

Statistique descriptive

Visualisation des données

Introduction

- Une fois qu'on a modélisé l'entrepôt, collecté les données, il faut **restituer** ces données (i.e. les **analyser** et **visualiser** ces analyses) afin d'offrir une aide à la décision.
- Aujourd'hui, nous allons voir les notions statistiques de base pour pouvoir faire de l'analyse de données.
- Nous verrons aussi comment visualiser ces informations statistiques.
 - En R → fichier disponible sur moodle + copier/coller
 - En Python → Notebook disponible sur Moodle

Introduction

La statistique générale se divise en 2 parties :

1. **La statistique descriptive** : description des unités statistiques qui composent une population, par des tableaux, graphiques, stat résumées (min, max, moyenne, écart type, médiane, ...), recherche de corrélations
2. **La statistique mathématique** : formulation de lois à partir de l'observation d'échantillons. Elle intervient dans les enquêtes, les sondages et permet de faire des prévisions grâce à l'extrapolations de résultats d'échantillons. Elle s'appuie sur la stat descriptive et sur le calcul de probabilités.

Terminologie

- Une **population** est un ensemble d'**unités statistiques**

Exemple : étude des pays de l'UE. 1 unité par pays.

- Un **échantillon** est un sous-ensemble de la population
- On peut découper une population en utilisant des **variables** (ou dimensions, ou caractéristiques)

Type de variable

- Variables qualitatives ou catégorielles.
 - Ex.: couleur des yeux, type d'engrais, méthode d'enseignement, catégorie grammaticale...
 - Deux types: *nominal* ou *ordinal*.
 - On appelle “niveaux” ou “modalités” les valeurs que peuvent prendre une variable qualitative.
- Variables quantitatives ou numériques
 - Elles peuvent être *discrètes* (à valeurs dans les entiers; exemple: comptage) ou *continues* (à valeurs dans les réels).
 - Deux types: *intervalle* (seule la différence à un sens, ex: heure) ou *ratio* (le rapport à un sens, ex: vitesse).
 - Ex.: taille, production en maïs, temps de réaction...
- Les procédures statistiques diffèrent en fonction des types des variables.

Classification du type de données

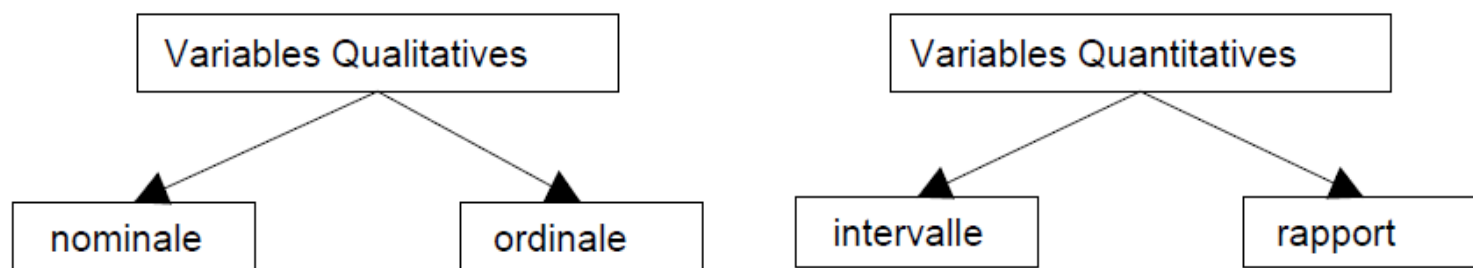


Fig. 1.1: Les deux grandes classes de variables.

Variables qualitatives

- Les variables qualitatives sont toutes les variables à valeur non numérique (e.g. bleu, blanc, rouge) et codées par les factor()).
- Les variables qualitatives peuvent être :
 - nominales sans ordre particulier : un simple nom.
 - ordonnées avec un ordre : un peu, beaucoup, passionnément, à la folie.

La règle est simple : s'il y a un ordre, vos graphiques doivent impérativement le respecter.



Exemple `library(MASS)`

- **Description**

- This data frame contains the responses of 237 Statistics I students at the University of Adelaide to a number of questions.

- **Usage**

- `survey`

- **Format**

The components of the data frame are:

- **Sex**: The sex of the student. (Factor with levels "Male" and "Female".)
- **Wr.Hnd**: span (distance from tip of thumb to tip of little finger of spread hand) of writing hand, in centimetres.
- **NW.Hnd**: span of non-writing hand.
- **W.Hnd**: writing hand of student. (Factor, with levels "Left" and "Right".)
- **Fold**: "Fold your arms! Which is on top?" (Factor, with levels "R on L", "L on R", "Neither".)
- **Pulse**: pulse rate of student (beats per minute).
- **Clap**: "Clap your hands! Which hand is on top?" (Factor, with levels "Right", "Left", "Neither".)
- **Exer**: how often the student exercises. (Factor, with levels "Freq" (frequently), "Some", "None".)
- **Smoke**: how much the student smokes. (Factor, levels "Heavy", "Regul" (regularly), "Occas" (occasionally), "Never".)
- **Height**: height of the student in centimetres.
- **M.I** whether the student expressed height in imperial (feet/inches) or metric (centimetres/metres) units. (Factor, levels "Metric", "Imperial".)
- **Age**: age of the student in years.

Exemple : summary (survey)



Sex	Wr.Hnd	NW.Hnd	W.Hnd	Fold	Pulse
Female:118	Min. :13.00	Min. :12.50	Left : 18	L on R : 99	Min. : 35.00
Male :118	1st Qu.:17.50	1st Qu.:17.50	Right:218	Neither: 18	1st Qu.: 66.00
NA's : 1	Median :18.50	Median :18.50	NA's : 1	R on L :120	Median : 72.50
	Mean :18.67	Mean :18.58			Mean : 74.15
	3rd Qu.:19.80	3rd Qu.:19.73			3rd Qu.: 80.00
	Max. :23.20	Max. :23.50			Max. :104.00
	NA's :1	NA's :1			NA's :45

Clap	Exer	Smoke	Height	M.I	Age
Left : 39	Freq:115	Heavy: 11	Min. :150.0	Imperial: 68	Min. :16.75
Neither: 50	None: 24	Never:189	1st Qu.:165.0	Metric :141	1st Qu.:17.67
Right :147	Some: 98	Occas: 19	Median :171.0	NA's : 28	Median :18.58
NA's : 1		Regul: 17	Mean :172.4		Mean :20.37
		NA's : 1	3rd Qu.:180.0		3rd Qu.:20.17
			Max. :200.0		Max. :73.00
			NA's :28		



Description des variable :

Hmisc::describe(survey)

```
> Hmisc::describe(survey)
```

```
survey |
```

```
12 Variables      237 Observations
```

```
-----
```

Sex

n	missing	distinct
236	1	2

Value	Female	Male
Frequency	118	118
Proportion	0.5	0.5

```
-----
```

Wr.Hnd

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
236	1	60	0.997	18.67	2.09	16.00	16.50	17.50	18.50	19.80	21.15	22.05

lowest : 13.0 14.0 15.0 15.4 15.5, highest: 22.5 22.8 23.0 23.1 23.2

```
-----
```

NW.Hnd

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
236	1	68	0.998	18.58	2.184	15.50	16.30	17.50	18.50	19.72	21.00	22.22

lowest : 12.5 13.0 13.3 13.5 15.0, highest: 22.7 23.0 23.2 23.3 23.5

```
-----
```

W.Hnd

n	missing	distinct
236	1	2

Value	Left	Right
Frequency	18	218
Proportion	0.076	0.924

```
-----
```

Fold

n	missing	distinct
237	0	3

Value	L on R	Neither	R on L
Frequency	99	18	120

Voir la structure du jeu de données: str()



```
library(MASS)
data("survey")
newsurvey <- surveydf <- na.omit(newsurvey) ##remove any NA values
str(df)
```

```
'data.frame':  168 obs. of  12 variables:
 $ Sex      : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 1 2 2 1 1 ...
 $ Wr.Hnd   : num  18.5 19.5 20 18 17.7 17 20 18.5 17 19.5 ...
 $ NW.Hnd   : num  18 20.5 20 17.7 17.7 17.3 19.5 18.5 17.2 20.2 ...
 $ W.Hnd    : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold     : Factor w/ 3 levels "L on R","Neither",...: 3 3 2 1 1 3 3 3 1 1 ...
 $ Pulse    : int   92 104 35 64 83 74 72 90 80 66 ...
 $ Clap     : Factor w/ 3 levels "Left","Neither",...: 1 1 3 3 3 3 3 3 3 2 ...
 $ Exer     : Factor w/ 3 levels "Freq","None",...: 3 2 3 3 1 1 3 3 1 3 ...
 $ Smoke    : Factor w/ 4 levels "Heavy","Never",...: 2 4 2 2 2 2 2 2 2 2 ...
 $ Height   : num   173 178 165 173 183 ...
 $ M.I      : Factor w/ 2 levels "Imperial","Metric": 2 1 2 1 1 2 2 2 1 2 ...
 $ Age      : num   18.2 17.6 23.7 21 18.8 ...
 - attr(*, "na.action")= 'omit' Named int [1:69] 3 4 12 13 15 16 19 25 26 29
...
..- attr(*, "names")= chr [1:69] "3" "4" "12" "13" ...
```

