Statistiques descriptives avec R

Christophe Ambroise

2024-04-26

Table of contents

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

1 Le langage R de base (R base)

1.1 Qu'est-ce que R ?

1.1.1 En bref

R est un logiciel de développement scientifique spécialisé dans le calcul et l'**analyse statistique**

1.1.2 Mais aussi

- un langage interprété,
- un environnement de développement,
- un projet open source (projet **GNU**),
- un logiciel multi-plateforme (Linux, Mac, Windows),
- la 18ième lettre de l'alphabet

1.2 Principales fonctionnalités

1. Gestionnaire de données - Lecture, manipulation, stockage.
2. Algèbre linéaire - Opérations classiques sur vecteurs, tableaux et matrices
3. Statistiques et analyse de données - Dispose d'un *grand* nombre de méthodes d'analyse de données (des plus anciennes et aux plus récentes)
4. Moteur de sorties graphiques - Sorties écran ou fichier
5. Système de modules - Alimenté par la communauté (+ de 2000 extensions!)
6. Interface facile avec C/C++, Fortran

1.3 Historique

1.1.1 Chronologie

1970s développement de S au Bell labs.
1980s développement de S^+ au AT&T. Lab
1993 développement de R sur le modèle de par Robert Gentleman et Ross Ihaka au département de statistique de l'université d'Auckland.
1995 dépôts des codes sources sous licence GNU/GPL
1997 élargissement du groupe
2002 la fondation dépose ses statuts sous la présidence de Gentleman et Ihaka
2010 Les débuts de Rstudio

1.4 Se tenir informé

1. La page web de la **fondation**

- `les statuts, des liens, des références.`

- `<http://www.r-project.org/>`

2. La page web du **CRAN** (Comprehensive R Arxiv Network)

- `binaires d'installation, packages, documentations, ...`
- `<http://cran.r-project.org/>`

3. La **conférence** des utilisateurs de R:
    - annuelle, prochaine édition à Toulouse
    - http://www.user2019.fr/
4. **The R journal** propose des articles sur
    - de nouvelles extensions, des applications, des actualités.
    - http://journal.r-project.org/

1.5 Qualités et défauts de R

1.5.1 Plus

- Libre et gratuit,
- Richesse des modules (en statistique),
- Rapidité d'exécution,
- Développement rapide (langage de scripts),
- Syntaxe intuitive et compact,
- Nombreuses possibilités graphiques.

1.5.2 Moins

- Aide intégrée succincte,
- Debugger un peu sec,
- Code parfois illisible (compacité),
- Personnalisation des graphiques un peu lourde.

1.6 Les concurrents plus ou moins directs

Les logiciels de développement scientifique sont spécialisés en

1. algèbre linéaire
    - Matlab de Mathworks, une référence,
    - Scilab de l'INRIA, l'alternative libre de matlab,

- Octave de GNU, l'alternative open source ,

2. statistiques
   - SAS (SAS Inc.), la référence,
   - S-PLUS (TIBCO), le concurrent,
3. calcul symbolique
   - Mathematica (Wolfram), la référence,
   - Mapple (Maplesoft), la référence aussi,
   - Maxima (GNU), l'alternative open source .
4. Autres
   - Julia
   - Python, le concurrent le plus sérieux

1.7 Pléthore de livres sur R

- Anciens ouvrages de référence
- R base, un livre de base
- Nouvelle mode: Wickam et co

2 R Studio comme interface

2.1 Les fonctionnalités principales

- Il fournit un éditeur intégré,
- fonctionne sur toutes les plates-formes (y compris sur des serveurs) et
- offre de nombreux avantages tels que l'intégration avec la version contrôle et gestion de projet.

2.2 Les panneaux de Rstudio

Lorsque vous ouvrez RStudio pour la première fois, vous serez accueilli par trois panneaux:

- La console interactive R (entier gauche)
- Environnement / Histoire (onglet en haut à droite)
- Files / Plots / Packages / Help / Viewer (onglet en bas à droite)

Une fois que vous ouvrez des fichiers, tels que des scripts R, un panneau d'éditeur s'ouvre également en haut à gauche.

2.3 Interagir avec R

Il existe deux manières principales d'interagir avec R:

- en utilisant la console
- ou en utilisant des fichiers de script (fichiers texte contenant votre code).

2.4 La fenêtre de la console (dans RStudio, le panneau en bas à gauche)

- endroit où R vous attend pour lui dire quoi faire et où il montrera les résultats d'une commande
- Vous pouvez taper des commandes directement dans la console,
- mais elles seront oubliées lorsque vous fermerez la session.

2.5 Editeur de script

- Il est préférable d'entrer les commandes dans l'éditeur de script
- et de sauvegarder le script
- envoyer la ligne actuelle ou le texte sélectionné à la console R à l'aide du raccourci Ctrl-Entrée

2.5.1 Raccoucis console - éditeur

Ctrl-1 et Ctrl-2 qui vous permettent de sauter entre le script et les fenêtres de la console.

2.6 Commentaires

- Utilisez `#` signes pour commenter.
- Commentez libéralement dans vos scripts R.
- Tout ce qui se trouve à droite d'un # est ignoré par R.

2.7 Astuce: Exécution de segments de votre code

- RStudio vous offre une grande souplesse dans l'exécution du code depuis l'éditeur fenêtre. Il existe des boutons, des choix de menus et des raccourcis clavier. Pour exécuter la ligne en cours, vous pouvez

1. cliquez sur le bouton `Run` situé au-dessus du panneau de l'éditeur,
2. sélectionnez "Run Lines" dans le menu "Code", ou
3. appuyez sur Ctrl-Entrée dans Windows ou Linux ou sur Commande-Entrée sur OS X. (Ce raccourci peut également être vu en survolant la souris sur le bouton).

2.8 Bloc de code

- Pour lancer un bloc de code, sélectionnez-le, puis Exécuter.
- Si vous avez modifié une ligne de code dans un bloc de code que vous venez d'exécuter, il n'est pas nécessaire de resélectionner la section et `Run`, vous pouvez utiliser le bouton suivant,
- Ré-exécuter la région précédente Cela exécutera le bloc de code précédent inculquer les modifications que vous avez apportées.

2.9 Utiliser R comme calculatrice

La chose la plus simple que vous puissiez faire avec R est l'arithmétique:

```
1 + 100
```

```
[1] 101
```

Et R imprimera la réponse, avec un précédent "[1]". Ne t'inquiète pas pour ça pour l'instant, nous l'expliquerons plus tard. Pour l'instant, considérez-le comme indiquant une sortie.

2.10 Commande incomplète

Comme bash, si vous tapez une commande incomplète, R vous attendra pour complèter l'entrée:

```
> 1 +

+
```

Chaque fois que vous appuyez sur Entrée et que la session R affiche un "+" au lieu d'un ">", signifie qu'il attend que vous complétiez la commande. Si vous voulez annuler une commande que vous pouvez simplement appuyer sur "Esc" et RStudio vous rendra le ">" rapide.

2.11 Astuce: Annulation des commandes

- Si vous utilisez R depuis la ligne de commande plutôt que depuis RStudio,
- vous devez utiliser `Ctrl + C` au lieu de`Esc` pour annuler la commande. Ce s'applique également aux utilisateurs de Mac!
- Annuler une commande n'est pas seulement utile pour tuer des commandes incomplètes:
- vous pouvez aussi l'utiliser pour dire à R d'arrêter d'exécuter du code (par exemple si
- en prenant beaucoup plus de temps que prévu, ou pour se débarrasser du code que vous êtes en train d'écrire.

2.12 Ordre des opérations

Lorsque vous utilisez R comme calculatrice, l'ordre des opérations est le même que vous aurait appris à l'école.

De la plus haute à la plus basse préséance:

- Parenthèses: `(`, `)`
- Exposants: `^` ou `**`
- Diviser: `/`
- Multiplier: `*`
- Ajouter: `+`
- Soustraire: `-`

```
3 + 5 * 2
```

```
[1] 13
```

Utilisez des parenthèses pour regrouper les opérations afin de forcer l'ordre de évaluation si elle diffère du défaut, ou pour préciser ce que vous avoir l'intention

```
(3 + 5) * 2
```

```
[1] 16
```

Cela peut devenir compliqué lorsque cela n'est pas nécessaire, mais clarifie vos intentions. Rappelez-vous que d'autres peuvent lire votre code ultérieurement.

```
(3 + (5 * (2 ^ 2))) # difficile à lire
3 + 5 * 2 ^ 2 # si vous vous souvenez des règles
3 + 5 * (2 ^ 2) # Si vous oubliez certaines règles, cela pourrait vous aider
```

Le texte après chaque ligne de code s'appelle un "commentaire". Tout ce qui suit le symbole de hachage (ou octothorpe) "#" est ignoré par R lorsqu'il exécute du code.

Les nombres vraiment petits ou grands ont une notation scientifique:

```
2/10000
```

```
[1] 2e-04
```

Ce qui est un raccourci pour "multiplié par`10 ^ XX`". Donc `2e-4` est un raccourci pour `2 * 10 ^ (- 4)`.

Vous pouvez aussi écrire des nombres en notation scientifique:

```
5e3 # Notez l'absence de moins ici
```

```
[1] 5000
```

2.13 Fonctions mathématiques

R a beaucoup de fonctions mathématiques intégrées. Pour appeler une fonction, nous tapons simplement son nom, suivi par des parenthèses ouvertes et fermantes. Tout ce que nous tapons à l'intérieur des parenthèses s'appelle la fonction arguments:

```r
sin (1) # fonctions trigonométriques
```

```
[1] 0.841471
```

```r
log (1) # logarithme naturel
```

```
[1] 0
```

```r
log10 (10) # base-10 logarithme
```

```
[1] 1
```

```r
exp (0.5) # e ^ (1/2)
```

```
[1] 1.648721
```

2.14 Se souvenir

Ne vous souciez pas d'essayer de vous souvenir de toutes les fonctions de R:

- rechercher sur Google,
- ou si vous vous souvenez de la début du nom de la fonction, utilisez *complétion* dans RStudio.

C'est un avantage que RStudio sur R , il dispose de *capacités d'auto-complétion* qui vous permettent plus facilement rechercher des fonctions, leurs arguments et les valeurs qu'ils prendre.

Taper un `?` Avant le nom d'une commande ouvrira la page d'aide pour cette commande. En plus de fournir

- une description la page d'aide affichera généralement
- une collection d'exemples

2.15 Comparer les choses

Nous pouvons également faire la comparaison en R:

```r
1 == 1 # égalité (noter deux signes égaux, lire comme "est égal à")
```

```
[1] TRUE
```

```r
1 != 2 # inégalité (lire comme "n'est pas égal à")
```

```
[1] TRUE
```

```r
1 <2 # moins que
```

```
[1] TRUE
```

```r
1 <= 1 # inférieur ou égal à
```

```
[1] TRUE
```

```r
1> 0 # plus grand que
```

```
[1] TRUE
```

```r
1 >= -9 # supérieur ou égal à
```

```
[1] TRUE
```

2.16 Astuce: Comparer les nombres

- Un mot d'avertissement sur la comparaison des chiffres: vous devriez ne jamais utiliser `==` pour comparer deux nombres à moins qu'ils ne soient entiers (un type de données pouvant représenter spécifiquement uniquement des nombres entiers).
- Les ordinateurs ne peuvent représenter que des nombres décimaux avec un un certain degré de précision, donc deux nombres qui semblent le même lorsqu'il est imprimé par R, peut effectivement avoir différentes représentations sous-jacentes et donc être différent par une petite marge d'erreur (appelée Machine tolérance numérique).
- Au lieu de cela, vous devez utiliser la fonction `all.equal`.
- Lectures complémentaires: http://floating-point-gui.de/

2.17 Variables et affectation

Stocker des valeurs dans des variables en utilisant l'opérateur d'affectation `<-`

```
x <- 1/40
```

```
x
```

```
[1] 0.025
```

Plus précisément, la valeur stockée est une approximation * décimale * de cette fraction appelée [nombre à virgule flottante] (http://en.wikipedia.org/wiki/Floating_point).

Recherchez l'onglet `Environment` dans l'un des volets de RStudio, et vous verrez que `x` et sa valeur est apparu. Notre variable `x` peut être utilisée à la place d'un nombre dans tout calcul qui attend un nombre:

```
log (x)
```

```
[1] -3.688879
```

Notez également que les variables peuvent être réaffectées:

```
x <- 100
```

`x` contenait la valeur 0.025 et a maintenant la valeur 100.

Les valeurs d'affectation peuvent contenir la variable affectée à:

```
x <- x + 1 #notice comment RStudio met à jour sa description de x en haut à droite
```

Le côté droit de l'affectation peut être toute expression R valide. Le côté droit est *entièrement évalué* avant que l'affectation ait lieu.

2.18 Noms de variables

Les noms de variables peuvent contenir des lettres, des chiffres, des traits de soulignement et des points. Ils ne peut pas commencer avec un nombre ni contenir des espaces du tout. Différentes personnes utilisent différentes conventions pour les noms de variables longues, celles-ci comprennent

- periods.between.words
- souligne_entre_mots
- camelCaseToSeparateWords

Ce que vous utilisez dépend de vous, mais **soyez cohérent**.

2.19 Affectation (suite)

Il est également possible d'utiliser l'opérateur `=` pour l'affectation:

```
x = 1/40
```

Mais c'est beaucoup moins courant parmi les utilisateurs de R.

Donc, la recommandation est d'utiliser `<-`.

2.20 Gérer votre environnement

Il existe quelques commandes utiles que vous pouvez utiliser pour interagir avec la session R.

`ls` listera toutes les variables et fonctions stockées dans l'environnement global (votre session de travail):

```
ls()
```

```
[1] "has_annotations" "x"
```

2.21 Astuce: objets cachés

- Comme dans le shell, `ls` cachera toutes les variables ou fonctions commençant avec un "." par défaut.
- Pour lister tous les objets, tapez `ls (all.names = TRUE)`

2.21.1 Remarque

Notez ici que nous n'avons donné aucun argument à `ls`, mais nous avons quand même nécessaire de donner aux parenthèses de dire à R d'appeler la fonction.

2.22 Contenu d'une fonction

Si vous tapez `ls` par lui-même, R imprimera le code source de cette fonction!

```
ls
```

```
function (name, pos = -1L, envir = as.environment(pos), all.names = FALSE,
    pattern, sorted = TRUE)
{
    if (!missing(name)) {
        pos <- tryCatch(name, error = function(e) e)
        if (inherits(pos, "error")) {
            name <- substitute(name)
            if (!is.character(name))
                name <- deparse(name)
            warning(gettextf("%s converted to character string",
                sQuote(name)), domain = NA)
            pos <- name
        }
    }
    all.names <- .Internal(ls(envir, all.names, sorted))
    if (!missing(pattern)) {
        if ((ll <- length(grep("[", pattern, fixed = TRUE))) &&
            ll != length(grep("]", pattern, fixed = TRUE))) {
            if (pattern == "[") {
                pattern <- "\\["
                warning("replaced regular expression pattern '[' by  '\\\\['")
            }
            else if (length(grep("[^\\\\]\\[<-", pattern))) {
                pattern <- sub("\\[<-", "\\\\\\[<-", pattern)
                warning("replaced '[<-' by '\\\\[<-' in regular expression pattern")
            }
        }
        grep(pattern, all.names, value = TRUE)
    }
    else all.names
}
<bytecode: 0x7fda093d2ce8>
<environment: namespace:base>
```

2.23 Supression

Vous pouvez utiliser `rm` pour supprimer des objets dont vous n'avez plus besoin:

```
rm (x)
```

Si vous avez beaucoup de choses dans votre environnement et souhaitez les supprimer toutes, vous pouvez transmettre les résultats de `ls` à la fonction `rm`:

```
rm(list=ls())
```

Dans ce cas, nous avons spécifié que les résultats de `ls` devraient être utilisés pour le L'argument `list` dans `rm`. Lorsque vous attribuez des valeurs aux arguments par nom, vous *devez* utilisez l'opérateur `=` !!

Si au lieu de cela nous utilisons `<-`, il y aura des effets secondaires inattendus, ou vous pourriez avoir un message d'erreur:

```
rm(list <- ls ())
```

```
Error in rm(list <- ls()): ... must contain names or character strings
```

2.24 Astuce: Avertissements vs erreurs

- Quand R fait quelque chose d'inattendu! Les erreurs, comme ci-dessus, sont émises lorsque R ne peut pas procéder à un calcul.
- En revanche, les avertissements signifient généralement que la fonction a été exécutée, mais cela n'a probablement pas fonctionné comme prévu.
- Dans les deux cas, le message que R imprime vous donne généralement des indices sur la manière de résoudre un problème.

2.25 R Packages

Il est possible d'ajouter des fonctions à R en écrivant un paquet, ou par obtenir un paquet écrit par quelqu'un d'autre. Au moment de l'écriture, il y a sont plus de 7 000 paquets disponibles sur CRAN (l'archive complète de R réseau). R et RStudio ont des fonctionnalités pour gérer les paquets:

- Vous pouvez voir quels paquets sont installés en tapant

```
installed.packages ()
```

- Vous pouvez installer des paquets en tapant

`install.packages (" packagename ")`, où `packagename` est le nom du package, entre guillemets.

