.02298851 0.74712644

round(table(enqL3APA$mention)/sum(table(enqL3APA$mention)), 4)
          AB      B      P
0.2299  0.0230  0.7471

```

Sous R, il y a souvent plusieurs chemins pour répondre à une question. Voici une autre possibilité.

```

mention <- na.omit(enqL3APA$mention)
summary(mention)

AB   B   P
20   2   65

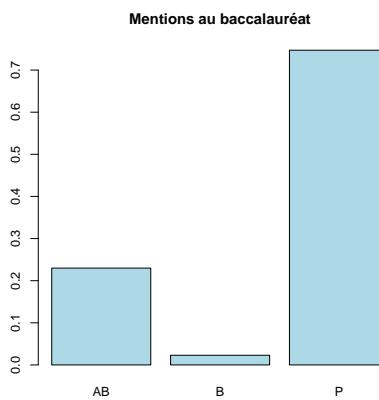
summary(mention)/length(mention)
          AB      B      P
0.22988506 0.02298851 0.74712644

levels(mention)
[1] "AB" "B"  "P"

```

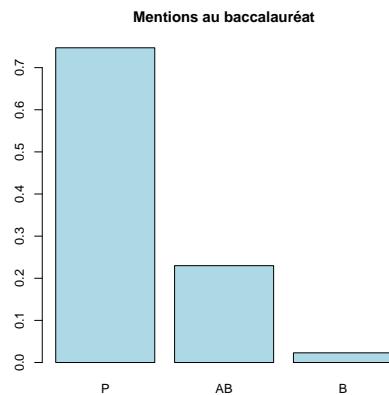
La variable qualitative 'mention' a ses modalités ordonnées. La représentation en secteurs - ou camembert - est interdite.

```
barplot(summary(mention)/length(mention), main = "Mentions au baccalauréat",
       col = "lightblue")
```



La représentation réalisée ne convient pas car les modalités sont ordonnées par ordre alphabétique et non par ordre croissant du mérite. Il faut donc modifier l'ordre des modalités.

```
mention.fac <- factor(mention, levels = c("P", "AB", "B"))
barplot(summary(mention.fac)/length(mention.fac), main = "Mentions au baccalauréat",
       col = "lightblue")
```



La répartition des mentions obtenues au baccalauréat par les étudiants de L3 APA reflète les résultats que l'on trouve au niveau national.

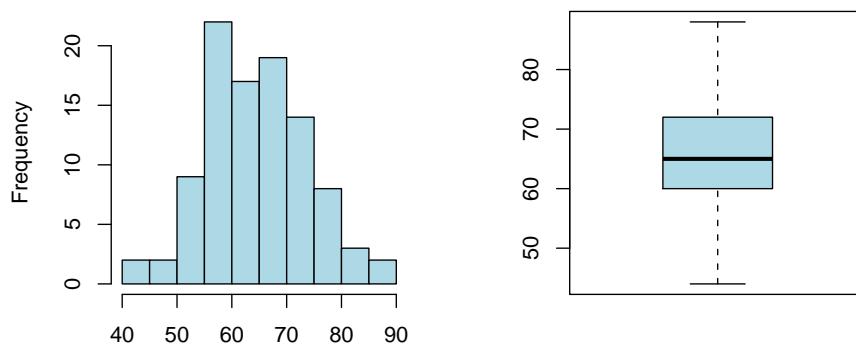
- 5) Calculer le 'rythme cardiaque' moyen d'un étudiant en L3 APA. Utiliser la fonction `summary(x)` et la fonction `mean(x)`. Qu'observe-t-on ?

```
summary(enqL3APA$rythmcard) [4]
Mean
65.62
mean(na.omit(enqL3APA$rythmcard))
[1] 65.62245
```

- 6) Construire l'histogramme et la boîte à moustaches du 'rythme cardiaque'. Commenter.

