Fiche TD avec le logiciel : tdr203

Introduction à la statistique univariée. Les représentations graphiques

A.B. Dufour & M. Royer

Cette fiche comprend des exercices portant à la fois sur les paramètres descriptifs et les représentations graphiques liés aux variables qualitatives et quantitatives.

Table des matières

1 Introduction **1**
1.1 Variables qualitatives 1
1.1.1 La représentation en secteurs ou camembert 2
1.1.2 La représentation en bâtons 2
1.2 Variables quantitatives 4
1.2.1 L'histogramme 5
1.2.2 Le graphe de Cleveland 5
1.2.3 La boîte à moustaches 6

2 Exercices **7**

1 Introduction

La statistique d'aujourd'hui ne se conçoit plus sans graphique. Nous résumons dans ce paragraphe les représentations de base associées aux variables qualitatives et quantitatives.

1.1 Variables qualitatives

Quelle que soit la représentation associée à une variable qualitative, elle est associée au résumé statistique (fréquences absolues ou fréquences relatives). Prenons, à titre d'exemple, les postes de jeu observées dans un échantillon de 64 handballeurs de haut niveau.

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

```
 [1] "TAD"    "TAA"    "DBI"    "ENV"    "HAU"    "LMS"    "LMI"    "EMP"    "POSTE"
[10] "NIVEAU"
```

```
postes <- handball$POSTE
summary(postes)
```

```
ailier arrcent  arrlat gardien   pivot
    16       5      19      13      11
```

```
summary(postes)/length(postes)
```

```
  ailier  arrcent   arrlat  gardien    pivot
0.250000 0.078125 0.296875 0.203125 0.171875
```

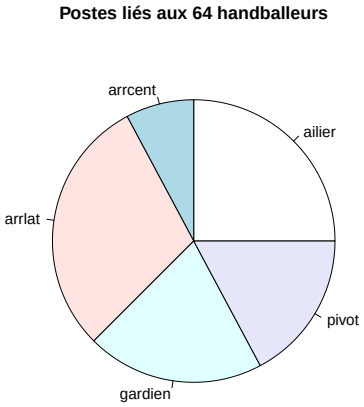
1.1.1 La représentation en secteurs ou camembert

La représentation en secteurs `pie` est associée aux fréquences relatives. Un cercle représente 360 degrés. A une modalité k de la variable, on associe un angle θ_k défini par $\theta_k = f_k \times 360$. Mais il n'est pas besoin de les calculer pour faire la représentation graphique.

```
freqrel <- summary(postes)/length(postes)
360 * freqrel
```

```
ailier arrcent  arrlat gardien   pivot
90.000  28.125 106.875  73.125  61.875
```

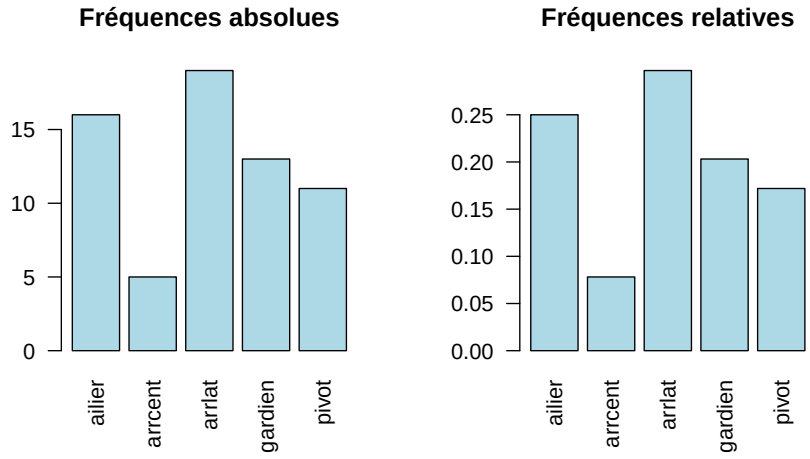
```
pie(summary(postes), main = "Postes liés aux 64 handballeurs")
```



1.1.2 La représentation en bâtons

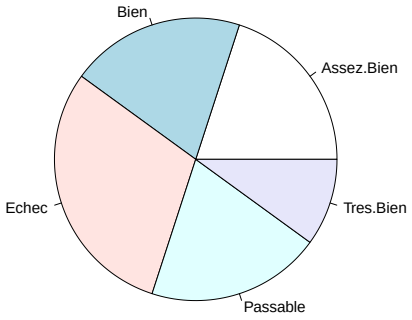
La représentation en bâtons `barplot` est réalisée soit sur les fréquences absolues, soit sur les fréquences relatives.

```
par(mfrow = c(1, 2))
barplot(summary(postes), main = "Fréquences absolues", col = "lightblue",
    las = 2)
barplot(summary(postes)/length(postes), main = "Fréquences relatives",
    col = "lightblue", las = 2)
```