Distribution variable qualitative

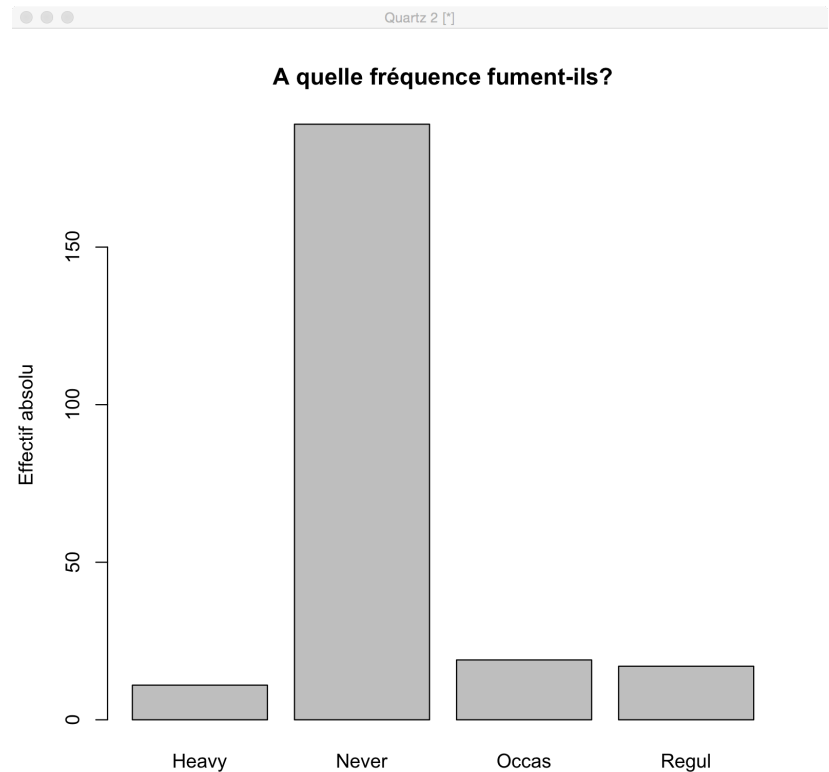
- La variable peut être décrite par la **distribution** des valeurs qu'elle prend : c'est à dire en associant un effectif à chaque valeur de la variable.
- Exemple distribution de la variable

Heavy	Never	Occas	Regul
11	189	19	17

Distribution variable qualitative - *graphique en barres*



Pour une variable discrète, une bonne représentation est le *graphique en barres* qui représente les effectifs ou les fréquences par modalité.





Fonction barplot

Le diagramme précédent s'obtient avec R par la commande :

```
barplot( table(survey$Smoke), main="A  
quelle fréquence fument-ils?", ylab=  
"Effectif absolu" )
```

Ce graphique n'est pas très lisible parce qu'on n'a pas l'information précise (pas de chiffre, échelle peu précise, ...)



Enrichissement d'un graphique

```
res <- table(survey$Smoke)/nrow(survey)*100
```

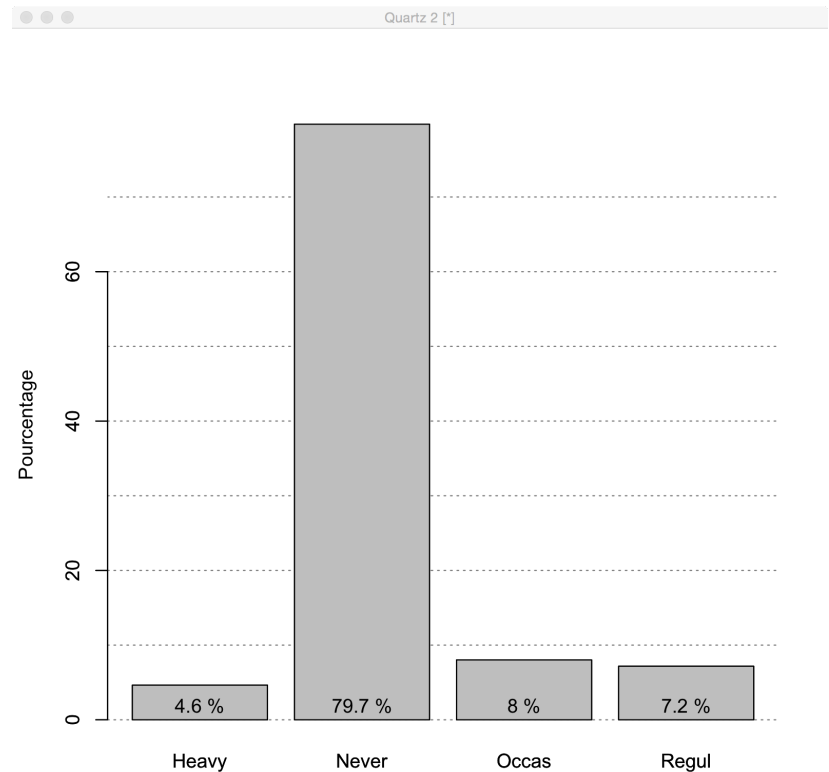
```
Heavy    Never    Occas    Regul  
4.641350 79.746835 8.016878 7.172996
```

```
coord.x <- barplot( res )
```

```
abline( h=seq(0,300,10), lty=3, col="gray50" )
```

```
barplot( res, ylab="Pourcentage", add=T )
```

```
text(  
coord.x,  
0 + strheight("B"),  
paste( round(res,1), "%" )  
)
```





2 variables sur le même graphique

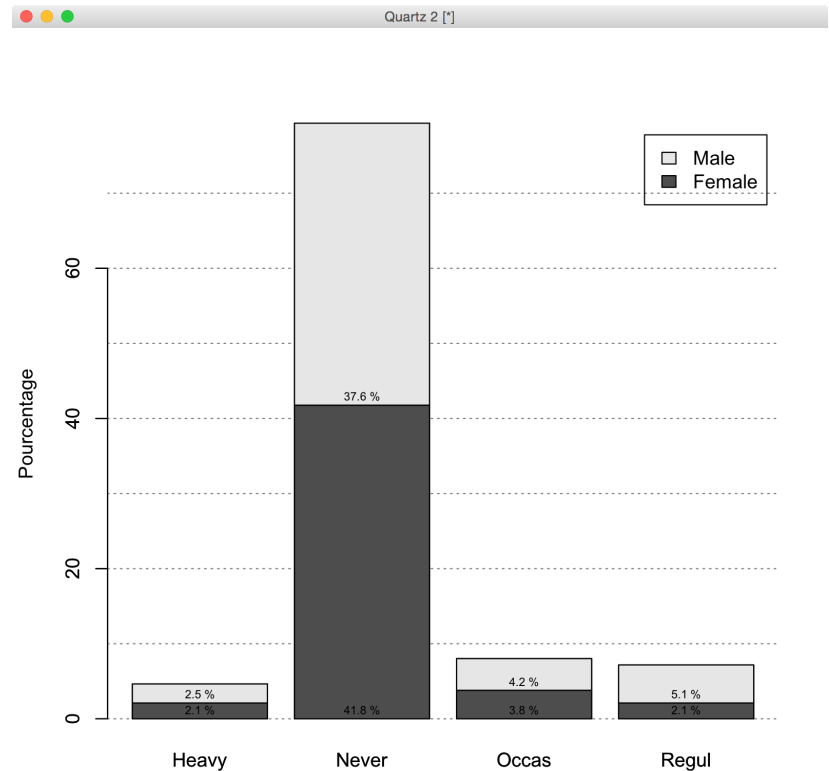
```
res <-  
table(survey$Sex,survey$Smoke)/nrow(survey)*100
```

```
coord.x <- barplot( res )
```

