Apprentissage non supervisé
Chapitre 0 : Introduction

Master "Machine Learning for Data Science", Paris V

Allou Samé
allou.same@ifsttar.fr

2017/2018

Objectifs de la classification non supervisée

Objectifs

- Obtenir une représentation simplifiée d'un ensemble de données (analyse exploratoire)
- Organiser les données en groupes (ou classes) homogènes
- Réduire les données (chaque groupe étant remplacé par un représentant)

Origines, histoire

- Classification des genres naturels en trois groupes : animaux, plantes, minéraux
- Classification (nomenclature) des espèces animales et végétales de Von Carl Linné (17^{e} siècle)

Applications courantes

- Informatique : webmining, regroupement de pages web, compression de données
- Traitement d'image : quantification vectorielle, segmentation en zones homogènes
- Ingénierie : reconnaissance de la parole
- Neurosciences : classification des potentiels d'action ("spike sorting")
- Médecine & Bio-informatique : classification des maladies, regroupement de gènes
- Astronomie, géographie : regroupement d'étoiles, de planètes, partitionnement de régions et de villes
- Marketing : segmentation de la clientèle en classes homogènes
- Sociologie : typologie des cultures, des langues, analyse des réseaux sociaux

Classification automatique Vs. Classement

Classification automatique

- organisation des données en groupes homogènes (les groupes sont inconnus)
- apprentissage non supervisé

Classement

- rangement des données dans des groupes connus à l'avance
- apprentissage supervisé

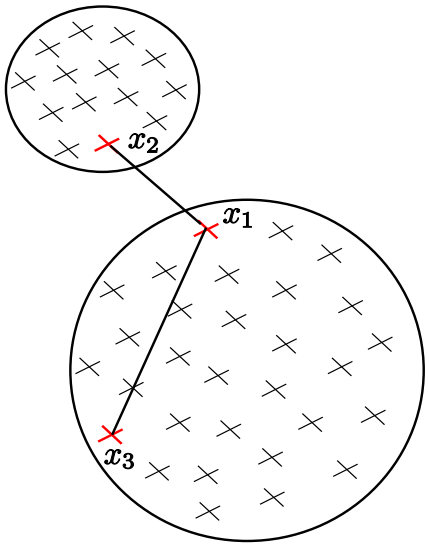
Terminologie anglais-français

français	anglais
classification	clustering, cluster analysis
classement	classification

Notion de classe

Plusieurs définitions ont été proposées :

- Les classes sont des groupes d'objets homogènes : les données appartenant à un même classe se ressemblent
- Les classes sont des groupes d'objets bien séparés : les données provenant de classes différentes sont dissemblables
- Une classe est un aggrégat de points de l'espace tel que la distance entre deux points d'une même classe est plus petite que la distance entre deux points de classes différentes