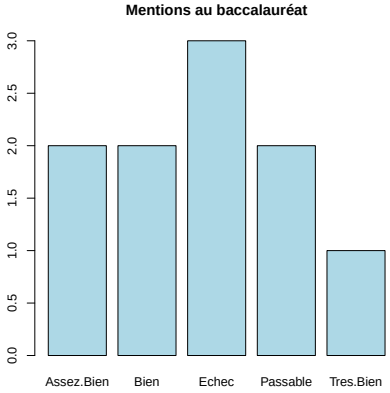
Remarque Fondamentale. Bien noter que dans le cas d'une variable qualitative nominale, observée sur **un** seul groupe, les deux représentations graphiques sont possibles. Dans le cas d'une variable qualitative ordinale, seule la représentation en bâtons est acceptable ; l'ordre des modalités doit être respectée. Reprenons par exemple la mention (`tdr202.pdf`).

```
mention <- c("Bien", "Echec", "Assez.Bien", "Echec", "Tres.Bien",
    "Assez.Bien", "Echec", "Bien", "Passable", "Passable")
mention.fac <- factor(mention)
pie(summary(mention.fac))
```



Cette représentation graphique n'a pas de sens car les modalités sont ordonnées. La représentation en bâtons est plus appropriée.

```
barplot(summary(mention.fac), col = "lightblue", main = "Mentions au baccalauréat")
```

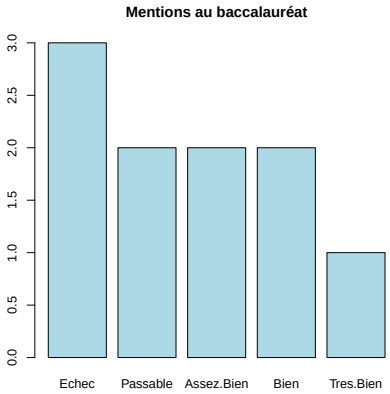


Celle-ci n'est pas plus appropriée car l'ordre des modalités n'est pas respecté (les modalités sont classées par ordre alphabétique). Il faut donc réajusté celui-ci avant de faire la représentation.

```
mention2.fac <- factor(mention, levels = c("Echec", "Passable",
    "Assez.Bien", "Bien", "Tres.Bien"))
summary(mention2.fac)
```

```
    Echec   Passable Assez.Bien       Bien  Tres.Bien
       3          2          2          2          1
```

```
barplot(summary(mention2.fac), col = "lightblue", main = "Mentions au baccalauréat")
```



1.2 Variables quantitatives

Prenons, à titre d'exemple, les tailles mesurées sur chacun des 64 handballeurs du paragraphe précédent.

```
tailles <- handball$TAD
summary(tailles)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  170.0   178.1   182.8   183.2   188.0   198.5
```

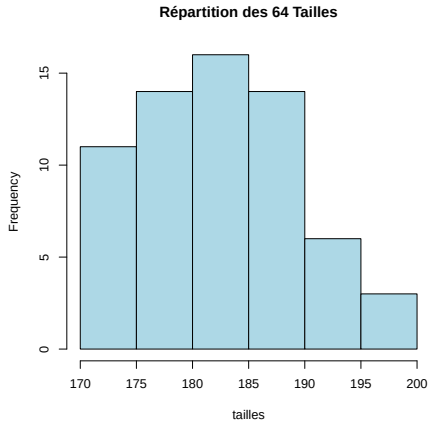
1.2.1 L'histogramme

Cette représentation, notée `hist(x)`, est une des plus classiques. La variable quantitative est découpée en intervalles d'amplitude constante (forme la plus courante). L'axe horizontal définit les intervalles ; l'axe vertical donne le nombre d'individus appartenant à chaque intervalle.

Noter, en utilisant le `help(hist)` que différents arguments de la fonction permettent de réaliser le découpage en classes.

- `breaks` donne les valeurs du découpage ;
- `include.lowest = TRUE` signifie que la valeur la plus petite est incluse dans la première classe ;
- `right = TRUE` signifie que les intervalles sont ouverts à gauche et fermés à droite.

```
hist(tailles, col = "lightblue", main = "Répartition des 64 Tailles")
```



Nous notons par exemple que 14 handballeurs ont une taille comprise entre 175 et 180 cm.