- Vous pouvez mettre à jour les paquets installés en tapant `update.packages ()`
- Vous pouvez supprimer un paquet avec `remove.packages (" packagename ")`
- Vous pouvez rendre un paquet disponible pour être utilisé avec `library (packagename)`

2.26 Exercices

1. Nommer toutes les panels de R studio
2. Réaliser une addition et une multiplication dans la console
3. Sauver un script R avec votre code d'addition

3 Chercher de l'aide

3.1 Lire les fichiers d'aide

R, et chaque paquet, fournissent des fichiers d'aide pour les fonctions. Pour chercher de l'aide sur un fonction d'une fonction spécifique qui est dans un package chargé

```
?nom_fonction
help(nom_fonction)
```

Cela chargera une page d'aide dans RStudio (ou du texte brut dans R seul).

3.2 Opérateurs spéciaux

Pour demander de l'aide sur des opérateurs spéciaux, utilisez des guillemets:

```
?"+"
```

3.3 Obtenir de l'aide sur les packages

De nombreux paquets contiennent des "vignettes": des tutoriels et des exemples de documentation. Sans aucun argument, `vignette()` listera toutes les vignettes pour tous les paquets installés;

`vignette (package =" nom-du-paquet ")`

listera toutes les vignettes disponibles pour

`package-name` et `vignette ("vignette-name")`

ouvriront la vignette spécifiée.

Si un paquet ne contient aucune vignette, vous pouvez généralement trouver de l'aide en tapant

`help("nom-du-paquet")`.

3.4 Quand vous vous souvenez de la fonction

Si vous ne savez pas exactement dans quel paquet est une fonction ou comment elle est spécifiquement orthographiée, vous pouvez effectuer une recherche floue:

```
??nom_fonction
```

4 Mélanger code et texte avec `knitr`

4.1 Motivations: reproductibilité de la recherche

4.1.1 Principe de Claerbout (Géophysicien, Stanford)

> An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

4.2 En bref

4.2.1 `knitr`

pour générer des rapports qui combinent le texte, le code et les résultats.

4.2.2 Markdown

pour mettre en forme le texte

> Markdown est un langage de balisage léger créé par John Gruber en 2004. Son but est d'offrir une syntaxe facile à lire et à écrire. Un document balisé par Markdown peut être lu en l'état sans donner l'impression d'avoir été balisé ou formaté par des instructions particulières. — Wikipedia

Lien Wikipedia

4.2.3 Code Chunks

code dans des blocs délimités par des guillemets triples suivis de `{r}`.

4.3 Référence

Rmardown est un univers extensible, si vous voulez continuez, lisez

https://rmarkdown.rstudio.com/

4.4 Installer `knitr`

4.4.1 Dans la console

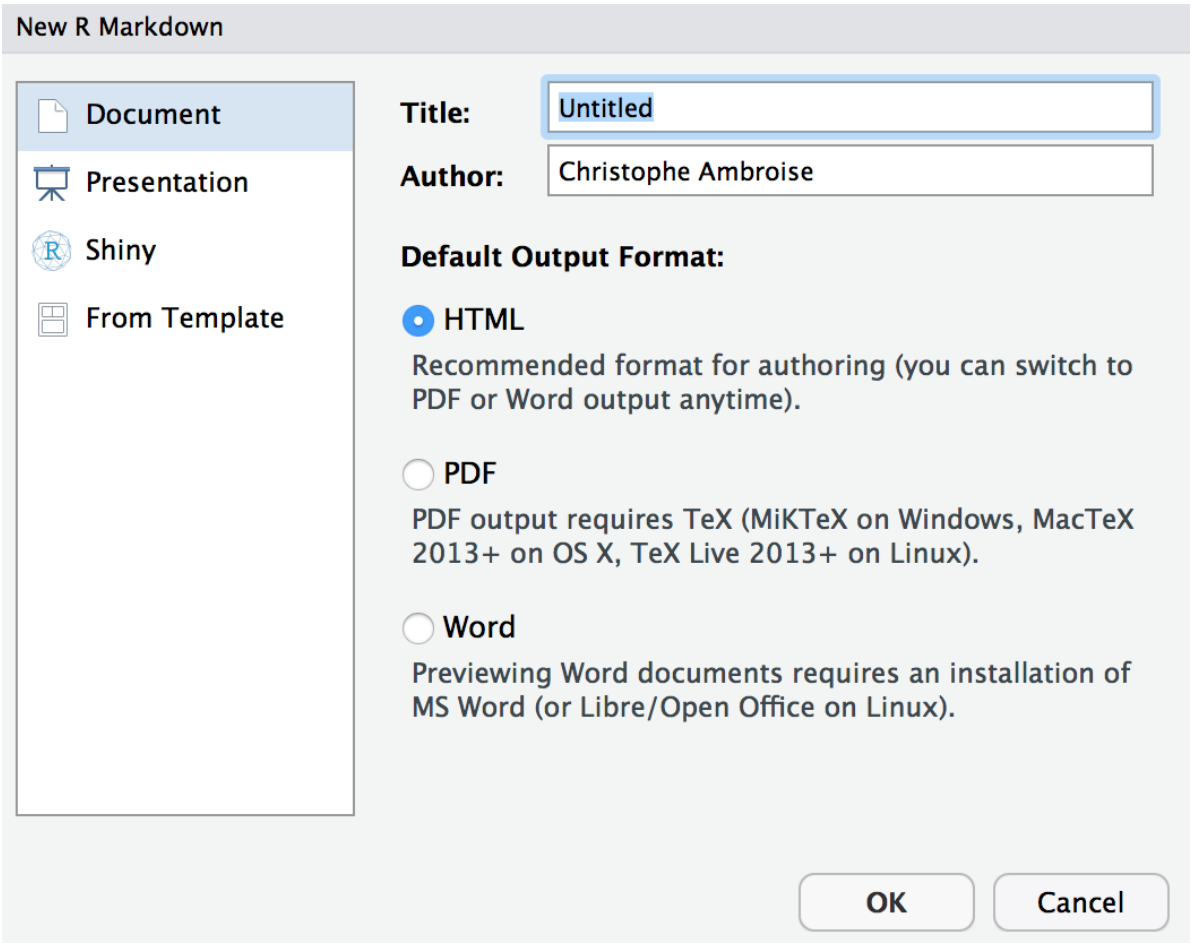
```
install.packages('knitr')
```

4.4.2 Par le menu

Tools -> Install Packages

4.5 Créer un de type fichier Rmarkdown (.Rmd)

Dans R Studio, cliquez sur Fichier → Nouveau fichier → R Markdown et vous obtiendrez une boîte de dialogue du type



4.6 YAML

Vous créez un fichier avec une entête dite yaml du type

```
---
title: "Initial R Markdown document"
author: "Karl Broman"
date: "April 23, 2015"
output: html_document
---
```

qui précise comment peut être transformé le fichier

4.7 Markdown

Markdown est un langage à balise

- écrivez **en gras** en utilisant deux astérisques, comme ceci: `**gras**`,
- écrivez en *italics* en utilisant des traits de soulignement, comme ceci: `_italics_`.

Vous pouvez créer une liste à puces en écrivant une liste avec des tirets ou astérisques, comme ceci:

```
* gras avec double astérisque
* italiques avec des soulignés
* police de type code avec backticks
```

ou comme ça:

```
- gras avec double astérisque
- italiques avec des soulignés
- police de type code avec backticks
```

Vous pouvez créer une liste numérotée en utilisant simplement des chiffres. Vous pouvez utiliser le même nombre encore et encore si vous voulez:

```
1. gras avec double astérisque
1. italiques avec des soulignés
1. police de type code avec backticks
```

Cela apparaîtra comme:

1. gras avec double astérisque
2. italiques avec des soulignés
3. police de type code avec backticks

Vous pouvez créer des en-têtes de section de différentes tailles en initiant une ligne avec un certain nombre de symboles `#`:

```
# Titre
## Section principale
### Sous-section
#### Sous-sous-section
```

4.8 Compilation

Vous *compilez* le document R Markdown en cliquant sur le "Knit HTML" en haut à gauche.

4.9 Un peu plus de Markdown

- Vous pouvez créer un hyperlien comme celui-ci:

```
[texte à afficher] (http://the-web-page.com).
```

- Vous pouvez inclure un fichier image comme ceci:

```
![Caption](http://url/for/file)
```

Vous pouvez faire des indices (par exemple, F_2) avec `F~2~` et des exposants (par exemple, F^2) avec `F^2^`.

Si vous savez écrire des équations dans [LaTeX] (http://www.latex-project.org/), vous serez heureux de savoir que vous pouvez utiliser `$ $` et `$$ $$` pour insérer des équations mathématiques, comme `$E = mc^2$` $E = mc^2$ et

```
$$y = \mu + \sum_ {i = 1}^p \beta_i x_i + \epsilon$$
```

$$y = \mu + \sum_{i=1}^{p} \beta_i x_i + \epsilon$$

4.10 Morceaux de code

Markdown est intéressant et utile, mais le plus grand intérêt vient du mélange du texte balisé avec des morceaux de code R.

- Quand traité, le code R sera exécuté; s'ils produisent des valeurs, figures, ceux-ci seront insérés dans le document final.

Les morceaux de code ressemblent à ceci:

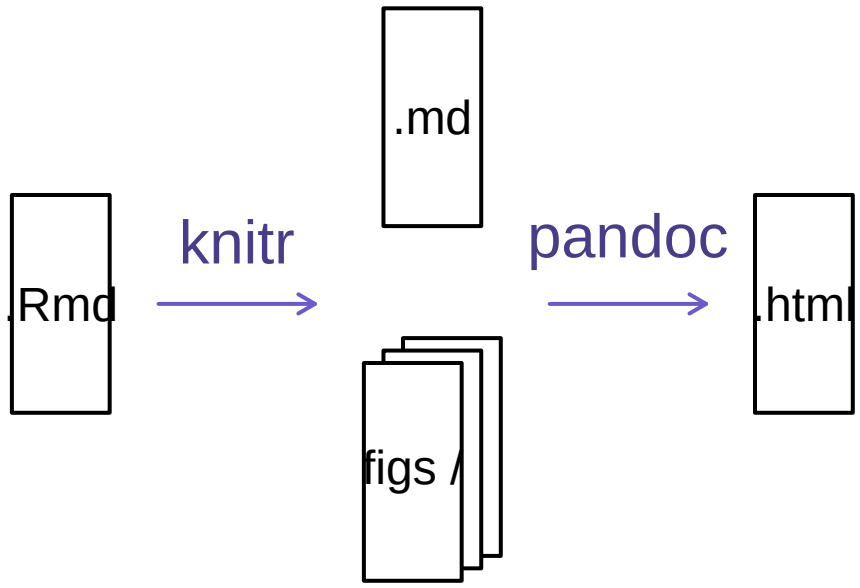
````
```{r load_data}
gapminder <- read.csv("~/Desktop/gapminder.csv")
```
````

C'est une bonne idée de donner chaque morceau un nom, car ils vous aideront à corriger les erreurs et, si des graphiques sont produit, les noms de fichiers sont basés sur le nom du bloc de code les a produites.

Lorsque vous appuyez sur le bouton "Knit HTML", le document R Markdown est traité par [knitr] (http://yihui.name/knitr) et un simple Markdown document est produit (ainsi que, potentiellement, un ensemble de fichiers de figure): le code R est exécuté et remplacé à la fois par l'entrée et la sortie; si les chiffres sont produits, des liens vers ces chiffres sont inclus.

Les documents Markdown et figure sont ensuite traités par l'outil pandoc, qui convertit le fichier Markdown en fichier fichier html, avec les chiffres incorporés.

4.11 Knitr en diagramme



4.12 Options de blocs

Il y a une variété d'options pour affecter la façon dont les morceaux de code sont traités.

- Utilisez `echo = FALSE` pour éviter que le code lui-même ne soit affiché.
- Utilisez `results ="hide"` pour éviter d'imprimer des résultats.
- Utilisez `eval = FALSE` pour que le code soit affiché mais pas évalué.
- Utilisez `warning = FALSE` et `message = FALSE` pour masquer les avertissements ou messages produits.
- Utilisez `fig.height` et `fig.width` pour contrôler la taille des figures produit (en pouces).

4.13 Resources

- Knitr in a knutshell tutorial
- Dynamic Documents with R and knitr (book)
- R Markdown documentation
- R Markdown cheat sheet

4.14 Exercice

1. Créer un fichier Rmarkdown qui produira une sortie html avec un code chunk de votre choix

5 Structures de données

5.1 Vecteurs: définition

5.1.1 Propriétés

- objet le plus *élémentaire* sous R,
- collection d'entités *de même nature*,
- (ou type) défini par la nature des entités qui le composent.

5.1.2 Les modes possibles

1. numérique
2. caractère
3. logique

5.2 Quelques vecteurs typés

1. Numérique

```
x0 <- 0
x1 <- c(-1,23,98.7)
mode(x0)
```

```
[1] "numeric"
```

2. Caractère

```r
y0 <- "bonjour"
y1 <- c("Pomme","Flore","Alexandre")
mode(y1)
```

```
[1] "character"
```

3. Logique

```r
z0 <- TRUE
z1 <- c(FALSE,TRUE,FALSE,TRUE,TRUE)
z2 <- c(T,F,F)
mode(z2)
```

```
[1] "logical"
```

5.3 Remarques sur l'affectation

5.3.1 Affectation

C'est l'opération qui consiste à *attribuer une valeur* à une variable.

En R, plusieurs choix sont possibles:

5.3.2 l'opérateur usuel est <- (signe inférieur suivi du signe moins)

```r
jo <- "l'indien"
jo
```

```
[1] "l'indien"
```

5.3.3 l'opérateur = peut être utilisé la plupart du temps

```r
nb.max.d.annees.pour.faire.une.these = 3
nb.max.d.annees.pour.faire.une.these
```

```
[1] 3
```

5.3.4 la commande `assign` permet cette opération (d'où l'anglicisme *assignation*)

```
assign("x", c(8,9,-pi,sqrt(2)))
```

5.4 Valeurs spéciales

5.4.1 Variables réservées par R

- `NA` est le code R pour les valeurs manquantes (absentes des données),
- `NaN` est le code de R pour signifier un résultat numérique aberrant ,
- `Inf` et `-Inf` sont les valeurs réservées pour plus et moins ∞,
- `NULL` est l'objet nul.

```
c(4,2,NA,5)
```

```
[1]  4  2 NA  5
```

```
0/0
```

```
[1] NaN
```

```
1/0
```

```
[1] Inf
```

```
names(1)
```

```
NULL
```

5.5 Opérations arithmétiques

Les opérations sur les vecteurs s'effectuent terme-à-terme

Soient `x,y` tels que

```r
x<-c(1,2,-3,-4)
y<-c(-5,-6,9,0)
```

5.5.1 +

addition des éléments de deux vecteurs

```r
x+y
```

```
[1] -4 -4  6 -4
```

5.5.2 -

soustraction des éléments de deux vecteurs

```r
x-y
```

```
[1]   6   8 -12  -4
```

5.5.3 *

multiplication des éléments de deux vecteurs

```r
x*y
```

```
[1]  -5 -12 -27   0
```

5.5.4 \

division des éléments de deux vecteurs

```r
x/y
```

```
[1] -0.2000000 -0.3333333 -0.3333333       -Inf
```

5.6 Recyclage des éléments du vecteur

Lors d'une opération entre vecteurs, les vecteurs trop courts sont ajustés pour atteindre la taille du plus grand vecteur en recyclant les données.

⇝ souvent pratique mais **attention aux effets de bords!**

5.7 Opérateurs mathématiques

5.7.1 Fonctions numériques élémentaires

`floor,ceiling,round`

```
floor(2/3)
```

```
[1] 0
```

```
ceiling(2/3)
```

```
[1] 1
```

```
round(2/3,3)
```

```
[1] 0.667
```

:::

5.8 Fonctions arithmétiques élémentaires

`` `\^{},\%\%,\%/\%,abs,log,exp,log10,sqrt,cos,tan,sin...` ``

s'appliquent toutes terme-à-terme.

```
log10(c(10,100,1000))
```

```
[1] 1 2 3
```

```
cos(c(pi/2,pi))^2 + sin(c(pi/2,pi))^2
```

```
[1] 1 1
```

5.9 Fonctions spécifiques à un vecteur

prod, sum, max, min, range, which.min, which.max, length

```
x <- c(-8,1.5,3)
prod(x)
```

```
[1] -36
```

```
sum(x)
```

```
[1] -3.5
```

```
length(x)
```

```
[1] 3
```

```
max(x)
```

```
[1] 3
```

```
min(x)
```

```
[1] -8
```

```
range(x)
```

```
[1] -8  3
```

```
which.max(x)
```

```
[1] 3
```

```
which.min(x)
```

```
[1] 1
```

Pour le minimum / maximum terme-à-terme: `pmin,pmax`.

