

ISV51: Programmation sous R

Analyse de données élémentaire

L3 GBI – Université d'Evry

semestre d'automne 2015

http://julien.cremeriefamily.info/teachings_L3BI_ISV51.html

Plan

Entrées/sorties

Statistiques descriptives

Plan

Entrées/sorties

- Charger des données

- Bases des graphiques sous R

Statistiques descriptives

Plan

Entrées/sorties

- Charger des données

- Bases des graphiques sous R

Statistiques descriptives

Saisir des données

Alternative à la concaténation

commande `scan`

Une utilisation élémentaire de `scan` permet une saisie plus agréable que la saisie directe des éléments d'un vecteur.

```
> x<-scan()  
1: 1  
2: 2  
3: 3  
4: 4  
5: 5  
6:  
Read 5 items  
>  
> x  
[1] 1 2 3 4 5
```

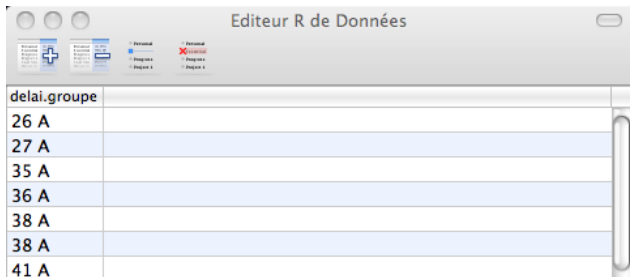
↪ valable pour les jeux de données d'au plus quelques dizaines d'éléments. . .

Éditer des données

commande `edit`

Permet d'éditer des données existantes à l'aide d'un mini-tableur. Utile pour faire de petites modifications.

```
> new.data <- edit(old.data)
```



delai.groupe	
26 A	
27 A	
35 A	
36 A	
38 A	
38 A	
41 A	

Figure: Éditeur Mac OS 10.6 / R 2.10 (obsolète !)

Fichiers binaires

commandes save et load

save sauvegarde un sous ensemble des variables de l'espace de travail dans un fichier binaire ; load permet de les recharger.

```
x <- rnorm(125)
y <- 1 + x + x^2
save(file="mes_simus",x,y)
rm(list=ls())
objects()

## character(0)

load(file="mes_simus")
objects()

## [1] "x" "y"
```

Jeux de données prédéfinies

commande data

R dispose d'une **collection de données prédéfinies** directement utilisables.
La commande `data()` permet de les lister puis de les charger.

```
data(iris)  
head(iris)
```

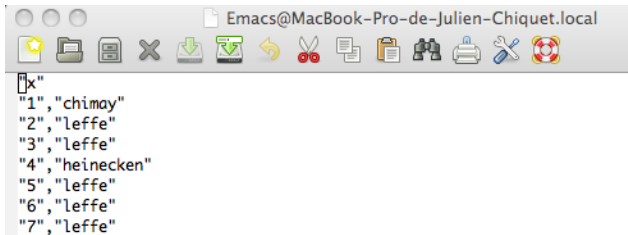
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1         5.1         3.5         1.4         0.2   setosa  
## 2         4.9         3.0         1.4         0.2   setosa  
## 3         4.7         3.2         1.3         0.2   setosa  
## 4         4.6         3.1         1.5         0.2   setosa  
## 5         5.0         3.6         1.4         0.2   setosa  
## 6         5.4         3.9         1.7         0.4   setosa
```

- ▶ La description d'un jeu de données est accessible dans l'aide.
- ▶ L'installation d'un nouveau package rend souvent disponibles de nouveaux jeux de données accessibles par `data`.

Lecture de fichiers

Méthodologie

Un bon éditeur permet de constater le formatage d'un fichier texte et comment en « attaquer » l'importation.

A screenshot of an Emacs editor window on a Mac. The title bar reads "Emacs@MacBook-Pro-de-Julien-Chiquet.local". The toolbar contains various icons for file operations. The main text area displays a CSV file with the following content:

```
"x"  
"1", "chimay"  
"2", "leffe"  
"3", "leffe"  
"4", "heinecken"  
"5", "leffe"  
"6", "leffe"  
"7", "leffe"
```

Figure: Fichier au formatage "csv"

Lecture de fichiers I

La fonction générique

commande `read.table`

Permet de lire un fichier formaté sous forme de table et de le convertir sous la forme du objet `data.frame`. Parmi les nombreuses options, les plus importantes sont

- ▶ `header` : présence ou pas d'une ligne nommant les colonnes du tableau
- ▶ `sep` : la chaîne de caractère définissant le séparateur (par défaut, un ou plusieurs espaces).

```
vignes <- read.table("data/baies_raisin.txt")
```

```
## Error in scan(file, what, nmax, sep, dec, quote, skip, nlines, na.strings, :  
la ligne 2 n'avait pas 20 éléments
```

Lecture de fichiers II

La fonction générique

```
vignes <- read.table("data/baies_raisin.txt", sep='\t')
head(vignes)
```

```
##           V1      V2           V3           V4
## 1 Population variete nbre pepin/baie 2008 poids pulpe/baie (g) 2008
## 2          CE    1784           1.00           0.89
## 3          CE     124           1.00           1.14
## 4          CE     210           1.20           1.26
## 5          CE    1805           1.20           0.66
## 6          CE    1303           1.20           0.83
##           V5           V6           V7
## 1 volume baie (cm3) 2008 nbre pepin/baie 2009 poids pulpe/baie (g) 2009
## 2                   7.70
## 3                   8.82
## 4                   10.20
## 5
## 6
```

Lecture de fichiers III

La fonction générique

```
vignes <- read.table("data/baies_raisin.txt", sep='\t', header=TRUE)
head(vignes)
```

##	Population	variete	nbre.pepin.baie.2008	poids.pulpe.baie..g..2008
## 1	CE	1784	1.0	0.89
## 2	CE	124	1.0	1.14
## 3	CE	210	1.2	1.26
## 4	CE	1805	1.2	0.66
## 5	CE	1303	1.2	0.83
## 6	CE	284	1.3	0.54
##	volume.baie..cm3..2008	nbre.pepin.baie.2009	poids.pulpe.baie..g..2009	
## 1	7.70	NA	NA	
## 2	8.82	NA	NA	
## 3	10.20	NA	NA	
## 4	NA	NA	NA	
## 5	NA	NA	NA	
## 6	4.61	NA	NA	

Lecture de fichiers

`read.csv`, `read.delim`

Commandes `read.csv` et `read.delim`

Raccourcis pour la fonction `read.table`, spécialisés dans l'importation des données « `.csv` » (*comma-separated value*) ou tabulées (le séparateur est la tabulation).

```
vignes <- read.delim(file="data/baies_raisin.txt", header=TRUE)
head(vignes)
```

```
##   Population variete nbre.pepin.baie.2008 poids.pulpe.baie..g..2008
## 1         CE      1784             1.0                0.89
## 2         CE       124             1.0                1.14
## 3         CE       210             1.2                1.26
## 4         CE      1805             1.2                0.66
## 5         CE      1303             1.2                0.83
## 6         CE       284             1.3                0.54
##   volume.baie..cm3..2008 nbre.pepin.baie.2009 poids.pulpe.baie..g..2009
## 1                   7.70                 NA                 NA
## 2                   8.82                 NA                 NA
## 3                  10.20                 NA                 NA
## 4                   NA                 NA                 NA
## 5                   NA                 NA                 NA
## 6                   4.61                 NA                 NA
```

Écriture dans un fichiers I

`write.table`

commandes `write.table`, `write.csv` et `write.delim`

La fonction `write.table` permet d'imprimer les données issues d'un `data.frame` dans un fichier texte externe. `write.csv` et `write.delim` sont des raccourcis pour les données csv ou tabulée.

```
vignes2008 <- vignes[, 1:5]
write.table(vignes2008, file="data/baies_raisin2008.txt")
rm(vignes2008)
```

Écriture dans un fichiers II

write.table