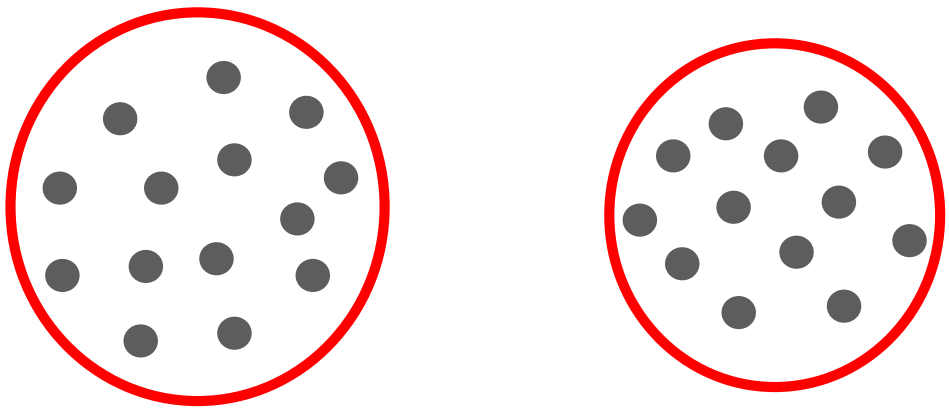
contre exemple

Exemple de configurations de classes

2 classes bien séparées

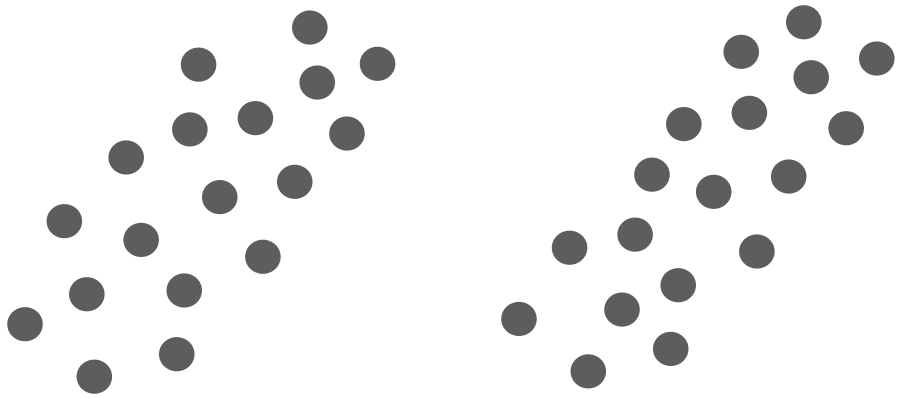
Exemple de configuration de classes

2 classes bien séparées



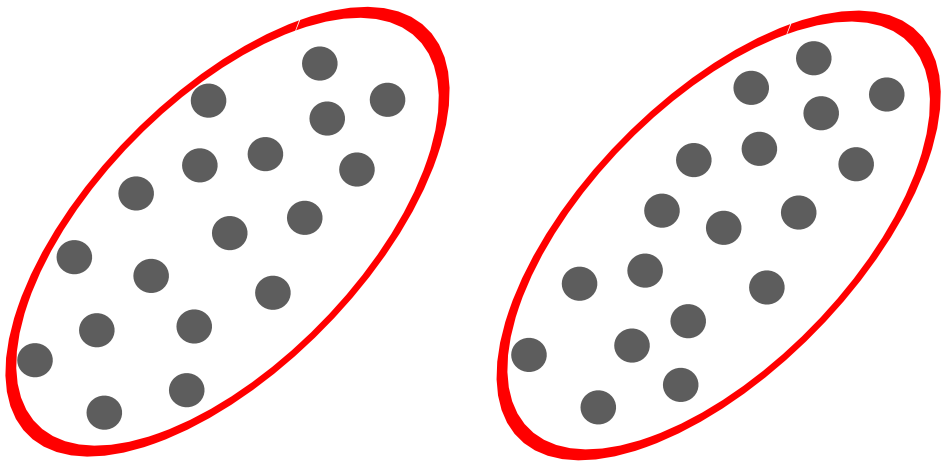
Exemple de configuration de classes

2 classes allongées



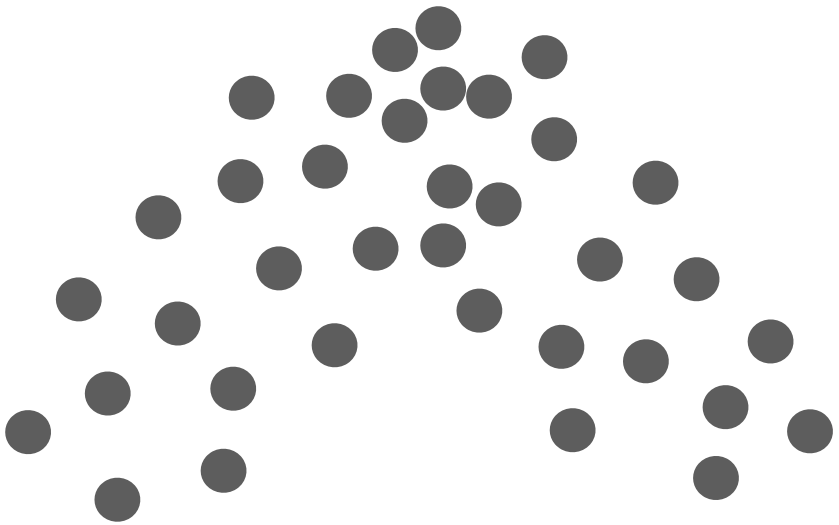
Exemple de configuration de classes

2 classes allongées



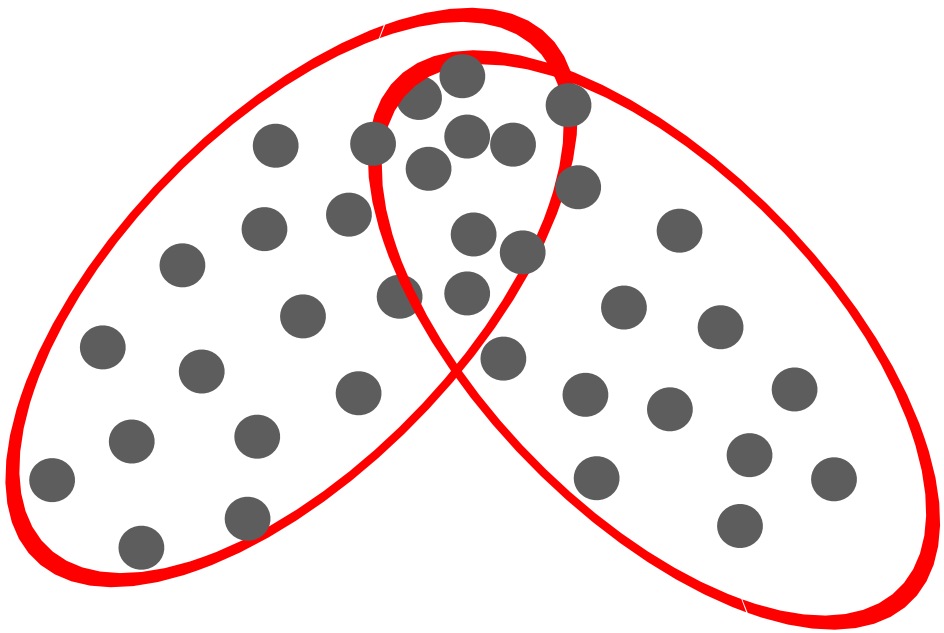
Exemple de configuration de classes

2 classes qui se chevauchent



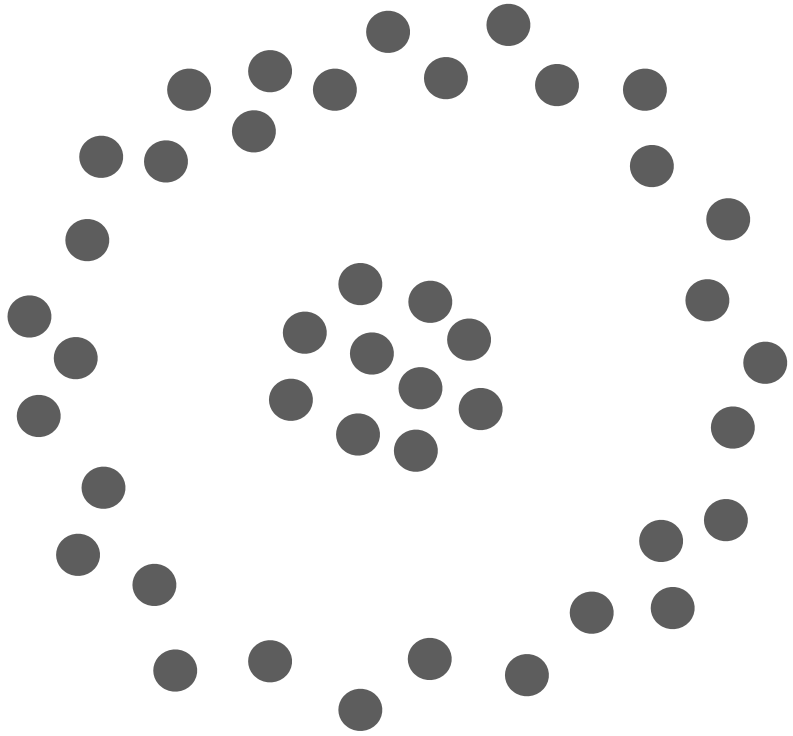
Exemple de configuration de classes

2 classes qui se chevauchent