5.10 Fonctions appliquées le long du vecteur

`` `cumsum, cumprod, cummin, cummax` ``

```
x <- c(-2,1,-3,2)
cumprod(x)
```

```
[1] -2 -2  6 12
```

```
cumsum(x)
```

```
[1] -2 -1 -4 -2
```

```
cummax(x)
```

```
[1] -2  1  1  2
```

```
cummin(x)
```

```
[1] -2 -2 -3 -3
```

5.11 Opérateurs ensemblistes

Fonctionnent pour tous les modes

`unique,intersect,union,setdiff,setequal,is.element`

```r
unique(c("banane","citron","banane"))
```

```
[1] "banane" "citron"
```

```r
intersect(c("banane","citron"),c("orange","banane"))
```

```
[1] "banane"
```

```r
union(c("banane","citron"),c("orange","banane"))
```

```
[1] "banane" "citron" "orange"
```

```r
setequal(c("banane","citron"),c("orange","banane"))
```

```
[1] FALSE
```

```r
is.element(1,sample(c(1,2,3),2))
```

```
[1] TRUE
```

5.12 Génération de vecteurs

Il existe de nombreuse manière de générer des vecteur de manière plus ou moins automatique

R étant un langage vectoriel savoir générer le vecteur que l'on veut représente un atout important pour programmer en R

5.13 L'opérateur :

```
from:to
```

```
Génère une séquence par pas de un depuis le nombre `from`
jusqu'à `to` (si possible).
```

```
-5:5
```

```
 [1] -5 -4 -3 -2 -1  0  1  2  3  4  5
```

```
5:-5
```

```
 [1]  5  4  3  2  1  0 -1 -2 -3 -4 -5
```

```
pi:6
```

```
[1] 3.141593 4.141593 5.141593
```

```
1:6/2
```

```
[1] 0.5 1.0 1.5 2.0 2.5 3.0
```

```
1:(6/2)
```

```
[1] 1 2 3
```

5.14 La commande seq

5.14.1 Plusieurs schémas possibles

- seq(from,to)
- seq(from,to,by=)
- seq(from,to,length.out=)

5.14.2 Exemple

```
seq(1,10)
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```r
seq(2,10,by=2)
```

```
[1]  2  4  6  8 10
```

```r
seq(2,10,length.out=6)
```

```
[1]  2.0  3.6  5.2  6.8  8.4 10.0
```

5.15 La commande `rep`

5.15.1 Fonctionne pour tous les modes

- `rep(x,times)`, où `times` peut être un vecteur,
- `rep(x,each)`.

5.15.2 Exemple

```r
rep(1,3)
```

```
[1] 1 1 1
```

```r
rep("Mercy",3)
```

```
[1] "Mercy" "Mercy" "Mercy"
```

```r
rep(c("A","B","C"),c(3,2,4))
```

```
[1] "A" "A" "A" "B" "B" "C" "C" "C" "C"
```

```r
rep(c(TRUE,FALSE), each=2)
```

```
[1]  TRUE  TRUE FALSE FALSE
```

5.16 Génération de vecteurs logiques

5.16.1 Obtenus par conditions avec

- les opérateurs logiques `'<'`, `'<='`, `'>'`, `'>='`, `'=='` `'!='`
- le ET, le OU, NON, OU exclusif: `'\&'` (intersection), `'|'` (union), `'!'` (négation), `xor`.

```
note1 <- c(8,9,14,3,17.5,11)
note2 <- c("C","B","A","B","E","B")
admis <- (note1 >= 10) & (note2 == "A" | note2 == "B")
mention <- (note1 >= 15) & (note2 == "A")
admis
```

```
[1] FALSE FALSE  TRUE FALSE FALSE  TRUE
```

```
sum(admis)
```

```
[1] 2
```

```
sum(mention)
```

```
[1] 0
```

5.17 Par concaténation

5.17.1 Avec `'c()'`

```
L'opérateur `'c()'` peut s'appliquer à n'importe quoi pourvu que
l'on concatène des vecteurs de même type.
```

```
c( c(1,2), c(3,4))
```

```
[1] 1 2 3 4
```

```
round(c(seq(-pi,pi,len=4),rep(c(1:3),each=2),0),2)
```

```
 [1] -3.14 -1.05  1.05  3.14  1.00  1.00  2.00  2.00  3.00  3.00  0.00
```

5.17.2 remarque

Dans le second exemple, les entiers composants `c(1:3)` ont été forcés au typage flottant.

5.17.3 Avec `paste`

`Concaténation de chaînes de caractères. Convertit en caractères les éléments passés en argument avant toute opération.`

```
paste("R","c'est","bien")
```

```
[1] "R c'est bien"
```

```
paste(2:4,"ieme")
```

```
[1] "2 ieme" "3 ieme" "4 ieme"
```

```
paste("A",1:5, sep="")
```

```
[1] "A1" "A2" "A3" "A4" "A5"
```

```
paste("A",1:5, sep="",collapse="")
```

```
[1] "A1A2A3A4A5"
```

5.18 Indexation des vecteurs

5.18.1 Principe

- Permet la **sélection d'un sous-ensemble** du vecteur `x`.
- Le sous-ensemble est spécifié **entre crochets** `x[subset]`.

5.18.2 L'objet `subset` peut prendre 4 types différents

1. **un vecteur logique**, qui doit être de la même taille que le vecteur `x`;
2. **un vecteur numérique aux composantes positives**, qui spécifie les valeurs à inclure;
3. **un vecteur numérique aux composantes négatives**, qui spécifie les valeurs à exclure;
4. **un vecteur de chaînes de caractères**, qui spécifie les noms des éléments de `x` à conserver.

5.19 Indexation des vecteurs: exemples

Vecteurs logiques

```
x <- c(3,6,-2,9,NA,sin(-pi/6))
x[x > 0]
```

```
[1]  3  6  9 NA
```

```
x[!is.na(x)]
```

```
[1]  3.0  6.0 -2.0  9.0 -0.5
```

```
x[!is.na(x) & x>0]
```

```
[1] 3 6 9
```

```
mean(x,na.rm=TRUE)
```

```
[1] 3.1
```

```
x[x <= mean(x,na.rm=TRUE)]
```

```
[1]  3.0 -2.0   NA -0.5
```

Vecteurs aux composantes positives ou négatives

```
x
```

```
[1]  3.0  6.0 -2.0  9.0   NA -0.5
```

```
x[2]
```

```
[1] 6
```

```
x[1:5]
```

```
[1]  3  6 -2  9 NA
```

```
x[-c(1,5)]
```

```
[1]  6.0 -2.0  9.0 -0.5
```

```
x[-(1:5)]
```

```
[1] -0.5
```

Vecteurs de chaînes de caractères

```
names(x) <- c("var1","var2","var3","var4","var5","var6")
x
```

```
var1 var2 var3 var4 var5 var6
 3.0  6.0 -2.0  9.0   NA -0.5
```

```
x[c("var1","var3")]
```

```
var1 var3
   3   -2
```

5.20 Autres commandes d'indexation et de sélection

1. **Classer**

- `sort` renvoie le vecteur classé par ordre croissant ou décroissant,
- `order` renvoie les indices d'ordre des éléments par ordre croissant ou décroissant,

2. **Extraire**

- `which` renvoie les indices de `x` vérifiant une condition;

3. **Échantillonner**

- `sample` échantillonne aléatoirement dans un vecteur `x`, avec ou sans remise.

5.21 Exemples

```
x <- -5:5
y <- sample(x)
sort(y)
```

```
 [1] -5 -4 -3 -2 -1  0  1  2  3  4  5
```

```
order(y)
```

```
 [1]  6  4  5  7  9  8  3  1 11  2 10
```

```
y[order(y)]
```

```
 [1] -5 -4 -3 -2 -1  0  1  2  3  4  5
```

```r
y[order(y,decreasing=TRUE)]
```

```
 [1]  5  4  3  2  1  0 -1 -2 -3 -4 -5
```

```r
which(sample(x,4) > 0)
```

```
[1] 1 3 4
```

5.22 Facteurs

5.22.1 Facteur

Un **facteur** est un vecteur de **variables catégorielles** .
Les **niveaux** du facteur peuvent être ordonnés ou pas.

5.22.2 Utilisation les facteurs s'utilisent pour

catégoriser les données d'un vecteur (ce qui s'avère très utile pour la gestion des variables qualitatives).

$\rightsquigarrow$ un facteur est souvent associé à d'autres vecteurs pour en définir une **partition**.

5.23 Création, manipulation

5.23.1 Création: la fonction `factor`

```r
factor(sample(1:3,10,replace=TRUE))
```

```
 [1] 3 1 2 2 2 1 2 3 2 1
Levels: 1 2 3
```

```r
factor(sample(1:3,10,replace=TRUE),levels=1:5)
```

```
 [1] 1 2 3 2 2 3 1 2 3 2
Levels: 1 2 3 4 5
```

5.23.2 Gestion: `nlevels,levels,table`

```
x <- factor(sample(c("thésard","CR","MdC"),15,replace=TRUE))
cat(nlevels(x),"niveaux:",levels(x))
```

```
3 niveaux: CR MdC thésard
```

```
table(x)
```

```
x
     CR     MdC thésard
      1       6       8
```

5.24 Un exemple de facteur associé à un vecteur

Un exemple de facteur associé à un vecteur

5.24.1 Données

`Chacun me donne son âge et son grade^[sauf un qui refuse :'( ]`

```
age <- c(25,35,32,27,32,40,26,25,26,28,30,NA,36,30,30)
grd <- c("thd","CR","MdC","thd","thd","MdC","MdC","thd","thd","MdC","CR","MdC","CR","thd",
```

5.24.2 Question : nombre d'individus par catégorie ?

```
table(grd)
```

```
grd
 CR MdC thd
  3   5   7
```

5.25 La fonction `tapply`

Un autre point fort de R

5.25.1 Utilisation

Applique une fonction sur un vecteur partitionné en groupes.

5.25.2 Question : âge moyen / écart-type par catégorie ?

```
tapply(age,grd,mean,na.rm=TRUE)
```

```
      CR      MdC      thd
33.66667 31.50000 27.85714
```

```
tapply(age,grd,sd,na.rm=TRUE)
```

```
      CR      MdC      thd
3.214550 6.191392 2.794553
```

5.26 Matrices (et tableaux)

Les matrices et tableaux sont la des structure de stockage très courantes

5.27 Tableau: définition

5.27.1 objet `array`

Un tableau est un vecteur muni d'un attribut dimension (`dim`),
lui même défini par un vecteur. Il est défini par la commande

`array(data,dim,dimnames=)`

5.27.2 Exemple

```
array(1:8,c(2,2,2))
```

```
, , 1

     [,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2

     [,1] [,2]
[1,]    5    7
[2,]    6    8
```

5.28 Matrice: définition

5.28.1 objet `matrix`

Une matrice est un tableau à deux dimensions. Elle est définie par la commande

`` `matrix(data,nrow=,ncol=,byrow)` ``

En conséquence

- Un objet `array` à deux dimensions est automatiquement converti en `matrix`
- Un vecteur auquel on ajoute un atribut dimension est automatiquement converti en `matrix`

5.28.2 Exemple

```
class(array(1:4,c(2,2)))
```

```
[1] "matrix" "array"
```

```
x <- c(1,2,3,4)
dim(x) <- c(2,2)
class(x)
```

```
[1] "matrix" "array"
```

5.29 Remarques importantes

- R range les éléments d'une matrice par défaut par **colonne**.

```
matrix(1:6,nrow=2)
```

```
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
matrix(1:6,nrow=2,byrow=TRUE)
```

```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

- Lors de la création d'une matrice, R **recycle** les éléments jusqu'à ce que les contraintes de dimension soient vérifiées.

5.30 Matrices: opérateurs élémentaires

Étant donné qu'une matrice est un vecteur pourvu d'une dimension, on a la proposition suivante:

```
**La   plupart   des   opérateurs   vectorielles   s'appliquent**
(arithmétiques/mathématiques, ensemblistes, d'indexation).
```

```
a <- matrix(sample(-4:4,9),3,3)
cat(max(a),sum(a),prod(a))
```

```
4 0 0
```

```
which(a > 0)
```

```
[1] 4 6 7 9
```

```r
cumsum(a[a > 0])
```

```
[1]  4  6  7 10
```

```r
order(a)
```

```
[1] 1 5 8 3 2 7 6 9 4
```

```r
round(exp(a),4)
```

```
       [,1]    [,2]    [,3]
[1,] 0.0183 54.5982  2.7183
[2,] 1.0000  0.0498  0.1353
[3,] 0.3679  7.3891 20.0855
```

5.31 Manipulation de matrices

5.31.1 Opérateurs matriciels usuels

- `+,/,*,\^{}` sont les opérateurs usuels terme-à-terme,
- `\%*\%` est le produit matriciel,
- `crossprod()` est le produit scalaire,
- `t()` transpose une matrice,
- `diag()` extrait / spécifie la diagonale.

5.31.2 Exemples

```r
a <- matrix(sample(-4:4,9),3,3)
b <- matrix(sample(a),3,3)
diag(a)
```

```
[1] 4 0 1
```

```r
diag(a) <- diag(b) <- 1
diag(a)
```

```
[1] 1 1 1
```

```
a + t(b) %*% b
```

```
     [,1] [,2] [,3]
[1,]   19  -18  -11
[2,]  -16   18    4
[3,]   -4    0    7
```

5.32 Concaténation de matrices

5.32.1 Trois fonctions selon l'effet voulu:

- `c()` concatène les éléments de plusieurs matrices en un vecteur,
- `cbind()` empile **horizontalement** plusieurs matrices,
- `rbind()` empile **verticalement** plusieurs matrices.

5.32.2 Exemples

```
a <- matrix(1,2,3)
b <- matrix(2,2,3)
c(a,b)
```

```
 [1] 1 1 1 1 1 1 2 2 2 2 2 2
```

```
cbind(a,b)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    1    1    2    2    2
[2,]    1    1    1    2    2    2
```

```
rbind(a,b)
```

```
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    1    1    1
[3,]    2    2    2
[4,]    2    2    2
```

5.33 Algèbre linéaire élémentaire

5.33.1 Résolution de systèmes linéaires, inversion matricielle

La commande `solve` résout

$$\mathbf{A}x = \mathbf{b},$$

```
A <- matrix(c(4,2,8,-3),2,2)
b <- c(2,3)
solve(A,b)
```

```
[1]  1.0714286 -0.2857143
```

ou inverse une matrice:

```
round(solve(A) %*% A,8)
```

```
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

5.34 Commandes avancées d'algèbre linéaire

Utile pour l'analyse numérique

R dispose des outils classiques d'algèbre linéaire

- `det`: calcule le **déterminant** d'une matrice;
- `chol`: factorisation de **Cholesky** ($A = C^tC$, avec A symétrique, C triangulaire supérieure);
- `qr`: factorisation **QR** ($A = QR$ avec Q orthogonale, R triangulaire supérieure);
- `eigen`: calcule valeurs propres et **vecteurs propres** d'une matrice;
- `svd`: calcule la décomposition en **valeurs singulières**.
- ...

5.35 Liste: définition

5.35.1 objet `list`

```
Une  liste est une  **collection d'objets  hétérogènes**. Elle
est  définie  par la  commande  `list(el1=,  el2=,  ...)`.  Les
éléments d'une liste peuvent posséder un nom.
```

```
list(c(1,2,3),c("robert","johnson"),matrix(rnorm(4),2,2))
```

```
[[1]]
[1] 1 2 3

[[2]]
[1] "robert"  "johnson"

[[3]]
           [,1]       [,2]
[1,]  0.515591  0.7454637
[2,] -0.164255 -1.4384202
```

```
list(numero = c(1,2,3), noms = c("robert","johnson"), mat = matrix(rnorm(4),2,2))
```

```
$numero
[1] 1 2 3

$noms
[1] "robert"  "johnson"

$mat
           [,1]      [,2]
[1,] 0.2909921 0.7667546
[2,] 0.4816714 1.9838393
```

5.36 Accéder aux éléments

5.36.1 Deux situations

- Les éléments de la liste **ne sont pas nommés**: on accède au ieme élément par indexation `nom_liste[[i]]` uniquement.

- Les éléments de la liste **sont nommés**: on peut y accéder comme ci-dessus ou en utilisant le nom de l'élément `nom_liste$nom_elt`.

```
maliste <- list(numero = c(1,2,3), noms = c("robert","johnson"), mat = matrix(rnorm(4),2,2
maliste$nom
```

```
[1] "robert"  "johnson"
```

```
maliste$nom[2]
```

```
[1] "johnson"
```

```
maliste[[2]]
```

```
[1] "robert"  "johnson"
```

```
maliste[[2]][2]
```

```
[1] "johnson"
```

[containsverbatim,allowframebreaks] ## Manipulation de listes

5.36.2 Sélectionner des éléments

`Fonctionne (presque) comme pour les vecteurs`

```
l1 <- list(1:2,c("a","c","g","t"))
l1[[-2]]
```

```
[1] 1 2
```

5.36.3 Commande `lapply`

`Applique une fonction à chaque élément d'une liste`

```
lapply(maliste,length)
```

```
$numero
[1] 3

$noms
[1] 2

$mat
[1] 4
```

5.36.4 Commande `c()`

Permet de concaténer deux listes.

```
c(list(1:2,c("a","c","g","t")),list(rnorm(3),"yop"))
```

```
[[1]]
[1] 1 2

[[2]]
[1] "a" "c" "g" "t"

[[3]]
[1]  0.5127267 -0.1453505  2.8462857

[[4]]
[1] "yop"
```

5.37 Tableau de données: définition

Un autre point fort de R

5.37.1 objet `data.frame`

C'est une liste à laquelle on impose certaines contraintes^[que je vous épargne], afin de rassembler vecteurs et facteurs sous la forme d'un tableau de données.