```
head(read.table(file="data/baies_raisin2008.txt", header=TRUE))
```

```
##   Population variete nbre.pepin.baie.2008 poids.pulpe.baie..g..2008
## 1          CE      1784              1.0                0.89
## 2          CE       124              1.0                1.14
## 3          CE       210              1.2                1.26
## 4          CE      1805              1.2                0.66
## 5          CE      1303              1.2                0.83
## 6          CE       284              1.3                0.54
##   volume.baie..cm3..2008
## 1                      7.70
## 2                      8.82
## 3                     10.20
## 4                      NA
## 5                      NA
## 6                      4.61
```

Pour aller plus loin...

Beaucoup de choses sur l'importation des données dans



R Data Import /Export.

<http://cran.r-project.org/doc/manuals/R-data.pdf>

- ▶ Exemples avancés avec `read.table`,
- ▶ communication avec les bases de données (SQL),
- ▶ importation de données Excel,
- ▶ ...

Plan

Entrées/sorties

Charger des données

Bases des graphiques sous R

Statistiques descriptives

Paramètres récurrents

Forme générique

La plupart des fonctions graphique s'utilisent par un appel du type

1. `nom.fonction(object, options),`
2. `nom.fonction(x, y , options).`

Parmi les options les plus courantes, on trouve :

- ▶ `type="p"` ; spécifie le type de tracé : "p" pour points, "l" pour lignes, "b" pour points liés par des lignes, "o" pour lignes superposées aux points. . .
- ▶ `xlim=` ; `ylim=`, spécifie les limites de axes x et y
- ▶ `xlab=` ; `ylab=`, annotation des axes x et y
- ▶ `main=` ; titre du graphe en cours
- ▶ `sub=` ; sous-titre du graphe en cours
- ▶ `add=FALSE` ; si TRUE superpose le graphe au précédent
- ▶ `axes=TRUE` ; si FALSE ne trace pas d'axes

Paramètres récurrents

Forme générique

La plupart des fonctions graphique s'utilisent par un appel du type

1. `nom.fonction(object, options),`
2. `nom.fonction(x, y , options).`

Parmi les options les plus courantes, on trouve :

- ▶ `type="p"` ; spécifie le type de tracé : "p" pour points, "l" pour lignes, "b" pour points liés par des lignes, "o" pour lignes superposées aux points. . .
- ▶ `xlim=` ; `ylim=`, spécifie les limites de axes x et y
- ▶ `xlab=` ; `ylab=`, annotation des axes x et y
- ▶ `main=` ; titre du graphe en cours
- ▶ `sub=` ; sous-titre du graphe en cours
- ▶ `add=FALSE` ; si TRUE superpose le graphe au précédent
- ▶ `axes=TRUE` ; si FALSE ne trace pas d'axes

Représenter un objet graphiquement I

commande `plot`

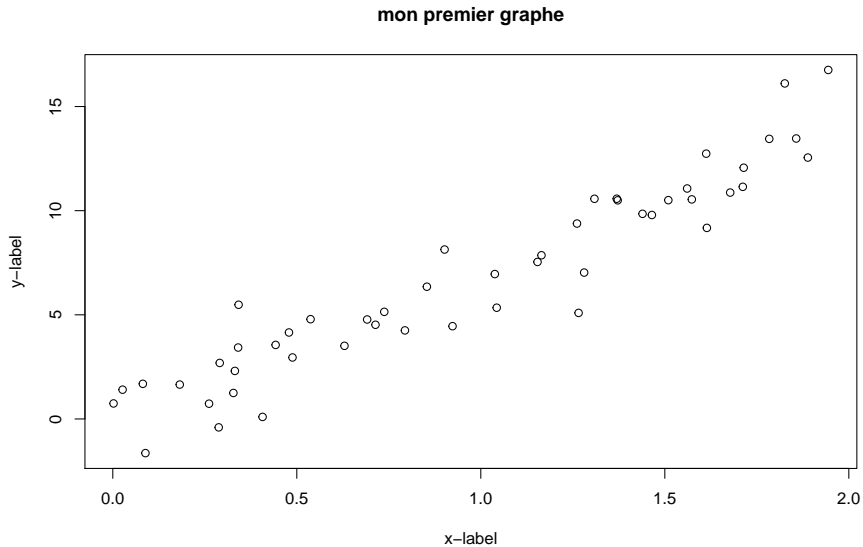
Fonction élémentaire de représentation graphique.

- ▶ `plot(vect)` représente le graphe des valeurs de `vect` sur l'axe des y .
- ▶ `plot(vect1,vect1)` représente le graphe des valeurs de `vect2` en fonction de `vect1`. `plot(object,...)` appelle la méthode `plot.class` si elle est définie pour l'objet de class `class`.

Par exemple, avec deux vecteurs :

```
x <- runif(50,0,2)
y <- 3 * x + 2 * x^2 + 1 + rnorm(50,sd=1.5)
plot(x, y, xlab="x-label",ylab="y-label",main="mon premier graphe")
```

Représenter un objet graphiquement II



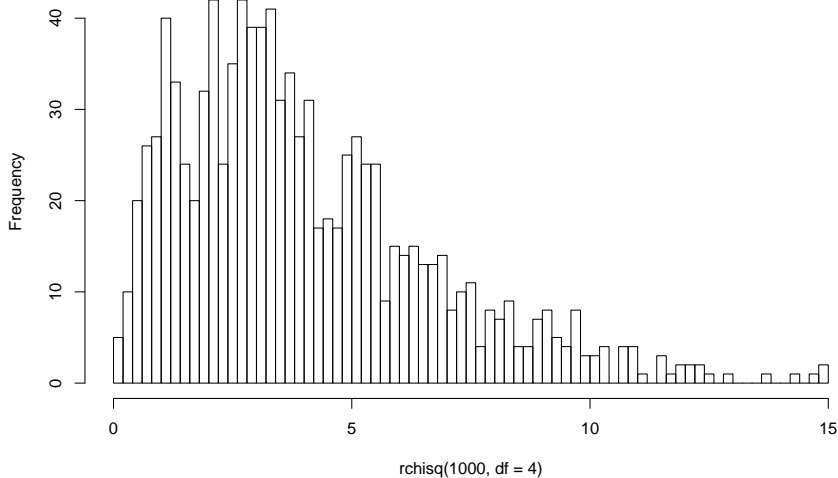
Autres exemples d'utilisation de plot (I) I

Beaucoup d'objet R accepte la commande `plot` ! En particulier, les histogrammes :

```
mon_histo <- hist(rchisq(1000,df=4),nclass=75, plot=FALSE)
plot(mon_histo,main="distribution empirique du Khi-2")
```

Autres exemples d'utilisation de `plot` (I) II

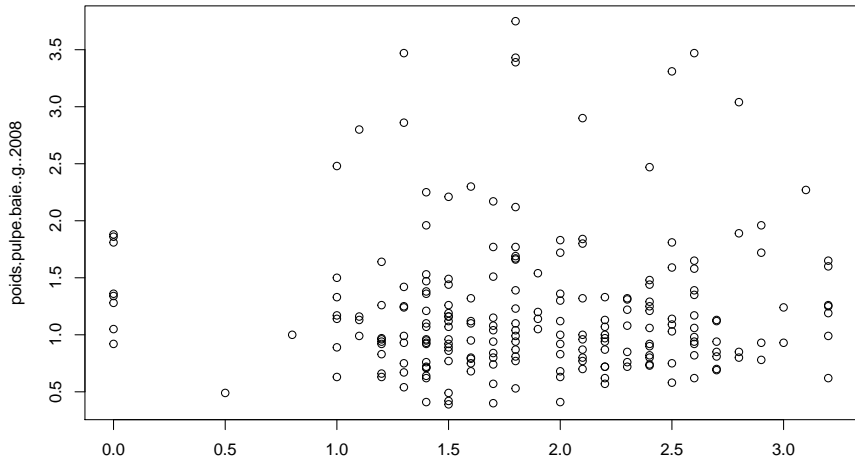
distribution empirique du Khi-2



Autres exemples d'utilisation de `plot` (II)

Objet "formule" entre variables numériques : graphe de dispersion

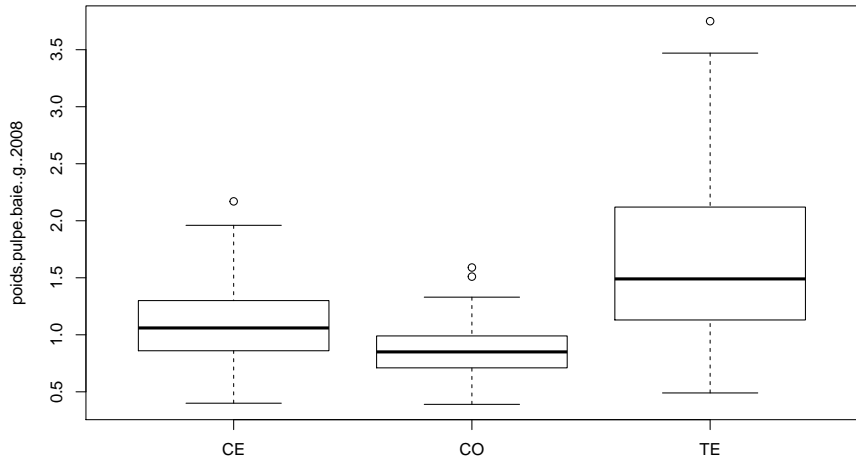
```
plot(poids.pulpe.baie..g..2008~nbre.pemin.baie.2008, vignes)
```



Autres exemples d'utilisation de plot (III)

Objet "formule" entre variables numérique et catégorielle : boxplot

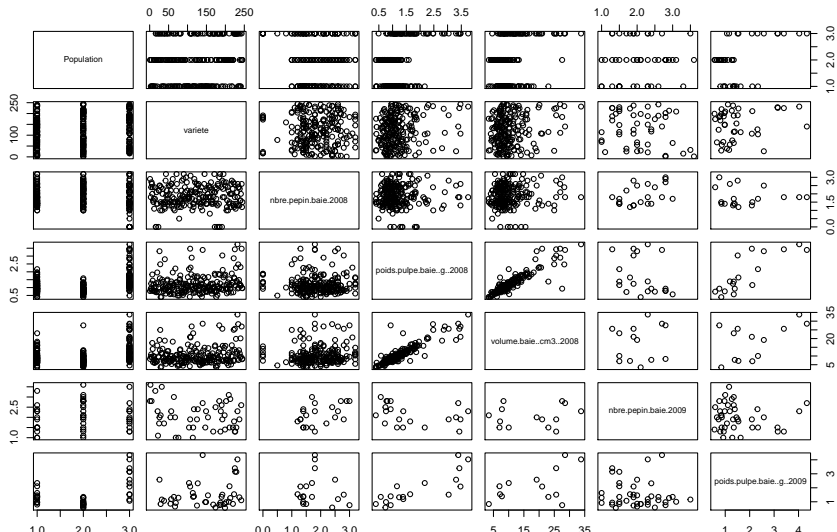
```
plot(poids.pulpe.baie..g..2008~Population, vignes)
```



Autres exemples d'utilisation de plot (IV)

Objet "data.frame" : graphes pair à pair