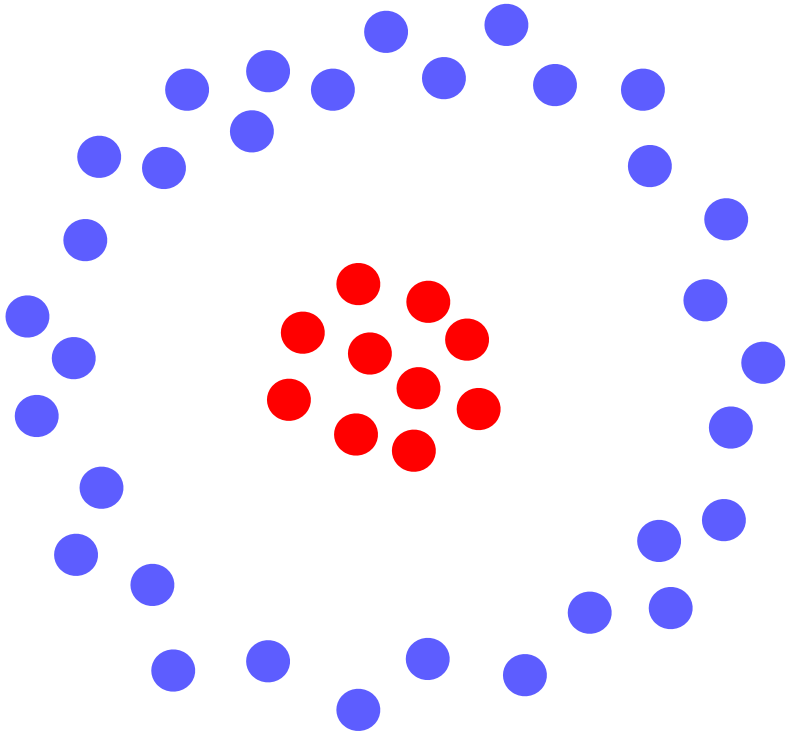
Exemple de configuration de classes

2 Classes imbriquées



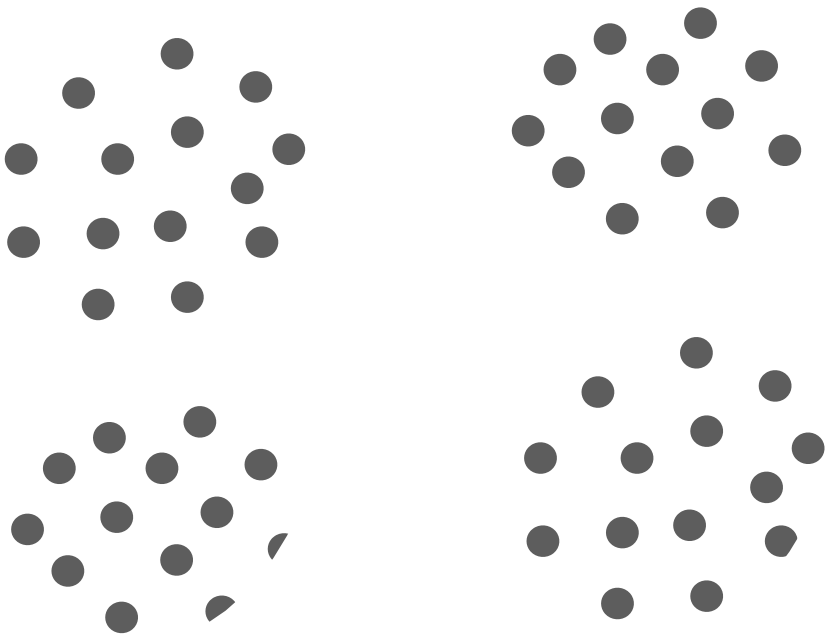
Exemple de configuration de classes

2 Classes imbriquées



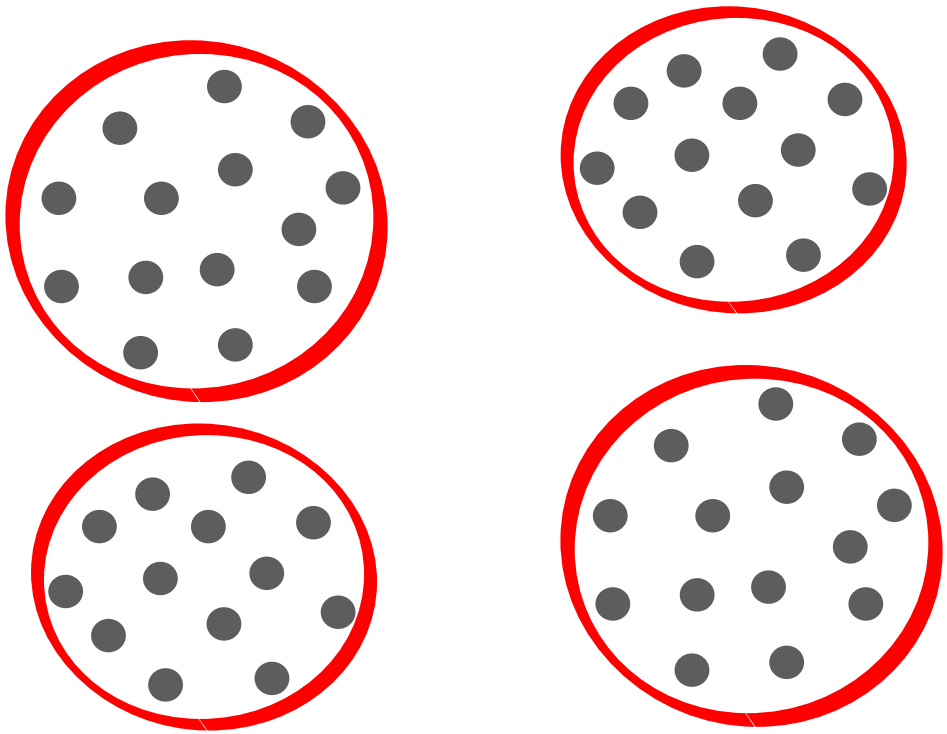
Exemple de configuration de classes

Combien de classes ?



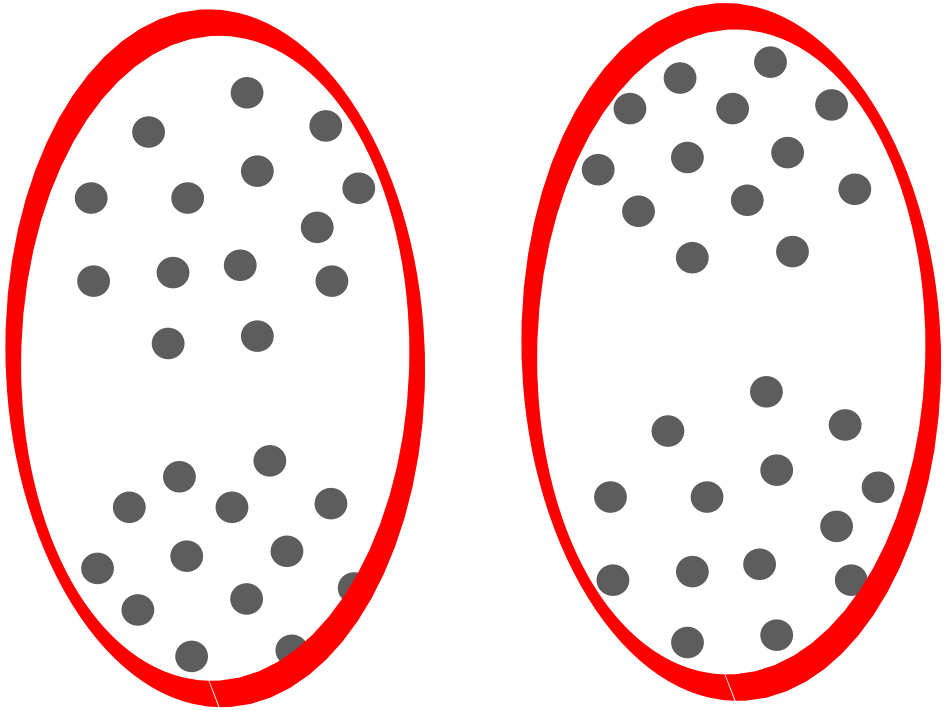
Exemple de configuration de classes

4 classes ?

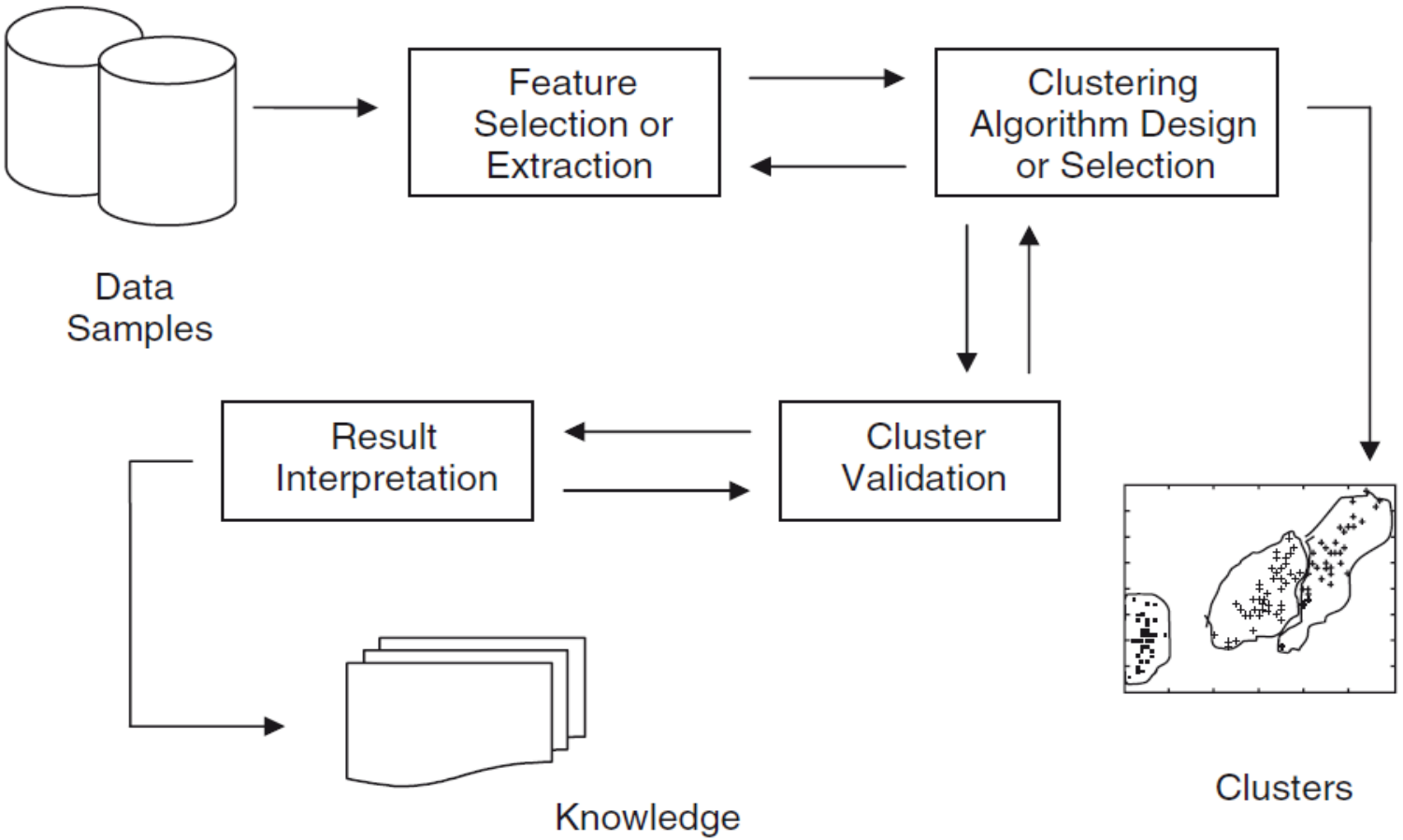


Exemple de configuration de classes

2 classes ?



Différentes étapes du processus de classification automatique



Source : Wiley IEEE Press 2009 ; Clustering ; Xu and Wunsch.