- Pratiquement, **un tableau de données est une matrice dont les colonnes sont de mode différent**,
- C'est l'objet idéal pour la **manipulation de données** (**forcez-vous** à l'utiliser).

5.38 Création de tableau de données

5.38.1 Syntaxe

On peut spécifier le nom des colonnes par le vecteur `row.names`
ou directement comme pour une liste:

`data.frame(e1=,e2=, ...,row.names=)`

5.38.2 Exemple

```
age <- c(25,35,32,27,32,40,26,25,26,28,30,NA,36,30,30)
grd <- c("thd","CR","MdC","thd","thd","MdC","MdC","thd","thd","MdC","CR","MdC","CR","thd",
sex <- factor(sample(c(rep("M",3),rep("F",12))))
donnees <- data.frame(age=age,grade=grd,sexe=sex)
head(donnees)
```

```
  age grade sexe
1  25   thd    F
2  35    CR    M
3  32   MdC    M
4  27   thd    F
5  32   thd    F
6  40   MdC    F
```

5.39 Manipulation des éléments du tableau de données

- Comme une liste !
- les commandes `attach()` `detach` placent ôtent les éléments du tableaux de données dans l'itinéraire de recherche.

```
donnees$age
```

```
 [1] 25 35 32 27 32 40 26 25 26 28 30 NA 36 30 30
```

```
attach(donnees,warn.conflicts=FALSE)
grade
```

```
 [1] "thd" "CR"  "MdC" "thd" "thd" "MdC" "MdC" "thd" "thd" "MdC" "CR"  "MdC"
[13] "CR"  "thd" "thd"
```

```
detach(donnees)
```

5.40 Travailler avec les tableaux de données

- beaucoup de fonctions prédéfinies
- penser aux fonctions `tapply` (ou `by`)

```
summary(donnees)
```

```
      age            grade            sexe
 Min.   :25.00   Length:15          F:12
 1st Qu.:26.25   Class :character   M: 3
 Median :30.00   Mode  :character
 Mean   :30.14
 3rd Qu.:32.00
 Max.   :40.00
 NA's   :1
```

```
attach(donnees,warn.conflicts=FALSE)
by(age,sexe,mean,na.rm=TRUE)
```

```
sexe: F
[1] 29.54545
------------------------------------------------------------
sexe: M
[1] 32.33333
```

```
by(age,grade,mean,na.rm=TRUE)
```

```
grade: CR
[1] 33.66667
------------------------------------------------------------
grade: MdC
[1] 31.5
------------------------------------------------------------
grade: thd
[1] 27.85714
```

```
detach(donnees)
```

5.41 Structures de contrôle

Les structure de contrôle sont les blocs d'un programme

5.42 Regrouper les expressions

5.42.1 Syntaxe

{expr_1; expr_2; . . . ; expr_n }

ou

{
expr_1 . . .
expr_n

}

5.42.2 Remarques sur les groupes

- La dernière valeur du groupe est retournée;
- un groupe d'expressions peut être passé à une fonction, réutilisé dans une expression plus grande, etc.

5.43 Exécution conditionnelle: `if,if/else,ifelse`

5.43.1 Syntaxe

```
if (condition) {
  expr_1
 else {
  expr_2
```

ou {

```
ifelse(condition, a, b)
```

5.43.2 Remarques

- `condition` est une valeur logique: penser à `\&,|,!`, ...
- le `else` est optionnel,
- `elseif` permet d'imbriquer les conditionnements.

5.44 Exécution répétée: boucle `for`

5.44.1 Syntaxe

for (var in set) { expr(var) }

ou

for (var in set) expr(var)

à fuir pour éviter les effets de bords sournois!

5.44.2 Remarques sur la boucle `for`

- `var` est la variable incrémentée,
- `set` est un vecteur définissant les valeurs successives,
- **lente** comparée aux opérateurs matriciels.

5.45 Exécution répétée: boucles `while` et`repeat`‘

5.45.1 Syntaxe

while (condition) { expr

}

ou

repeat { expr }

5.45.2 Remarque

- Comme pour `for`, éviter les imbrications sources de lenteur.

5.46 Contrôle des boucles: `break, next`

5.46.1 Exemples d’utilisation

repeat { expr if (condition) {break} }

ou

while (condition1){ expr_1 if (condition2) {next} expr_2

}

5.46.2 Remarque

- `break` est la seule manière d’interrompre une boucle `repeat`.

5.47 Les fonctions

Permettent de factoriser des lignes de codes

5.48 Définir une fonction

5.48.1 Syntaxe

nom_de_la_fonction <- function(arg1,arg2, . . .) { expression

return(var) }

5.48.2 Remarques

- `return` peut être omis (à éviter): dans ce cas la dernière valeur calculée est renvoyée.
- peut être tapée directement dans l'interpréteur ou dans un fichier externe `functions.R`, chargé par `source`.

5.49 Un exemple simple

5.49.1 Moyenne empirique d'un vecteur

Avec suppression des valeurs manquantes !

```
moyenne <- function(x){
  ## suppression des valeurs manquantes
  x.not.na <- x[!is.na(x)]
  ## moyenne empirique
  resultat <- sum(x.not.na) / length(x.not.na)
  return(resultat)
}
```

5.49.2 Tests

```
moyenne(rnorm(100))
```

```
[1] 0.1654921
```

```
moyenne(c(1,-5,3,NA,8.7))
```

```
[1] 1.925
```

5.50 Les arguments, leurs valeurs par défaut

Encore un point fort de R

5.50.1 Propriétés

```
-    les   arguments    peuvent   être   passés   dans   le
  **désordre** s'ils sont **nommés**: `var=object`,
-   on peut définir  une valeur  par défaut  pour n'importe
  quel argument lors de la définition de la fonction: `var=10`.
-  en cas de **sorties multiples**, les sorties doivent être
  renvoyées sous forme de liste.
```

5.50.2 Remarques

- Les valeurs par défaut rendent la lecture des fonctions beaucoup plus aisée pour l'utilisateur: **imposer peu d'arguments obligatoires}**.
- Les noms des éléments de la liste définie dans la fonction sont conservés à l'extérieur de la fonction.

5.51 Un exemple (un tout petit peu) plus avancé

5.51.1 Résumé numérique d'un vecteur

```
resume <- function(x,na.rm=TRUE,affiche=FALSE) {
  mu <- mean(x,na.rm=na.rm)
  s2 <- var(x,na.rm=na.rm)
  if (affiche) {
    cat("\nMoyenne:",mu,"et variance:",s2)
  }
  return(list(moyenne = mu, variance = s2))
}
```

```
out <- resume(rnorm(100))
out$variance
```

```
[1] 0.9047296
```

```
out <- resume(affiche=TRUE,x=rexp(100,0.5))
```

```
Moyenne: 2.161461 et variance: 5.23698
```

5.52 Les packages

Les packages sont des codes qui peuvent être inclus pour étendre les fonctionnalités de R

5.53 Motivations: reproductibilité de la recherche

5.53.1 Principe de Claerbout (Géophysicien, Stanford)

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

5.53.2 Démarche

- Proposer une méthode et exposer dans un article ses propriétés,
- Écrire et déposer un package sur CRAN,
- Publier un article dans « journal of statistical software » ou une note dans « Bioinformatics ».

5.54 Simplicité de la création d'un package

5.54.1 Définir un objectif

```
Par exemple, `SIMoNe` : construire un graphe des interactions
significatives entre gènes à partir de données du transcriptome.
```

5.54.2 Organisation type

- Fichier DESCRIPTION
- Répertoire R : fonctions R (fonction `inferGraph(data)`)
- Répertoire man : documentation des fonctions
- Répertoire data : données
- (Répertoire src : pour les fichiers à compiler, header etc.)

6 De l'intérêt des statistiques descriptives: coefficient de Gini

6.1 Le capital au 21ème siècle

- Livre d'économie publié en 2013 et écrit par Thomas Piketty, paru aux éditions du Seuil
- Bestseller aux états unis

Figure 1: Le capital au 21ème siècle

6.2 Quelle thèse ?

- les inégalités de revenu, les inégalités de patrimoine et le rapport capital/revenu dans les pays développés suivent chacun une courbe en U
- on retrouve au début du xxie siècle des niveaux d'inégalités comparables aux niveaux d'inégalités du xixe siècle et du début du xxe siècle.

6.3 Quelles données, quel résumé ?

- Si l'on se concentre sur le revenu des personnes appartenant à des ménages fiscaux
- A partir des impôts on peut accéder revenu de chaque ménage du pays avant et après impôt.
- L'INSEE utilise une mesure du revenu par unité de consommation, à l'aide d'une échelle d'équivalence. L'échelle la plus utilisée actuellement consiste à décompter
  - 1 unité de consommation (UC) pour le premier adulte du ménage,
  - 0,5 UC pour les autres personnes de 14 ans ou plus,
  - 0,3 UC pour les enfants de moins de 14 ans.

6.4 Coefficient de Gini

- Le coefficient ou indice de Gini porte le nom du statisticien et démographe italien Corrado Gini (1884–1965).

- indicateur de dispersion permettant principalement d'apprécier les inégalités dans la distribution des richesses d'un territoire
- varie entre zéro et un,
- zéro étant la situation d'égalité parfaite (chaque citoyen est exactement aussi riche que son voisin),
- un étant la situation d'inégalité parfaite (un citoyen possède toutes les richesses, les autres aucune).

Les pays du monde s'ordonnent ainsi entre 0,25 (pays d'Europe scandinave et centrale) et 0,70 (pays émergents d'Amérique latine, d'Afrique centrale)

6.5 Coefficient de Gini

6.5.1 Courbe de Lorenz

$$L(F(x) = p) = \frac{\int_{-\infty}^{x} t\, f(t)\, dt}{\int_{-\infty}^{\infty} t\, f(t)\, dt} = \frac{\int_{-\infty}^{x} t\, f(t)\, dt}{\mu}$$

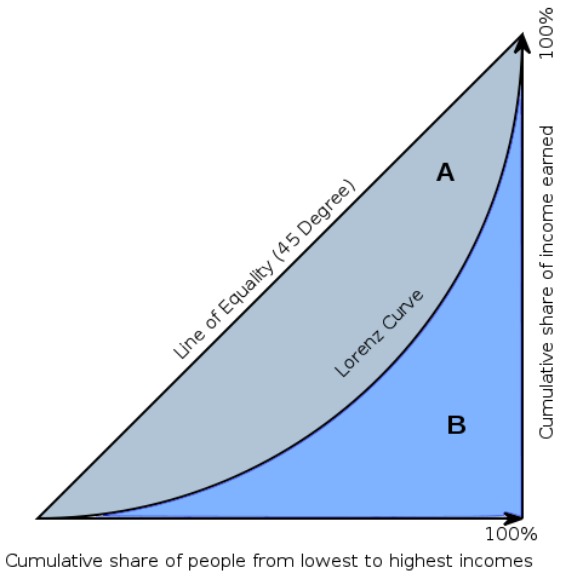


Figure 2: Coefficient de Gini

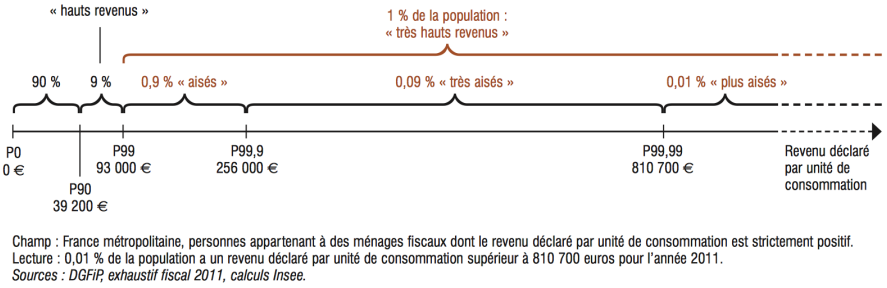
$$G = 2A = 1 - 2B$$

6.6 Coefficient de Gini

Si x_i est la richesse ou le revenu de la personne i, et qu'il y a n personnes, alors le coefficient de Gini G est donné par :

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2 \sum_{i=1}^{n} \sum_{j=1}^{n} x_j} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}$$

6.7 Revenu en France en 2011



Champ : France métropolitaine, personnes appartenant à des ménages fiscaux dont le revenu déclaré par unité de consommation est strictement positif.
Lecture : 0,01 % de la population a un revenu déclaré par unité de consommation supérieur à 810 700 euros pour l'année 2011.
Sources : DGFiP, exhaustif fiscal 2011, calculs Insee.

Figure 3: Revenu en France en 2011

6.8 Evolution du coefficient de Gini en France

6.9 Coefficient de Gini dans le monde

7 Concepts fondamentaux

7.1 Qu'est ce que la statistique ?

Activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation issues d'une population:

- une branche des mathématiques appliquées
- une méthode

- un ensemble de techniques

	1996	1999	2002	2005	2008	2009	2010	2010[1]	2011[1]
Seuils de niveau de vie (en milliers d'euros 2011)									
Niveau de vie médian (D5)	16,7	17,3	18,5	18,7	19,7	19,8	19,7	19,6	19,6
Premier décile de niveau de vie (D1)	8,9	9,5	10,3	10,3	10,9	10,8	10,6	10,6	10,5
Neuvième décile de niveau de vie (D9)	31,2	32,8	35,1	34,6	36,9	37,2	37,0	36,7	37,5
Rapports interdéciles									
D9/D1	3,4	3,5	3,4	3,4	3,4	3,4	3,5	3,5	3,6
D9/D5	1,9	1,9	1,9	1,9	1,9	1,9	1,9	1,9	1,9
D5/D1	1,9	1,8	1,8	1,8	1,8	1,8	1,8	1,8	1,9
Masses de niveau de vie détenues									
S20 (en %)	9,0	9,1	9,3	9,0	9,0	8,9	8,7	8,7	8,6
S50 (en %)	31,0	30,9	31,1	31,0	30,9	30,7	30,2	30,1	29,8
S80 (en %)	63,0	62,3	62,3	62,0	61,6	61,8	61,0	60,7	60,5
(100-S80)/S20	4,1	4,1	4,1	4,2	4,3	4,3	4,5	4,5	4,6
Indice de Gini	**0,279**	**0,284**	**0,281**	**0,286**	**0,289**	**0,290**	**0,299**	**0,303**	**0,306**

1. À partir de 2010, les estimations de revenus financiers mobilisent l'enquête Patrimoine 2010 (*encadré 1*).
Champ : France métropolitaine, personnes vivant dans un ménage dont le revenu déclaré au fisc est positif ou nul et dont la personne de référence n'est pas étudiante.
Lecture : les 20 % les plus modestes disposent en 2011 de 8,6 % de la somme des revenus disponibles par UC (S20), les 20 % les plus aisés perçoivent 39,5 % de la somme des revenus disponibles par UC (complément à 100 de S80), soit 4,6 fois plus.
Sources : Insee, enquêtes Revenus fiscaux et sociaux rétropolées de 1996 à 2004, enquêtes Revenus fiscaux et sociaux 2005 à 2011 ; DGI ; DGFiP ; Cnaf ; Cnav ; CCMSA.