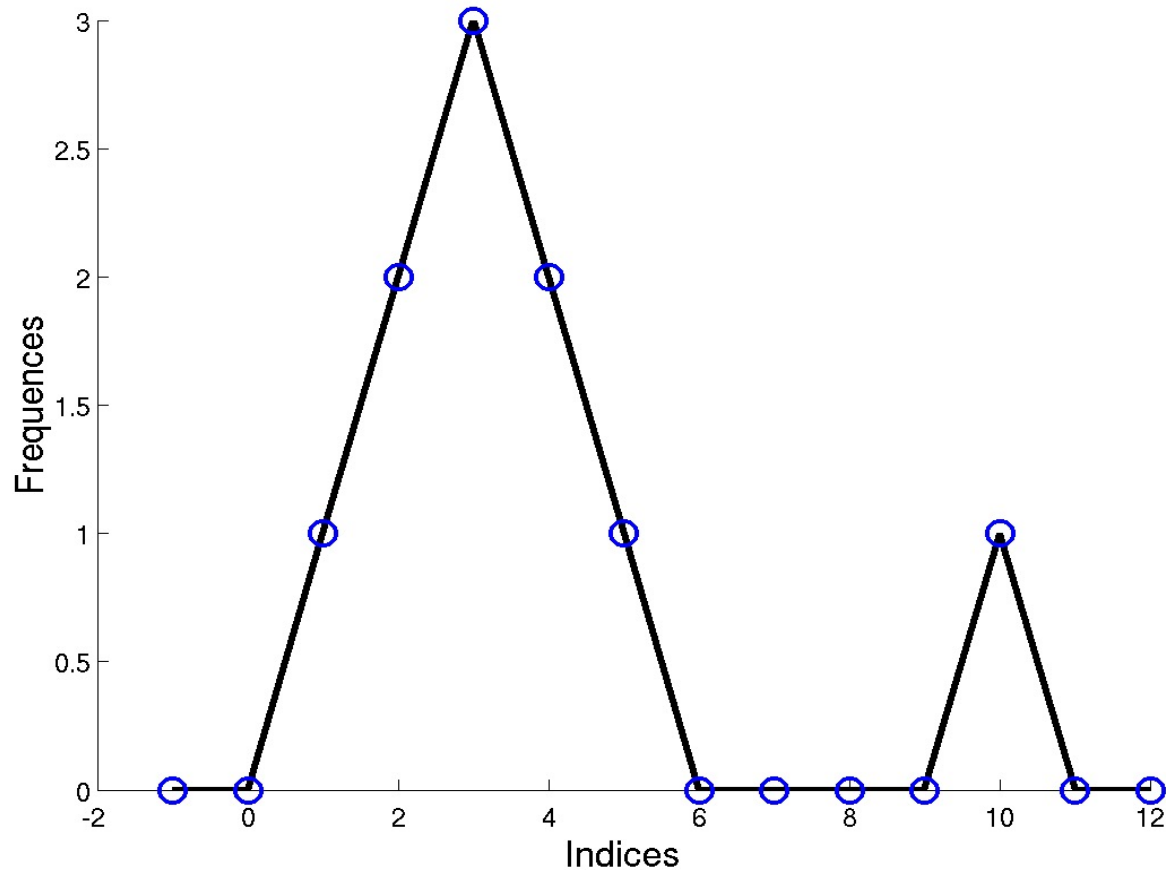
```
abline( h=seq(0,300,10), lty=3, col="gray50" )
```

```
barplot( res, ylab="Pourcentage", legend.text=T, add=T  
)
```

```
text(  
c( coord.x, coord.x),  
c(  
rep(0,ncol(res)) + strheight("B",cex=0.6),  
res[1,]+ strheight("B",cex=0.6) ),  
paste( round(c(res[1,],res[2,]),1), "%" ),  
cex=0.6  
)
```



Distribution -



Préférentiellement pour des variables discrètes

Les graphiques en secteurs/Diagramme circulaire



- Utiles quand on veut représenter la relation entre une partie et un tout.

- En R : fonction `pie`

```
res <-  
table(survey$Smoke)/nr  
ow(survey)*100
```

```
pie(res,col=rainbow(le  
ngth(res)))
```

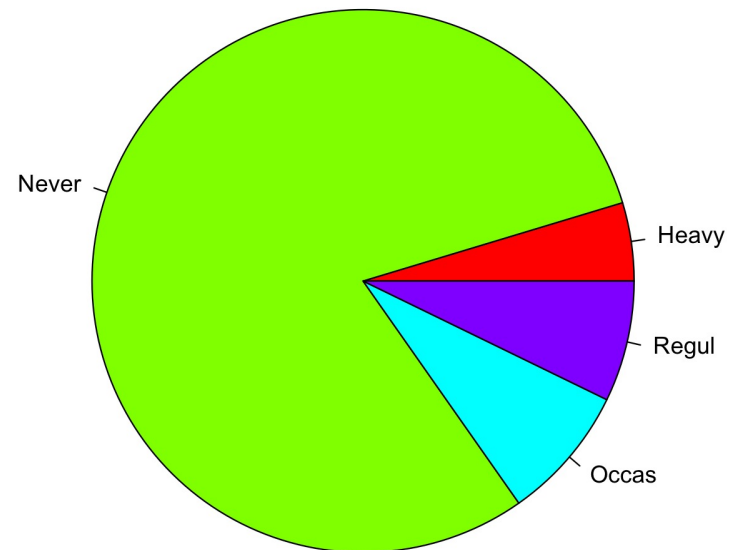




Tableau de contingence

- Un **tableau de contingence** est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères.
- Avec R, on peut utiliser `table` avec 2 variables :

```
table(survey$Sex, survey$Smoke)
```

	Heavy	Never	Occas	Regul
Female	5	99	9	5
Male	6	89	10	12



Tableau de contingence avec probabilité

```
> mos <- table(survey$Sex,survey$Smoke)
> mos.proba <- mos/rowSums(mos)
> mos.proba
```

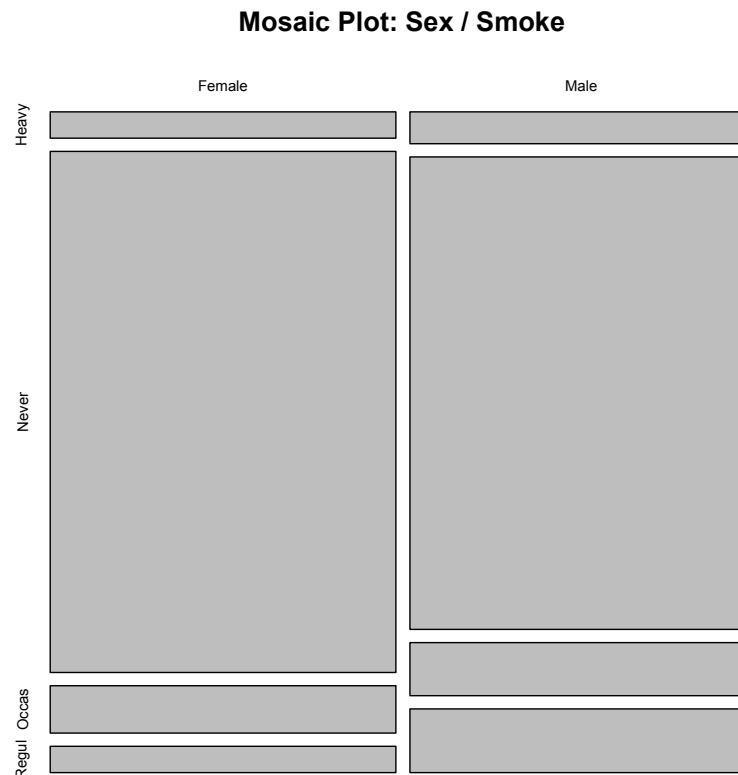
	Heavy	Never	Occas	Regul
Female	0.04237288	0.83898305	0.07627119	0.04237288
Male	0.05128205	0.76068376	0.08547009	0.10256410



Mosaic plot

Pour visualiser les écarts par rapport à l'hypothèse d'indépendance, on peut utiliser la fonction `mosaicplot()`

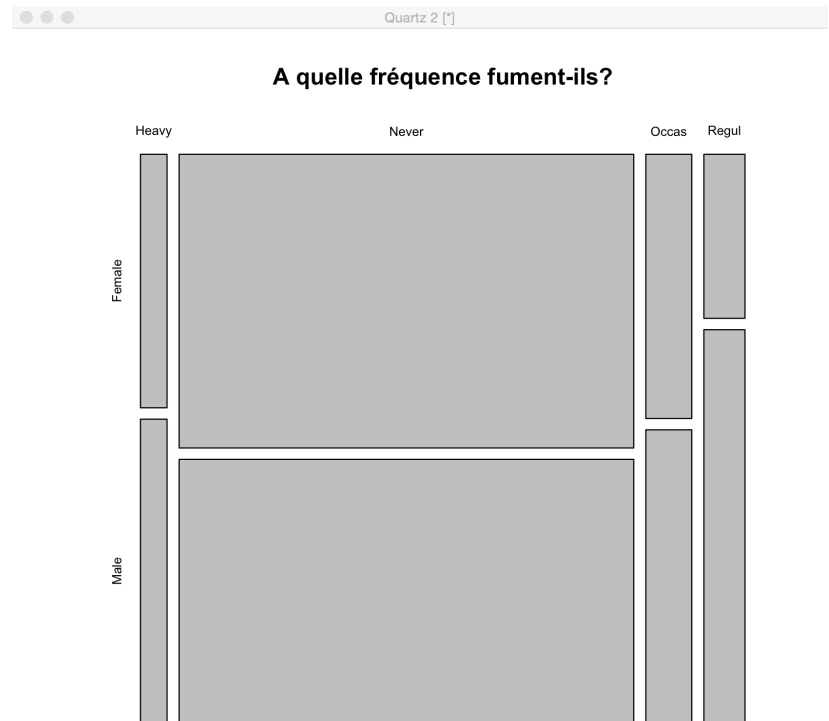
```
> mosaicplot(mos, main = "Mosaic Plot: Sex / Smoke")
```




Graphique « mosaïque »

Pour représenter graphiquement un tableau de contingence, on peut utiliser une « mosaïque »

```
mosaicplot(table(  
  survey$Smoke,  
  survey$Sex),  
  main="A quelle  
  fréquence fument-  
  ils?" )
```



Fonction

 est un très bon environnement pour produire de façon reproductible des graphiques statistiques de haute qualité.