Format des données

- Ensemble de n individus décrits par p variables (caractères)

$$\Omega = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$$

avec $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$

- Cet ensemble peut également être représenté sous la forme d'un tableau $\mathbf{X}$ de n lignes et p colonnes

$\mathbf{X} =$

	var 1	...	var j	...	var p
indiv $\mathbf{x}_1$	x_{11}	...	x_{1j}	...	x_{1p}
indiv $\mathbf{x}_i$	x_{i1}	...	x_{ij}	...	x_{ip}
indiv $\mathbf{x}_n$	x_{n1}	...	x_{nj}	...	x_{np}

Exemple : données Iris (R. A. Fisher, 1936)

- Exemple classique étudié en statistique
- Proposé par Fisher pour illustrer les méthodes de classification
- 150 iris provenant de 3 familles différentes : Virginia, Versicolor et Setosa
- Données : longueur et largeur du sépale et du pétale

long-sp	larg-sp	long-pt	larg-pt	espece
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
5,5	2,4	3,7	1,0	Iris-versicolor
5,8	2,7	3,9	1,2	Iris-versicolor
6,0	2,7	5,1	1,6	Iris-versicolor
5,4	3,0	4,5	1,5	Iris-versicolor
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
6,3	3,3	6,0	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica
7,1	3,0	5,9	2,1	Iris-virginica
6,3	2,9	5,6	1,8	Iris-virginica

Autre format de données : tableau de proximités

Tableau de proximités

Tableau carré de nombres mesurant une ressemblance ou une dissemblance entre les elements d'un ensemble : distances géographiques, distances routières, durées de trajets, corrélations entre variables

Exemple : distances croisées entre des villes européennes

	Lond	Stoc	Lisb	Madr	Par	Amst	Berl	Prag	Rome	Dubl
Londres	0	569	667	530	141	140	357	396	569	190
Stockholm	569	0	1212	1043	617	446	325	423	787	648
Lisbonne	667	1212	0	201	596	768	923	882	714	714
Madrid	530	1043	201	0	431	608	740	690	516	622
Paris	141	617	596	431	0	177	340	337	436	320
Amst	140	446	768	608	177	0	218	272	519	302
Berlin	357	325	923	740	340	218	0	114	472	514
Prague	396	423	882	690	337	272	114	0	364	573
Rome	569	787	714	516	436	519	472	364	0	755
Dublin	190	648	714	622	320	302	514	573	755	0

Variables

Quantitative : si la variable est à valeurs dans $\mathbb{R}$

discrète : si la variable est à valeurs dans un sous ensemble dénombrable de $\mathbb{R}$
ex : age en années, nombre d'enfants à charge

continue : si la variable est à valeurs dans un intervalle de $\mathbb{R}$
ex : taille, poids, consommation, revenu, montant facture d'électricité, diamètre d'une pièce en sortie d'usine...

Qualitative : variable à valeur dans un ensemble fini

nominale : s'il n'y a pas de relation d'ordre entre les modalités
ex : sexe, situation familiale

ordinale : s'il y a une relation d'ordre entre les modalités
ex : réponse à un sondage ayant pour modalités : "très bon", "bon", "moyen", "mauvais", "très mauvais"

Exemples de variables

Numéro	Libellé	Nature
1	Sexe	
2	age en années	
3	Situation familiale (C, M, D, ME, CE, MS, CS, ...)	
4	Ancienneté en mois	
5	Catégorie socio-professionnelle (codée)	
6	Taux d'endettement	
7	Moyenne des mouvements créditeurs en euros	
8	Nombre de paiements par carte bancaire	
9	Nombre de jours à débit dans le mois	
10	Possession de la carte VISA Premier	

Homogénéisation et transformation de variables quantitatives

Centrer-réduire

Permet d'uniformiser l'échelle de grandeur des variables

- Centrer : soustraire de chaque valeur la moyenne de la variable
- Réduire : diviser chaque valeur par l'écart-type de la variable
- Centrer et réduire : faire les deux opérations

Discrétiser

Transformer une variable quantitative en variable qualitative

- discrétisation définie a priori : ex. remplacer l'âge par une des valeurs 1, 2, 3, 4 suivant les intervalles : (1) 0-18 ans, (2) 19-40 ans, (3) 41-65 ans, (4) $>$ 65 ans
- discrétisation en intervalles de même longueur

Transformation d'une variable qualitative en variable binaire

- Cas d'une variable nominale : codage disjonctif complet (on remplace la variable qualitative par les indicatrices de chaque modalité)

$$\begin{array}{cc} 1 & \\ 2 & \\ 1 & \Longleftrightarrow \\ 3 & \\ 3 & \end{array} \quad \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}$$

- Cas d'une variable ordinale : codage additif

$$\begin{array}{cc} 1 & \\ 2 & \\ 1 & \Longleftrightarrow \\ 3 & \\ 3 & \end{array} \quad \begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{array}$$

Statistiques élémentaires sur une variable

Notations

- Données : tableau numérique $\mathbf{X} = (x_{ij})$ de taille $n \times p$
- Individu : $\mathbf{x}_i = (x_{ij}, \ldots, x_{ip})$: i^{e} ligne du tableau
- Variable : $\mathbf{x}^j = (x_{1j}, \ldots, x_{nj})$: j^{e} colonne du tableau

Statistiques élémentaires sur une variable quantitative

Statistiques

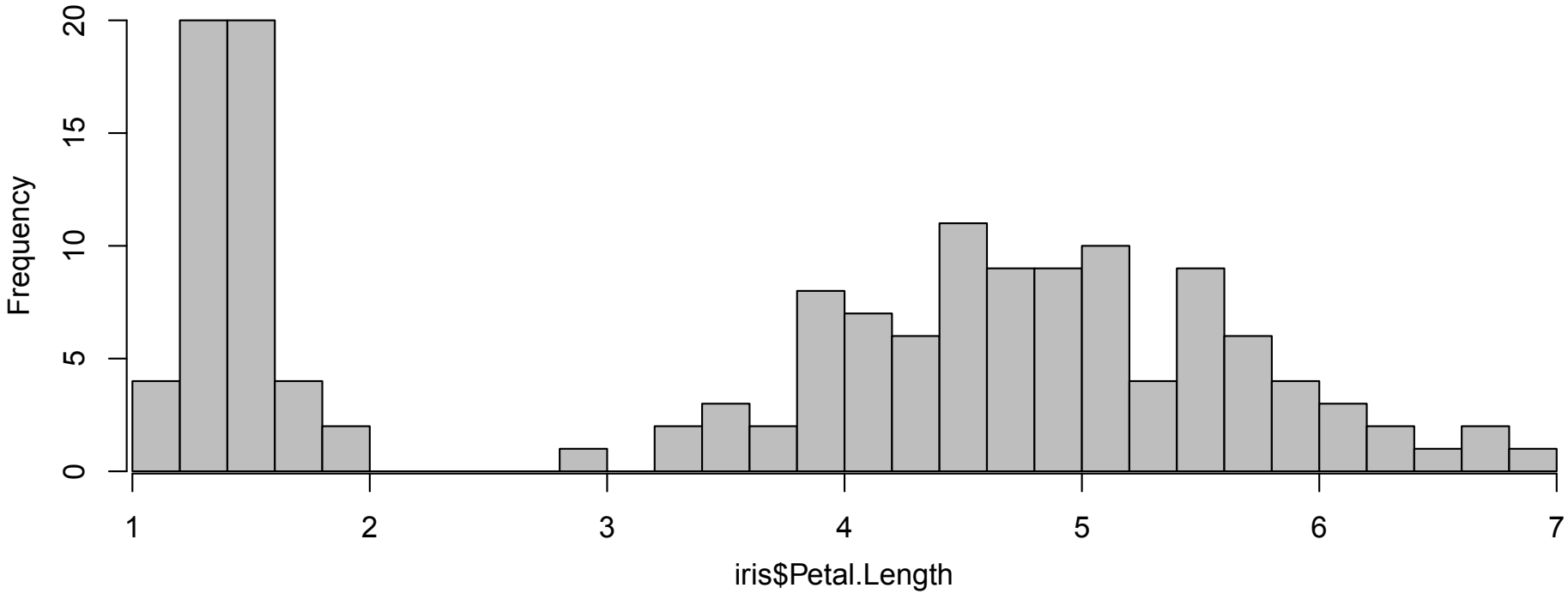
- Minimum et maximum
- Moyenne empirique : $\overline{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
- Variance empirique : $(s^j)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}^j)^2$
- Écart-type empirique $s^j = \sqrt{(s^j)^2}$
- Quartiles :
  - premier quartile q_1 : valeur qui partage l'échantillon en 25% à gauche et 75% à droite
  - deuxième quartile q_2 qui correspond à la médiane : valeur qui partage l'échantillon en 50% à gauche et 50% à droite
  - troisième quartile q_3 : valeur qui partage l'échantillon en 75% à gauche et 25% à droite

