TD Exploration des données avec 

Magali Berland
magali.berland@uvsq.fr
Option "Introduction à la Modélisation en Biologie" M2 IHPS

20 novembre 2017

Table des matières

1 Outils de travail

- Installez R : https ://cran.r-project.org/
- Installez RStudio : https ://www.rstudio.com/products/rstudio/download/
- Créez un nouveau projet (File > New project)
- Mettez en place la gestion des versions avec git (Tools > Project Options > Git/SVN)
- Commencez un nouveau fichier au format "Notebook"
- Vous rédigerez votre compte rendu de TD sous forme d'un Notebook.

2 Statistique descriptive unidimensionnelle

Toute étude sophistiquée d'un corpus de données doit être précédée d'une étude exploratoire à l'aide d'outils, certes rudimentaires mais robustes, en privilégiant les représentations graphiques. C'est la seule façon de se familiariser avec des données et de dépister les sources de problèmes :

- valeurs manquantes, erronées ou atypiques, biais expérimentaux,
- modalités trop rares,
- distributions "anormales" (dissymétrie, multimodalité, épaisseur des queues),
- incohérences, liaisons non linéaires.

C'est ensuite la recherche de pré-traitements des données afin de corriger les sources de problèmes et les rendre exploitables par des techniques plus sophistiquées :

- transformation des variables : logarithme, puissance, réduction, rangs,...
- codage en classe ou recodage de classes,
- imputations ou non des données manquantes.

Ensuite, les techniques exploratoires multidimensionnelles permettent des

- représentations graphiques synthétiques,
- réductions de dimension pour la compression ou le résumé des données,
- recherches et représentations de typologies des observations.

Chargez (et installez si nécéssaire) les `library` suivantes :

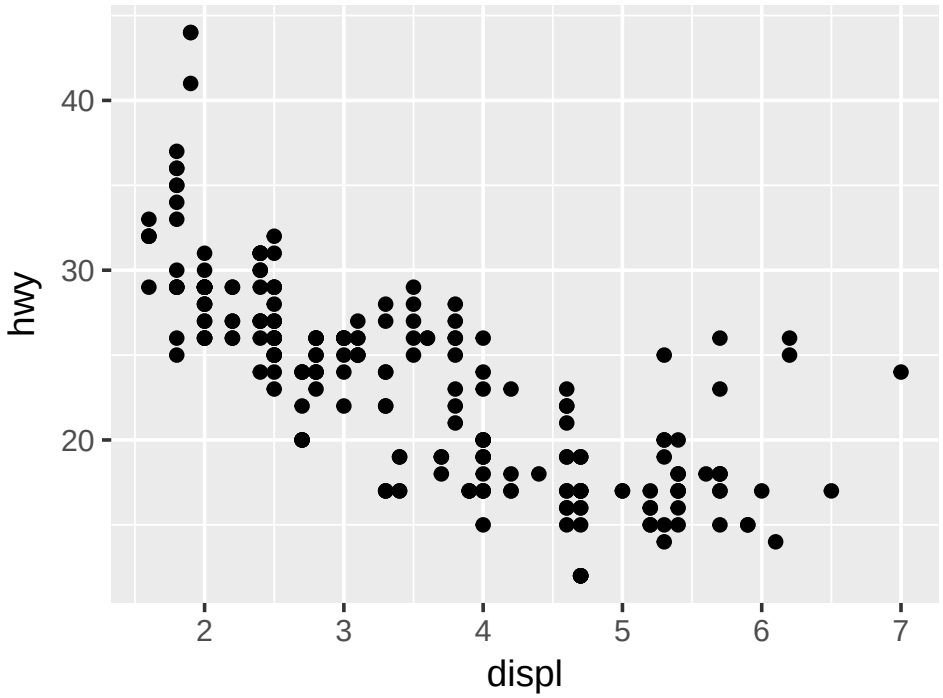
```
library(tidyverse)
library(ade4)
```

2.1 Variable quantitative

Une variable est quantitative si elle reflète une notion de grandeur, c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesure qui sert de référence.