```
par(mfrow = c(1, 2))
hist(enqL3APA$rythmcard, col = "lightblue", main = "Fréquence cardiaque",
     xlab = "")
boxplot(enqL3APA$rythmcard, col = "lightblue")
```

Fréquence cardiaque



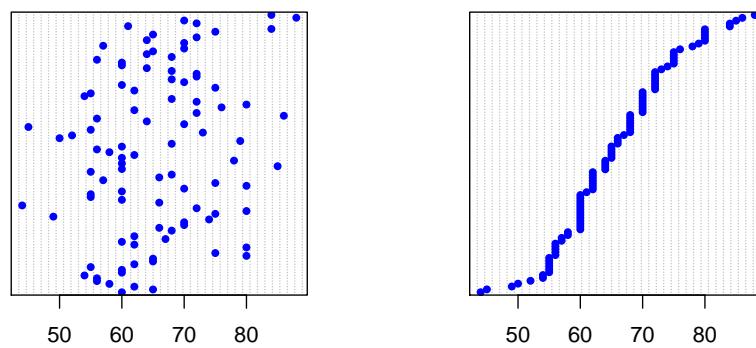
La fréquence cardiaque des étudiants varie entre 40 et 90 pulsations par minute. 50% des étudiants varient entre 60 et 72 pulsations. Les moustaches

sont symétriques ce qui signifie que les valeurs extrêmes sont réparties de la même façon. Aucun étudiant ne se détache des autres par une fréquence cardiaque très élevée ou très basse.

- 7) Toujours pour le rythme cardiaque, construire le graphe de Cleveland sur les données brutes et sur les données ordonnées. Conclure.

```
par(mfrow = c(1, 2))
dotchart(enqL3APA$rythmcard, col = "blue", main = "Fréquence cardiaque",
          pch = 20)
dotchart(sort(enqL3APA$rythmcard), col = "blue", pch = 20)
```

Fréquence cardiaque



Exercice 2

On considère 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/>

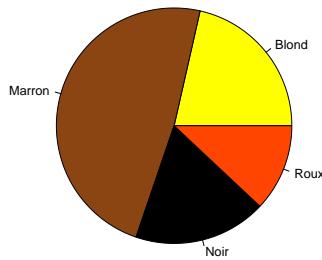
- 1) Donner l'instruction permettant d'importer le fichier dans R. Appeler le fichier **qualnom**.

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

- 2) Représenter les données sur 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 étudiants")
```

Couleur des cheveux de 592 étudiants

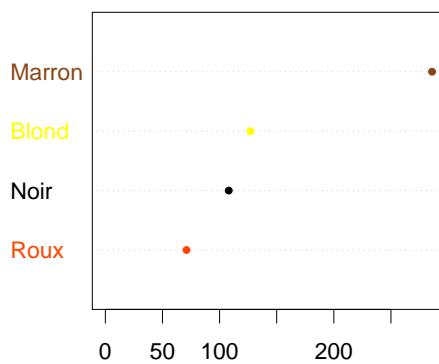


- a) Quelle est la couleur de cheveux dominante ? Dans quel ordre peut-on classer les couleurs ?
- b) Y a-t-il plus d'écart entre les proportions de cheveux correspondant aux couleurs Noir et Roux ou entre les proportions correspondant aux cheveux de couleur Blond et Noir ?
- 3) La représentation en secteurs n'est pas la représentation optimale ; il faut s'en méfier ! Consulter en effet la documentation de la fonction `pie()` à l'aide de la commande `help(pie)`.

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 = 1.5, color = c("orangered", "black", "yellow",
        "chocolate4"), main = "Couleur des cheveux de 592 étudiants")
```

Couleur des cheveux de 592 étudiants



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 ?"

Le graphe de Cleveland donne une représentation agréable pour les variables qualitatives avec de nombreuses modalités, et aussi pour les variables qualitatives ordonnées.

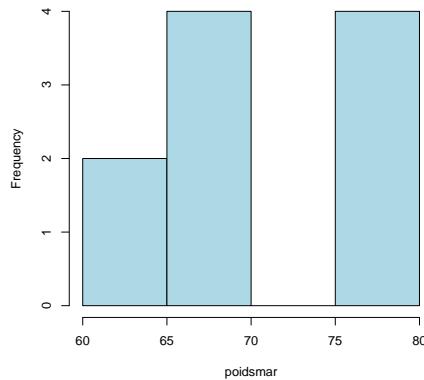
Exercice 3

Soit la série statistique ordonnée des poids (en kg) de 10 marathoniens :
 61 62 67 67 68 69 76 77 78 79.

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

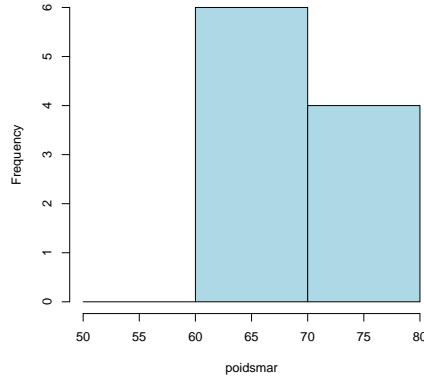
- 1) Construire l'histogramme de base associé à cette série statistique.