Important Plots	Important Functions
Scatter Plot	<code>geom_point()</code> , <code>geom_smooth()</code> , <code>stat_smooth()</code>
Bar Chart	<code>geom_bar()</code> , <code>geom_errorbar()</code>
Histogram	<code>geom_histogram()</code> , <code>stat_bin()</code> , <code>position_identity()</code> , <code>position_stack()</code> , <code>position_dodge()</code>
Box Plot	<code>geom_boxplot()</code> , <code>stat_boxplot()</code> , <code>stat_summary()</code>
Line Plot	<code>geom_line()</code> , <code>geom_step()</code> , <code>geom_path()</code> , <code>geom_errorbar()</code>
Pie Chart	<code>coord_polar()</code>

Exemple d'utilisation pour sauvegarder un graphique dans un fichier au format PDF :

```
pdf("monfichier.pdf")
```

```
plot(0)
```

```
dev.off()
```


Variables quantitatives

Variables quantitatives / numériques

- On distingue :
 - les variables **quantitatives discrètes**, ne pouvant prendre qu'un nombre fini de valeurs (par exemple le nombre de jambes d'un individu).
 - les variables **quantitatives continues**, pouvant prendre un nombre infini de valeurs (par exemple la taille d'un individu).

Cette distinction est un peu artificielle puisque les variables continues stricto sensu n'existent pas à cause de la précision limitée des instruments de mesure

Statistiques pour « résumer »

- Résumé de la tendance centrale : moyenne, médiane
- Résumé de la dispersion : intervalle de variation, intervalle interquartile
- Certaines statistiques résument la dispersion ET prennent en compte la tendance centrale : écart-type, variance, coefficient de variation

Tendance centrale

- Le **mode** : c'est la valeur la plus fréquente d'une série.

Par exemple, pour la variable Smoke, le mode est « Never »

- La **moyenne arithmétique** :

La moyenne s'applique à des variables quantitatives

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

- La **médiane** : partage la population en 2 groupes de même taille, ceux qui ont une valeur inférieure/supérieure à la médiane

La médiane n'a de sens que si l'on peut ordonner les valeurs

Moyenne ou Médiane ?

La moyenne n'est pas toujours un bon indicateur :

Exemple :

Salaires de l'entreprise 1 : {1,1,1,2,2,3,3,3,5,10,15}

Moyenne 4.18, Médiane 3

Salaires de l'entreprise 2 : {1,1,1,1,1,2,2,2,3,5,100}

Moyenne 10.82, Médiane 2

Dispersion

- Minimum et maximum
- **Intervalle de variation** (ou étendue)
= max – min
- **Intervalle interquartile**

Les quartiles sont les 3 valeurs Q1, Q2, Q3 qui partagent la population en 4 sous-populations de même taille (donc Q2 est la médiane)

L'intervalle Interquartile **IIQ = Q3 – Q1**

Variance et écart type

- **Variance** : Il exprime la dispersion de n valeurs par rapport à leur moyenne m .
 - Une variance de 0 veut dire que toutes les valeurs sont identiques
 - Si elle est faible, les valeurs sont assez concentrées autour de la moyenne
 - Si elle est élevée, les valeurs sont dispersées autour de la moyenne

$$V(X) = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2.$$

- **Ecart-type** (ou, en anglais, standard deviation) : c'est la racine carrée de la variance



Statistiques

Fonction	Opération
sum(x)	somme
mean(x)	moyenne
var(x)	variance
sd(x)	écart-type
min(x)	minimum
max(x)	maximum
median(x)	médiane
quantile(x)	quantiles à 0, 25%,50%,75% et 100%
length(x)	Nombre d'observations pour la variable

Jeu de données : Iris de Fisher

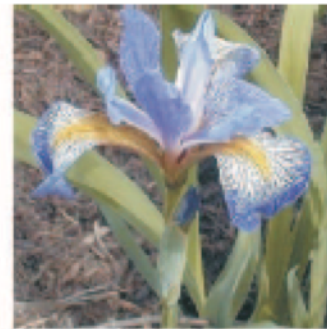
- SLong: Longueur des sépales
- Slarg: Largeur des sépales
- PLong: Longueur des pétales
- Plarg: Largeur des pétales
- Indic: Espèces (0: setosa, 1: versicolor, 2: virginica)



Setosa



Versicolor



Virginica



Présentation des données – statistiques descriptives

```
> summary(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

Présentation des données – statistiques descriptives



```
Entrée [7]: 1 print(dataset.describe())
```

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