Nous allons dans un premier temps travailler sur le jeu de données `mpg` (n'hésitez pas à lire la page d'aide). Répondez aux questions suivantes :

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```



1. Identifiez les variables quantitatives de ce jeu de données
2. Que décrives les variables `displ`, `hwy` et `drv` ?
3. Que représente le graphique ci-dessus ?
4. Exécutez la commande `ggplot(data = mpg)`. Que voyez-vous ?
5. Faites le nuage de points de `hwy` en fonction de `cyl`. Que représente ce graphique ?
6. Sur le premier graphique, une poignée de points semblent en dehors de la tendance. Quelles hypothèses pouvez-vous formuler pour expliquer cela ?
7. Proposez une représentation graphique qui valide ou infirme votre hypothèse
8. Traces un histogramme de la variable `hwy`. Que remarquez-vous ?

2.2 La médiane et les quartiles : la notion de quantile

La médiane est le quantile d'ordre 1/2 : elle partage la série des observations en deux ensembles d'effectifs égaux.
Le premier quartile est le quantile d'ordre 1/4 , le troisième quartile celui d'ordre 3/4 (le second

quartile est donc confondu avec la médiane).

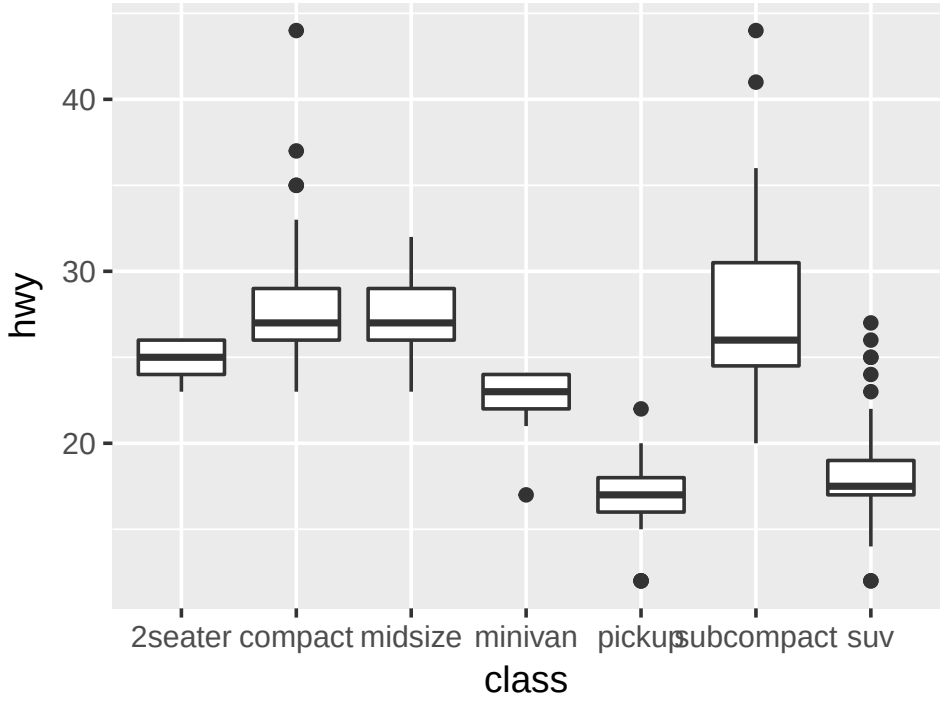
Calculez les quantiles de la variable `hwy` et explicitez leur signification :

```
  0%  25%  50%  75% 100%
  12   18   24   27   44
```

```
30%
 19
```

```
  99%
39.68
```

Tracez des diagrammes boîte à moustache représentant la variable `hwy` en fonction de la variable `class`. Il s'agit d'un graphique très simple qui résume la ou les variables quantitatives à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane.



Caractéristiques numériques

Les caractéristiques (ou résumés) numériques introduites ici servent à synthétiser les variables quantitatives au moyen d'un petit nombre de valeurs numériques. On distingue essentiellement les caractéristiques de tendance centrale (ou encore de position ou de localisation) et les caractéristiques de dispersion.

Tendance centrale Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est–à–dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont

— la médiane `median()`,
— la moyenne `mean()`

Dispersion Elles servent à préciser la variabilité de la série, c'est–à–dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

— L'étendue `range()`,
— l'intervalle interquartiles ($x3/4 - x1/4$),
— l'écart–type `sd()` (racine carrée de la variance)
Calculez ces caractéristiques numériques pour les variables quantitatives du jeu de données.