```
hist(poidsmar, col = "lightblue", main = "")
```



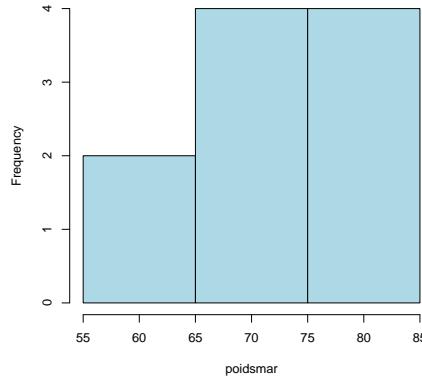
- 2) Réaliser les histogrammes en choisissant respectivement les séries d'intervalles suivants :
 a) les intervalles [50, 60]]60, 70] et]70, 80] ;

```
hist(poidsmar, col = "lightblue", main = "", breaks = c(50, 60,
70, 80))
```



- b) les intervalles $[55, 65]$ $[65, 75]$ et $[75, 85]$.

```
hist(poidsmar, col = "lightblue", main = "", breaks = c(55, 65,
75, 85))
```



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

Exercice 4

Revenons à l'enquête réalisée au près des étudiants de la filière L3 APA ou IGAPA. Les variables 'poids' et 'taille' ne peuvent être étudiées qu'hommes et femmes séparément. L'information qualitative est contenue dans la variable **sexé**. Nous rappelons que la variable poids de l'ensemble des étudiants (unique-ment les hommes) s'obtient avec la commande :

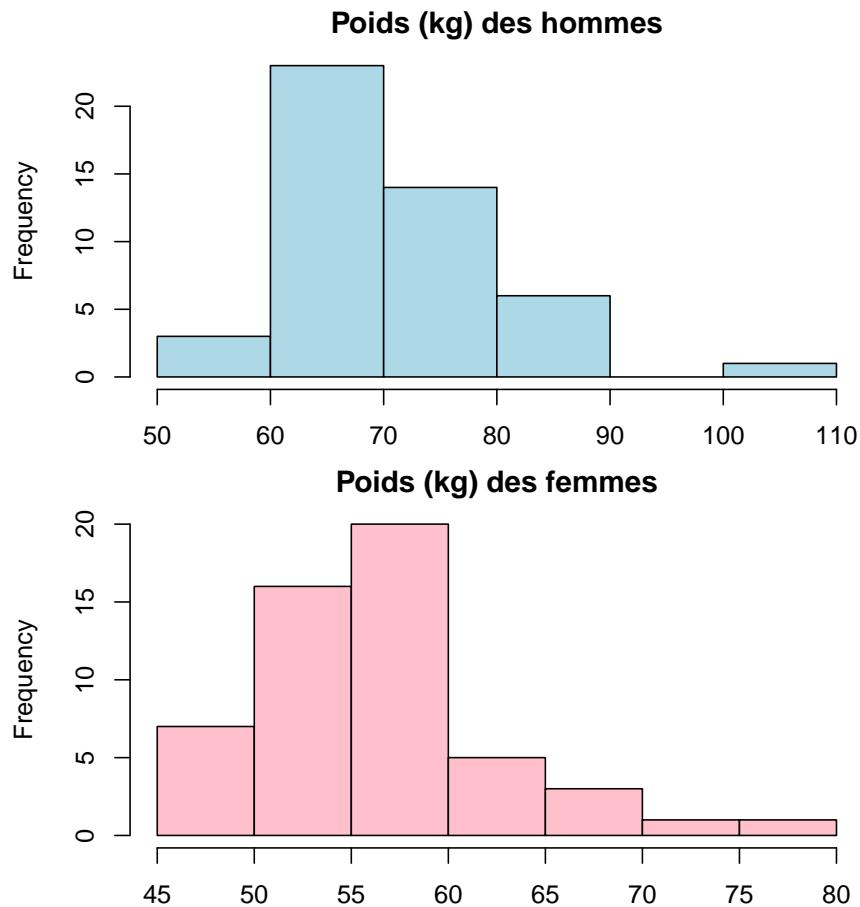
```
poidsHom <- enqL3APA$poids[enqL3APA$sexé=="masculin"]
```

- 1) Construire les deux vecteurs **poidsHom** et **poidsFem**.