```
plot(vignes)
```



Tracer une fonction symbolique I

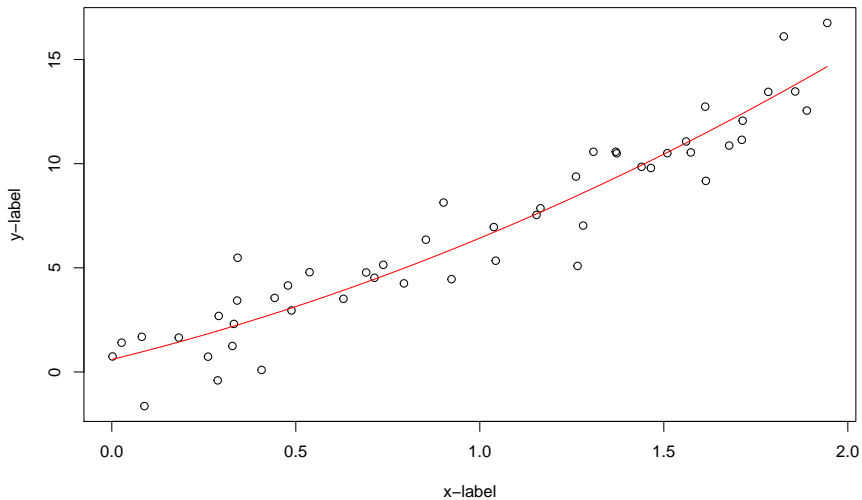
commande `curve`

Elle permet de tracer une fonction définie par une expression de x .

```
plot(x, y, main="données + modèle ajusté",  
      xlab="x-label", ylab="y-label")  
a <- coefficients(lm(y~1+x+I(x^2)))  
curve(a[1] + a[2]*x + a[3]*x^2, add=TRUE, col="red")
```

Tracer une fonction symbolique II

données + modèle ajusté



Ajouter une légende I

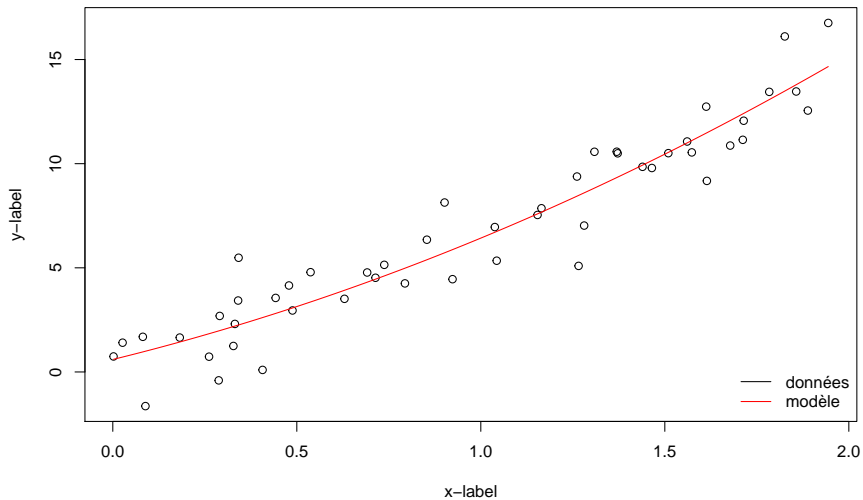
commande legend

Pour ajouter une légende. Attention aux options, assez nombreuses !

```
plot(x, y, main="données + modèle ajusté",  
      xlab="x-label", ylab="y-label")  
a <- coefficients(lm(y~1+x+I(x^2)))  
curve(a[1] + a[2]*x + a[3]*x^2,add=TRUE,col="red")  
legend("bottomright",c("données", "modèle"),lty=c(1,1),col=c("black", "red"),bty="n")
```

Ajouter une légende II

données + modèle ajusté



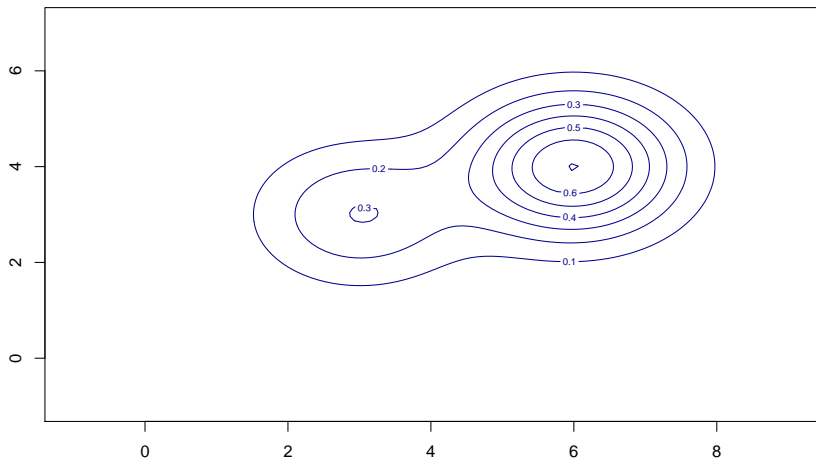
Représentation 3D (courbe de niveaux) I

commande `contour`

`contour(x,y,z)` permet de tracer des courbes de niveaux : `x` et `y` sont des vecteurs et `z` une matrice telle que les dimensions de `z` soient `length(x)`, `length(y)`.

```
x<-seq(-1,9,length=100)
y<-seq(-1,7,length=100)
z<-outer(x,y,function(x,y) 0.3*exp(-0.5*((x-3)^2 +(y -3)^2)) +
      0.7*exp(-0.5*((x-6)^2 +(y -4)^2)))
contour(x,y,z,col="blue4")
```

Représentation 3D (courbe de niveaux) II



Ajout de droites l

commande `abline`

`abline` permet d'ajouter à un graphe courant

- ▶ des droites de décalage `a` et de coefficient directeur `b` avec `abline(a,b)`,
- ▶ des droites verticales avec `abline(v=)`,
- ▶ des droites horizontales avec `abline(h=)`.

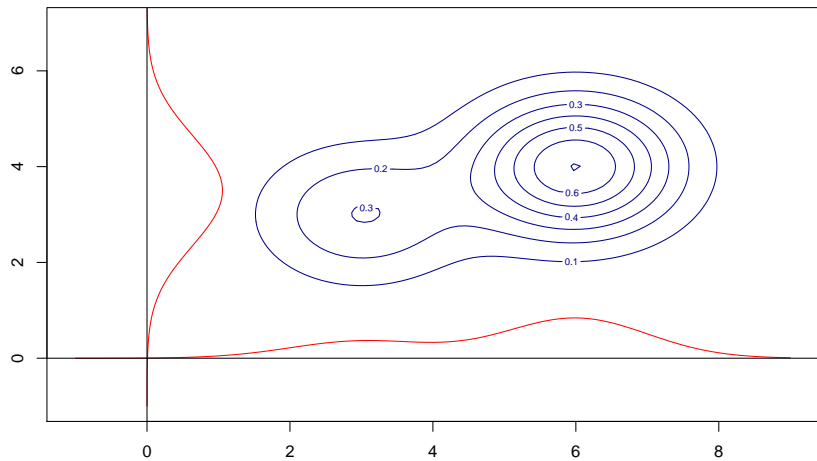
commandes `lines` et `points`

Pour ajouter une courbe ou des points : s'utilisent de manière similaire à `plot`.

Ajout de droites II

```
contour(x,y,z,col="blue4")
curve((0.3*dnorm(x,mean=3) + 0.7*dnorm(x,mean=6))*3,-1,9,col="red",ylim=c(-1,7),add=TRUE)
x<-seq(-1,9,length=100)
lines((0.5*dnorm(x,mean=3) + 0.5*dnorm(x,mean=4))*3,x,col="red")
abline(h=0)
abline(v=0)
```

Ajout de droites III



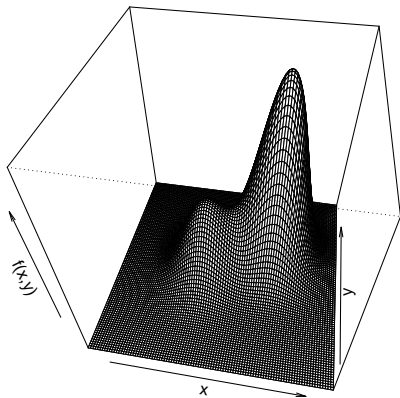
Graphe en 3D I

commande `persp`

Fonctionne comme la fonction `contour` en proposant une représentation en perspective.