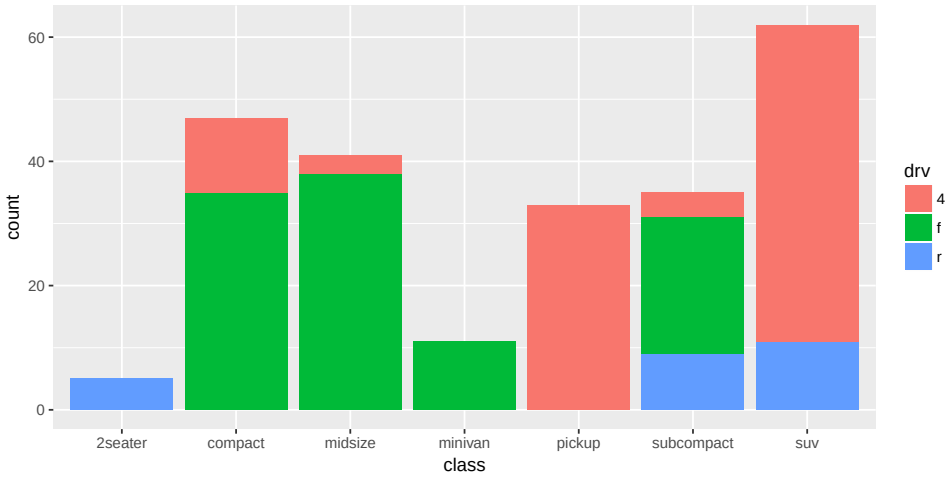
2.3 Variables qualitatives

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées modalités. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite ordinale. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite nominale.

Identifiez les variables qualitatives ordinales et nominales du jeu de données.

Traitement statistique : Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques. Les notions d'effectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).

Tracez ce diagramme en barres et commentez-le



3 Statistique descriptive bidimensionnelle

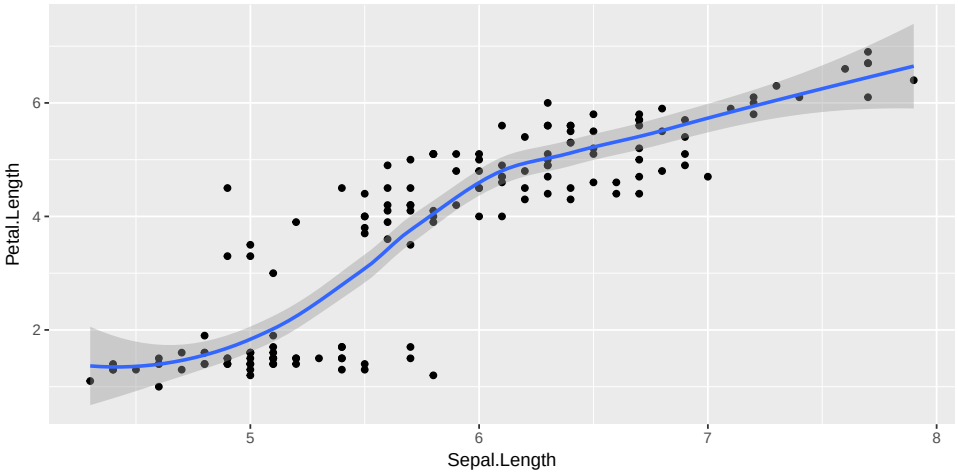
L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors liaison. Dans certains cas, cette liaison peut être considérée a priori comme causale, une variable X expliquant l'autre Y ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations et une liaison n'entraîne pas nécessairement une causalité.

Le nuage de points (ou diagramme de dispersion)

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y, puis à représenter chaque individu observé

par les coordonnées des valeurs observées. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé nuage.

Tracez le graphe suivant (jeu de données `iris`).



Indice de liaison

Le coefficient de corrélation linéaire est un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Il est défini à partir de la covariance qui généralise à deux variables la notion de variance. La covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque". C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (coefficient de Pearson) comme le rapport entre la covariance et le produit des écarts-types.

```
cor(iris$Sepal.Length, iris$Petal.Length)
```

```
[1] 0.8717538
```

Calculez les corrélations pour toutes les variables quantitatives deux à deux (jeu de données `mpg`). Proposez une manière de représenter cette information graphiquement.