

Premiers pas vers l'analyse de données ...


A.B. Dufour & D. Clot

Cette fiche comprend des exercices portant sur les paramètres descriptifs principaux et les représentations graphiques des variables quantitatives et qualitatives. L'objectif est de se placer dans un contexte d'Analyse de Données.

Table des matières

1	Les Données	1
1.1	Les crimes violents aux U.S.A.	2
1.2	La sécurité routière dans les départements français	2
1.3	Les clients d'une banque	3
1.4	Quelques conseils	3
2	Variables et Descriptions Générales	3
2.1	Variable quantitative	3
2.1.1	Paramètres descriptifs	4
2.1.2	Les graphiques	6
2.2	Variable qualitative	8
2.2.1	Les paramètres statistiques	9
2.2.2	Les graphiques	9
3	La sécurité routière en France en 2009 et 2010	11


1 Les Données

Les données peuvent être créées sous  mais elles proviennent généralement de tableurs (Excel, OpenOffice) ou existent déjà dans les différentes librairies. Nous allons donc présenter les trois modes de lecture de jeux de données les plus couramment rencontrés.

Les fichiers sont accessibles sur le site <http://pbil.univ-lyon1.fr/R/enseignement.html>, dans le menu **Données**, sous-menu **Dossier de fichiers**.

1.1 Les crimes violents aux U.S.A.

Les données ont été rentrées sous OpenOffice et sauvegardées avec l'extension `csv`. Le tableau comprend quatre colonnes : l'état, l'année de recueil de l'information, le nombre d'habitants et le nombre de crimes violents.

Si le jeu de données `CrimeStateDate.csv` a été importé dans le répertoire de travail, sa lecture sous  s'effectue par la fonction `read.csv` :

```
CSD <- read.csv("CrimeStateDate.csv", header=TRUE)
```

```
head(CSD)
```

```
  Etat Date Population Crime_Violent
1 Alabama 1960   3266740         6097
2 Alabama 1961   3302000         5564
3 Alabama 1962   3358000         5283
4 Alabama 1963   3347000         6115
5 Alabama 1964   3407000         7260
6 Alabama 1965   3462000         6916
```

Pour pouvoir bénéficier de tous les arguments ou de toutes les informations sur la fonction :


```
args(read.csv)
help("read.csv")
```

Si le jeu de données est lu directement à partir du site de `pbil`, il faut entrer l'adresse url complète :

```
http://pbil.univ-lyon1.fr/R/donnees/CrimeStateDate.csv
```

1.2 La sécurité routière dans les départements français

Les données ont été rentrées sous Excel et proviennent du site officiel <http://www.securite-routiere.gouv.fr/>. Elles ont été sauvegardées sous l'extension `txt` avec comme séparateur de colonne la tabulation.

Si le jeu de données `SecRoutiere0910.txt` a été importé dans le répertoire de travail, sa lecture sous  s'effectue par la fonction `read.table` :

```
SR0910 <- read.table("SecRoutiere0910.txt", header=TRUE)
```

```
names(SR0910)
```

```
[1] "departement" "numdep" "region" "numregion" "acc.corps.10"
[6] "acc.corps.09" "tues.10" "tues.09" "blessees.10" "blessees.09"
[11] "population" "ratio"
```

Le fichier contient 10 colonnes : le nom du département français, le numéro du département, la région auquel appartient le département, le numéro de la région, le nombre d'accidents corporels en 2010, le nombre d'accidents corporels en 2009, le nombre de tués sur les routes en 2010, le nombre de tués sur les routes en 2009, le nombre de blessés en 2010, le nombre de blessés en 2009, le nombre d'habitants estimé au 1er janvier 2009 (INSEE) et le nombre de tués par million d'habitants en 2010.


1.3 Les clients d'une banque

Les données font partie de la librairie `ade4` et leur lecture en est très simple.


```
library(ade4)
data(banque)
names(banque)
[1] "csp"      "duree"    "oppo"     "age"      "sexe"     "interdit" "cableue"
[8] "assurvi" "soldevu" "eparlog"  "eparliv"  "credhab"  "credcon"  "versesp"
[15] "retresp" "remiche" "preltre"  "prelfin"  "viredeb"  "virecre"  "porttit"
dim(banque)
[1] 810 21
```

Le fichier comprend 21 variables qualitatives dont la description en anglais est disponible grâce à l'aide en ligne `?banque`. Une description en français se trouve dans la fiche `pps049.pdf` dans le menu **Données**, sous-menu **Données socio-économiques** du site pédagogique de pbil.