```
poidsHom <- enqL3APA$poids[enqL3APA$sexé == "masculin"]
poidsFem <- enqL3APA$poids[enqL3APA$sexé == "feminin"]
```

- 2) Construire les histogrammes du poids des étudiants et des étudiantes.

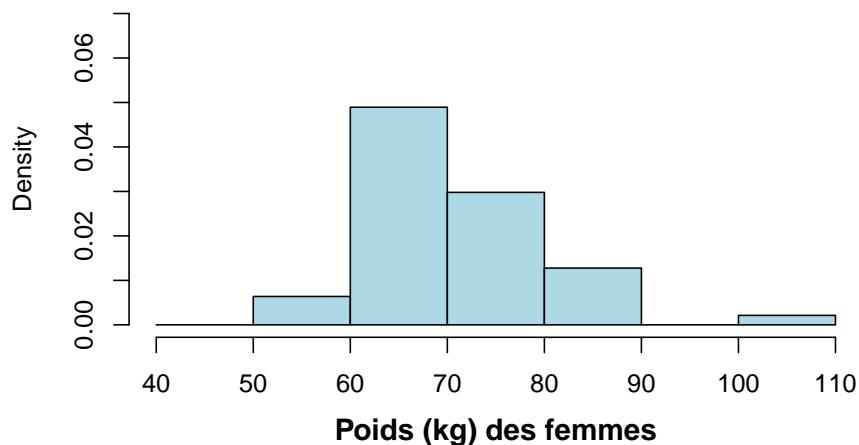
```
par(mar = c(2, 4, 2, 2))
par(mfrow = c(2, 1))
hist(poidsHom, col = "lightblue", main = "Poids (kg) des hommes",
     xlab = "")
hist(poidsFem, col = "pink", main = "Poids (kg) des femmes", xlab = "")
```



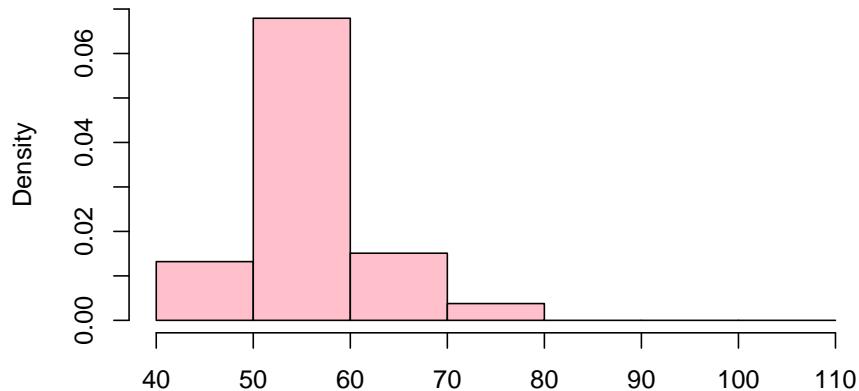
- 3) Afin de pouvoir comparer les 2 groupes, construire à nouveau les histogrammes en imposant les mêmes classes grâce à la commande `breaks`, et en travaillant avec les fréquences relatives grâce à la commande `freq = FALSE`. Commenter.

```
par(mar = c(2, 4, 2, 2))
par(mfrow = c(2, 1))
hist(poidsHom, col = "lightblue", main = "Poids (kg) des hommes",
     breaks = seq(40, 110, by = 10), freq = FALSE, ylim = c(0, 0.07))
hist(poidsFem, col = "pink", main = "Poids (kg) des femmes", breaks = seq(40,
110, by = 10), freq = FALSE, ylim = c(0, 0.07))
```

Poids (kg) des hommes



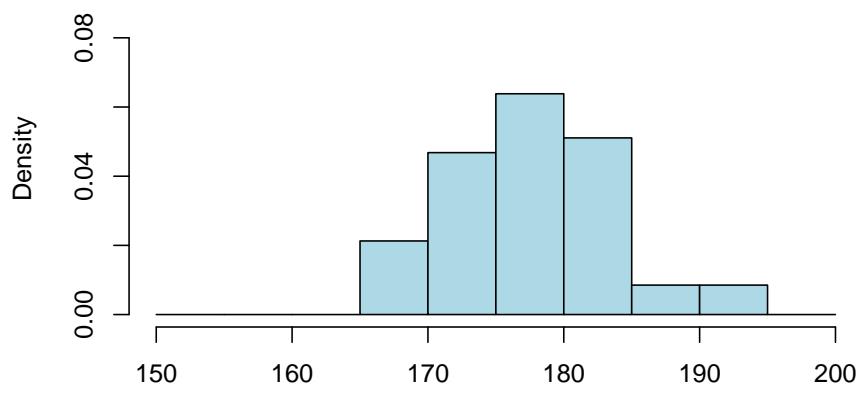
Poids (kg) des femmes



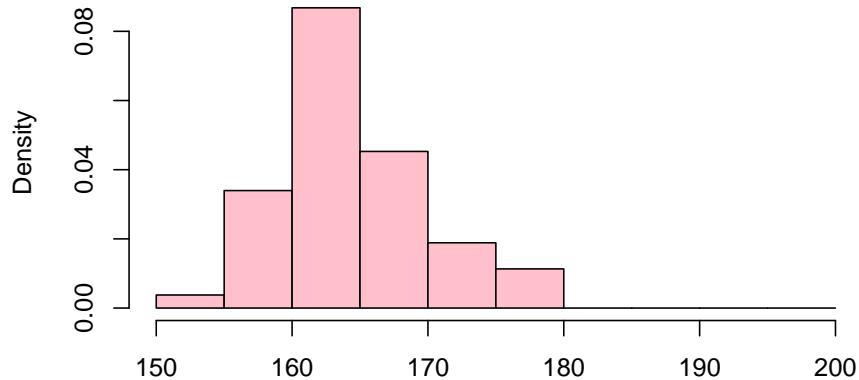
- 4) Faire la même étude avec la variable 'taille'.