```
persp(x,y,z, box=TRUE,theta = 10, phi = 45,xlab = "x", ylab = "y", zlab = "f(x,y)")
```

Graphe en 3D II



Rediriger la sortie graphique

Par défaut, R envoie les graphiques sur la sortie *écran*. De nombreuses

Exportation de graphes

Se réalise en encadrant les fonctions graphiques par les commandes `format_export(file="nom_fichier")` et `dev.off()`, où `format_fichier` peut prendre les valeurs `pdf`, `postscript`, `png`,

```
pdf(file="ma_sortie.pdf")  
plot(runif(20),runif(20))  
dev.off()
```

Graphes multiples

Ouverture d'une nouvelle fenêtre graphique

Se fait, selon les plateformes, avec les commandes

- ▶ `x11()` pour Linux,
- ▶ `quartz()` ou `x11()` pour Mac OS,
- ▶ `windows()`.

Découpage d'une fenêtre

Plusieurs possibilités :

- ▶ `layout(mat,width=,height=)`, qui s'utilise en découpant l'écran via la matrice `mat`.
- ▶ `par(mfrow=vect)` ou `par(mfcol=vect)` qui découpent en n lignes et m colonne spécifiées par le vecteur `vect`. Le remplissage se fait par ligne ou par colonne selon la fonction choisie.

Découpage du support graphique I

```
m1 <- matrix(1:4,2,2)
layout(m1)

m2 <- matrix(c(1:3,3),2,2)
layout(m2)

m3 <- matrix(0:3,2,2)
layout(m3,c(1,3),c(1,3))
```


Découpage du support graphique I

1

3

2

4

Découpage du support graphique II

1

2


3


Découpage du support graphique III



Pour aller plus loin

- ▶ La commande `par` gère les options graphiques,
- ▶ Le package `lattice`, pour des graphes multivariés,
- ▶ Le package `ggplot2`, dont nous verrons une introduction en fin de module

 `Lattice : Multivariate Data Visualization with R` Deepayan Sarkar
<http://lmdvr.r-forge.r-project.org/>

 `ggplot2 : Grammar of graphics`, Hadley Wickham
<http://ggplot2.org/>

↪ Au delà des mécanismes de représentation graphiques élémentaires, les possibilités graphiques de R sont **liées à la nature des résumés statistiques** opérés sur les données (cf. section suivante).

Plan

Entrées/sorties

Statistiques descriptives

- Généralités

- Statistique descriptive univariée

 - Variable qualitative

 - Variable quantitative

- Statistique descriptive multivariée

 - Croisement qualitatives/quantitatif

 - Couple de variables qualitatives

 - Couple de variables quantitatives

- Générateur aléatoire

Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Quelques définitions

- ▶ **Statistique** - activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation.
- ▶ **Population** - ensemble d'entités objet de l'investigation statistique.
- ▶ **Individu** - élément de la population d'étude
- ▶ **Variable/Attribut** - descripteur ou caractère des individus de la population d'étude.

Quelques définitions

- ▶ **Statistique** - activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation.
- ▶ **Population** - ensemble d'entités objet de l'investigation statistique.
- ▶ **Individu** - élément de la population d'étude
- ▶ **Variable/Attribut** - descripteur ou caractère des individus de la population d'étude.

Quelques définitions

- ▶ **Statistique** - activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation.
- ▶ **Population** - ensemble d'entités objet de l'investigation statistique.
- ▶ **Individu** - élément de la population d'étude
- ▶ **Variable/Attribut** - descripteur ou caractère des individus de la population d'étude.

Quelques définitions

- ▶ **Statistique** - activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation.
- ▶ **Population** - ensemble d'entités objet de l'investigation statistique.
- ▶ **Individu** - élément de la population d'étude
- ▶ **Variable/Attribut** - descripteur ou caractère des individus de la population d'étude.

Nature des variables

On distingue deux grandes familles de variable :

- ▶ **qualitative** ou factorielle : les valeurs prises sont les modalités
 - ▶ **ordinaire** : modalités intrinsèquement ordonnées (niveau de vie)
 - ▶ **nominale** : pas de structure d'ordre (sexe).
- ▶ **quantitative** : les valeurs prises sont des nombres
 - ▶ **discrète** : à valeurs dans un ensemble dénombrable (âge en année)
 - ▶ **continue** : à valeurs dans un ensemble indénombrable (taille, poids)

Nature des variables

On distingue deux grandes familles de variable :

- ▶ **qualitative** ou factorielle : les valeurs prises sont les modalités
 - ▶ **ordinaire** : modalités intrinsèquement ordonnées (niveau de vie)
 - ▶ **nominale** : pas de structure d'ordre (sexe).

- ▶ **quantitative** : les valeurs prises sont des nombres
 - ▶ **discrète** : à valeurs dans un ensemble dénombrable (âge en année)
 - ▶ **continue** : à valeurs dans un ensemble indénombrable (taille, poids)

Nature des variables

On distingue deux grandes familles de variable :

- ▶ **qualitative** ou factorielle : les valeurs prises sont les modalités
 - ▶ **ordinaire** : modalités intrinsèquement ordonnées (niveau de vie)
 - ▶ **nominale** : pas de structure d'ordre (sexe).

- ▶ **quantitative** : les valeurs prises sont des nombres
 - ▶ **discrète** : à valeurs dans un ensemble dénombrable (âge en année)
 - ▶ **continue** : à valeurs dans un ensemble indénombrable (taille, poids)

Mode d'étude d'une population

Échantillonnage

Processus de sélection d'individus dans la population d'étude.

↪ seule solution dans le cas d'une population infinie ou grande

Objectifs d'une étude statistique

À partir d'un échantillon,

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

Données

Soient n **individus** mesurés par p **variables**

Tableau de données

$$\mathbf{X} = (x_{ij}) = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & & x_{ij} & & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

- ▶ Chaque variable est représentée par la colonne $\mathbf{X}_{.j} = (x_{1j}, \dots, x_{nj})^\top$
- ▶ Chaque individu est représenté par la ligne $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})$

Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Contexte

On considère une seule **colonne à la fois** du tableau de données, soient n observations de la j^{e} variable :

$$\mathbf{X}_{.j} = (x_{1j}, \dots, x_{nj})^{\top}$$

Les résumés statistiques se regroupent selon la nature de la variable j , soientt

- ▶ qualitative ordinale (ou quantitative discrète)
- ▶ qualitative nominale
- ▶ quantitative continue

Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Variable quantitative discrète ou qualitative ordinale

La variable prend ses valeurs dans $E = \{\epsilon_1, \dots, \epsilon_K\}$ avec $\epsilon_1 < \dots < \epsilon_K$.

Tableau de fréquence

Les résumés statistiques naturels sont liés aux fréquences :

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

Fréquences et compagnie en R I

```
data(mtcars)
print(counts <- table(mtcars$gear))

##
## 3 4 5
## 15 12 5

print(frequencies <- counts/length(mtcars$gear))

##
##      3      4      5
## 0.46875 0.37500 0.15625

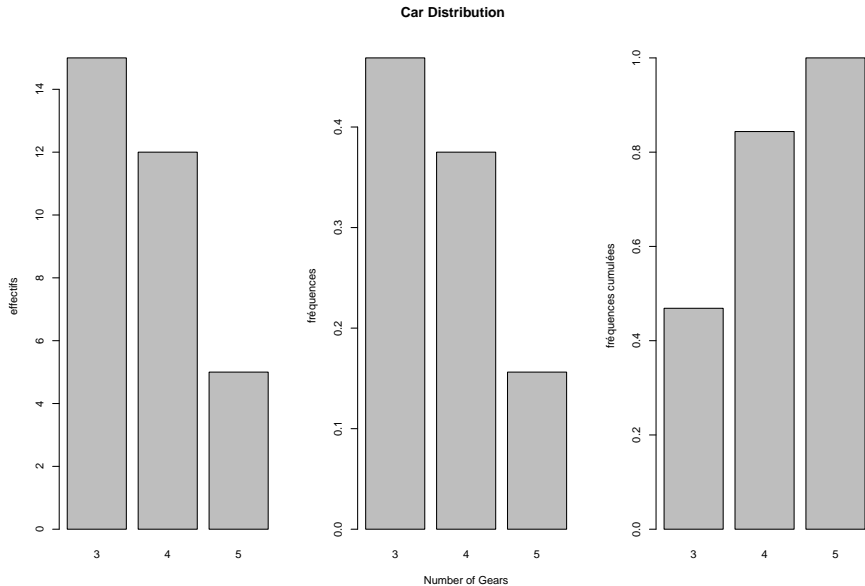
print(cumFreq      <- cumsum(frequencies))

##      3      4      5
## 0.46875 0.84375 1.00000
```

Fréquences et compagnie en R II

```
par(mfrow=c(1,3))
barplot(counts, ylab="effectifs", xlab="")
barplot(frequencies, ylab="fréquences", xlab="Number of Gears")
barplot(cumFreq, ylab="fréquences cumulées", xlab="")
title(outer=TRUE,main="\nCar Distribution")
```

Fréquences et compagnie en R III



Fréquences et compagnie en R I

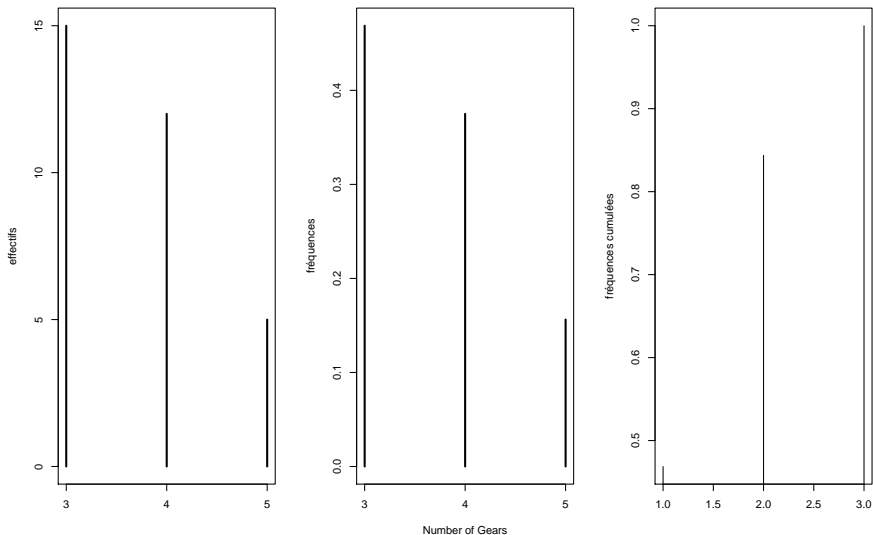
Une alternative avec plot