```
Entrée [8]: 1 print(dataset.groupby('class').size())
```

```
2  
  
class  
Iris-setosa      50  
Iris-versicolor  50  
Iris-virginica   50  
dtype: int64
```

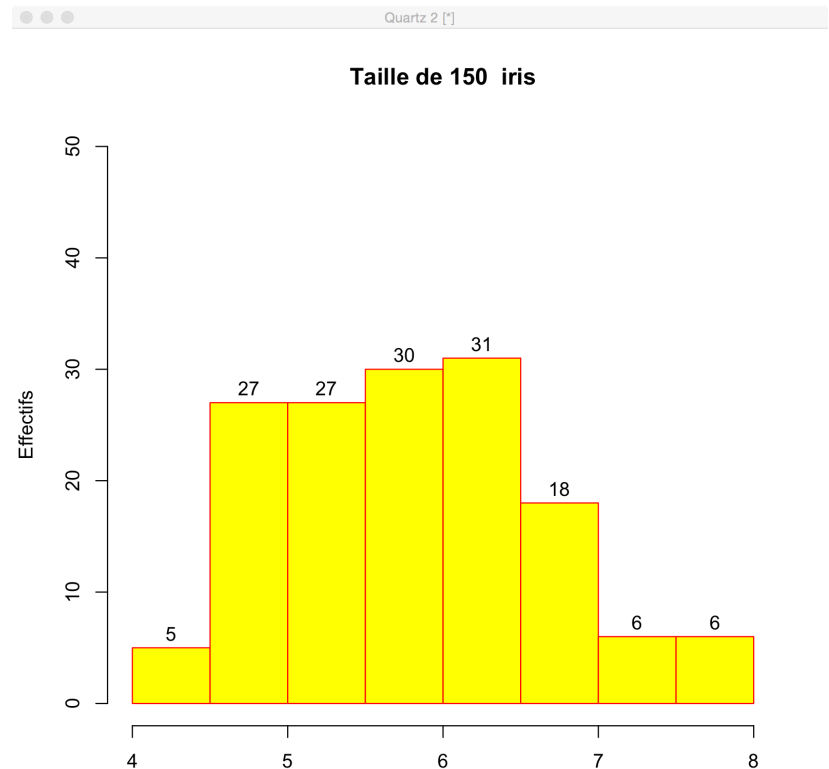
Visualisation d'une variable continue

- Pour avoir un aperçu d'une variable continue il est important de regarder quelle est la distribution des valeurs qu'elle peut prendre.
- Le graphique traditionnellement utilisé est un histogramme. Il donne pour des intervalles les effectifs ou la fréquence d'observation de ces valeurs.
- Simple à lire, la discrétisation des variables continues n'est pas sans poser des problèmes. En effet le graphique dépend en grande partie du nombre d'intervalles et/ou de la manière de les calculer.



Histogramme pour Sepal.length

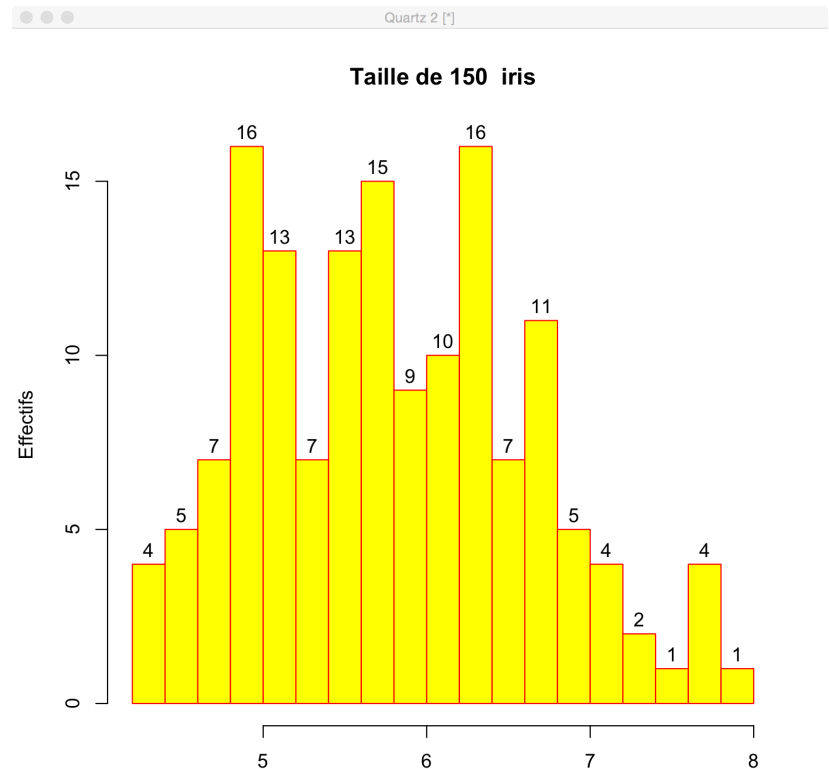
```
hist(iris[,1],  
col = "yellow", border  
= "red",  
main = paste("Taille  
de", nrow(iris), "  
iris"),  
ylab = "Effectifs",  
ylim = c(0, 50),  
labels = TRUE)
```





On fixe le nombre d'intervalles

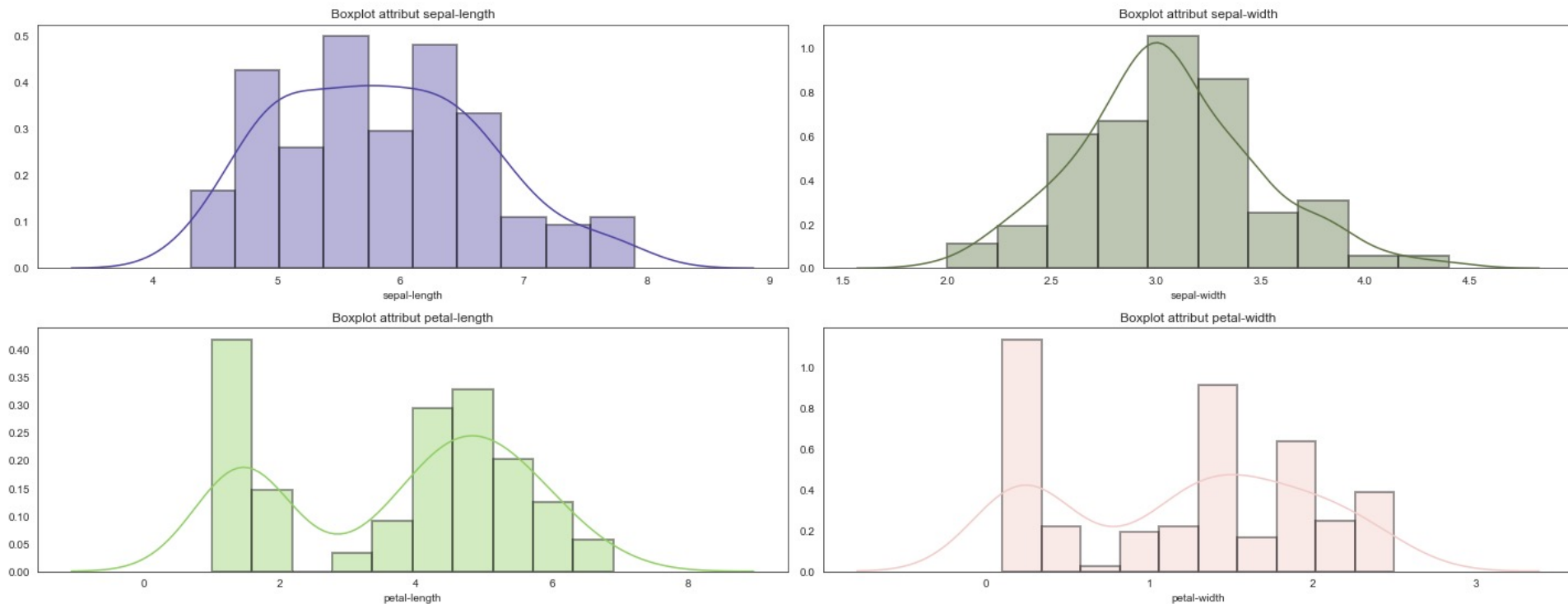
```
hist(iris[,1],  
     breaks=20,  
     col = "yellow", border  
     = "red",  
     main = paste("Taille  
de", nrow(iris), "  
iris"),  
     ylab = "Effectifs",  
     labels = TRUE)
```



Visualisation des variables



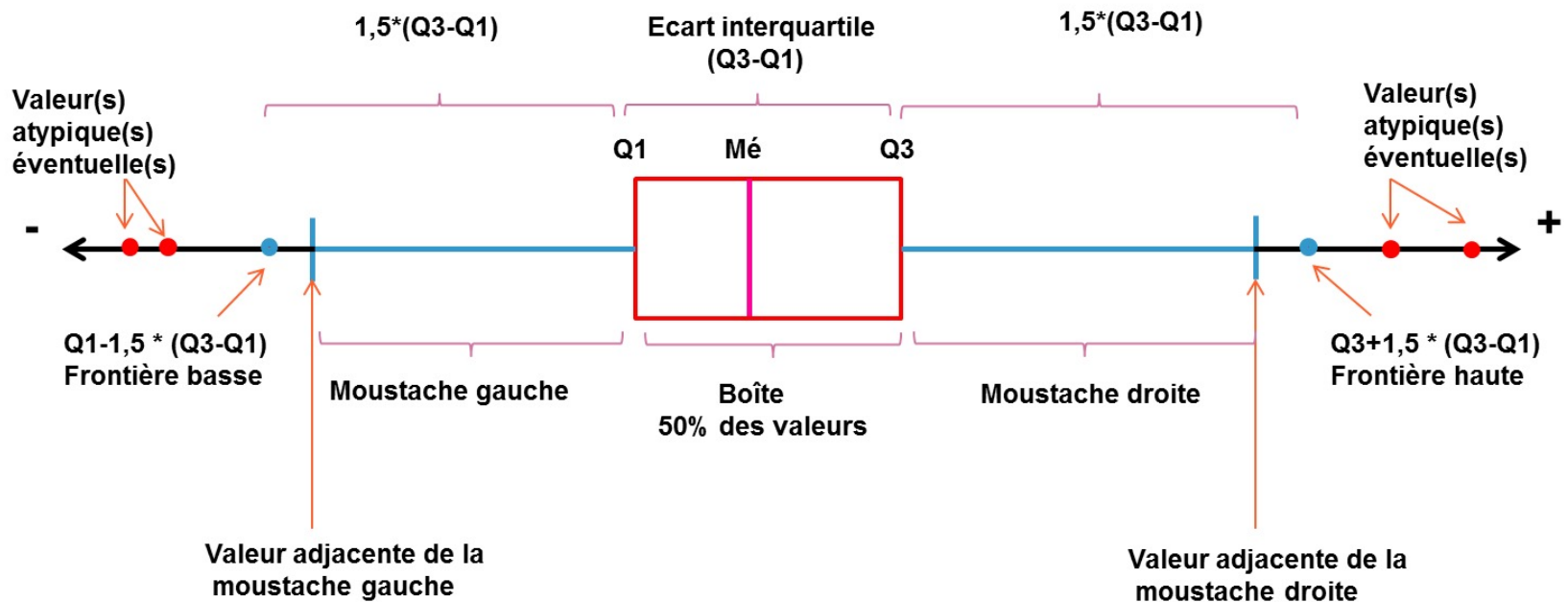
- Le problème des histogrammes est que le choix du découpage en intervalles est assez arbitraire.



Quartiles

- Quartile à 25% = valeur de la variable qui délimite 25% des premières données
- de la série statistique classée par ordre croissant
- Quartile à 75% = valeur de la variable qui sépare 75% des premières données