Figure 4: Evolution du coefficient de Gini en France

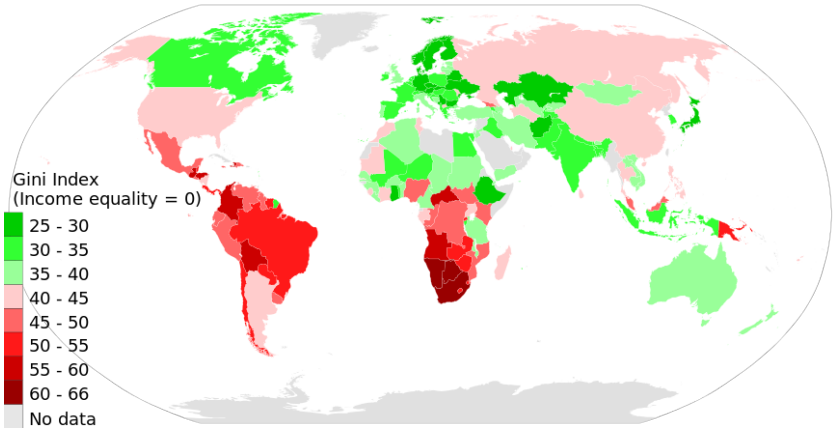


Figure 5: Coefficient de Gini dans le monde

7.2 Population

- population, ensemble d'entités objet de l'investigation statistique
- individus, définis comme les éléments d'une certaine population

7.2.1 Différentes notion de population

- dans certain cas la population de référence est **finie** et ses éléments peuvent être explicitement dénombrés
- la notion de population revêt parfois une signification plus abstraite (exemple population de malades)
- parfois la notion de population s'identifie avec celle de procédure de génération de données (données d'expression)

7.3 Variables, distributions

7.3.1 Variables

Chaque individu est décrit par un ensemble de variables:

- **qualitative** (sexe, nationalité, état matrimonial, ...): les valeurs prises par le caractère sont les modalités
  - ordinale (notion d'ordre): modalités intrinsèquement ordonnées
  - nominale : pas de structure d'ordre : par exemple le sexe.
- **quantitative** (taille, poids, . . .)
  - discrète
  - continue

7.4 Modes d'étude d'une population

- Étude exhaustive Dite par recensement. L'étude d'une population de grande taille est souvent difficile voire impossible
- Échantillon. Le processus de sélection d'un échantillon est l'échantillonnage. Seule solution d'une le cas d'une population **infinie**

7.5 Inférence statistique

Processus visant à

- déduire de conclusions générales relative à la population totale
- à partir de la connaissance partielle relative à un nombre de cas particulier

7.6 Objectifs d'une étude statistique

1. **Statistique Descriptive ou Exploratoire** : synthétiser, résumer, structurer l'information
2. **Statistique inférentielle** : formuler ou valider des hypothèses relatives à la population totale

7.7 Le tableau de données

- n mesurés par p
- Tableau

$$X = (x_i^j) = \begin{pmatrix} x_1^1 & x_1^j & x_1^p \\ & . & . \\ x_i^1 & x_i^j & x_i^p \\ & . & . \\ x_n^1 & x_n^j & x_n^p \end{pmatrix}$$

- Chaque variable est représentée par le vecteur $\boldsymbol{x}^j = (x_1^j, \ldots, x_n^j)'$
- Chaque individu est représenté par le vecteur $\boldsymbol{s}x_i = (x_i^1, \ldots, , x_i^p)'$
- X : réalisation d'un échantillon de taille n du de dimension p:

$$\boldsymbol{X} = (X^1, \ldots, X^p)'$$

8 Statistiques descriptives

8.1 Analyse statistique descriptive classique : des analyses univariées aux approches multivariées

1. Phase exploratoire et univariée
2. Approches bivariées
3. Approches multivariées

4. Interprétation des résultats

8.2 En pratique

- processus itératif :
  - analyses univariées
  - exploration de relations plus complexes entre les variables
- les méthodes dépendent des objectifs de l'analyse et des caractéristiques du jeu de données (nature des variables, relation entre variables, ...)
- interprétation prudente: variables manquantes,limitations des méthodes employées (linéaires, ...), ...

8.3 Phase exploratoire et univariée

8.3.1 Distribution des variables

- Moyenne, médiane, mode, écart-type, etc. pour résumer les caractéristiques principales
- Histogrammes, diagrammes en boîtes, etc. pour visualiser la forme de la distribution

8.3.2 Analyse des valeurs aberrantes

- Identification des points qui s'écartent significativement de la distribution.
- Étude de leur influence sur les résultats et décision de les supprimer ou non.

8.4 Approches bivariées

Détecter la **Dépendance entre paires de variables**

- représentation graphique
- statistiques
- tests

8.5 Approches multivariées

8.5.1 Analyse factorielle

- Réduction de la dimensionnalité d'un ensemble de variables corrélées.
- Identification des facteurs explicatifs principaux.

8.6 Clustering

- Affectation des individus à des groupes pré-définis en fonction de leurs caractéristiques.
- Utilisation d'algorithmes comme k-means

8.7 Interprétation des résultats

- Interpréter les résultats de manière claire et concise.
- Tenir compte des limitations des méthodes utilisées.
- Présenter les résultats de manière visuelle et pédagogique.
- L'interprétation des résultats doit être faite avec prudence et en tenant compte des limitations des méthodes employées

8.8 Observations

Nous considérerons dans un premier temps une seule colonne du tableau de données, soient n observations:

$$x_1, ..., x_n$$

8.9 Variable quantitative discrète ou qualitative ordinale

La variable prend ses valeurs dans

$$V_x = \{\epsilon_1, ..., \epsilon_K\}$$

avec

$$\epsilon_1 < ... < \epsilon_K$$

8.9.1 Tableau de fréquence

- ϵ_k, la modalité
- n_k, l'effectif des observations ayant la valeur ϵ_k
- $f_k = \frac{n_k}{n}$, la fréquence
- $F_k = \sum_{j=1}^{k} f_j$, la fréquence relative cumulée

8.10 Variable qualitative nominale

- Même représentation sans l'ordre.
- Pas de fréquence cumulée.

8.11 Camembert, diagramme en barres, radar

Couplés à la commande `table` Le diagramme en barres et le graphe en camembert permettent de visualiser le découpage d'une population en donnée catégorielle.

```
  pop pepin.08 poids.08 volcm3.08
1  CE      1.0     0.89      7.70
2  CE      1.0     1.14      8.82
3  CE      1.2     1.26     10.20
4  CE      1.2     0.66        NA
5  CE      1.2     0.83        NA
6  CE      1.3     0.54      4.61
```

```
attach(vigne)
par(mfrow=c(1,2))
pie(table(pop))
barplot(table(pop),las=3)
```

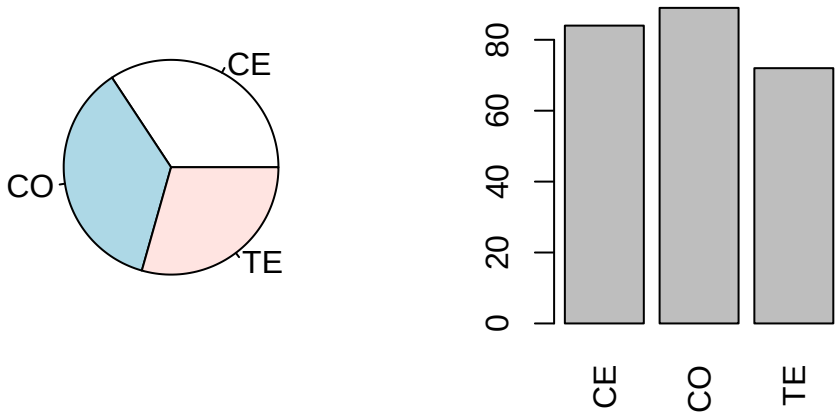
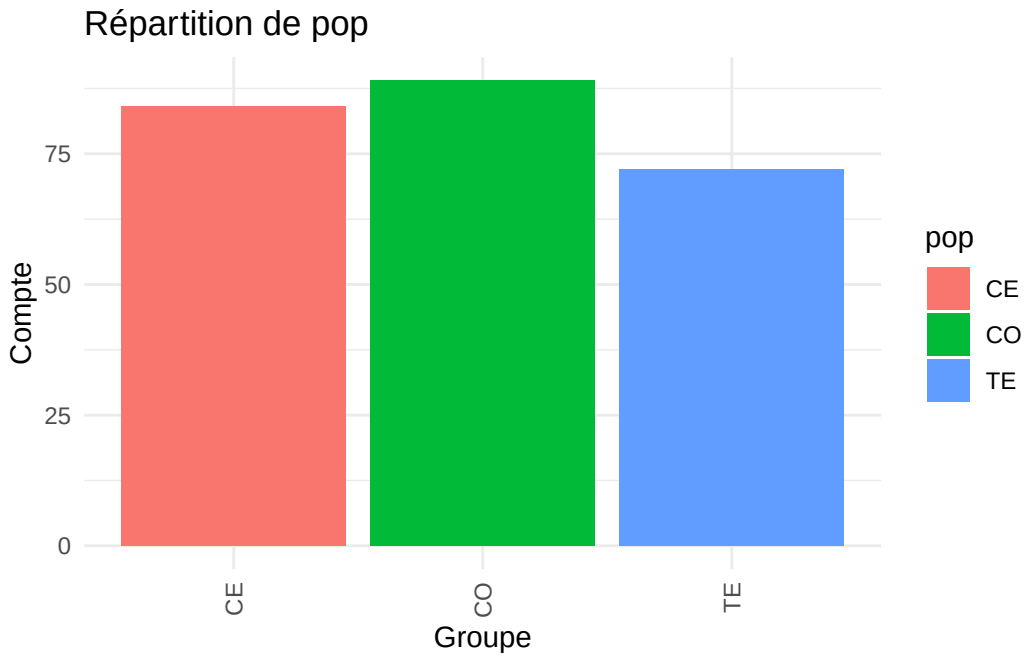


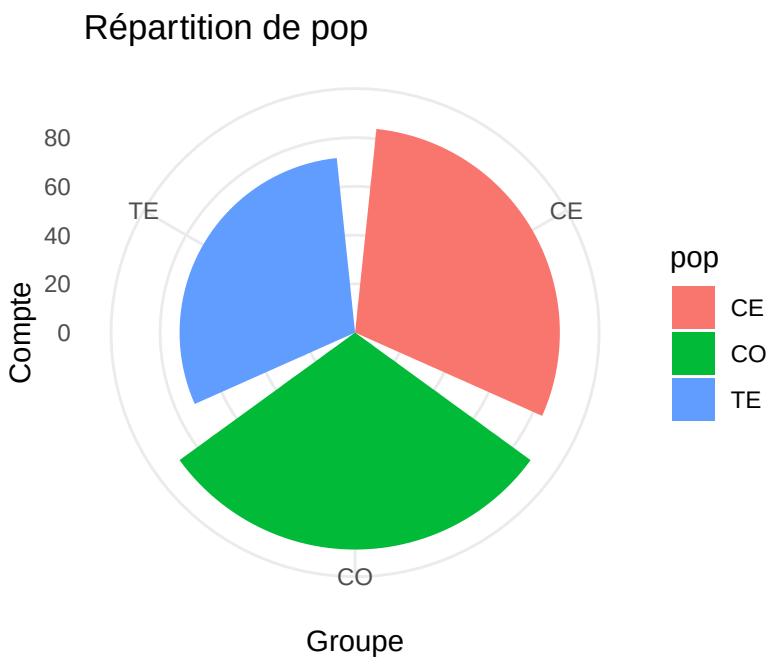
Figure 6: Camembert des différentes populations

8.12 En ggplot2

```
ggplot(vigne, aes(x = pop,fill=pop)) +
  geom_bar() + # geom_bar() compte les occurrences par défaut
  theme_minimal() +
  labs(title = "Répartition de pop", x = "Groupe", y = "Compte") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) # Pour faire pivoter
```




```
ggplot(vigne, aes(x = pop,fill=pop)) +
  geom_bar() + # geom_bar() compte les occurrences par défaut
  coord_polar()+
  theme_minimal() +
  labs(title = "Répartition de pop", x = "Groupe", y = "Compte")
```



8.13 Variable continue, résumés numériques

8.13.1 Tendance centrale

- Moyenne: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Remarque: la somme des écarts à la moyenne empirique est nulle

$$\sum_i (x_i - \bar{x}) = 0$$

- Inconvénient : problème des valeurs aberrantes
- Moyenne tronquée: $M_k = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}$ où $x_{(i)}$ est l'observation de rang i
- Médiane:

$$M = \begin{cases} x_{(n/2)} \text{ si } n \text{ est pair,} \\ x_{(\lfloor n/2 \rfloor + 1)} \text{ sinon} \end{cases}$$

- Fractile empirique d'ordre α

$$\hat{f}_\alpha = \begin{cases} x_{(n\alpha)} \text{ si } n\alpha \text{ est entier,} \\ x_{(\lfloor n\alpha \rfloor + 1)} \text{ sinon} \end{cases}$$

8.14 Indicateurs de dispersion

- la variance empirique: $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$
- la variance empirique corrigée: $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$
- l'étendue: $max_i x_i - min_i x_i$
- l'étendue interquartile $EI = Q_3 - Q_1 = \hat{f}_{3/4} - \hat{f}_{1/4}$

8.15 Reprenons l'exemple de la vigne

- Renommons les variables et considérons uniquement les données de 2008, pour une manipulation plus agréable.
- J'enlève également la colonne variété

8.16 Résumé statistique 'Commande summary

Le résumé numérique s'adapte selon la nature des variables (univariée, multivariée, factorielle)

```
 summary(pepin.08)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.400   1.800   1.858   2.400   3.200      27
```

```
summary(pop)
```

```
CE CO TE
84 89 72
```

```
knitr::kable(summary(vigne))
```

pop	pepin.08	poids.08	volcm3.08
CE:84	Min. :0.000	Min. :0.390	Min. : 3.44
CO:89	1st Qu.:1.400	1st Qu.:0.820	1st Qu.: 7.14
TE:72	Median :1.800	Median :1.060	Median : 8.91
NA	Mean :1.858	Mean :1.212	Mean :10.47
NA	3rd Qu.:2.400	3rd Qu.:1.360	3rd Qu.:11.90
NA	Max. :3.200	Max. :3.750	Max. :33.80
NA	NA's :27	NA's :28	NA's :44

8.17 Résumé statistique Package Hmisc

```
library(Hmisc)
```

```
Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

```
describe(vigne)
```

```
vigne

 4  Variables      245  Observations
--------------------------------------------------------------------------------
pop
       n  missing distinct
     245        0        3

Value         CE    CO    TE
Frequency     84    89    72
Proportion 0.343 0.363 0.294
--------------------------------------------------------------------------------
pepin.08
       n  missing distinct     Info     Mean      Gmd      .05      .10
     218       27       26    0.997    1.858    0.753      1.0      1.2
     .25      .50      .75      .90      .95
     1.4      1.8      2.4      2.7      2.9

lowest : 0   0.5 0.8 1   1.1, highest: 2.8 2.9 3   3.1 3.2
--------------------------------------------------------------------------------
poids.08
       n  missing distinct     Info     Mean      Gmd      .05      .10
     217       28      122        1    1.212   0.6078    0.578    0.676
```

```
     .25      .50      .75      .90      .95
   0.820    1.060    1.360    1.868    2.472

lowest : 0.39 0.4  0.41 0.42 0.49, highest: 3.31 3.39 3.43 3.47 3.75
---------------------------------------------------------------------------
volcm3.08
       n  missing distinct     Info     Mean      Gmd      .05      .10
     201       44      188        1    10.47    5.306     4.71     5.98
     .25      .50      .75      .90      .95
    7.14     8.91    11.90    16.58    23.06

lowest : 3.44  3.48  3.97  4.13  4.32 , highest: 26.97 27.59 28.29 28.54 33.8
---------------------------------------------------------------------------
```

8.18 Tableau de fréquences

Dans le cas où la variable x est continue, la réalisation d'un tableau de fréquence nécessite un partitionnement préalable du domaine de définition en K classes de largeur

- constante
- ou variable

8.19 Fonction de répartition empirique

$$\hat{F} : \mathbb{R} \mapsto [0,1],\ x \mapsto \frac{1}{n} card\{i : x_i \leq x\}$$

- Le graphe de la fonction de répartition est une fonction en escalier appelé diagramme cumulatif

8.20 Graphe en tiges et feuilles

Un graphe en tige est feuille permet de visualiser le tableau des fréquences

```
stem(pepin.08)
```

```
The decimal point is 1 digit(s) to the left of the |

 0 | 00000000
 2 |
 4 | 0
```

```
 6 |
 8 | 0
10 | 00000000000
12 | 0000000000000000000
14 | 00000000000000000000000000000000000000000
16 | 000000000000000000000
18 | 0000000000000000000000
20 | 000000000000000000000
22 | 0000000000000000000
24 | 00000000000000000000
26 | 0000000000000000000
28 | 00000000
30 | 000
32 | 0000000
```

8.21 Fonction de répartition empirique

Crée un objet qui peut être tracé avec .