```
par(mfrow=c(1,3))
plot(counts, ylab="effectifs", xlab="", type="h")
plot(frequencies, ylab="fréquences", xlab="Number of Gears", type="h")
plot(cumFreq, ylab="fréquences cumulées", xlab="", type="h")
title(outer=TRUE,main="\nCar Distribution")
```

Fréquences et compagnie en R II

Une alternative avec plot

Car Distribution



Variable qualitative nominale

La variable prend ses valeurs dans $E = \{\epsilon_1, \dots, \epsilon_K\}$.

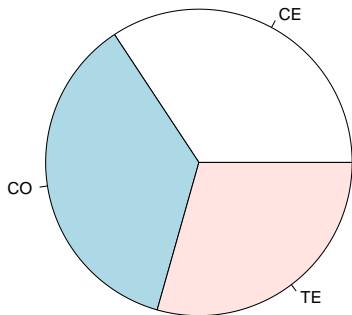
- ▶ mêmes représentations que pour les variables ordinales (i.e. construits sur les fréquences), mais sans ordre.
- ▶ Les diagrammes en barre restent appropriés mais **sans ordre naturel** en abscisse.

Camembert

Fournit une représentation non ordonné des effectifs.

À n'utiliser que pour un faible nombre de modalités (sinon illisible)

```
pie(table(vignes$Population))
```



Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Résumés numériques

Indicateurs de tendance centrale

- ▶ moyenne empirique : `mean`
- ▶ moyenne pondérée : `weighted.mean`
- ▶ médiane : `median`

Indicateurs de dispersion

- ▶ variance empirique (corrigée) : `var`
- ▶ écart-type : `sd`
- ▶ étendu : `range`
- ▶ fractiles empirique : `quantile`

`summary/fivenum` reprend ces indicateurs numériques élémentaires...

Résumés numériques I

Résumés numériques II

```
vol.cm3 <- vignes$volume.baiee..cm3..2008; vol.cm3 <- vol.cm3[!is.na(vol.cm3)]
mean(vol.cm3)

## [1] 10.46512

median(vol.cm3)

## [1] 8.91

var(vol.cm3) ## version corrigée !

## [1] 28.39179

sum((vol.cm3 - mean(vol.cm3))^2)/length(vol.cm3)

## [1] 28.25053

sum((vol.cm3 - mean(vol.cm3))^2)/(length(vol.cm3)-1)

## [1] 28.39179
```

Résumés numériques III

```
sd(vol.cm3) ## version corrigée

## [1] 5.328394

range(vol.cm3)

## [1] 3.44 33.80

library(stats)
quantile(vol.cm3)

##      0%      25%      50%      75%     100%
## 3.44  7.14  8.91 11.90 33.80

summary(vol.cm3)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 3.44    7.14    8.91   10.47   11.90   33.80

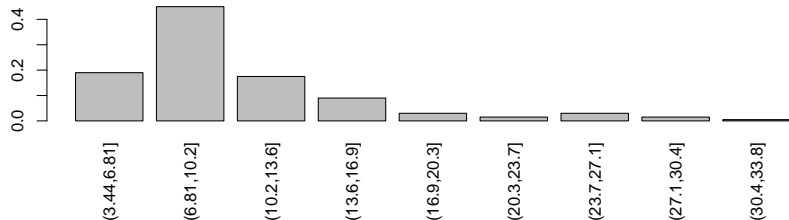
fivenum(vol.cm3) # correspond à summary pour un vecteur

## [1] 3.44 7.14 8.91 11.90 33.80
```

Tableau de fréquences

Pour une variable continue, nécessite un partitionnement préalable du domaine de définition en K classes (de largeur constante ou variable).

```
eff <- table(cut(vol.cm3, seq(min(vol.cm3),max(vol.cm3),len=10)))  
barplot(eff/sum(eff), las=3)
```



Graphe en tiges et feuilles

Alternative au diagramme en barres

Permet de visualiser le tableau des fréquences

```
stem(vol.cm3)
```

```
##  
## The decimal point is at the |  
##  
## 2 | 45  
## 4 | 01335667722245579  
## 6 | 000122334444555556789999900111112222333444555566677788889  
## 8 | 00112223444445555566677888990000223335566667799  
## 10 | 000001222345566779113455567899  
## 12 | 123489911234469  
## 14 | 01245685688  
## 16 | 012685  
## 18 | 2828  
## 20 | 170  
## 22 | 1  
## 24 | 0066  
## 26 | 906  
## 28 | 35  
## 30 |  
## 32 | 8
```

Boîte à moustaches ou boxplot I

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

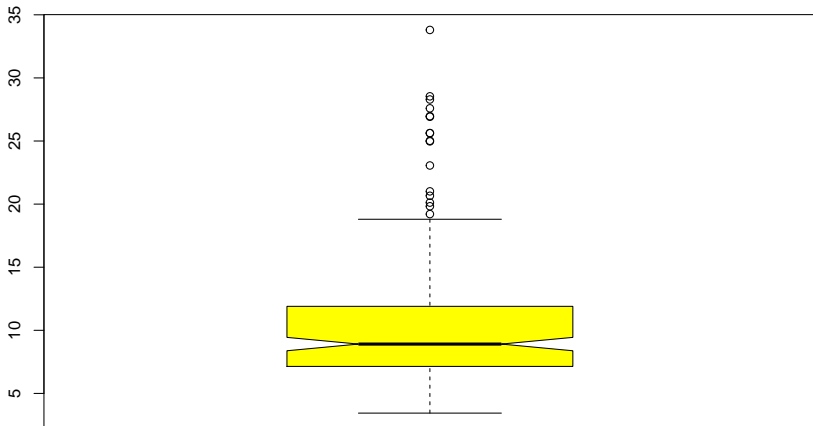
Définition

Graphique constitué

1. d'une **boîte** délimitée par les quartiles et la médiane
2. d'une **paire de moustaches** : minimum et maximum de l'échantillon auquel on a ôté les outliers.
 - ▶ *Les règles utilisées pour les outliers varient*
3. des **outliers** eux-même.

```
boxplot(vol.cm3,col="yellow",notch=T)
```

Boîte à moustaches ou boxplot II



Fonction de répartition empirique I

Définition

La version empirique de la fonction de répartition $F(x) = \mathbb{P}(X \leq x)$ s'écrit

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

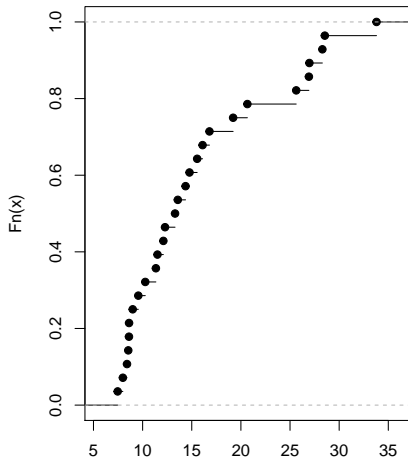
↪ le graphe de la fonction de répartition est une fonction en escalier appelé **diagramme cumulatif**