1.4 Quelques conseils

1. Il est préférable de travailler dans un dossier spécifique. Vérifiez, à l'aide de la commande `getwd()` que vous vous trouvez bien dans votre espace de travail. Si ce n'est pas le cas, allez dans le menu **Changer de répertoire courant**.
2. Copiez le fichier dans votre dossier de travail (par le clic droit de la souris)
3. Allez sous .

2 Variables et Descriptions Générales

Construire un cours de statistique par l'utilisation d'un logiciel comme , c'est bien sûr s'extraire du temps de calcul et faciliter la réalisation de graphiques. Mais cela introduit une complexité : la connaissance du vocabulaire et du sens des concepts liés à la statistique d'une part, liés au logiciel d'autre part.

Relation entre la terminologie statistique et les noms des objets de .

tableau	data frame
variable qualitative	factor
modalité	level
variable quantitative	numeric, integer

2.1 Variable quantitative

Une **série statistique** associée à une variable quantitative X est une liste de valeurs mesurées sur n individus. A chaque individu i est associée la valeur x_i .

Exemple. On étudie la variable 'nombre de crimes violents' dans les 51 états de l'Amérique du Nord en 2005.

```
cv2005 <- CSD[CSD$Date=="2005",]
dim(cv2005)
[1] 51 4
```

```

crime <- cv2005$Crime_Violent
class(crime)
[1] "integer"
length(crime)
[1] 51

```

2.1.1 Paramètres descriptifs

Les paramètres classiques associés à une variable quantitative sont :

- ★ le minimum [`min(x)`]

```

min(crime)
[1] 625

```

- ★ le maximum [`max(x)`]

```

max(crime)
[1] 190178

```

- ★ la moyenne [`mean(x)`]

```

mean(crime)
[1] 27268.529

```


- ★ Lorsque la série est ordonnée, on peut rechercher les paramètres qui coupent la distribution en plusieurs parties égales. Le cas le plus courant est de couper la distribution en 4 parties : on recherche alors les trois valeurs appelées **quartiles** telles que, à l'intérieur de chaque partie, on retrouve 25% des individus [`quantile(x)`].

```

quantile(crime)
      0%      25%      50%      75%     100%
625.0  5190.0 14659.0 31651.5 190178.0

```

La procédure de calcul des quantiles est définie comme suit.

- On note q une des trois valeurs suivantes 0.25, 0.5, 0.75.
 - On cherche la position i du quartile dans la série ordonnée : $i = q(n - 1) + 1$ où n est le nombre total d'individus.
 - On repère les deux valeurs de la distribution qui encadrent le quartile cherché.
 - On calcule ce dernier par une règle de trois.
- ★ La variance descriptive (en $1/n$) est à recalculer en fonction de la variance donnée par le logiciel  [`var(x)`] qui est l'estimation de la variance de la population à partir de l'échantillon. Pour ce faire, il suffit d'écrire une petite fonction.

```
vardes <- fonction(x) var(x)*(length(x)-1)/length(x)
vardes(crime)

[1] 1278094825

var(crime)

[1] 1303656721
```

★ l'écart-type descriptif c'est-à-dire la racine de la variance :

```
ecartype <- fonction(x) sqrt(vardes(x))
ecartype(crime)

[1] 35750.452
```

Les résultats liés aux paramètres de position s'obtiennent également en utilisant une seule fonction de \mathcal{R} [summary(x)].

```
summary(crime, digits=6)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 625.0  5190.0 14659.0 27268.5 31651.5 190178.0
```

L'argument `digits` est rajouté dans le résumé statistique car la variable est considérée par cette fonction comme entière. Le nombre de chiffres significatifs est alors modifié par la fonction.

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

Ce problème est essentiellement dû aux valeurs élevées de la variable. Si on considère une situation plus classique en centrant la variable, l'écart entre les données calculées une à une et le résumé statistique disparaît.

```
crimec <- scale(crime, center=TRUE, scale=FALSE)
min(crimec)
[1] -26643.529
max(crimec)
[1] 162909.47
quantile(crimec)
      0%      25%      50%      75%     100%
-26643.5294 -22078.5294 -12609.5294  4382.9706 162909.4706
summary(crimec)
  V1
Min.  :-26644
1st Qu.: -22079
Median :-12610
Mean   :      0
3rd Qu.:  4383
Max.   :162909
```

2.1.2 Les graphiques

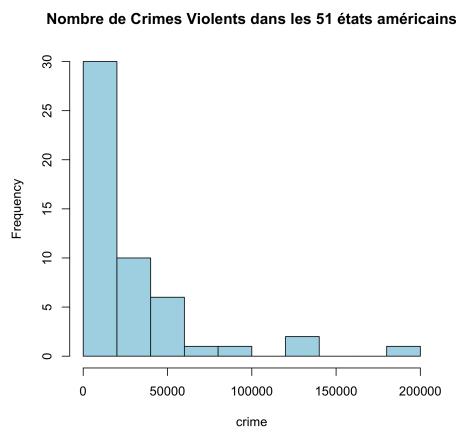
L'histogramme

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.