Statistiques élémentaires sur données Iris

```
> summary(iris)
```

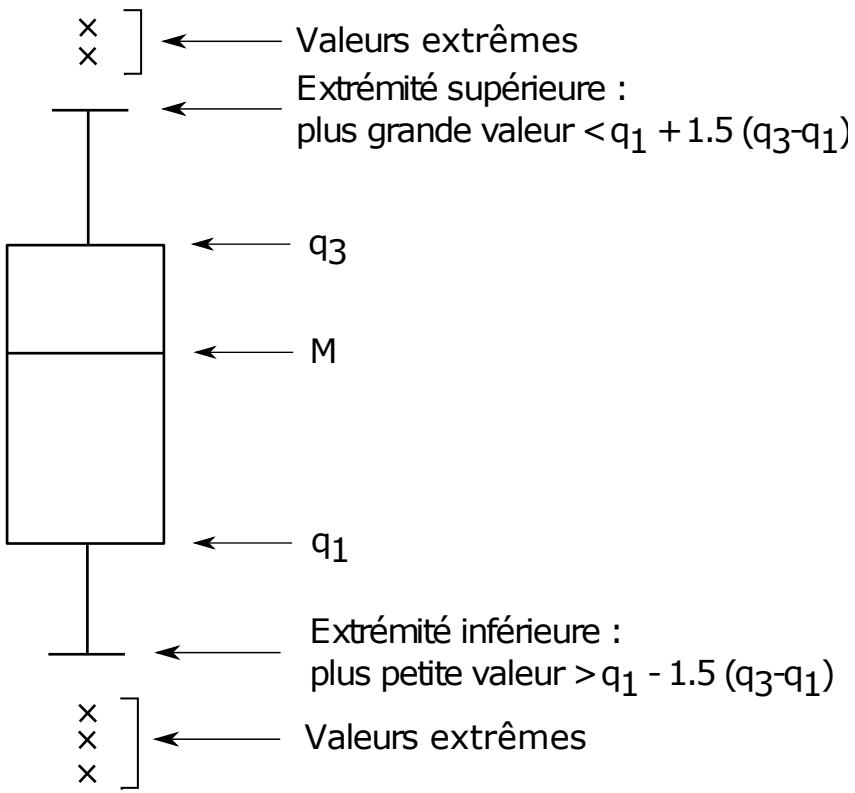
```
  sep_length      sep_width       pet_length      pet_width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :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
```

Histogramme

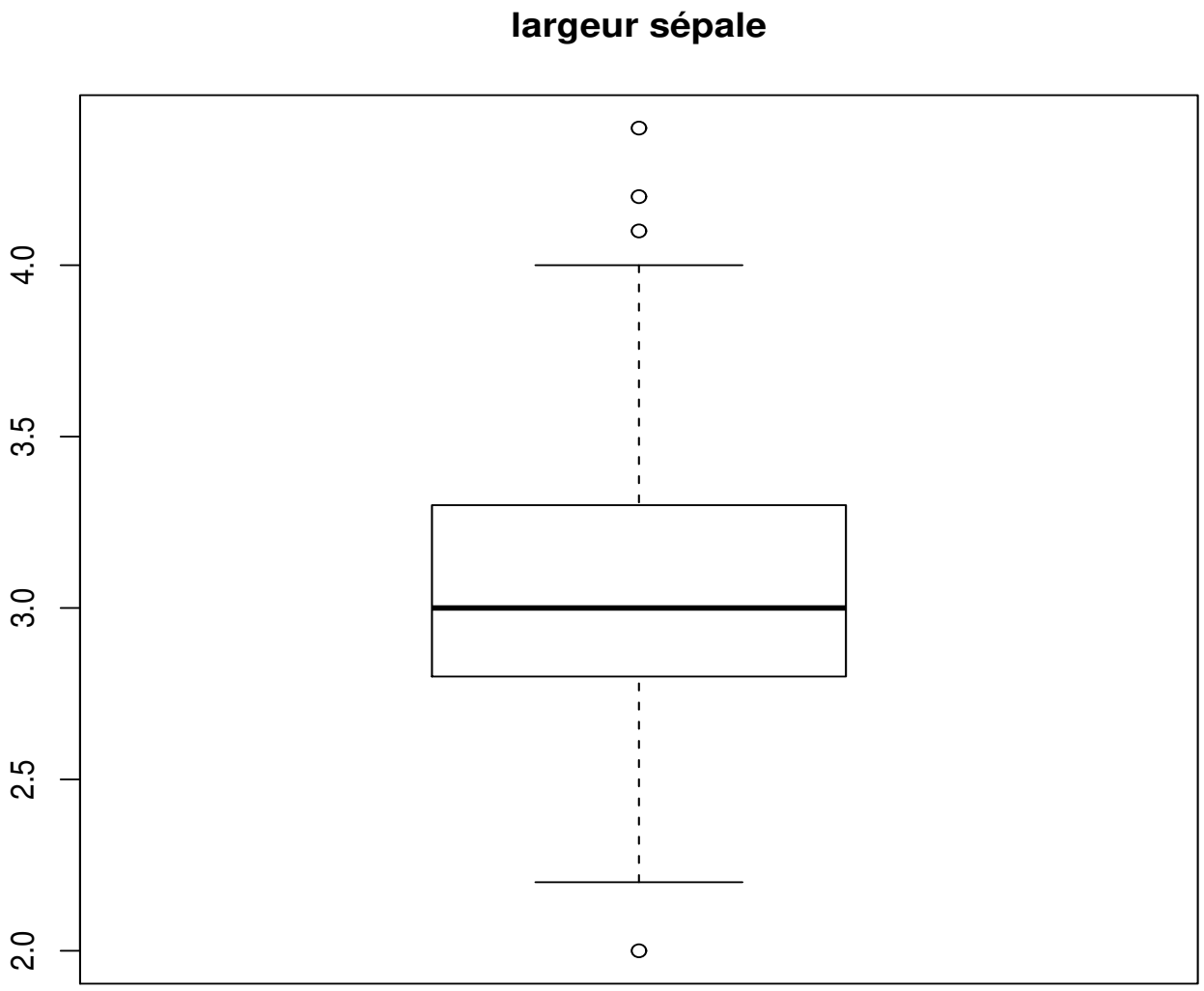


Boîte à moustache ou boxplot

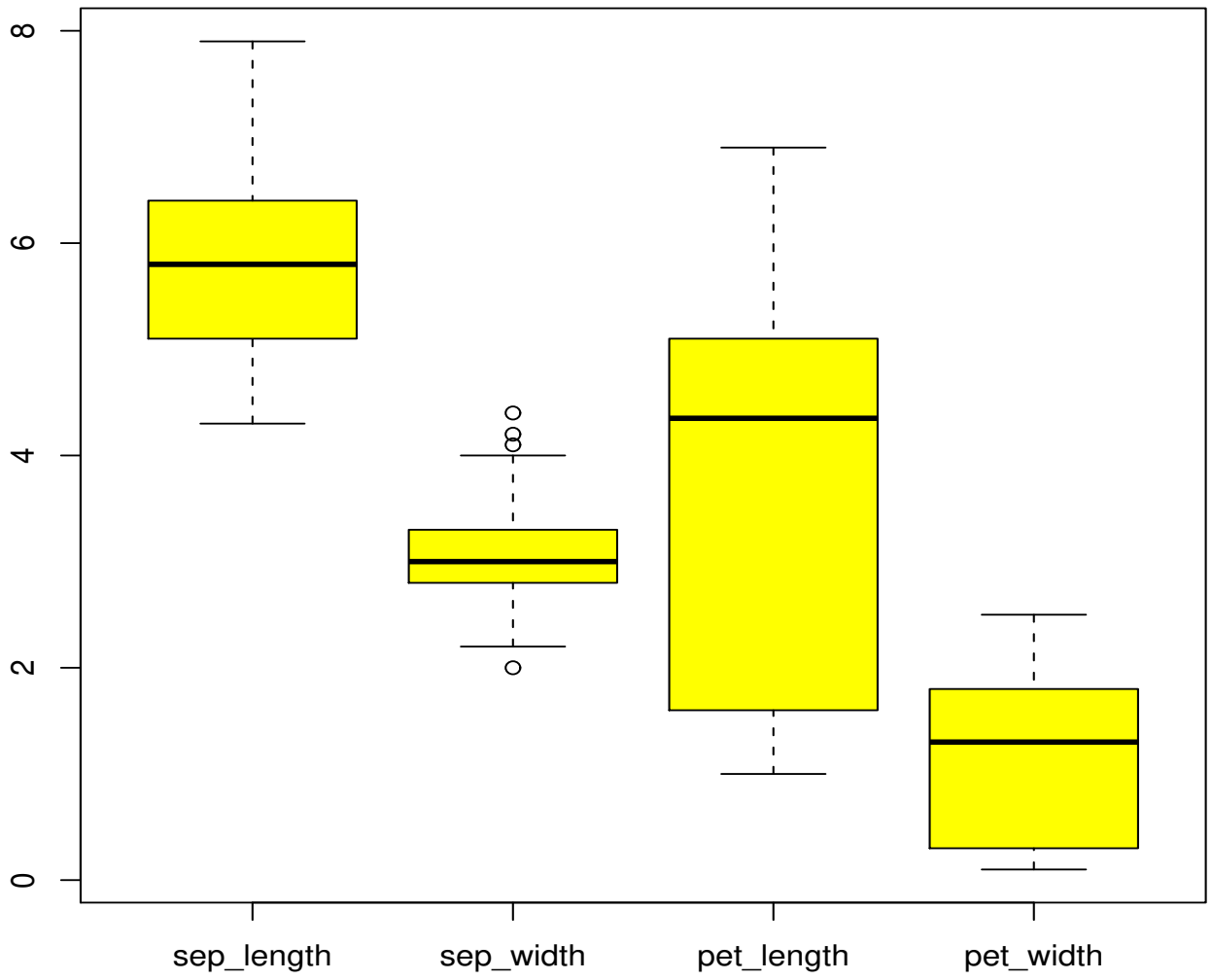
- Représente les principales caractéristiques d'une variable numérique
- Permet de repérer d'éventuelles valeurs aberrantes
- Facilite la comparaison de plusieurs distributions