```
par(mfrow=c(1,2))
plot(ecdf(vol.cm3[vignes$Population == "TE"]), main="Population TE", xlab="")
plot(ecdf(vol.cm3[vignes$Population != "TE"]), main="Autres Populations", xlab="")
title(outer=TRUE, main="\nF.d.r du volume des baies")
```

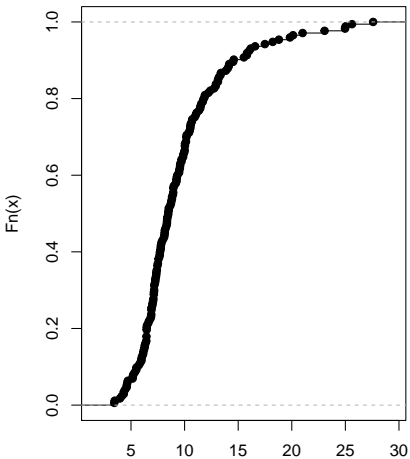
Fonction de répartition empirique II

F.d.r du volume des baies

Population TE



Autres Populations



Histogramme et estimateur à noyau I

Définition

Ce sont des estimateurs de la fonction de densité de x . On pose

$$\sum_i h_i \mathbf{1}_{[a_i, a_{i+1}[}(x) \text{ pour } a_1 < \dots < a_{k+1}.$$

On a $\sum_i h_i (a_{i+1} - a_i) = 1$ et $h_i (a_{i+1} - a_i) = \hat{\mathbb{P}}(X \in [a_i, a_{i+1}[)$.

Réalisation

1. Découpage en intervalles $[a_i, a_{i+1})$
2. Calcul de la fréquence f_i et de la hauteur h_i
3. Aire du rectangle proportionnel à la fréquence

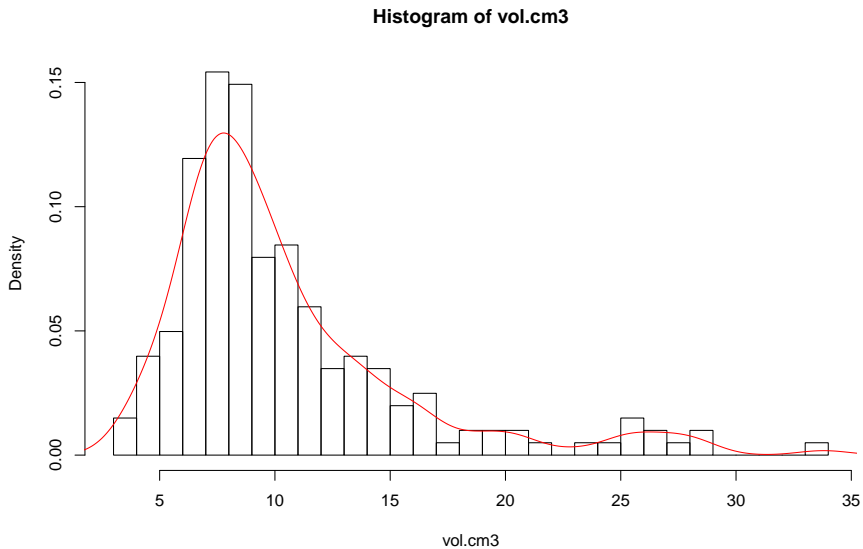
Histogramme et estimateur à noyau II

Remarques

- ▶ Attention : hauteur proportionnelle à la fréquence si et seulement si les intervalles ont tous la même largeur
- ▶ Nombre d'intervalles : Important, mais réglage difficile. . .

```
hist(vol.cm3,nclass=25,prob=TRUE)  
lines(density(vol.cm3), col="red")
```

Histogramme et estimateur à noyau III



Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

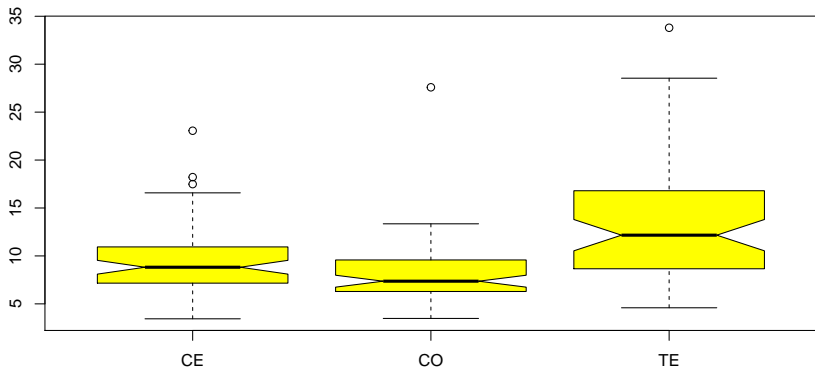
Couple de variables quantitatives

Générateur aléatoire

Représentation conditionnellement à un facteur

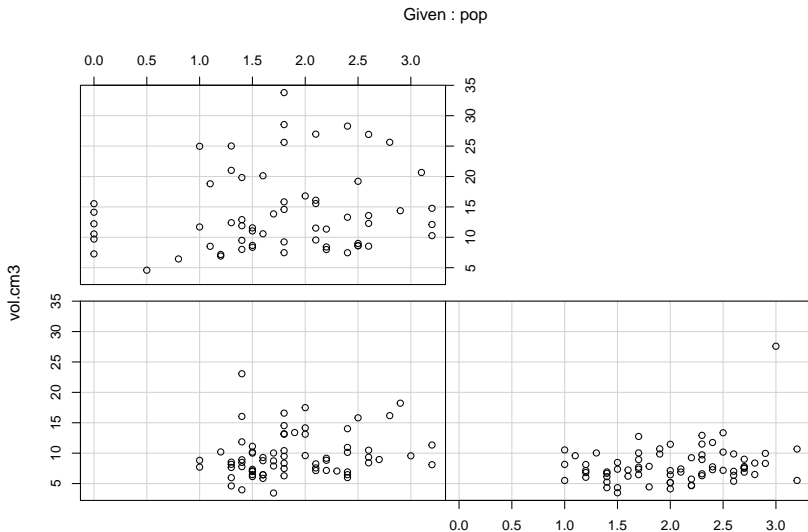
Les boîtes à moustaches se prettent bien à cet exercice

```
pop <- vignes$Population[!is.na(vignes$volume.baie..cm3..2008)]  
boxplot(vol.cm3~pop,col="yellow",notch=T)
```



Graphe conditionné par une variable

```
pepin <- vignes$nbre.pepin.baie.2008[!is.na(vignes$volume.baie.cm3.2008)]  
coplot(vol.cm3 ~ pepin | pop, show.given=FALSE)
```



Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Tableau croisé ou table de contingence I

Tableau de contingence

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

Marges

À 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

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

Tableau croisé ou table de contingence II

```
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         6   12   29  14
## Brown        20   24   44  18
## Green         1    2   14   2
## Hazel         2    4    7   1
```

Diagramme mosaïque I

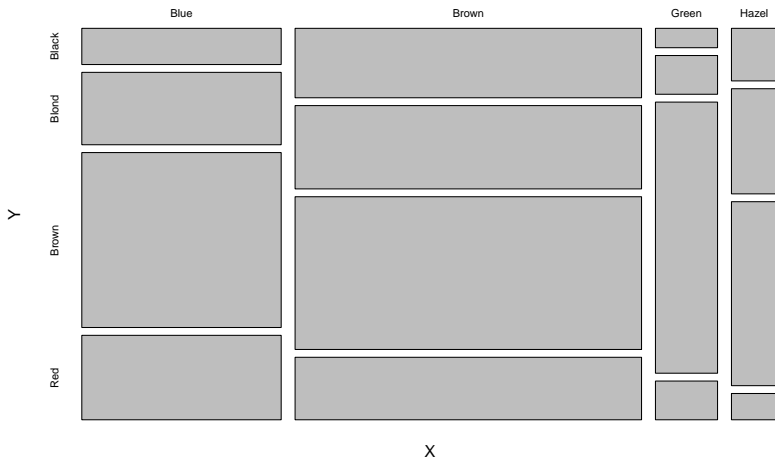
Représenter un tableau de contingence avec des informations sur ses marges

- ▶ 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}

```
plot(ContingencyTable)
```


Diagramme mosaïque II

ContingencyTable



Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

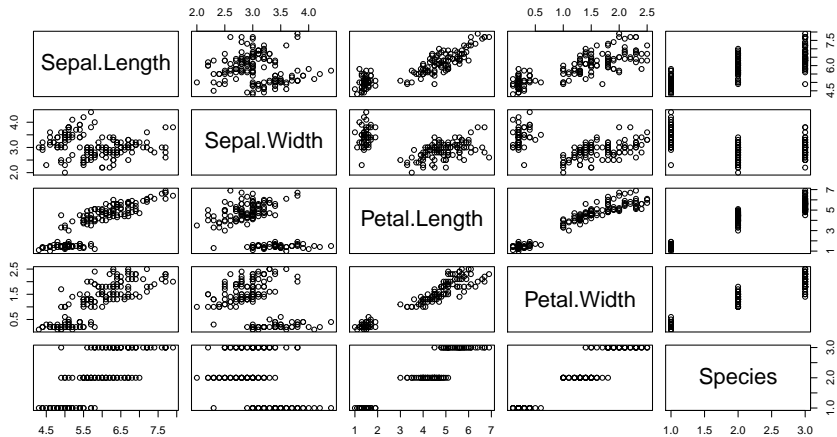
Couple de variables quantitatives

Générateur aléatoire

Graphes pair à pair

Représente toutes les paires de graphes naturels d'un tableau

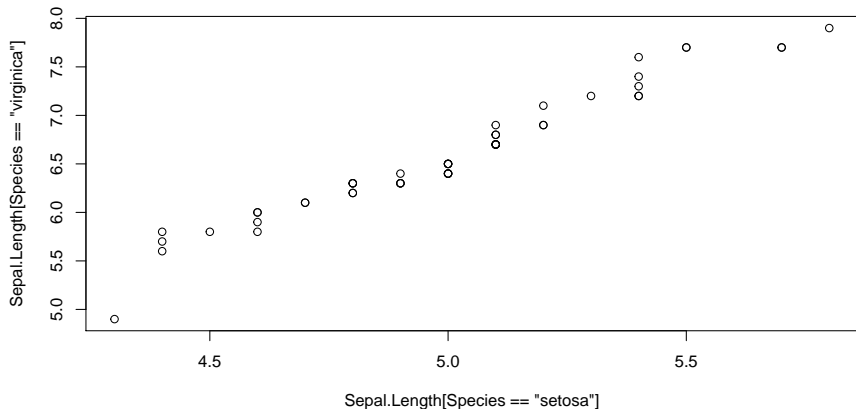
```
data(iris); pairs(iris)
```



Graphe quantile/quantile

Pour comparer visuellement les distributions de variables continues.

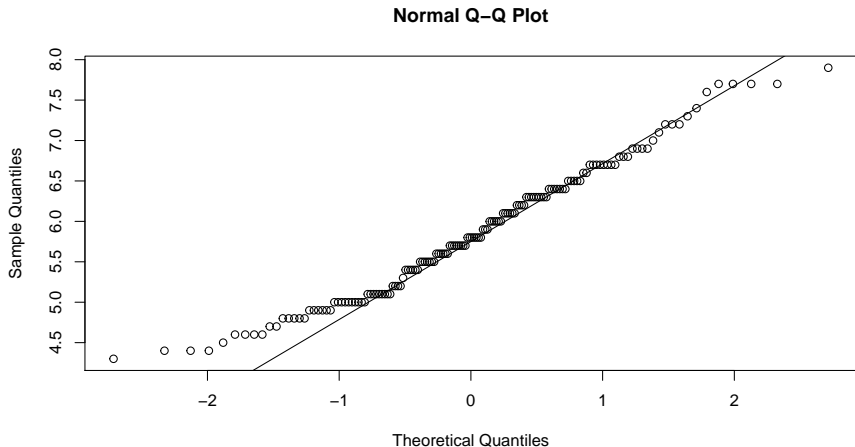
```
with(iris, qqplot(Sepal.Length[Species=="setosa"], Sepal.Length[Species=="virginica"]
```



Écart à la distribution normale

Une distribution est-elle normale? qqnorm/qqline donne une indication.