Notez, 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(crime, col="lightblue", main="Nombre de Crimes Violents dans les 51 états américains")
```



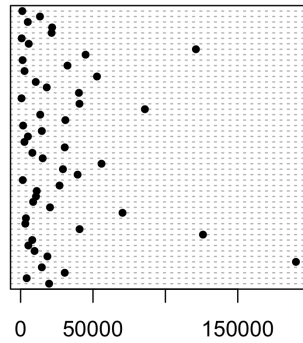
Exercice. Commenter l'allure de l'histogramme. Puis réaliser une transformation de type logarithme sur les données [$\log(x)$], représenter l'histogramme lié à la transformation et commenter à nouveau l'allure de la représentation.

Le graphe de Cleveland

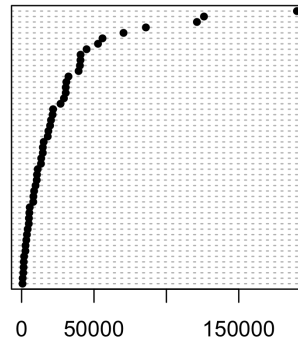
C'est une représentation où l'axe horizontal définit la variable quantitative étudiée et l'axe vertical les individus par ordre d'entrée dans la série statistique. Elle est notée `dotchart(x)`. Elle prend tout son sens sur la série statistique ordonnée.

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

Série Brute



Série Ordonnée



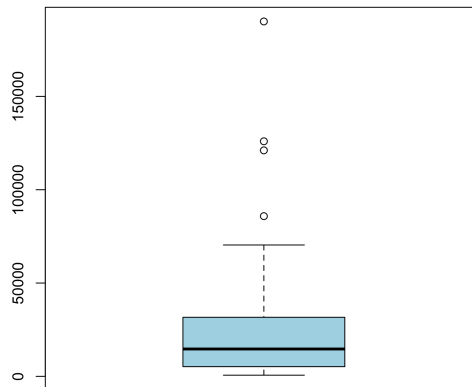
Exercice. Réaliser un graphe de Cleveland sur la transformation logarithme des données. Commenter.

La boîte à moustaches

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

```
boxplot(crime, col="lightblue", main="Nombre de Crimes Viloents dans les 51 états américains")
```

Nombre de Crimes Viloents dans les 51 états américains



La boîte est constituée par le premier quartile Q_1 (5190) et le troisième quartile Q_3 (31651.5). 50% des états d'Amérique du Nord ont un nombre de crimes violents compris entre ces deux valeurs. On représente à l'intérieur de la boîte la médiane appelée aussi deuxième quartile Q_2 (14659). 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(crime)[[4]] + 1.5 * (quantile(crime)[[4]] - quantile(crime)[[2]])
valmax
[1] 71343.75
```

ii- La moustache supérieure est la valeur de la distribution (classée) immédiatement inférieure à la valeur seuil.

```
sort(crime)[41:51]
[1] 40273 40650 40725 44891 52761 55877 70392 85839 121091 125957 190178
```

Dans notre exemple, il s'agit de la valeur 70392..

★ Pour la moustache inférieure,

i- On calcule une valeur seuil $val_{min} = Q_1 - 1.5 \times (Q_3 - Q_1)$.

```
valmin <- quantile(crime)[[2]] - 1.5 * (quantile(crime)[[4]] - quantile(crime)[[2]])
valmin
[1] -34502.25
```

ii- La moustache inférieure est la valeur de la distribution (classée) immédiatement supérieure à la valeur seuil. Dans notre exemple, comme `valmin` est négative, la moustache inférieure est définie par la valeur minimale de la distribution.

Une boîte 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. L'exemple en contient 4.

Exercice. Réaliser la boîte à moustaches sur la transformation logarithme des données et retrouver, à l'aide des quartiles, les différentes informations visualisées par le graphe.

2.2 Variable qualitative

Une **série statistique** associée à une variable qualitative A est une liste de valeurs observées sur n individus. Ces valeurs observées sont en nombre restreint et sont appelées les **modalités** de la variable qualitative. A chaque individu i est associée une et une seule modalité. On note p le nombre de modalités de la variable A .