```
par(mfrow=c(1,2))
plot(ecdf(poids.08[pop!="TE"]))
plot(ecdf(poids.08[pop=="TE"]))
```

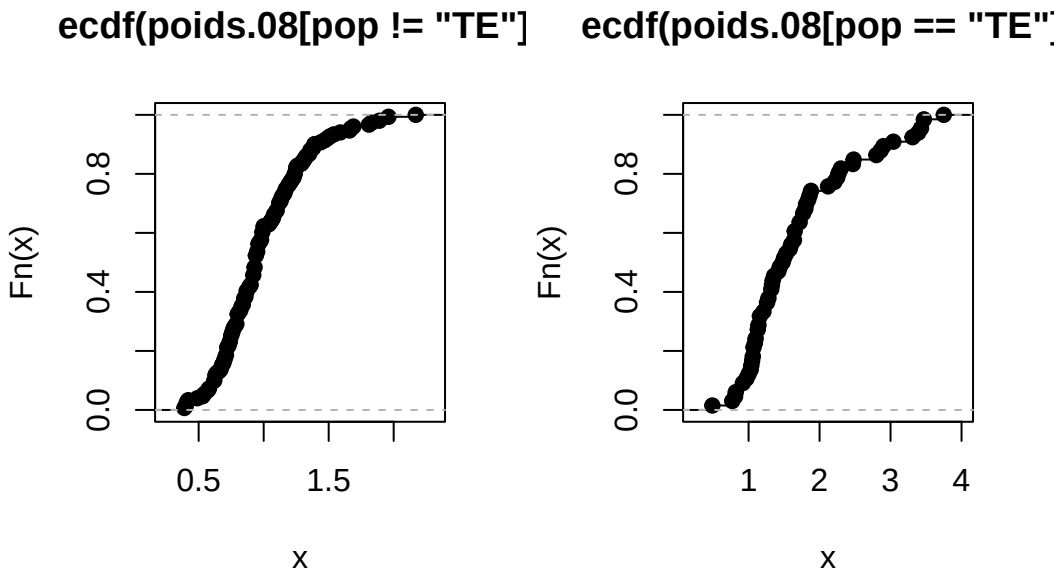


Figure 7: F.d.r empirique du poids pour les non TE et les TE

8.22 Comparaison de distribution

Pour comparer visuellemnt deux distributions, la manière la plus efficace est le graphe quantile/quantile (qui doivent correspondre si les distributions sont proches.)

```r
qqplot(poids.08[pop=="CE"],poids.08[pop=="CO"])
```

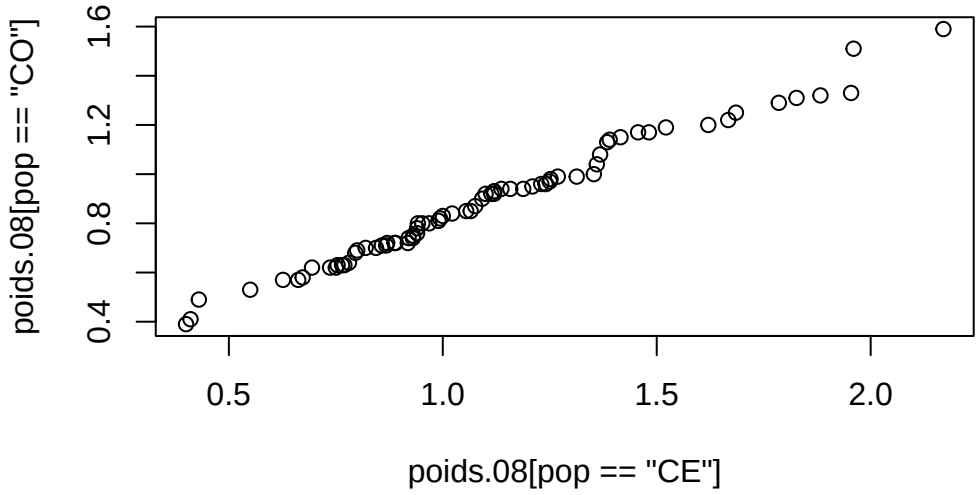


Figure 8: qqplot des distributions de poids de CE et CO

8.23 Comparaison de distribution en ggplot2

```r
library(ggplot2)
# Séparation des données
poids_CE <- vigne$poids.08[vigne$pop == "CE"]
poids_CO <- vigne$poids.08[vigne$pop == "CO"]

# Générer le QQ plot avec qqplot mais sans l'afficher, pour récupérer les points
qq <- qqplot(poids_CE, poids_CO, plot.it = FALSE)

# Créer un dataframe avec les points
qq_data <- data.frame(x = qq$x, y = qq$y)

# Utiliser ggplot2 pour afficher le QQ plot
ggplot(qq_data, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
```

```
labs(title = "QQ plot de poids.08 pour CE vs CO",
     x = "Quantiles de CE",
     y = "Quantiles de CO") +
theme_minimal()
```

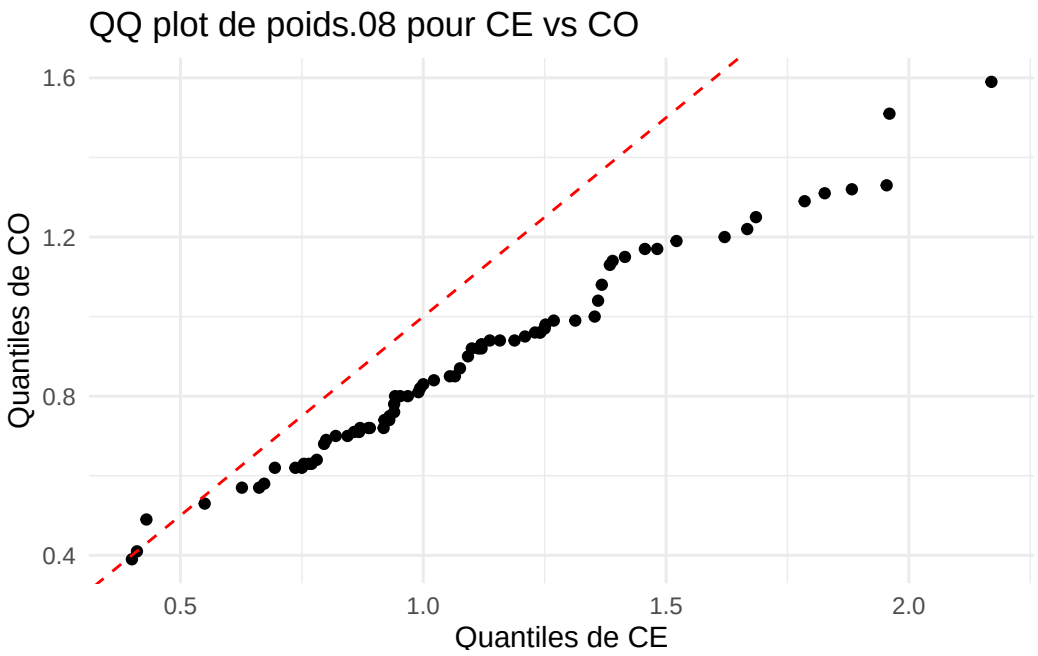


Figure 9: qqplot des distributions de poids de CE et CO

8.24 Histogramme et estimateur à noyaux

- Estimateur de la fonction de densité

$$\hat{f}_n(x) = \sum_i h_i \mathbb{I}_{[a_i, a_{i+1}[}(x) \quad a_1 < ... < a_{k+1}$$

- Découpage en intervalles
- Calcul de la fréquence
- Aire du rectangle proportionnel à la fréquence

$$\sum_i h_i(a_{i+1} - a_i) = 1 \text{ et } h_i(a_{i+1} - a_i) = \hat{P}_F(X \in [a_i, a_{i+1}[)$$

- Attention : hauteur proportionnelle à la fréquence si et seulement si les intervalles ont tous la même largeur

- Nombre d'intervalles :
  - Important
  - Réglage difficile
  - Règle empirique : règle de Sturges $1 + 10/3 * \log_{10}(n)$

```
hist(pepin.08,nclass=25,prob=TRUE)
lines(density(pepin.08[!is.na(pepin.08)]),col="red")
```

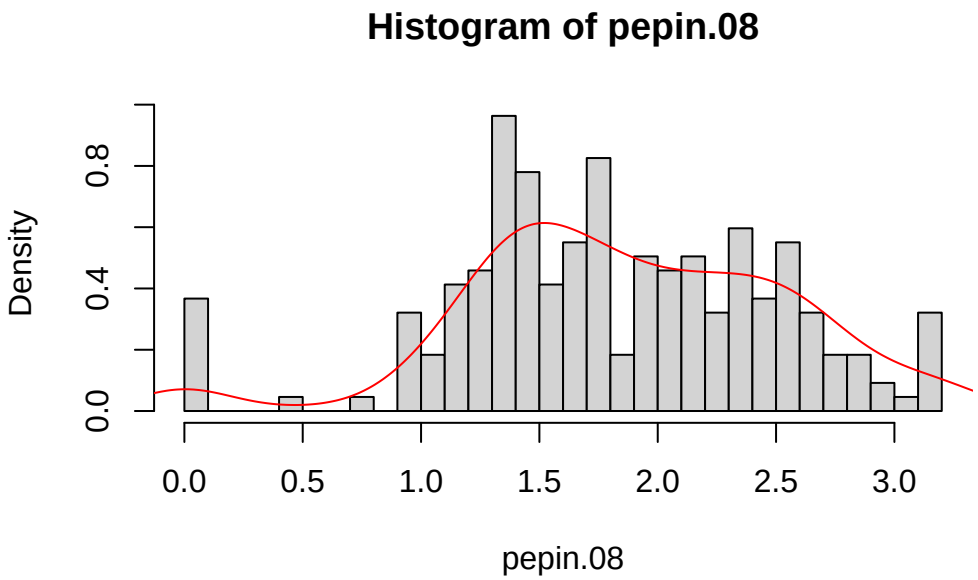


Figure 10: histogramme du nombre de pépins par baie en 2008

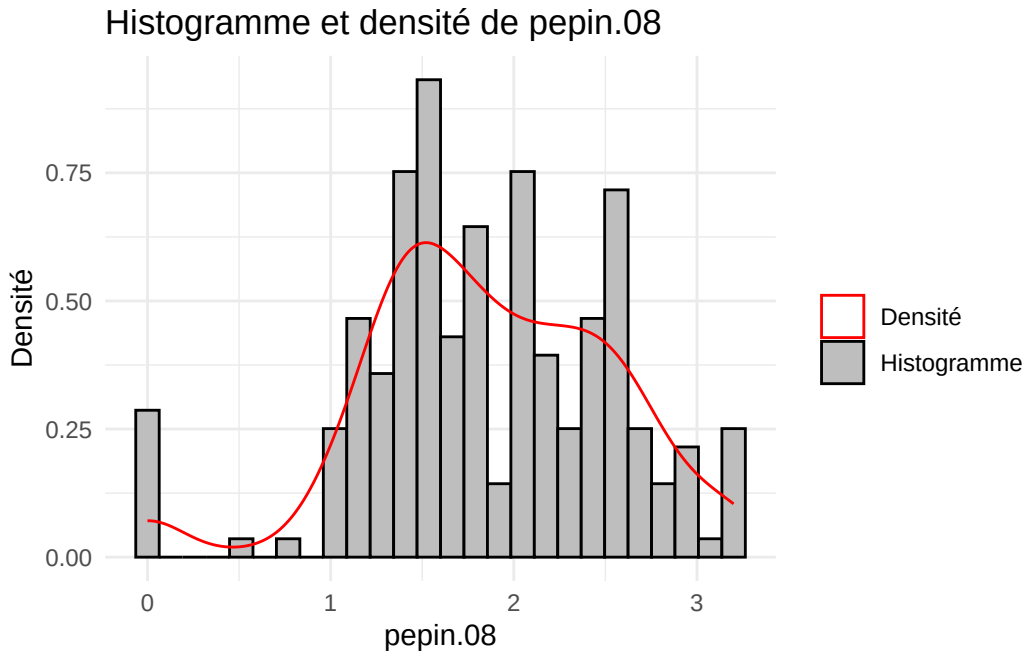
8.25 En ggplot2

```
ggplot(data = vigne, aes(x = pepin.08)) +
  geom_histogram(aes(y = ..density.., color = "Histogramme"), binwidth = diff(range(vigne$
  geom_density(aes(color = "Densité")) +
  scale_color_manual(name = "", values = c("Histogramme" = "black","Densité" = "red"), lab
  labs(title = "Histogramme et densité de pepin.08",
       x = "pepin.08",
       y = "Densité") +
  theme_minimal()
```

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```

```
Warning: Removed 27 rows containing non-finite outside the scale range
(`stat_bin()`).
```

```
Warning: Removed 27 rows containing non-finite outside the scale range
(`stat_density()`).
```



8.26 Boite à moustache

- Éléments atypiques (aberrants, *outliers*)
  - Notion arbitraire
  - Règle empirique assez souvent utilisée : valeurs situées à l'extérieur de $[q_1 - 1.5 \times Iqr, q_3 + 1.5 \times Iqr]$
- Définition : Graphique constitué
  - d'un rectangle délimité par les quartiles et partagé en deux par la médiane
  - d'une paire de moustaches : minimum et maximum de l'échantillon auquel on a ôté les éléments atypiques
  - des outliers eux-mêmes

8.27 Boîtes à moustaches

La boîte à moustache permet de visualiser les grands traits caractéristiques d'une distribution.

```
boxplot(poids.08~pop,col=c("red","green3","steelblue"),notch=T)
```

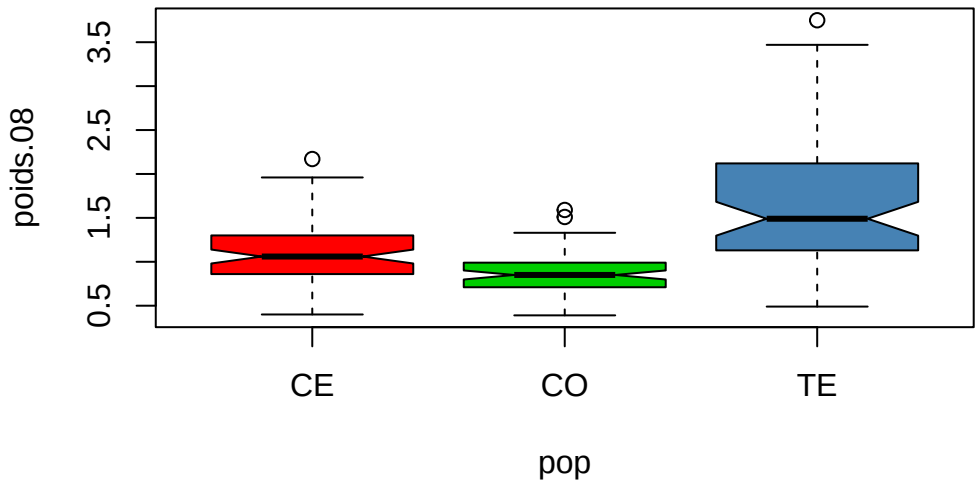
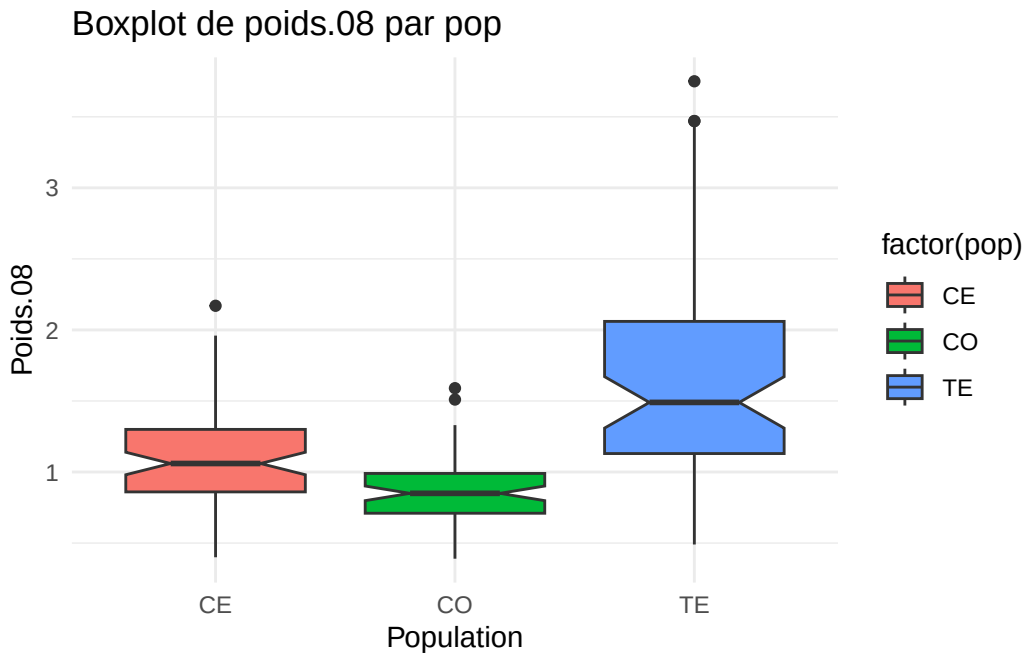


Figure 11: boîtes à moustaches du poids selon les populations

8.28 En ggplot

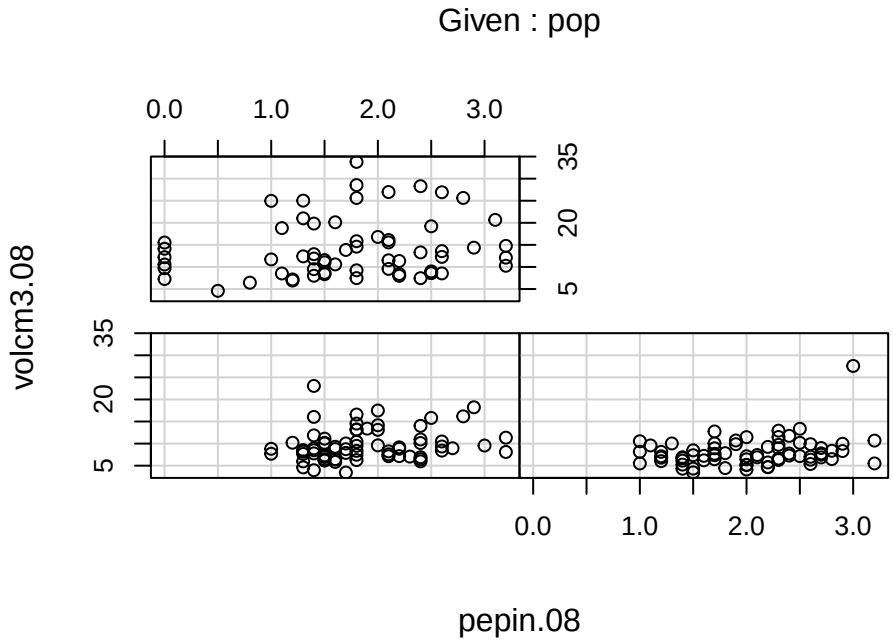
```
ggplot(vigne, aes(x = pop, y = poids.08, fill = factor(pop))) +
  geom_boxplot(notch = TRUE) +
  labs(title = "Boxplot de poids.08 par pop",
       x = "Population",
       y = "Poids.08") +
  theme_minimal()
```