Exemple de boîte à moustaches



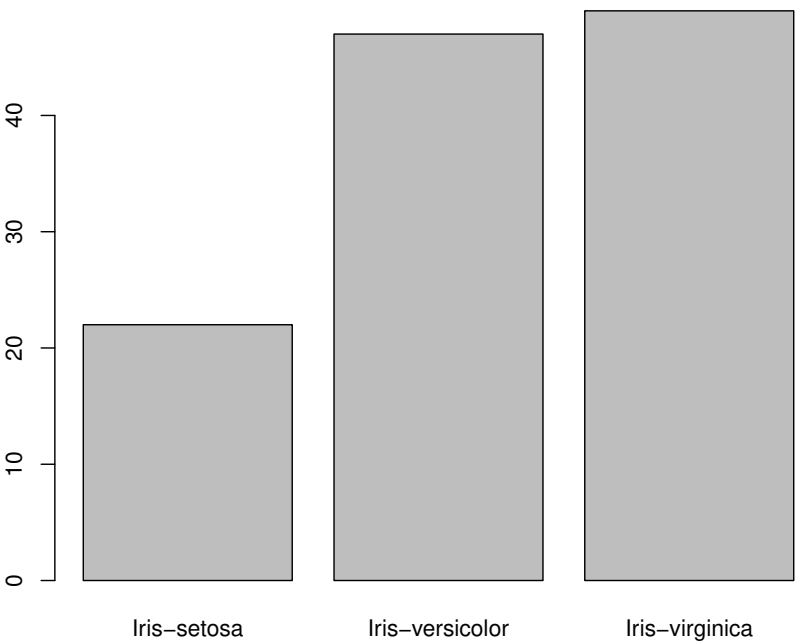
Boîtes à moustaches (pour les 4 variables quantitatives des données Iris)



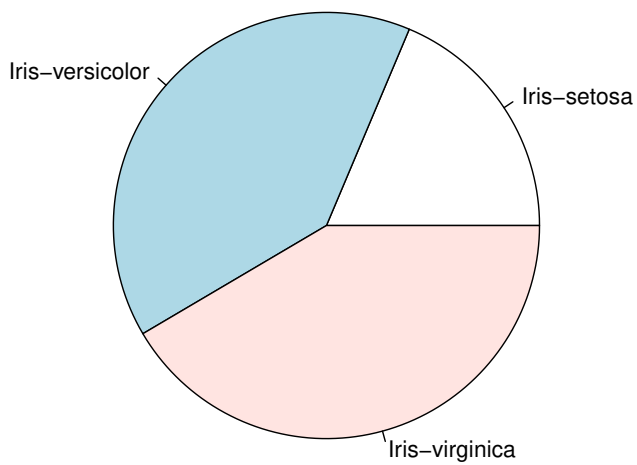
Description d'une variable qualitative

Variable « espèce » du jeu de données « Iris » pour les longueurs de sépale supérieures à 5