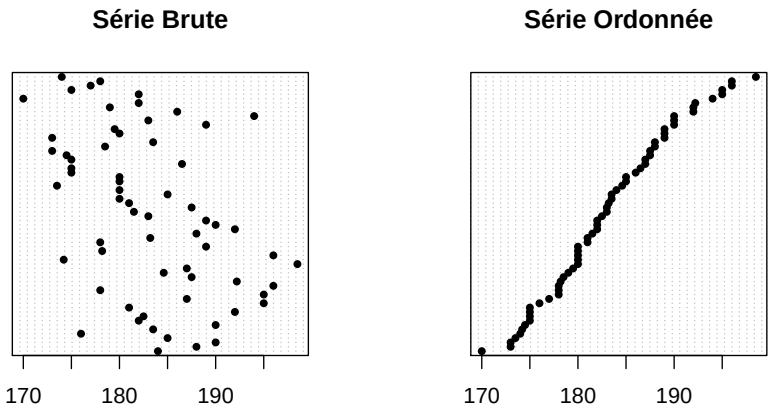
```
length(tailles[tailles > 175 & tailles <= 180])
```

```
[1] 14
```

1.2.2 Le graphe de Cleveland

C'est une représentation où l'axe horizontal définit la variable quantitative étudiée et l'axe horizontal les individus par ordre d'entrée dans la série statistique. Elle est notée `dotchart(x)`. Elle prend tout son sens sur la série statistique ordonnée.

```
par(mfrow = c(1, 2))
dotchart(tailles, main = "Série Brute", pch = 20)
dotchart(sort(tailles), main = "Série Ordonnée", pch = 20)
```



1.2.3 La boîte à moustaches

Cette représentation graphique est basée sur la série statistique ordonnée du minimum au maximum et sur son découpage en quartiles. Elle est notée `boxplot(x)`.

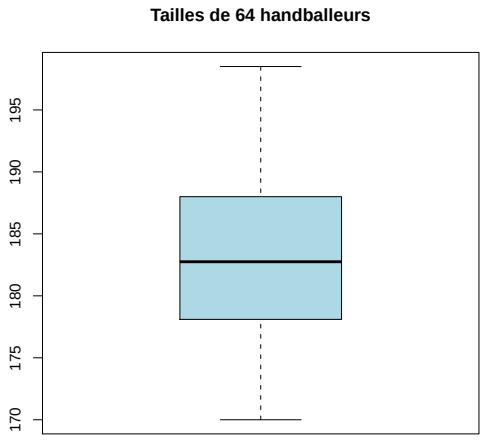
```
sort(tailles)
```

```
 [1] 170.0 173.0 173.0 173.5 174.0 174.2 174.5 175.0 175.0 175.0 175.0 176.0 177.0
[14] 178.0 178.0 178.0 178.2 178.5 179.0 179.5 180.0 180.0 180.0 180.0 180.0 181.0
[27] 181.0 181.5 182.0 182.0 182.0 182.5 183.0 183.0 183.2 183.5 183.5 184.0 184.6
[40] 185.0 185.0 186.0 186.5 187.0 187.0 187.5 187.5 188.0 188.0 189.0 189.0 189.0
[53] 190.0 190.0 190.0 192.0 192.0 192.2 194.0 195.0 195.0 196.0 196.0 198.5
```

```
quantile(tailles)
```

```
    0%    25%    50%    75%   100%
170.00 178.15 182.75 188.00 198.50
```

```
boxplot(tailles, col = "lightblue", main = "Tailles de 64 handballeurs")
```



La boite est constituée par le premier quartile Q_1 (178.15cm) et le troisième quartile Q_3 (188cm). 50% des handballeurs sont compris entre ces deux valeurs. On représente à l'intérieur de la boite la médiane ou deuxième quartile Q_2 (182.75cm). Sa position vers le bas ou vers le haut indique un tassement vers le bas ou vers le haut des valeurs ; une position centrale est plus liée à une répartition uniforme des valeurs.

Les moustaches se calculent ainsi.

- Pour la moustache supérieure,
  - i- on calcule une valeur seuil $val_{max} = Q_3 + 1.5 \times (Q_3 - Q_1)$.

```
valmax <- quantile(tailles)[[3]] + 1.5 * (quantile(tailles)[[3]] -
    quantile(tailles)[[1]])
valmax
```

```
[1] 201.875
```

  - ii- La moustache supérieure est la valeur de la distribution (classée) immédiatement inférieure à la valeur seuil. Dans notre exemple, il s'agit de la valeur maximale de la distribution.
- Pour la moustache inférieure,
  - i- On calcule une valeur seuil $val_{min} = Q_1 - 1.5 \times (Q_3 - Q_1)$.

```
valmin <- quantile(tailles)[[1]] - 1.5 * (quantile(tailles)[[3]] -
    quantile(tailles)[[1]])
valmin
```

```
[1] 150.875
```

  - ii- La moustache inférieure est la valeur de la distribution (classée) immédiatement supérieure à la valeur seuil. Dans notre exemple, il s'agit de la valeur minimale de la distribution.