```
Warning: Removed 28 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



8.29 Graphe conditionné par une variable

```
coplot(volcm3.08~pepin.08 | pop,show.given=FALSE)
```

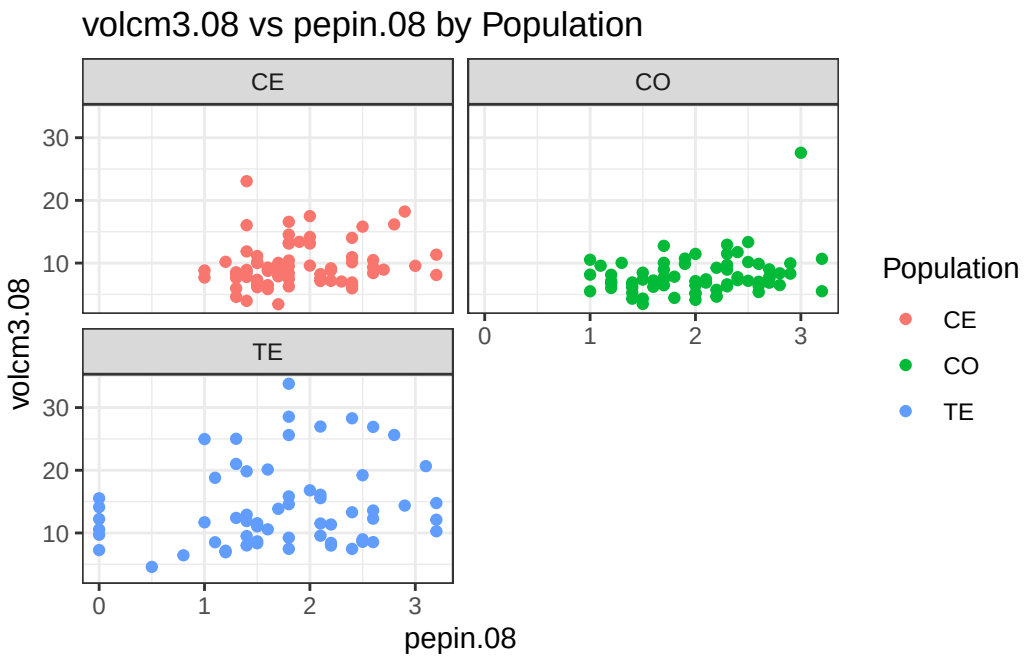


```
Missing rows: 4, 5, 11, 19, 20, 21, 49, 71, 78, 79, 80, 81, 82, 83, 84, 99, 112, 129, 139,
```

8.30 Graphe conditionné par une variable (ggplot)

```
ggplot(data = vigne,
       aes(x = pepin.08, y = volcm3.08, color = pop)) +
  geom_point() +
  facet_wrap(~ pop, nrow = 2) +
  labs(title = "volcm3.08 vs pepin.08 by Population",
       x = "pepin.08",
       y = "volcm3.08",
       color = "Population") +
  theme_bw()
```

```
Warning: Removed 44 rows containing missing values or values outside the scale range
(`geom_point()`).
```



8.31 Conclusion

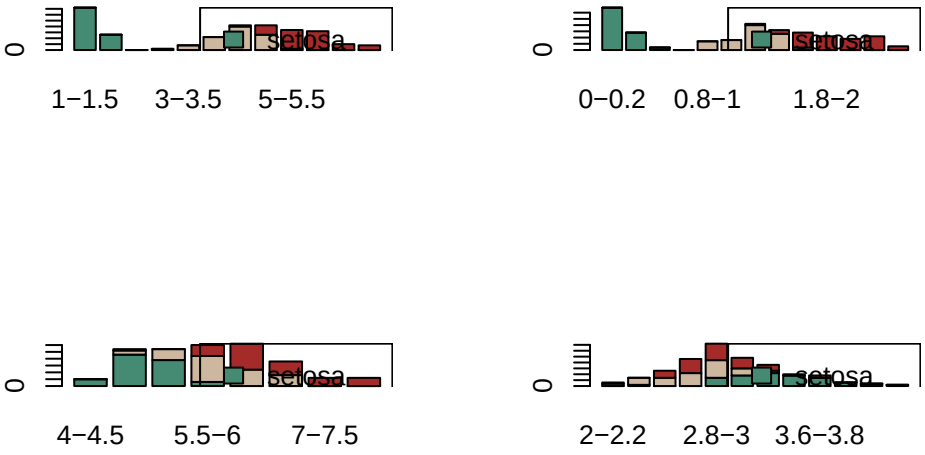
Mise en évidence de certaines caractéristiques :

- Présence de données atypiques
- Absence de symétrie de la distribution
- Présence de populations hétérogènes
- ...

9 Description bidimensionnelle

9.1 Histogrammes et variable qualitative

```
histy<-function(x,y){
    mycolors<-colors()[seq(2,600,by=10)]
    hist(x,plot=FALSE)$breaks->breaks
    length(levels(y))->nbLevels
    z<-matrix(0,nbLevels,length(breaks)-1)
    for (l in 1:nbLevels){
        xl<-x[as.numeric(y)==l]
        inter<-matrix(c(breaks[-length(breaks)],breaks[-1]),ncol=2)
        apply(inter,1,function(interi) sum((interi[1]<xl) & (xl<=interi[2]))) -> z[l,]
        }
    as.table(z)->z
    dimnames(z)[2]<-list(apply(inter,1,function(x) toString(paste(x[1],x[2],sep="-"))))
    dimnames(z)[1]<-list(levels(y))
    barplot(z,col=mycolors[2:(nbLevels+2)])
    legend("topright",levels(y),fill= mycolors[2:(nbLevels+2)])
    invisible(z)
    }
par(mfrow=c(2,2))
histy(iris$Petal.Length,iris$Species)
histy(iris$Petal.Width,iris$Species)
histy(iris$Sepal.Length,iris$Species)
histy(iris$Sepal.Width,iris$Species)
```



9.2 Statistiques associées à un vecteur aléatoire

- Rappels
  - X réalisation d'un échantillon de taille n du vecteur aléatoire $\boldsymbol{X}$
  - $\boldsymbol{x}_i$ réalisation de taille 1 de $\boldsymbol{X}$
  - $\boldsymbol{x}^j$ réalisation d'un échantillon de taille n de $\boldsymbol{X}^j$

9.3 Moyenne et variance

- Moyenne empirique

$$\overline{\boldsymbol{x}} = (\overline{x}^1, \ldots, \overline{x}^p)' \quad \text{où} \quad \overline{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

- Variance empirique

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x}^j)^2$$

9.4 Covariance vs Corrélation empirique

- Covariance empirique

$$s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x^j}).(x_i^j - \overline{x^{j'}})$$

- Corrélation empirique

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}$$

9.5 Matrice de variance empirique

$$S = (s_{jj'}) = \frac{1}{n}(X - 1_n\overline{\boldsymbol{x}})'(X - 1_n\overline{\boldsymbol{x}}) = \frac{1}{n}Y'Y$$

où 1_n est la matrice de dimension $(n, 1)$ remplie de 1 et Y est la matrice centrée associée à X.

- empirique

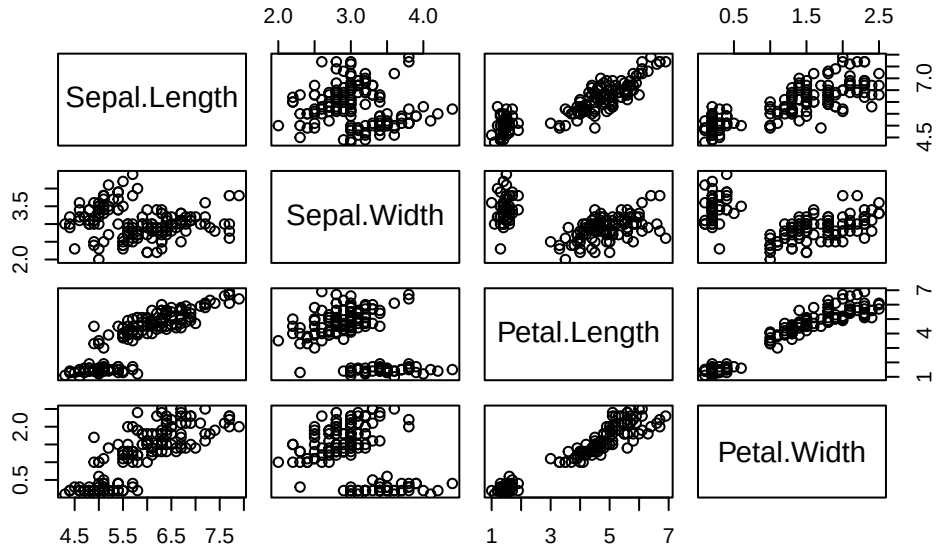
$$R = (r_{jj'}) = D_{1/s_j} S D_{1/s_j}$$

9.6 Graphique de dispersion

- Représentation de chaque individu i par le point du plan (x_i^1, x_i^2)
- Nuage de n points dans le plan
- Visualisation synthétique des données : permet de voir
  - les relations linéaires
  - les regroupements en classes homogènes

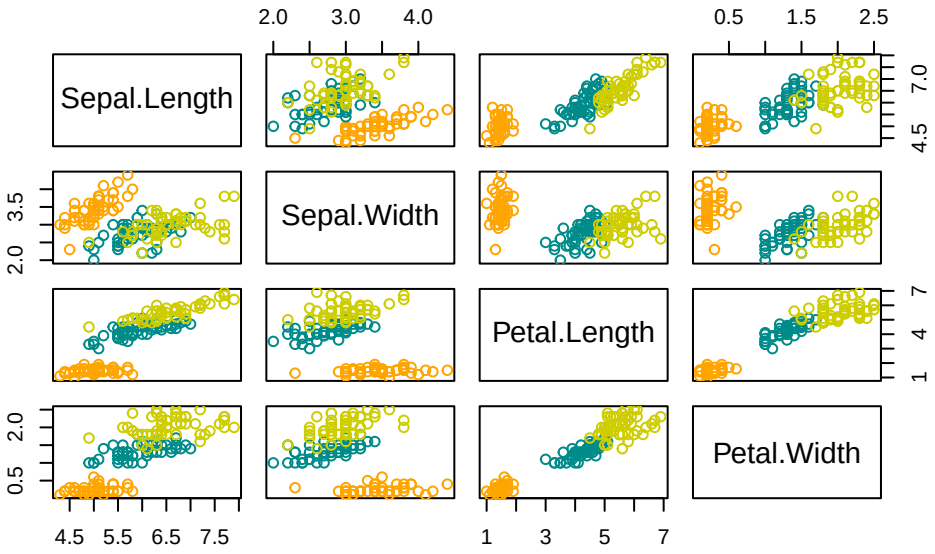
9.7 Les 5 variables des iris (pairs)

```
data(iris)
pairs(iris[,1:4])
```



9.8 Les 5 variables des iris en couleurs (pairs)

```
pairs(iris[,1:4],col=c('orange','cyan4','yellow3')[iris$Species])
```



9.9 Les 5 variables des iris en couleurs (ggpairs)

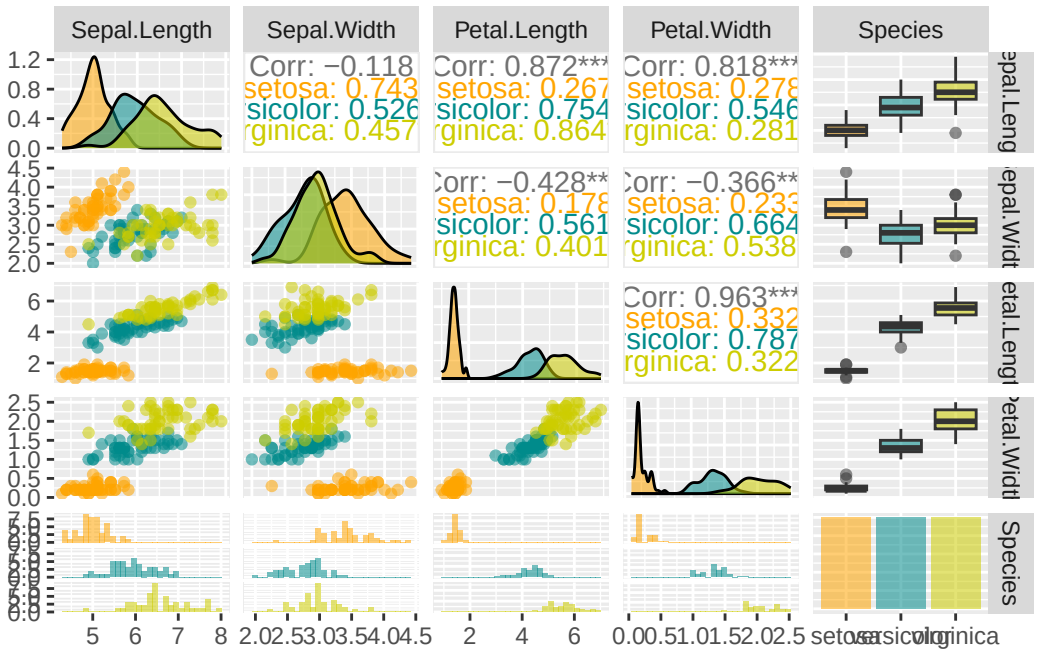
```
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
colors <- c("orange","cyan4",  "yellow3")
ggpairs(iris,aes(color = Species, alpha = 0.5))+
scale_color_manual(values = colors, aesthetics = c("colour", "fill"))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



9.10 Covariance et corrélation

- 2 variables : covariance et corrélation empirique
- > 2 variables : matrices de cov. et de corr. empiriques

9.10.1 Les iris

Table 2: Matrice de covariance

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal.Width	-0.0424340	0.1899794	-0.3296564	-0.1216394
Petal.Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal.Width	0.5162707	-0.1216394	1.2956094	0.5810063

Table 3: Matrice de corrélation

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal.Width	-0.0424340	0.1899794	-0.3296564	-0.1216394

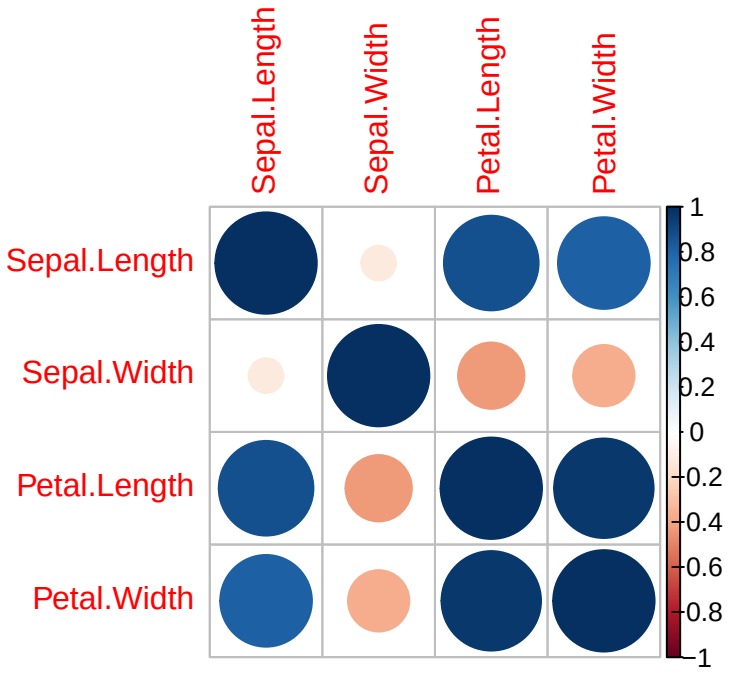
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Petal.Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal.Width	0.5162707	-0.1216394	1.2956094	0.5810063

9.11 Représentation graphique (corrplot)