```
with(iris, qqnorm(Sepal.Length))  
with(iris, qqline(Sepal.Length))
```



Statistique du couple : covariance

Définition

Décrit l'écart conjoint de 2 variables à leurs espérances respectives

$$\text{cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$$

```
cov.mat <- cov(iris[, -5])
```

	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

Statistique du couple : corrélation

Définition

Il s'agit de la version normalisé de la covariance

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

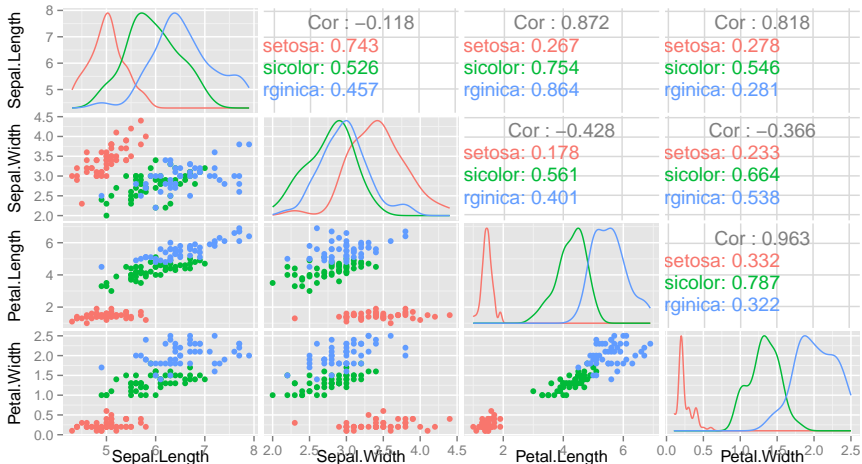
```
data(iris)
cor.mat <- cor(iris[, -5])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Graphe pair à pair et corrélation

La corrélation dit à quel point deux variables s'expliquent linéairement l'une l'autre.

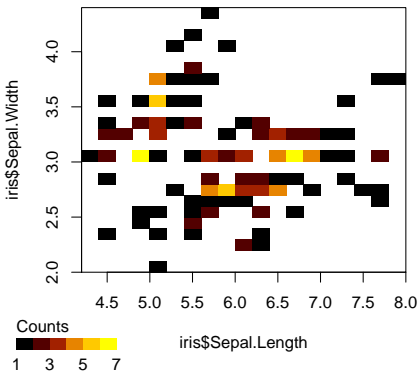
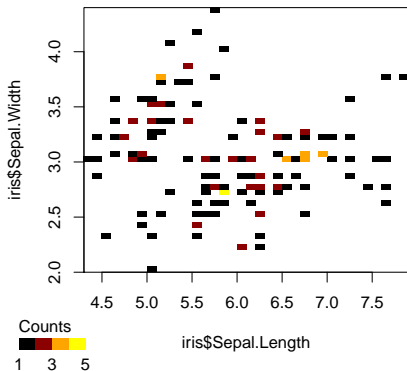
```
library(GGally); ggpairs(iris, columns = 1:4, color = "Species" )
```



Histogramme bidimensionnelle

Regroupe les points d'un graphe de dispersion par pavé bidimensionnels.

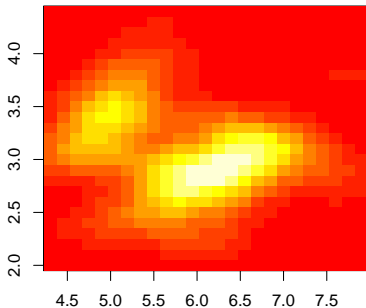
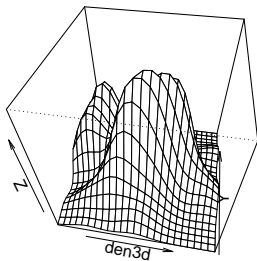
```
library(squash);  
par(mfrow=c(1,2))  
hist2(iris$Sepal.Length,iris$Sepal.Width)  
hist2(iris$Sepal.Length,iris$Sepal.Width, nx=20)
```



Histogramme bidimensionnelle lisse

Estime la densité associé à deux variables continues. Représentation 3D ou image possible

```
library(MASS)
den3d <- kde2d(iris$Sepal.Length, iris$Sepal.Width)
par(mfrow=c(1,2))
persp(den3d, box=TRUE,theta = 10, phi = 45)
image(den3d)
```



Plan

Entrées/sorties

Statistiques descriptives

Généralités

Statistique descriptive univariée

Variable qualitative

Variable quantitative

Statistique descriptive multivariée

Croisement qualitatives/quantitatif

Couple de variables qualitatives

Couple de variables quantitatives

Générateur aléatoire

Quelques distributions disponibles

Distribution	R	Paramètres
beta	beta	
binomiale	binom	size, prob
binomiale négative	nbinom	
Cauchy	cauchy	
Chi-deux	chisq	df
Exponentielle	exp	rate
Fisher	f	df1, df2
Gamma	gamma	
géométrique	geom	
hypergéométrique	hyper	
log-normal	lnorm	
logistique	logis	
normale	norm	mean, sd
normale multivariée	mvnorm	mean, sigma
Poisson	pois	
Student	t	df
uniforme	unif	min, max
Weibull	weibull	
Wilcoxon	wilcox	

Table: Principales distributions

Quelques distributions disponibles

Distribution	R	Paramètres
beta	beta	
binomiale	binom	size, prob
binomiale négative	nbinom	
Cauchy	cauchy	
Chi-deux	chisq	df
Exponentielle	exp	rate
Fisher	f	df1, df2
Gamma	gamma	
géométrique	geom	
hypergéométrique	hyper	
log-normal	lnorm	
logistique	logis	
normale	norm	mean, sd
normale multivariée	mvnorm	mean, sigma
Poisson	pois	
Student	t	df
uniforme	unif	min, max
Weibull	weibull	
Wilcoxon	wilcox	

Table: Principales distributions

Quelques distributions disponibles

Distribution	R	Paramètres
beta	beta	
binomiale	binom	size, prob
binomiale négative	nbinom	
Cauchy	cauchy	
Chi-deux	chisq	df
Exponentielle	exp	rate
Fisher	f	df1, df2
Gamma	gamma	
géométrique	geom	
hypergéométrique	hyper	
log-normal	lnorm	
logistique	logis	
normale	norm	mean, sd
normale multivariée	mvnorm	mean, sigma
Poisson	pois	
Student	t	df
uniforme	unif	min, max
Weibull	weibull	
Wilcoxon	wilcox	

Table: Principales distributions

Tirage aléatoire

Forme générique : `r+distrib(n,...)`

`r` pour « random » : `n` donne la taille de l'échantillon et `...` sont les paramètres requis selon la forme de `distrib`.

```
rexp(10,rate=1/5)
```

```
## [1] 2.6403409 0.9194465 3.8042244 3.9043327 1.1889369 12.2990216  
## [7] 0.1930232 6.8175360 16.5653974 4.7909241
```

```
rchisq(10,df=5)
```

```
## [1] 3.150175 11.666952 4.613559 5.786057 7.668031 2.983181 12.913080  
## [8] 1.855266 8.280931 1.390537
```

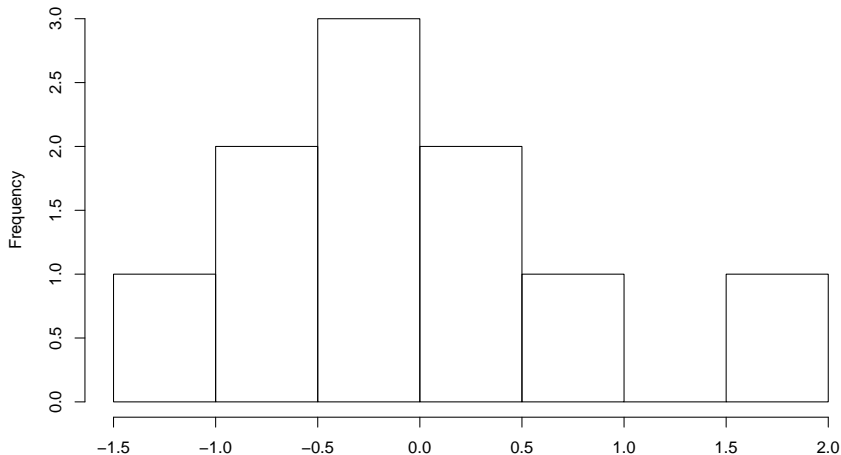
```
runif(10,min=-2,max=2)
```

```
## [1] -1.3300977 -1.9725736 1.9131699 0.7622626 -0.7475197 1.7300227  
## [7] 0.2942702 0.9741818 1.6627920 -1.4103824
```