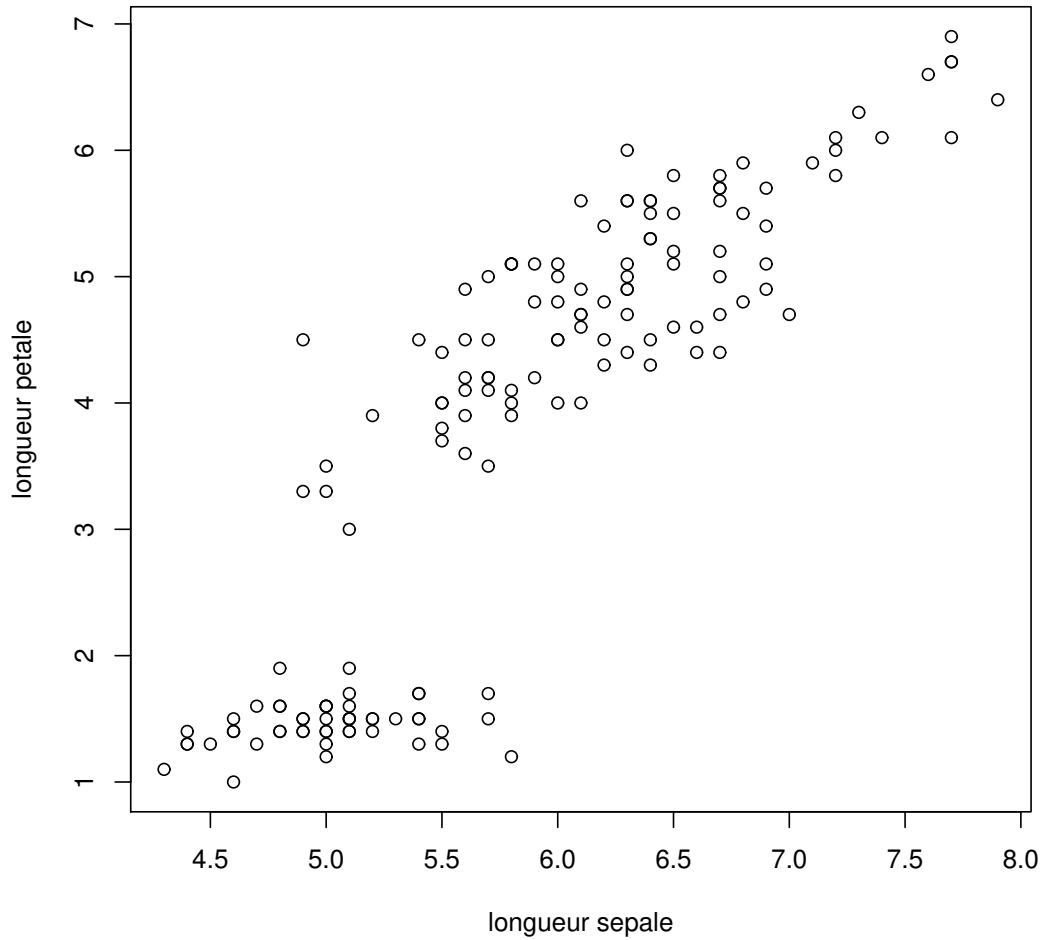
Diagramme en barre



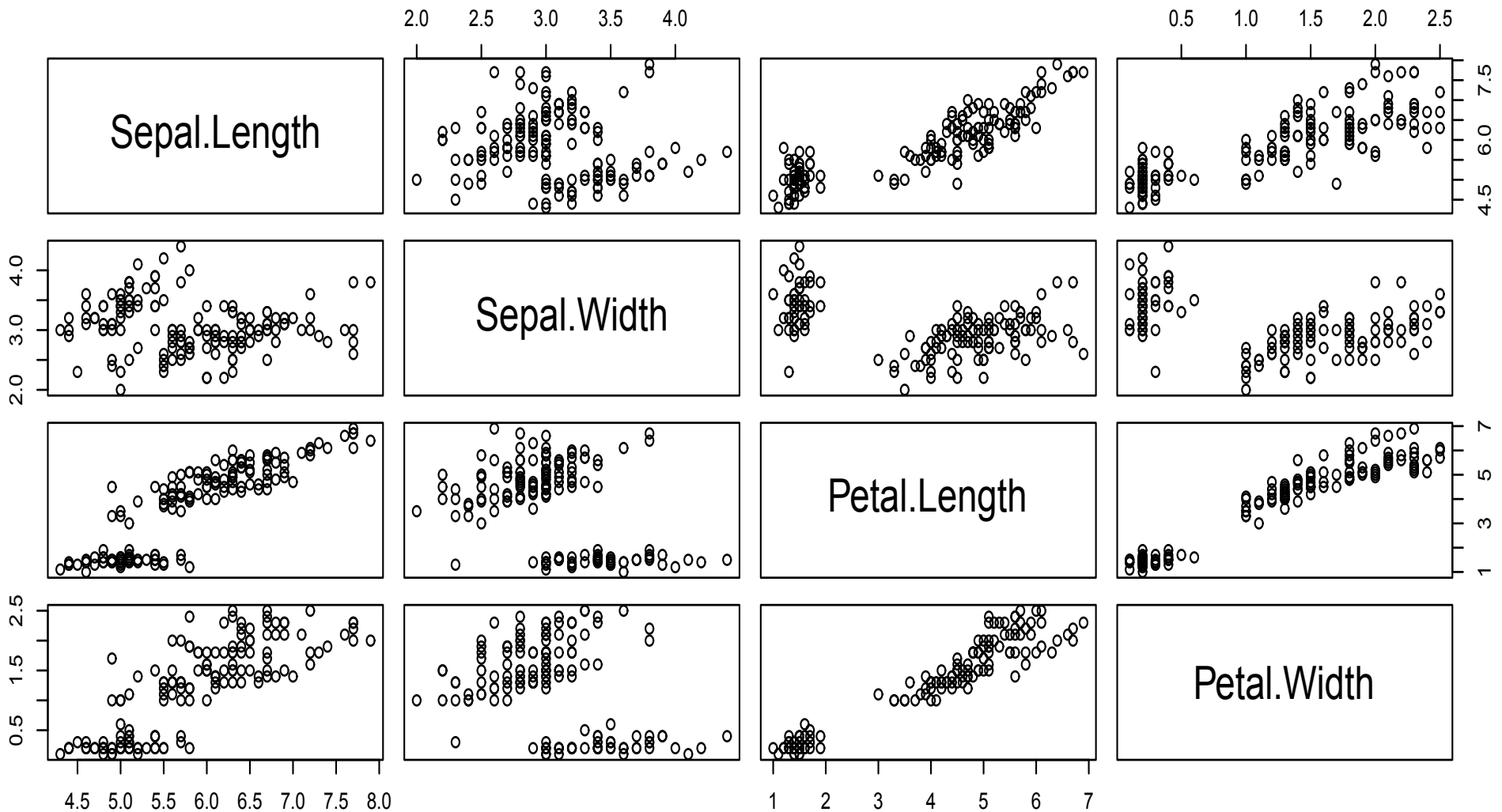
Camembert



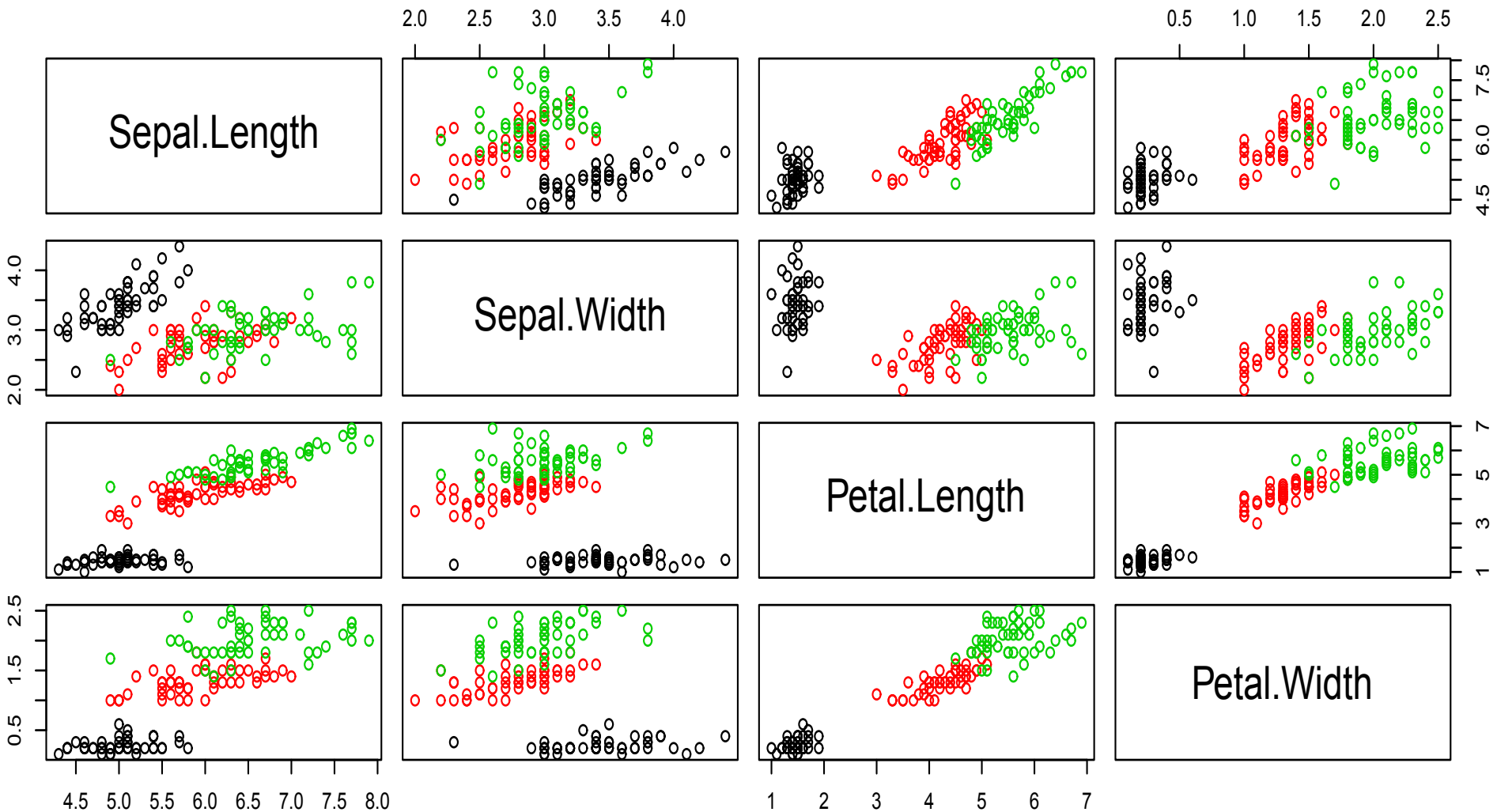
Description bidimensionnelle de deux variables quantitatives : nuage de points



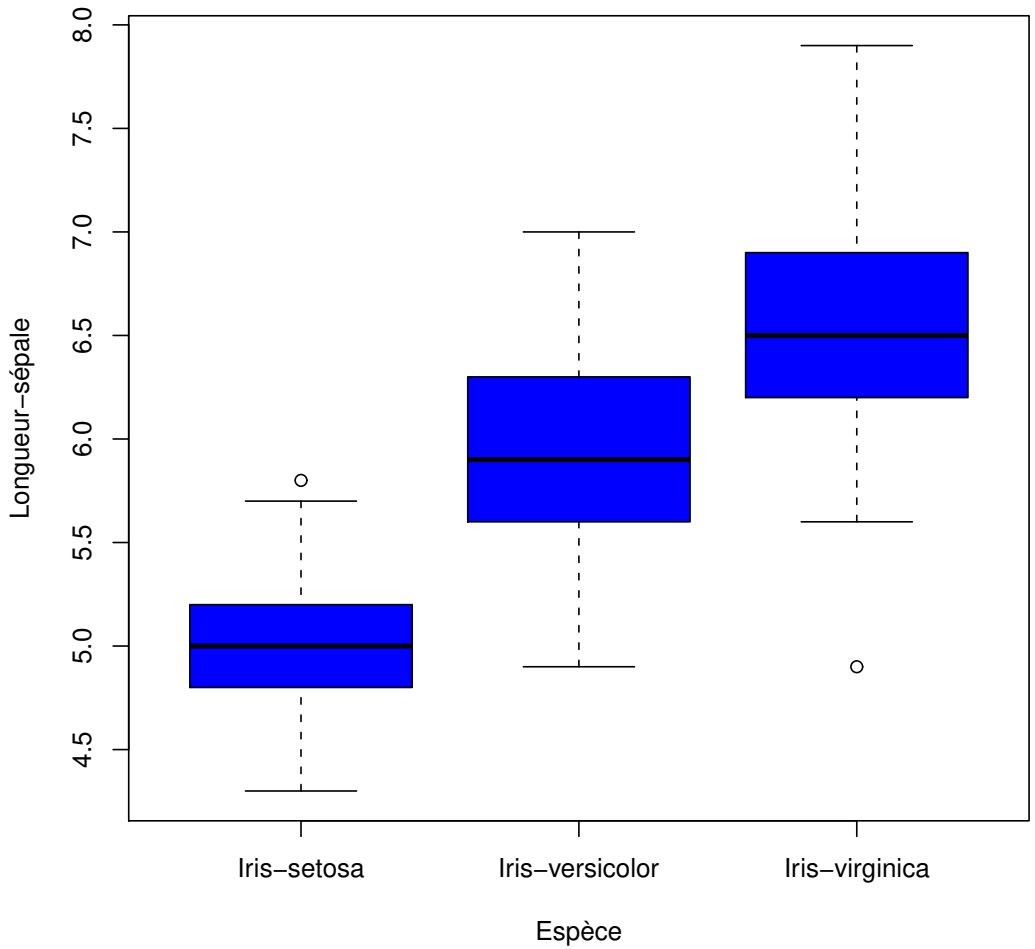
Description bidimensionnelle de plus de deux variables quantitatives : nuages de points multiples