```
library(corrplot)
```

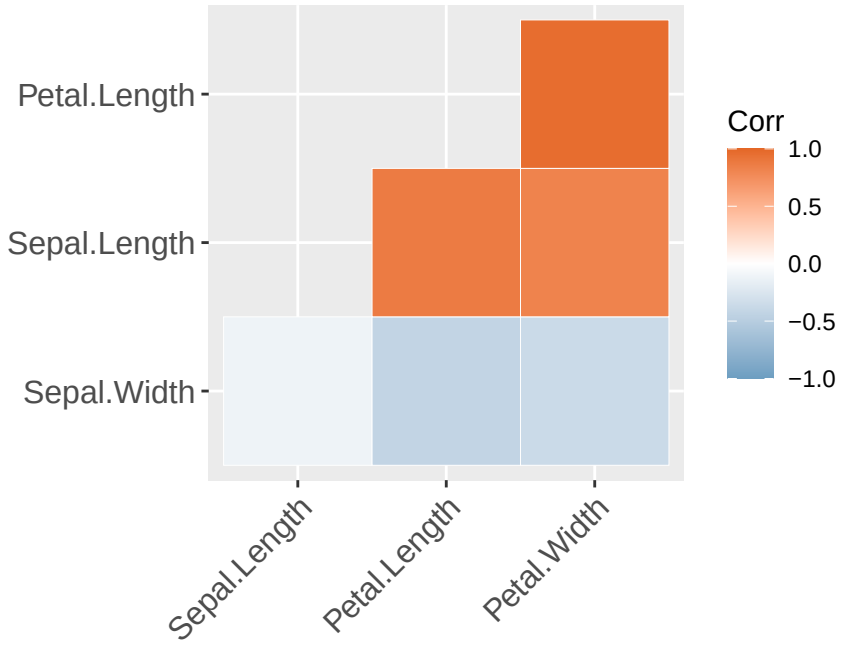
```
corrplot 0.92 loaded
```

```
data(iris)
corrplot(cor(iris[, 1:4]))
```



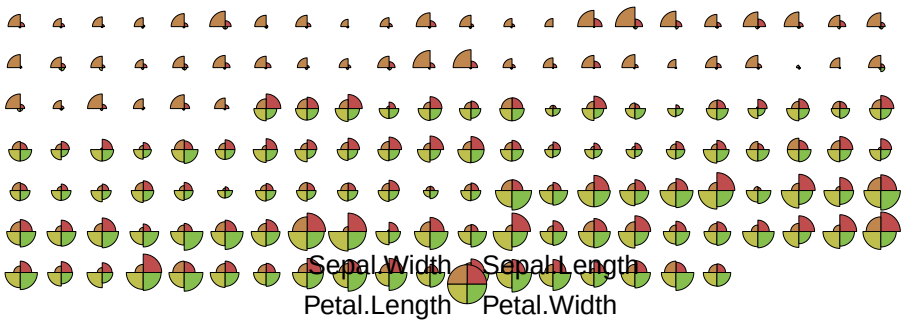
9.12 Représentation graphique (ggcorrplot)

```
library(ggcorrplot)
data(iris)
ggcorrplot(cor(iris[, 1:4]),
             hc.order = TRUE, type = "lower",
  outline.color = "white",
  ggtheme = ggplot2::theme_gray,
  colors = c("#6D9EC1", "white", "#E46726")
)
```



9.13 Description multidimensionnelle

9.13.1 Les diagrammes fleurs



9.14 Fléau de la dimension (curse of dimensionality)

- Espace de grande dimension
- Calculs similaires à ceux du plan
- Difficile de généraliser

9.14.1 Exemple 1 :

- Dans $\mathbb{R}$ - Pts uniformément répartis dans $[-1, +1]$ - % de points situées à 1 distance $\leq$ 0.75 de l'origine : 75%
- Dans $\mathbb{R}^{10}$ - Pts uniformément répartis dans $[-1, +1]^{10}$ - % de points situées à 1 distance $\leq$ 0.75 de l'origine : 5%

9.14.2 Exemple 2 :

on veut construire un histogramme en s'appuyant sur au moins une moyenne de 10 points par intervalle et 10 classes par variable

- $\mathbb{R}$: 10 classes $n = 100$
- $\mathbb{R}^2$: 100 classes $n = 1000$

- $\mathbb{R}^{10}$: 10^{10} classes $n = 10^{11} = 100 milliards$
- Si p assez grand, l'espace $\mathbb{R}^p$ est pratiquement vide et sauf si les données se situent au voisinage d'une variété de faible dimension, l'analyse des données n'apportera aucune information intéressante.
- Les points voisins d'un point donné sont tous très loin : difficultés dans l'emploi de méthodes du type k-plus proches voisins

10 Couple de variables discrètes

10.1 Echantillon de deux variables discrètes

Soit un échantillon où chaque individu est caractérisé par deux variables discrète X et Y distribuées suivant des loi multinomiales à respectivement p et q modalités:

$$((X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n))$$

10.1.1 Domaine de variation

- X_i prend ses valeurs sur $V_X = \{l_1, ..., l_p\}$
- Y_i prend ses valeurs sur $V_Y = \{m_1, ..., m_q\}$

10.1.2 Exemple

- X pourrait être la couleur des yeux:
  - $V_X = \{marron, bleu, vert, noisette\}$, $p = 4$
- Y pourrait être la couleur des cheveux:
  - $V_Y = \{noir, brun, roux, blond\}$, $q = 4$

10.2 Tableau de données initiales

Ce type d'échantillon est classiquement représenté par un tableau binaire T à

- n lignes (une ligne par individu)
- $p + q$ colonnes (une colonne par modalité)

où

10.2.1 Codage

$$T_{ij} = \begin{cases} 1, \text{ si l'individu i possède la modalité j,} \\ 0, \text{sinon.} \end{cases}$$

10.2.2 Exemple (suite)

En poursuivant avec l'exemple des couleurs de cheveux et d'yeux. Si

	mar.	bleu	vert	nois.	noir	brun	...
Individu 1	1	0	0	0	0	1	...
Individu 2	0	1	0	0	0	1	...
...	...	...	...	...	...	...	...
Individu n	0	0	1	0	1	0	...

10.3 Tableau croisé ou table de contingence

Ce type de tableau peut se représenter sous la forme d'un

- tableau de contingence
- ou tableau croisé

à

- p lignes
- q colonnes

Chaque case du tableau de contingence compte le nombre d'individus possédant la modilité i de la variable X et j de la variable Y: n_{ij}

10.3.1 Tableau de contigence

	m_1	m_2	...	m_q
l_1	n_{11}	n_{12}	...	n_{1q}
l_2	n_{21}	n_{22}	...	n_{2q}
...	...	...	...	...
l_p	n_{p1}	n_{p2}	...	n_{pq}

10.3.2 Marges

A ce tableau on peut rajouter une ligne et une colonne contenant les marges

- $n_{i\bullet} = \sum_j n_{ij}$ (marge en ligne)
- $n_{\bullet j} = \sum_i n_{ij}$ (marge en colonne)

Le nombre total d'individus de l'échantillon est bien sûr

$$n = \sum_{ij} n_{ij} = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$$

10.4 Tableau de contingence en R

Pour obtenir un tableau de contingence en R, il suffit d'utiliser l'instruction `table`:

```
X<-sample(c("Brown","Blue","Hazel","Green"),prob=c(6,3,1,1),replace=T,size=200)
Y<-sample(c("Black","Brown","Red","Blond"),prob=c(2,5,2,3),replace=T,size=200)
print(ContingencyTable<-table(X,Y))
```

```
       Y
X       Black Blond Brown Red
  Blue      5    15    24  10
  Brown    19    28    40  14
  Green     1     2    12   4
  Hazel     6     5    10   5
```

11 Dépendance versus indépendance

11.1 Probabilités jointes

Le tableau de contingence permet d'estimer la loi jointe du couple variables (X, Y):

$$P(X = l_i, Y = m_j) = \frac{n_{ij}}{n}$$

11.1.1 Exemple

La probabilité d'avoir les yeux bleus et les cheveux blonds peut être estimée par

```
ProbabilityBlueEyeBlondHair<-ContingencyTable[1,2]/sum(ContingencyTable)
print(ProbabilityBlueEyeBlondHair)
```

```
[1] 0.075
```

11.2 Probabilités marginales

A partir du tableau de contingence, on peut estimer les probabilités marginales, qui sont les probabilités des modalités:

- en lignes $P(X = l_i) = \frac{n_{i\bullet}}{n}$
- en colonnes $P(Y = m_j) = \frac{n_{\bullet j}}{n}$

11.2.1 Exemple

La probabilité d'avoir les yeux bleus peut être estimée par

```
ProbabilityBlueEye<-sum(ContingencyTable[1,])/sum(ContingencyTable)
print(ProbabilityBlueEye)
```

```
[1] 0.27
```

11.3 Probabilités conditionnelles

Ayant les loi marginales et jointes, il est aisée d'avoir les estimatations des probabilités conditionnelles

- en lignes $P(X = l_i|Y = m_j) = \frac{P(X=l_i,Y=m_j)}{P(Y=m_j)} = \frac{n_{ij}}{n_{\bullet j}}$
- en colonnes $P(Y = m_j|X = l_i) = \frac{P(X=l_i,Y=m_j)}{P(X=l_i)} = \frac{n_{ij}}{n_{i\bullet}}$

11.3.1 Exemple

La probabilité d'avoir les yeux bleus sachant que l'on a les cheveux blonds peut être estimée par

```
ProbabilityBlueEyeConditionnalBlond<-ContingencyTable[1,2]/sum(ContingencyTable[,2])
print(ProbabilityBlueEyeConditionnalBlond)
```

```
[1] 0.3
```

11.4 Hypothèse d'indépendance

Si les deux variables X et Y étaient indépendantes alors il suffirait des marges pour reconstruire le tableau.

Comme $P(X = l_i, Y = m_j) = P(X = l_i)P(Y = m_j)$, on pourrait estimer la probabilité jointe $P(X = l_i, Y = m_j)$ par

$$\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}.$$

et les n_{ij} théoriques seraient

$$nP(X = l_i, Y = m_j) = \frac{n_{i\bullet}n_{\bullet j}}{n}$$

12 Représentation d'un couple de variables qualitatives

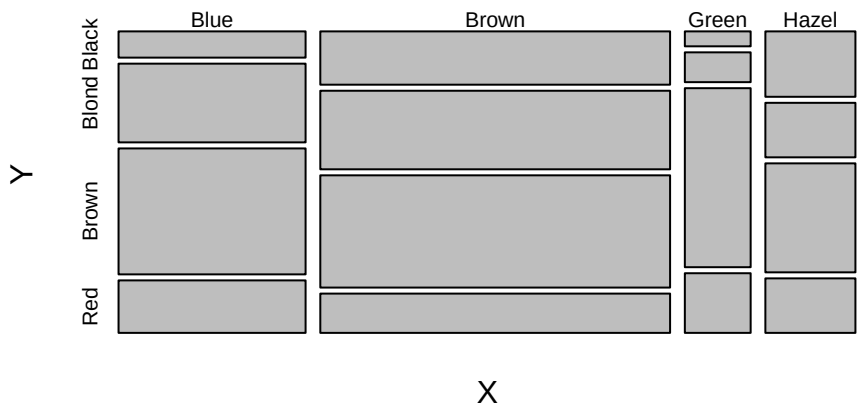
12.1 Diagramme mosaïque

Le tableau mosaique vise à représenter un tableau de contingence avec des informations sur ses marges:

- chaque colonne j possède
  - une largeur proportionnelle à sa marge $n_{\bullet j}$
- chaque case ij dans une colonne j possède
  - une hauteur proportionnelle à $\frac{n_{ij}}{n_{\bullet j}}$
- la surface de chaque case ij est donc proportionnelle à son effectif n_{ij}

12.2 Tableau mosaïque en R

```
plot(ContingencyTable)
```



12.3 Diagnostique de la représentation

Si les deux variables X et Y sont indépendantes, les hauteurs des cases $i\bullet$ sont toujours les mêmes (car proportionnelles à $n_{i\bullet}$).

Plus le diagramme mosaïque semble être traversé de lignes horizontales et plus l'hypothèse d'indépendance semble vraisemblable.

12.3.1 Aide au diagnostique

Pour aider le diagnostique, on peut ajouter de la couleur indiquant pour chaque case l'écart entre l'attendu et l'observé (résidus de Pearson)

12.3.2 Résidus de Pearson

$$d_{ij} = \frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

avec

- $f_{ij} = \frac{n_{ij}}{n}$ les fréquences observées
- $e_{ij} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n}$ les fréquences attendues sous hyp. d'indépendance

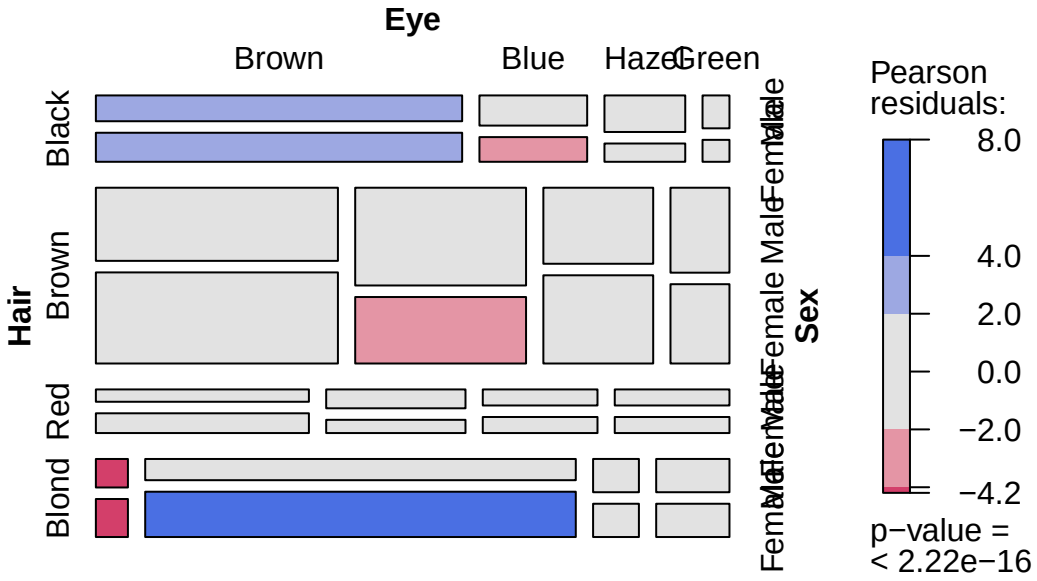
12.4 Exemple réel

592 hommes et femmes dont la couleur des yeux et des cheveux a été notée

```
library(vcd)
```

```
Loading required package: grid
```

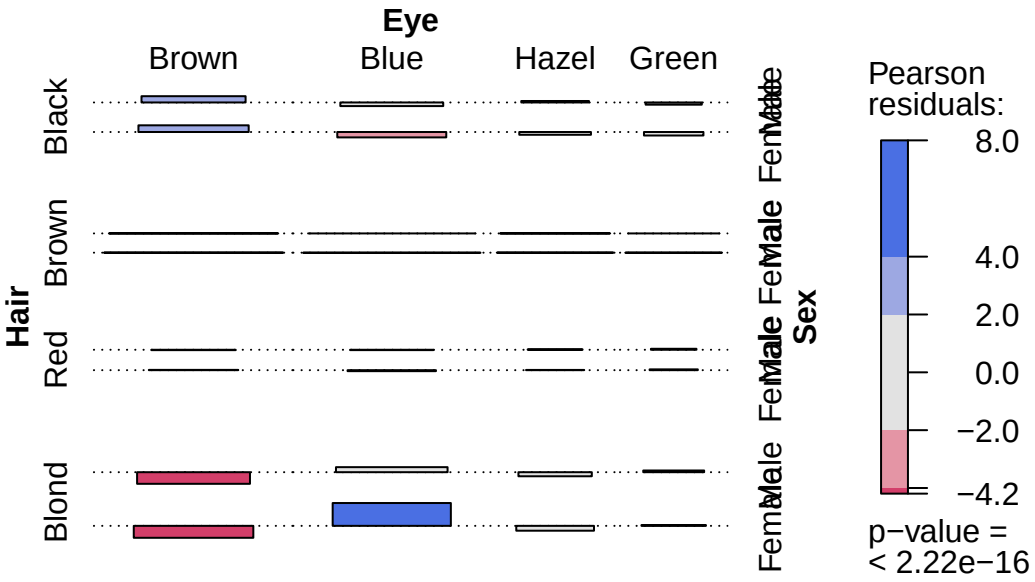
```
mosaic(HairEyeColor, shade=TRUE, legend=TRUE)
```



12.5 Diagrammes d'association

Dans les diagrammes d'association chaque case est représentée par une base dont la hauteur (positive ou négative) est directement le résidu de Pearson correspondant

```
assoc(HairEyeColor, shade=TRUE, legend=TRUE)
```


Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond
Sex
Male
Female
Male
Female
Male
Female
Male
Female
Pearson
residuals:
8.0
4.0
2.0
0.0
−2.0
−4.2
p−value =
< 2.22e−16