Exemple. On considère la catégorie socio-professionnelle des clients de la banque (données `ade4`).

```
class(banque$csp)
[1] "factor"
levels(banque$csp)
[1] "agric" "artis" "cadsu" "inter" "emplo" "ouvri" "retra" "inact" "etudi"
length(levels(banque$csp))
[1] 9
```


2.2.1 Les paramètres statistiques

Les paramètres classiques associés à une variable qualitative sont :

- ★ les fréquences absolues c'est-à-dire le nombre d'individus par modalité notées n_k avec k variant de 1 à p ,

```
summary(banque$csp)
agric artis cadsu inter emplo ouvri retra inact etudi
 29    48   103   102   151   183    52    85    57
```

- ★ les fréquences relatives $f_k = \frac{n_k}{n}$.

```
summary(banque$csp)/length(banque$csp)
      agric      artis      cadsu      inter      emplo      ouvri      retra
0.035802469 0.059259259 0.127160494 0.125925926 0.186419753 0.225925926 0.064197531
      inact      etudi
0.104938272 0.070370370
```

On peut si on le souhaite ne garder que 4 chiffres significatifs :

```
round(summary(banque$csp)/length(banque$csp),4)
      agric      artis      cadsu      inter      emplo      ouvri      retra      inact      etudi
0.0358 0.0593 0.1272 0.1259 0.1864 0.2259 0.0642 0.1049 0.0704
```

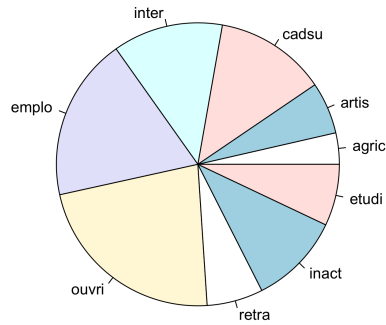
2.2.2 Les graphiques

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(banque$csp)/length(banque$csp)
360*freqrel
      agric      artis      cadsu      inter      emplo      ouvri      retra      inact
12.888889 21.333333 45.777778 45.333333 67.111111 81.333333 23.111111 37.777778
      etudi
25.333333
pie(summary(banque$csp), main="CSP présentes dans cette banque")
```

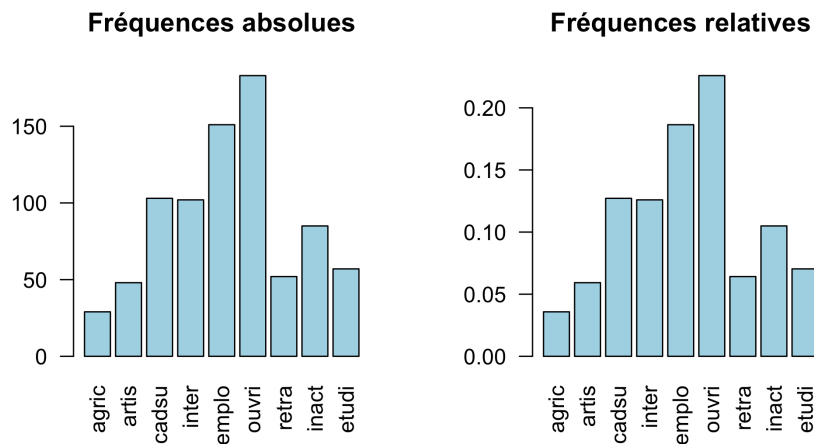
CSP présentes dans cette banque



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(banque$csp), main="Fréquences absolues", col="lightblue", las=2)
barplot(summary(banque$csp)/length(banque$csp), main="Fréquences relatives", col="lightblue", las=2)
```



Notez que lorsque la variable qualitative a ses modalités ordonnées, seule la représentation en bâtons a un sens.

Exercice. Réaliser l'étude de la variable 'solde sur le compte courant' `soldevu`. Discuter le résultat.

3 La sécurité routière en France en 2009 et 2010

A l'aide de toutes les informations précédentes, répondre aux différentes questions suivantes.

1. Les régions françaises contiennent-elles le même nombre de département ?
2. Analysez le nombre de tués sur les routes en 2009 (choisir les paramètres et les graphiques les plus informatifs).
3. Analysez le nombre de tués sur les routes en 2010 (choisir les paramètres et les graphiques les plus informatifs).
4. En utilisant vos connaissances du moment, essayez de structurer une information liant les nombres de tués sur les routes en 2009 et 2010.