- de la série statistique classée par ordre croissant
- Intervalle inter-quartile =
- [Quartile à 25% Quartile à 75%]

Boîte à moustaches



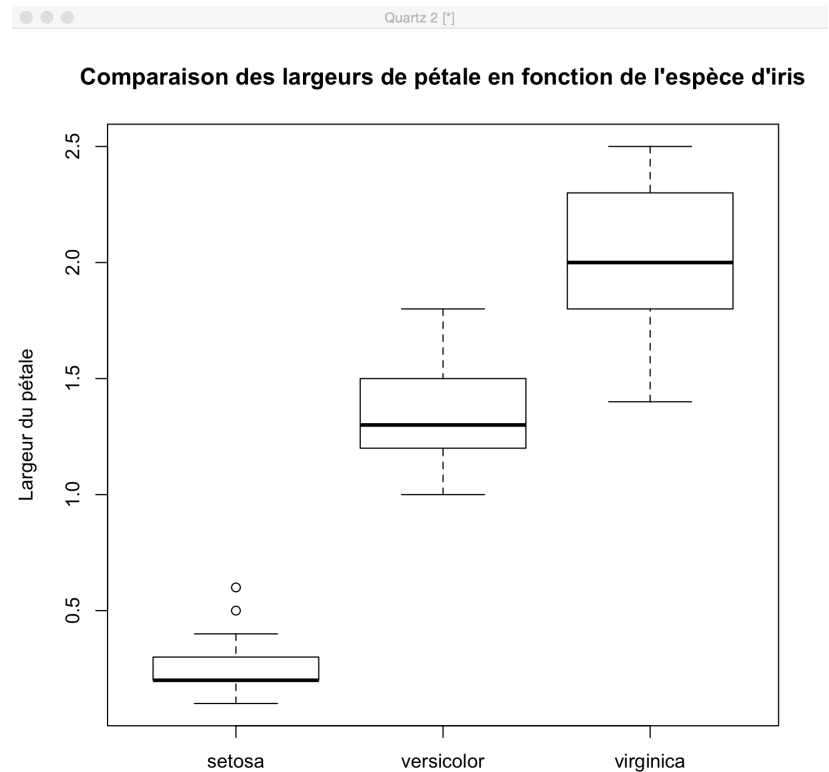
Boîte à Moustache

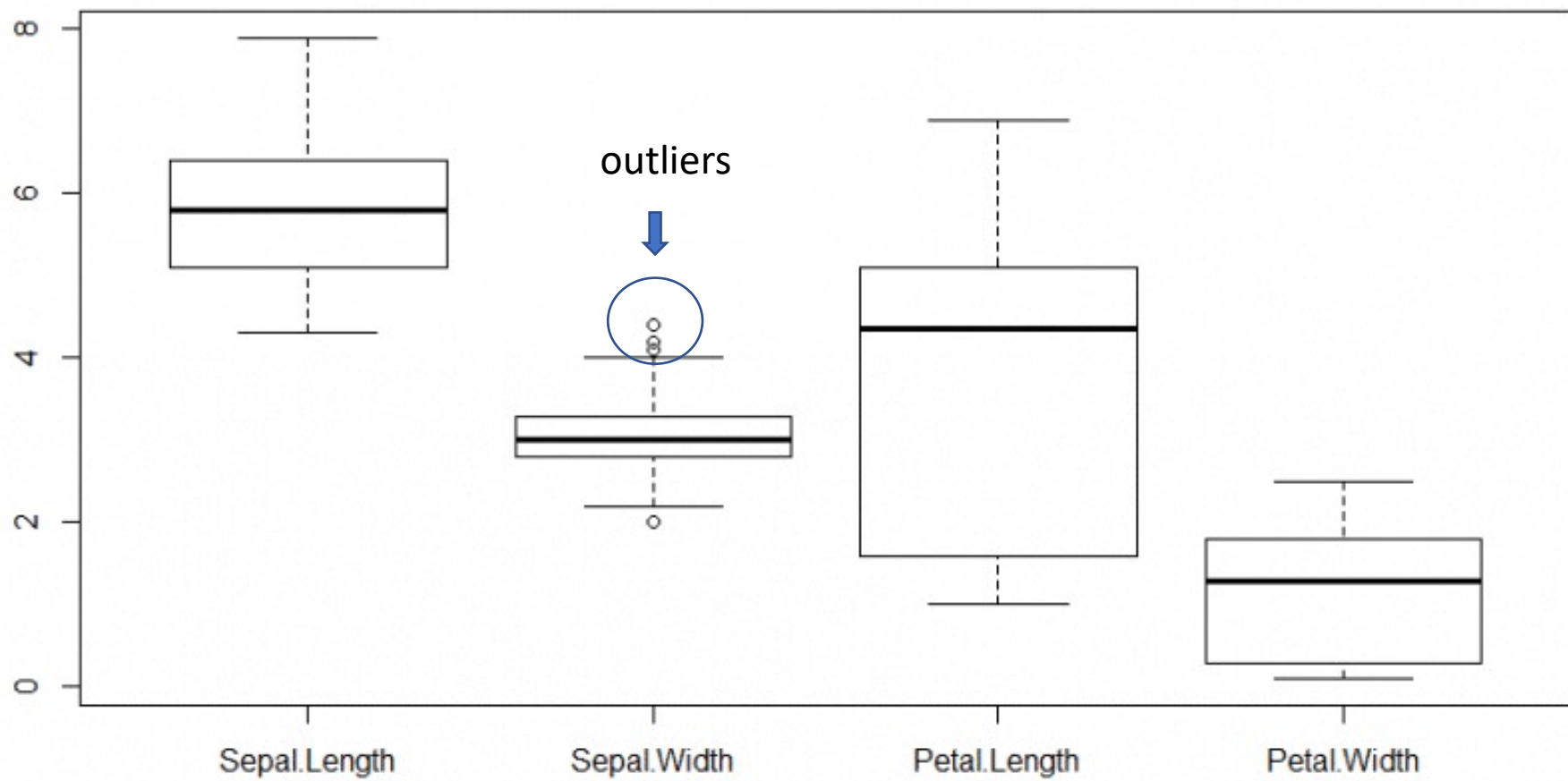
- Pour avoir un aperçu de la distribution de variables continues, ce type de graphique est très utile. Il permet d'un coup d'œil d'avoir une idée sur les quantiles, la moyenne et les points extrêmes.
- En R : fonction `boxplot`



boxplot

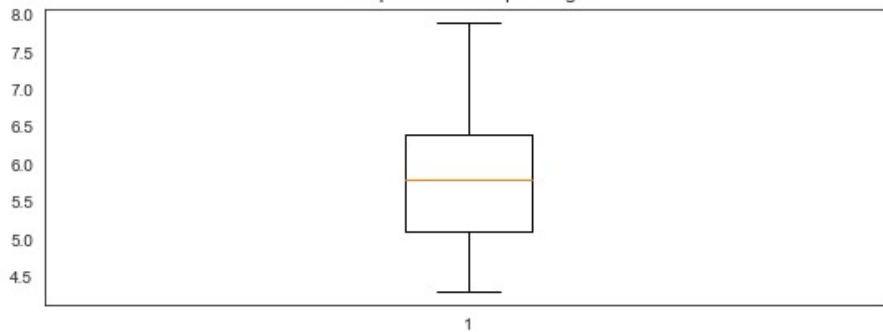
```
data(iris)
boxplot(
  iris[, "Petal.Width"]
  ~iris[, "Species"],
  ylab="Largeur du
  pétale",
  main="Comparaison
  des largeurs de
  pétale en fonction
  de l'espèce d'iris")
```



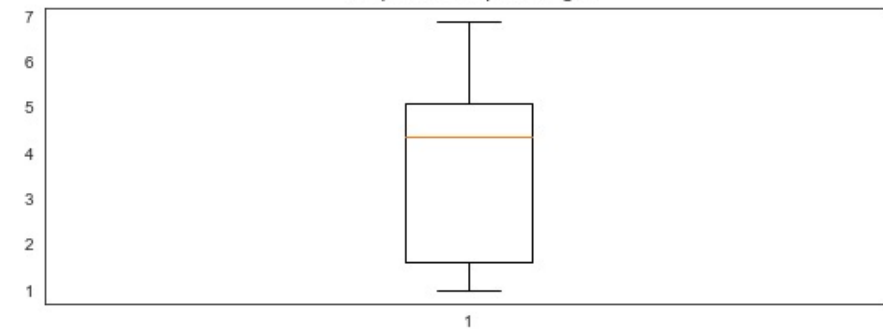




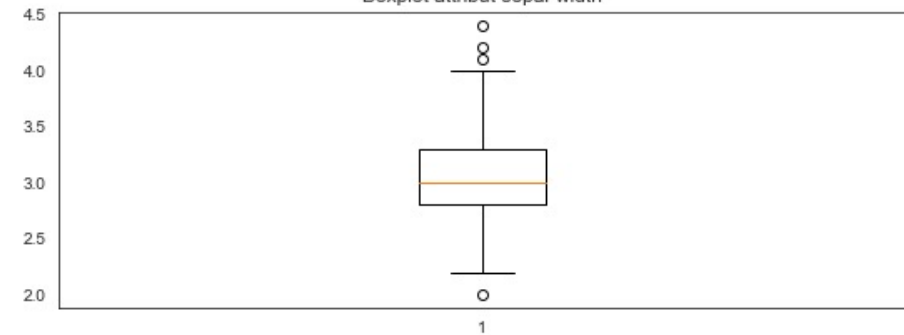
Boxplot attribut sepal-length



Boxplot attribut petal-length



Boxplot attribut sepal-width



Boxplot attribut petal-width

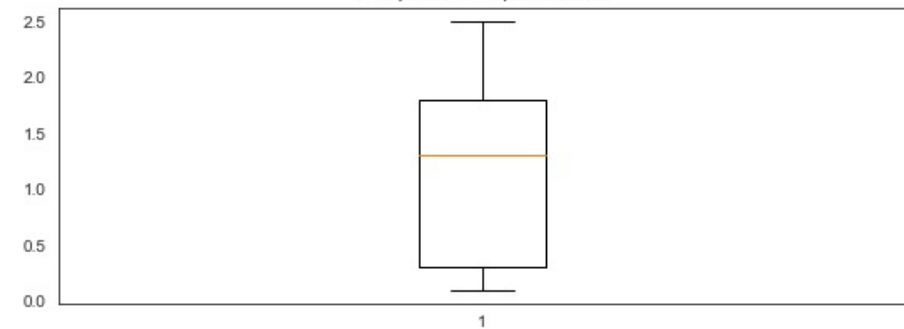
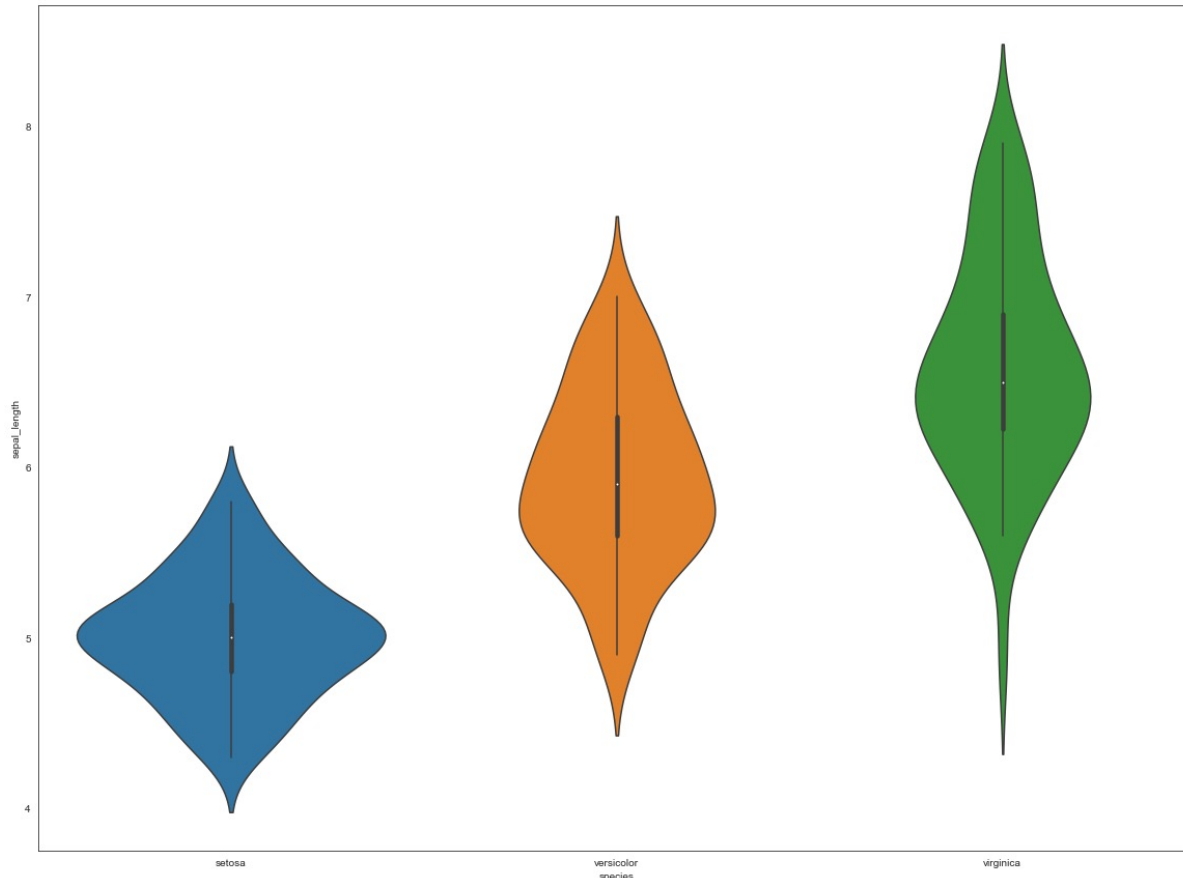


Diagramme en violon



- Les diagrammes en violon essaient de combiner les avantages des boîtes à moustaches et des estimateurs de la densité locale



Corrélation

- Il est primordial de regarder la corrélation des variables entre elles.
- La corrélation consiste à vérifier si les valeurs d'une des variables varient linéairement selon les valeurs prises par une autre variable.
- Attention corrélation ne signifie pas causalité.

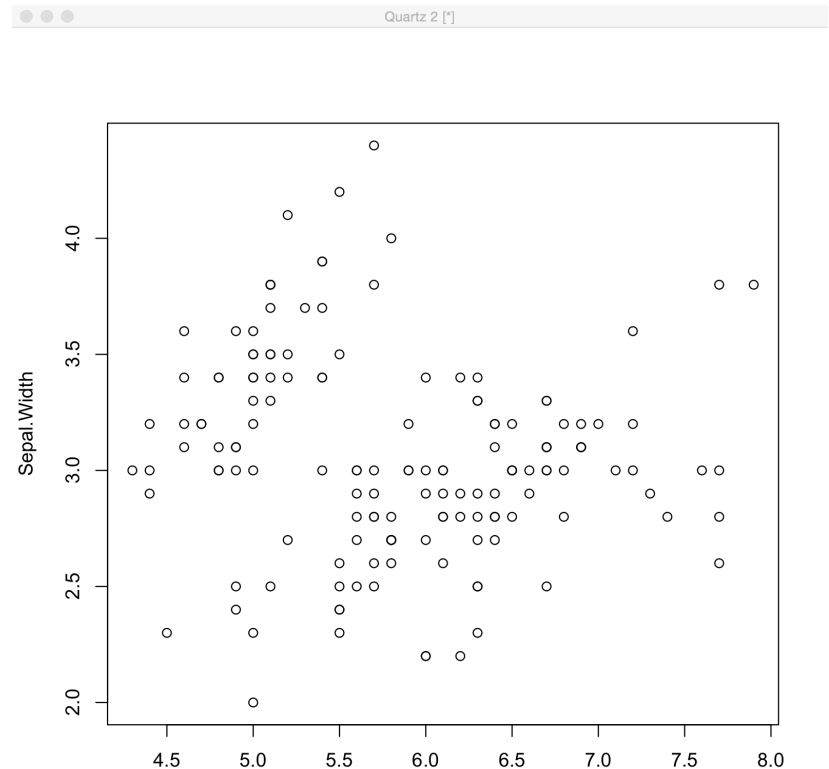


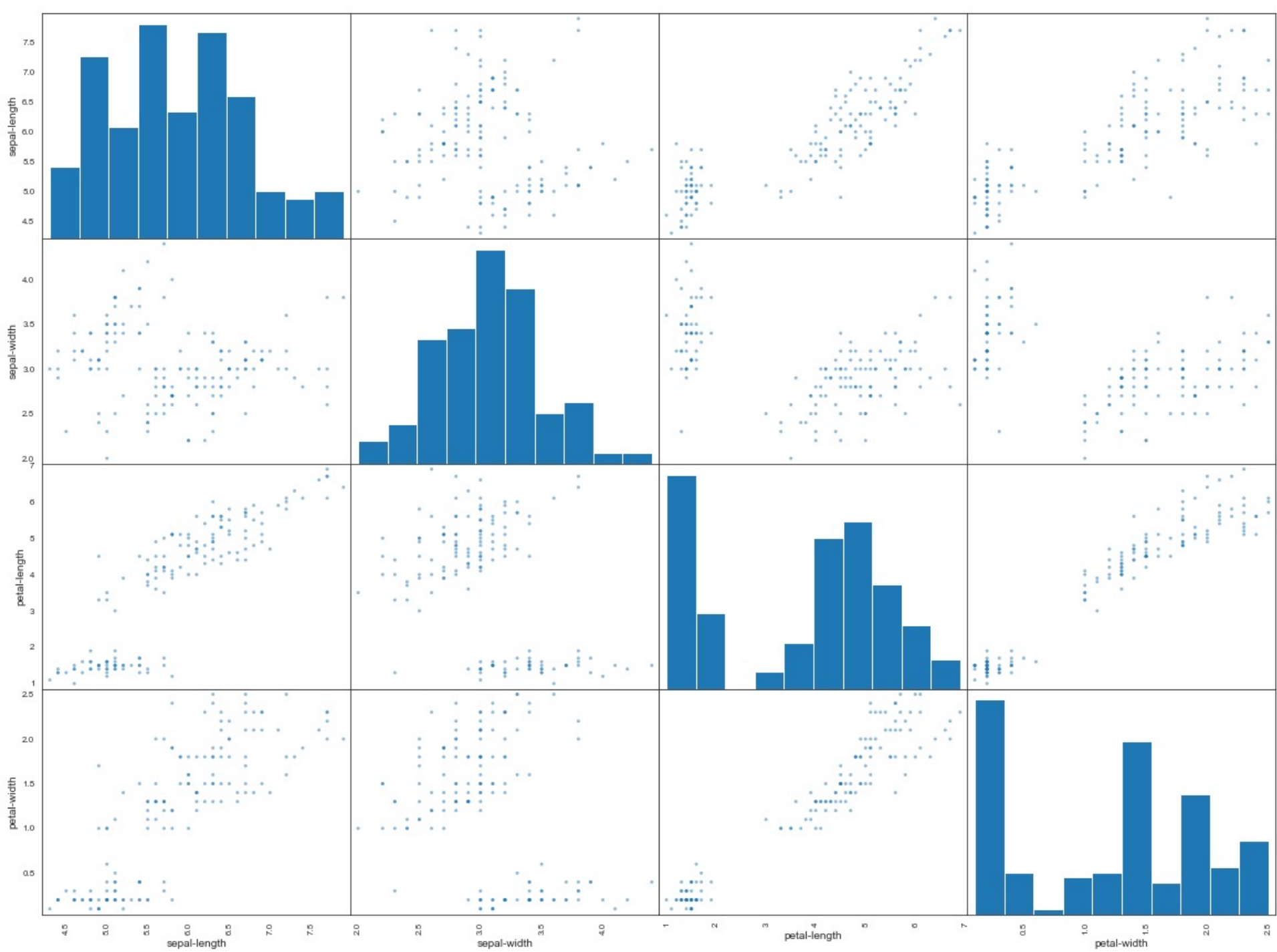
Recherche (graphique) de corrélation

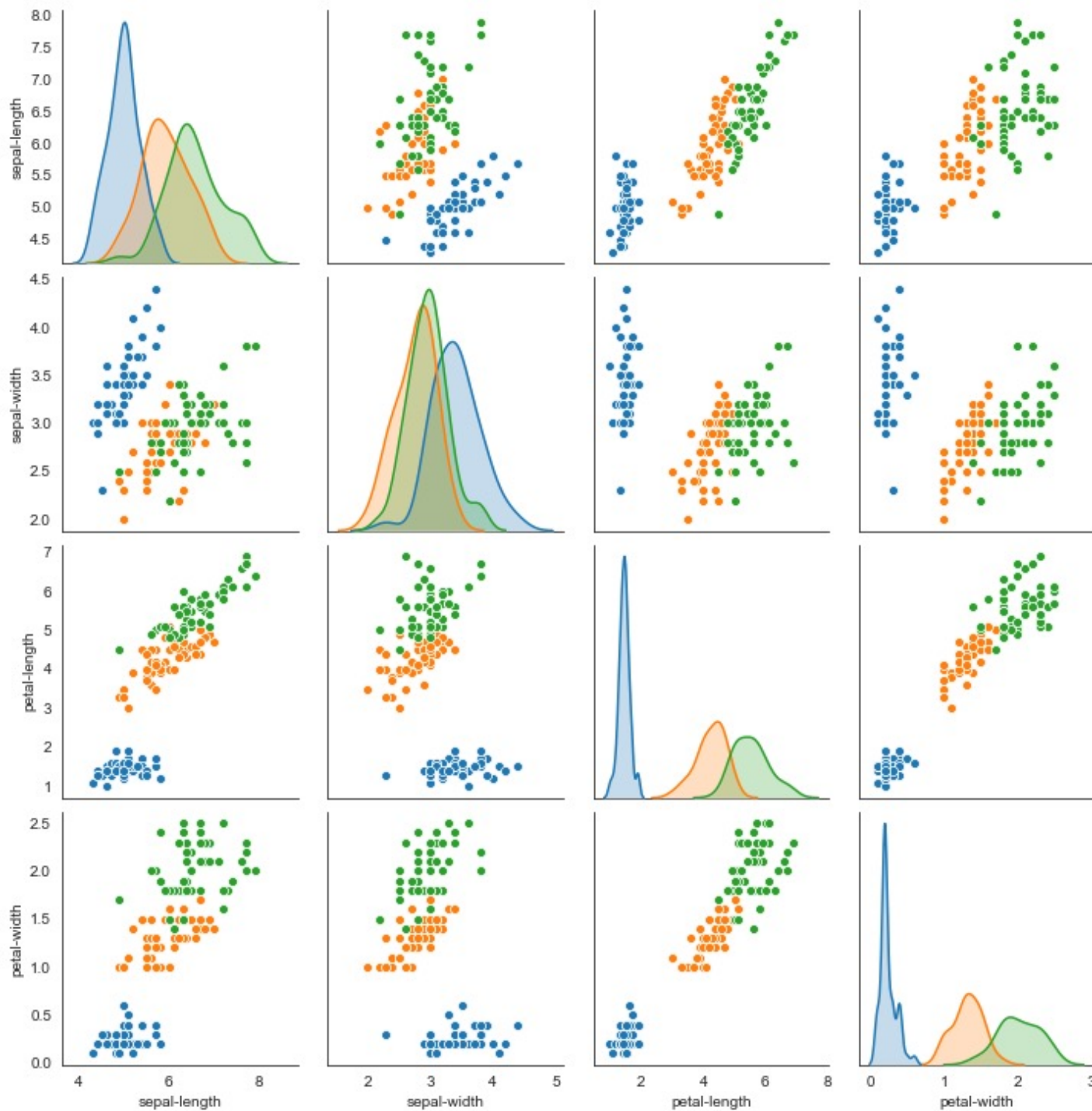
Exemple : en abscisse la longueur de sépale et en ordonnée la largeur.