Une boite à moustaches peut posséder des points en dehors des moustaches. Ce sont les individus qui ont des valeurs très basses ou très hautes par rapport à celles attendues dans l'échantillon. On les qualifie de points aberrants.

2 Exercices

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)
```

1) Donner les proportions d'hommes et de femmes en filière APA depuis 2006 ; faire une représentation graphique adéquate. Commenter.
2) Donner les pourcentages d'étudiants intéressés par les handicaps moteur, sensoriel, mental et par les problèmes sociaux. Commenter.
3) Faire une représentation graphique liée à la variable 'main d'écriture'.
4) Etudier la variable 'mention au baccalauréat'. Qu'observe-t-on ?

Remarque. Lorsque le data frame contient des données manquantes symbolisées par `NA`, les calculs et les graphiques considèrent celui-ci comme

une modalité. C'est intéressant pour repérer le nombre de `NA`. Si on veut enlever les données manquantes, deux instructions sont disponibles : `summary(na.omit(factor))` ou `table(factor)`.

Refaire l'étude en utilisant ces instructions. Commenter tous les résultats.

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 ?

Remarque. Lorsque le data frame contient des données manquantes symbolisées par `NA`, les calculs et les graphiques sont réalisés bien sûr sans tenir compte de ces dernières. Le nombre de données manquantes apparait dans le `summary`. Pour calculer la moyenne simple dans le cas de données manquantes, utiliser l'instruction `mean(na.omit(x))`.

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

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

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`.

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 etudiants")
```

   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 etudiants")
```

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.

1) Construire l'histogramme de base associé à cette série statistique.
2) Réaliser les histogrammes en choisissant respectivement les séries d'intervalles suivants :
   a) les intervalles $[50, 60]$ $]60, 70]$ et $]70, 80]$;
   b) les intervalles $[55, 65]$ $]65, 75]$ et $]75, 85]$.
   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 séparemment. L'information qualitative est contenue dans la variable `sexe`. Nous rappelons que la variable poids de l'ensemble des étudiants (uniquement les hommes) s'obtient avec la commande :

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

1) Construire les deux vecteurs `poidsHom` et `poidsFem`.
2) Construire les histogrammes du poids des étudiants et des étudiantes.
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.
4) Faire la même étude avec la variable 'taille'.

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

1) Représenter la boite à moustaches du poids en fonction du sexe. Commenter.
2) Représenter la boite à moustaches de la taille en fonction du sexe. Commenter.
3) Représenter la boite à moustaches du rythme cardiaque en fonction du sexe. Commenter.

Remarque. Dans le cas où deux variables sont croisées, les lignes contenant des données manquantes sont supprimées.
Dans l'exemple `boxplot(poids~sexe)`, un individu sera retiré de l'analyse soit parce ce que son poids est manquant, soit parce que son groupe d'appartenance -

le sexe - est manquant, soit parce que l'information sur le poids et l'appartenance sont manquants.

Pour aller plus loin...
Une variable `x` est attachée à un data frame `df`. Pour accéder à cette variable, on a écrit jusqu'à présent `df$x` comme par exemple pour le rythme cardiaque du data frame `enqL3APA$rythmcard`. Le résumé statistique s'obtient alors avec l'instruction :

```
summary(enqL3APA$rythmcard)
```

Pour simplifier les écritures, on peut attacher l'ensemble des variables en début de session mais il faut bien penser à détacher à la fin de l'étude pour éviter les interférences. Cela donne, sur notre exemple

```
attach(enqL3APA)
summary(rythmcard)
detach(enqL3APA)
```

. Il existe une commande qui permet de séparer directement un data frame en deux sous data frames :

```
neofich <- split(enqL3APA,sexe)
neofich$masculin
neofich$feminin
```

Pour plus d'information sur les représentations graphiques et notamment l'ensemble des arguments lié à chaque fonction, vous pouvez consulter les fiches suivantes :

- `LANG04` : Les graphiques de Base (Cours, Introductions)
- `tdr18` : (Ne pas) utiliser les couleurs (Fiches de TD, Le logiciel R)