```
tailleHom <- enqL3APA$taille[enqL3APA$sexe == "masculin"]
tailleFem <- enqL3APA$taille[enqL3APA$sexe == "feminin"]
par(mar = c(2, 4, 2, 2))
par(mfrow = c(2, 1))
hist(tailleHom, col = "lightblue", main = "Taille (cm) des hommes",
      breaks = seq(150, 200, by = 5), freq = FALSE, ylim = c(0, 0.09))
hist(tailleFem, col = "pink", main = "Taille (cm) des femmes", breaks = seq(150,
200, by = 5), freq = FALSE, ylim = c(0, 0.09))
```

Taille (cm) des hommes



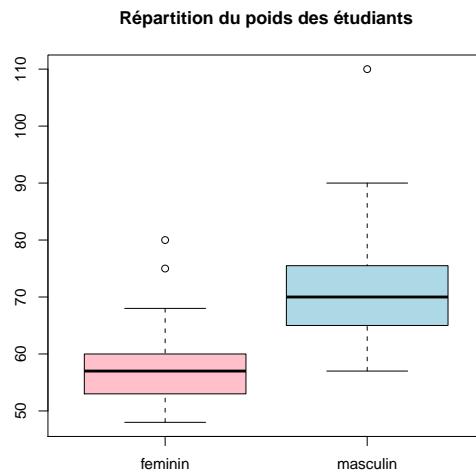
Taille (cm) des femmes



Pour représenter par exemple sur une boîte à moustaches, une variable quantitative liée à une variable qualitative, utiliser la commande :
boxplot(x~factor).

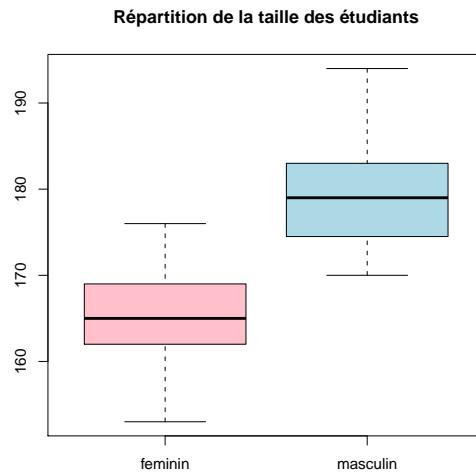
- 1) Représenter la boîte à moustaches du poids en fonction du sexe. Commenter.

```
boxplot(enqL3APA$poids ~ enqL3APA$sexe, col = c("pink", "lightblue"),
main = "Répartition du poids des étudiants")
```



- 2) Représenter la boîte à moustaches de la taille en fonction du sexe. Commenter.

```
boxplot(enqL3APA$taille ~ enqL3APA$sexe, col = c("pink", "lightblue"),
main = "Répartition de la taille des étudiants")
```



- 3) Représenter la boîte à moustaches du rythme cardiaque en fonction du sexe. Commenter.

```
boxplot(enqL3APA$rythmcard ~ enqL3APA$sex, col = c("pink", "lightblue"),
       main = "Répartition de la fréquence cardiaque des étudiants")
```

Répartition de la fréquence cardiaque des étudiants