```
plot(iris[,1:2])
```

Pas de corrélation visible
(d'ailleurs, $\rho = -0.1175698$)

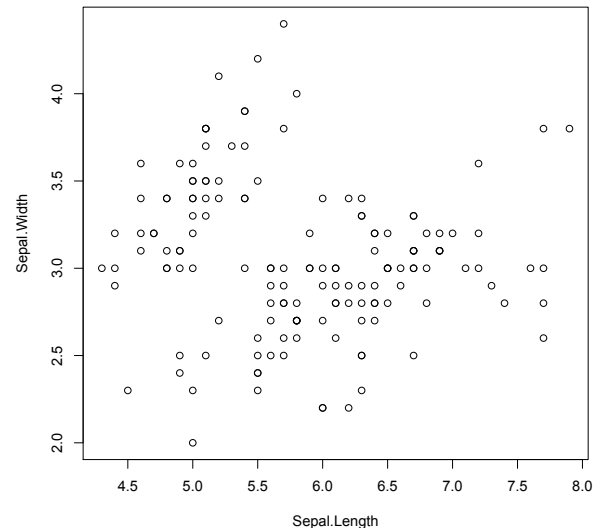






Visualisation de la densité des points

- Comme on peut le voir sur les graphiques permettant de visualiser des nuages de points, la lecture est difficile car les points se superposent.
- Il existe plusieurs méthodes pour visualiser la **densité** des points. on peut par exemple la représenter par des surfaces de taille variable (mais parfois, l'œil humain perçoit mal les légères différences de surface).

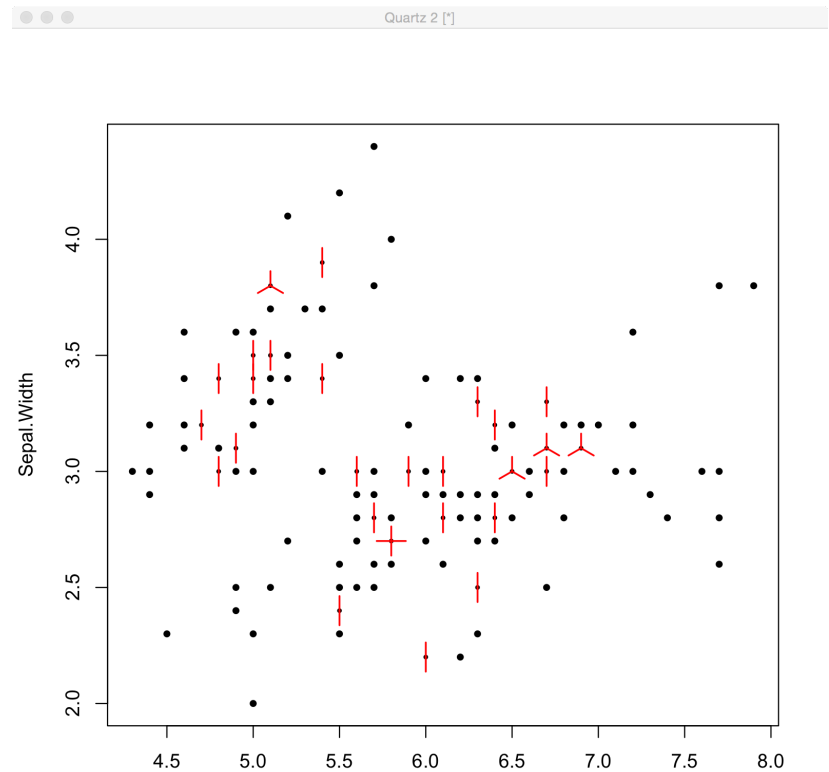


Sunflowerplot

- Le principe de la fonction est d'ajouter un rayon à un point central dès que le nombre d'observations atteint un certain seuil. Ainsi plus le nombre de rayons est élevé plus il y a d'observations.
- On peut modifier la différence à partir de laquelle les points sont considérés comme distincts à l'aide de l'argument **digits** qui précise l'arrondi à utiliser.

Sunflowerplot