Exemple avec la loi normale : histogramme

Avec $n = 10$

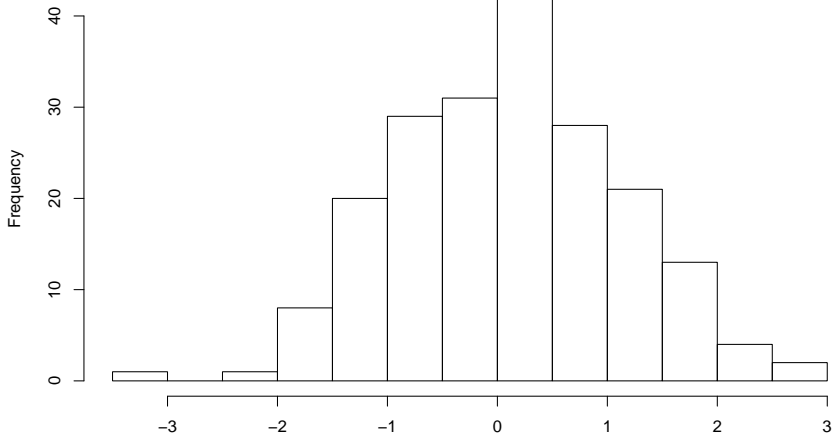
taille de l'échantillon = 10



Exemple avec la loi normale : histogramme

Avec $n = 200$

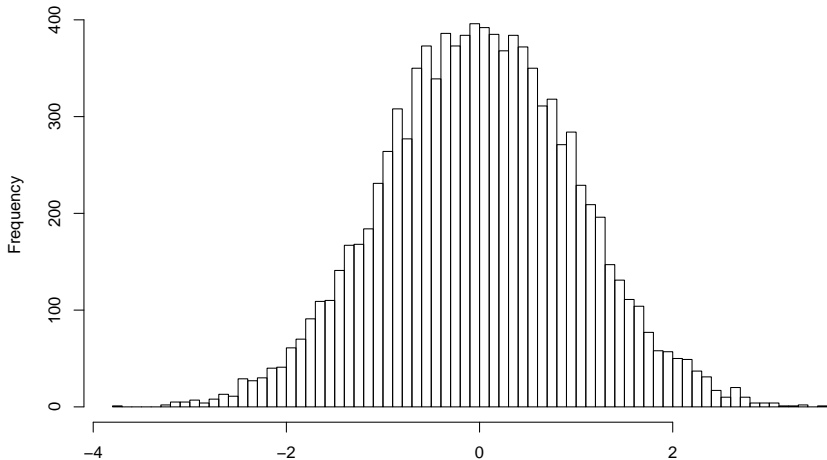
taille de l'échantillon = 200



Exemple avec la loi normale : histogramme

Avec $n = 10000$

taille de l'échantillon = 10000



Tirage aléatoire avec `sample`

Définir une distribution discrète

La fonction `sample(x, size, replace=FALSE, prob=NULL)` permet d'échantillonner les éléments de `x` : le tirage est de taille `size`, avec ou sans remise. Si `prob` est vide, chaque élément est équiprobable.

```
sample(1:5)
```

```
## [1] 5 2 4 3 1
```

```
sample(1:5,10,replace=TRUE)
```

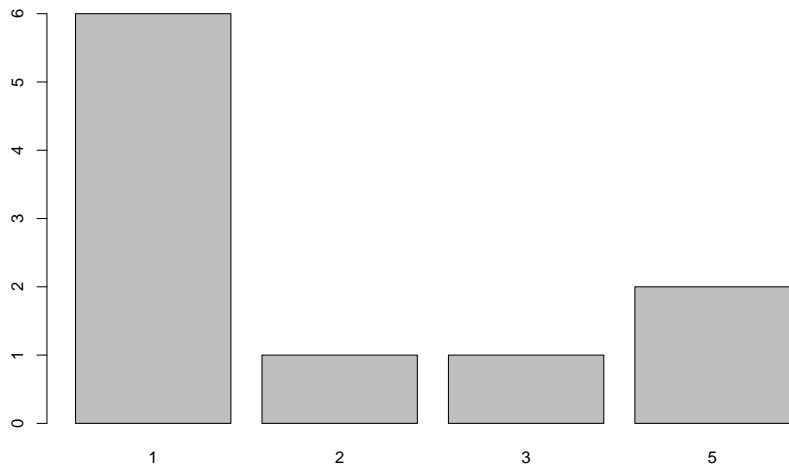
```
## [1] 3 3 1 3 2 5 2 1 5 4
```

```
sample(1:5,10,replace=TRUE,prob=c(.35,.1,.1,.1,0.35))
```

```
## [1] 4 3 1 1 5 1 5 3 1 2
```

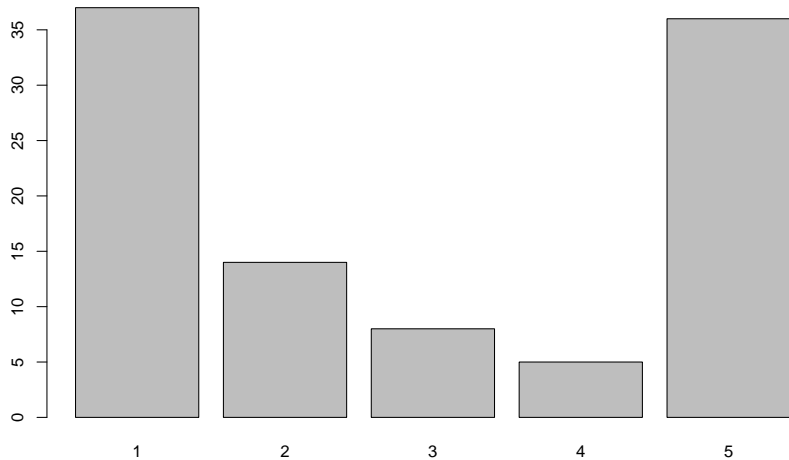
Avec $n = 10$

taille de l'échantillon = 10



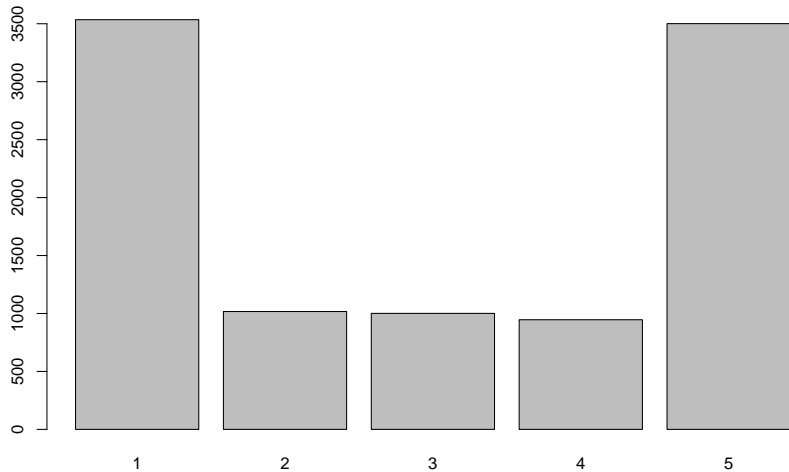
Avec $n = 100$

taille de l'échantillon = 100



Avec $n = 10000$

taille de l'échantillon = 10000



Fonction de répartition

Forme générique : `p+distrib(x,...)`

`p` pour « probability distribution function » : donne $\mathbb{P}(X \leq x)$, où X est une variable aléatoire de loi `distrib`.

```
pnorm(0.5)
```

```
## [1] 0.6914625
```

```
pnorm(0.5,mean=2,sd=3)
```

```
## [1] 0.3085375
```

```
pnorm((0.5-2)/3)
```

```
## [1] 0.3085375
```

```
pbinom(5,10,.25)
```

```
## [1] 0.9802723
```

Densité

Forme générique : `d+distrib(x,...)`

`d` pour « density » : donne la densité pour une variable aléatoire continue et $\mathbb{P}(X = \mathbf{x})$ pour X une variable aléatoire discrète.

```
dnorm(0.5)
```

```
## [1] 0.3520653
```

```
dexp(3,1/8)
```

```
## [1] 0.08591116
```

```
dbinom(5,10,.25)
```

```
## [1] 0.0583992
```

```
dpois(4,2)
```

```
## [1] 0.09022352
```


Fractiles

Forme générique : `q+distrib(alpha,...)`

`q` pour « quantile » : donne la valeur de x définie par

$$\mathbb{P}(X \leq x) = \alpha,$$

où X est une variable aléatoire de loi `distrib`.

```
qnorm(0.95)
## [1] 1.644854

qt(0.4,df=28)
## [1] -0.2557675

qchisq(0.05,df=6)
## [1] 1.635383
```