```
sunflowerplot(  
iris[, 1:2])
```



covariance

- La covariance σ_{xy} permet d'étudier les variations simultanées de deux variables quantitatives par rapport à leur moyenne respective.
- La covariance est utilisée pour étudier les coefficients de corrélation

covariance

Interprétation de la valeur de la covariance :

- Plus la covariance est élevée, plus la relation entre les deux variables est forte
- Si la covariance vaut zéro, cela signifie qu'il n'y pas de relation linéaire entre les variables
- Si la covariance est négative, cela signifie que les deux variables varient en sens inverse

covariance

$$\sigma_{xy} = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

X et Y représentent les moyennes des 2 variables, n est le nombre d'unités statistiques.

Coefficient de corrélation

- La covariance n'est pas normalisée, on a donc des problèmes pour la comparer avec d'autres mesures.
- Le coefficient de corrélation ρ est la standardisation du coefficient de covariance σ_{xy} . Cette standardisation se fait en divisant la covariance par l'écart-type de chacune des variables.
- Donc $\rho = \sigma_{xy} / (\sigma_x * \sigma_y)$
- La mise en évidence d'une corrélation entre deux facteurs ne démontre pas l'existence d'une relation de causalité entre ces facteurs.

Langage R

- `cov(x, y)` : **covariance** des 2 vecteurs
- `cor(x, y)` : **coefficient de corrélation** entre les 2 vecteurs (vaut $\text{cov}(x,y) / (\text{sd}(x) * \text{sd}(y))$).

Exemple :

```
cov(iris[, "Petal.Width"], iris[, "Petal.Length"] )  
[1] 1.295609
```

```
cor(iris[, "Petal.Width"], iris[, "Petal.Length"] )  
[1] 0.9628654
```

→ Coefficient élevé

Recherche de corrélation : en calculant

- **Coefficient de corrélation de Spearman**

$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Si la valeur est de 1, on dit qu'il y a une corrélation positive entre deux variables. Cela signifie que lorsqu'une variable augmente, l'autre variable augmente également.
- Si la valeur est -1, on dit qu'il y a une corrélation négative entre deux variables. Cela signifie que lorsqu'une variable augmente, l'autre diminue.
- Si la valeur est 0, il n'y a pas de corrélation entre deux variables. Cela signifie que les variables évoluent de manière aléatoire l'une par rapport à l'autre.
- On dit qu'il y a corrélation si le coefficient est ≥ 0.7 ou ≤ -0.7 ,