Description bidimensionnelle de plus de deux variables quantitatives : nuages de points multiples



Description bidimensionnelle : variable quantitative Vs. variable qualitative



Description bidimensionnelle : deux variables qualitatives

Données

$\mathbf{x}^j$	$\mathbf{x}^{j'}$
a	a
b	c
a	b
b	b
b	c
b	b

Tableau de contingence

	a	b	c
a	1	1	0
b	0	2	2

Degré de liaison entre deux variables quantitatives : covariance et corrélation

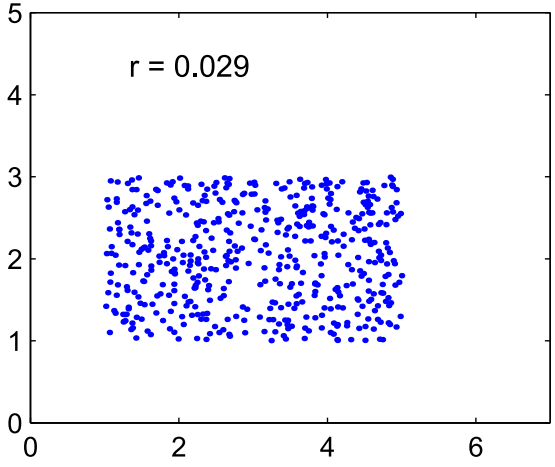
Covariance empirique

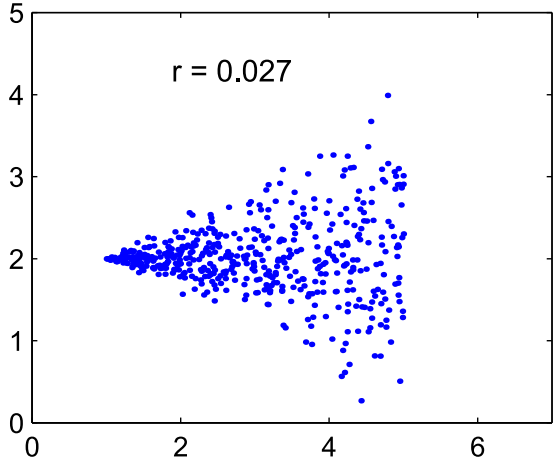
$$s_{jj'} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}^{j})(x_{ij'} - \overline{x}^{j'})}{n}$$

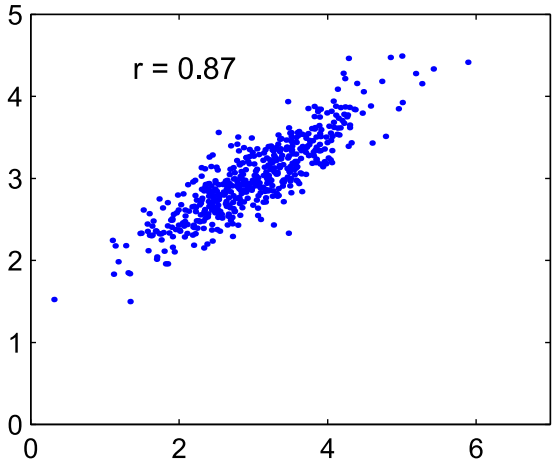
Coefficient de corrélation empirique

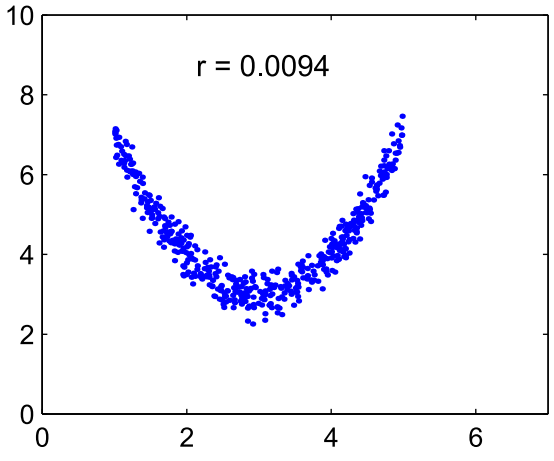
$$r_{jj'} = \frac{s_{jj'}}{s_j \, s_{j'}}$$

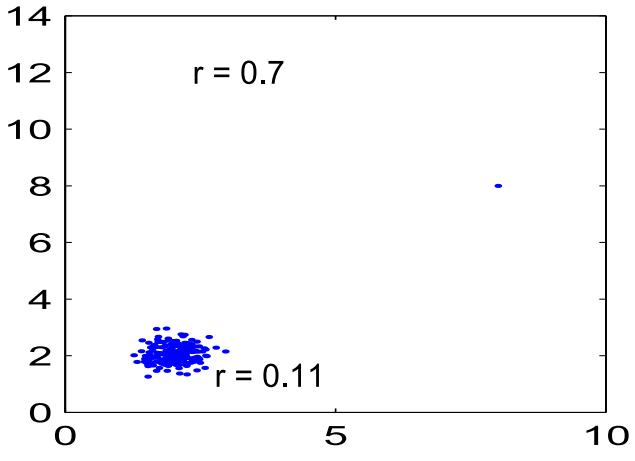
Exemples de coefficients de corrélation

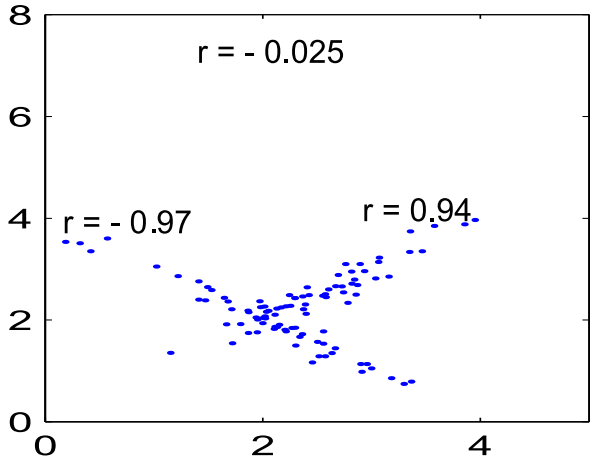












Indicateur numérique de liaison entre une variable quanti. $\mathbf{x}^j$ et une variable quali. $\mathbf{x}^{j'}$

On suppose que la variable qualitative $\mathbf{x}^{j'}$ prend L modalités

On note n_ℓ le nombre d'occurrences de chaque modalité $(1 \leq \ell \leq L)$

Pour la variable $\mathbf{x}^j$, on note $\overline{x}_\ell$ la moyenne des valeurs correspondant à la modalité ℓ et $\overline{x}$ la moyenne globale.

Indicateur numérique : rapport entre la variance inter-classes et la variance totale

$$\rho = \frac{\sum_{\ell=1}^{L} n_\ell (\overline{x}_\ell - \overline{x})^2}{\sum_{i=1}^{n} (x_{ij} - \overline{x})^2}$$

Distance et Norme

Distance

Une distance d sur un espace métrique E est une application de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ (séparation)

(ii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symétrie)

(iii) $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in E, \quad d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (inégalité triangulaire)

Norme

Une norme $\| \cdot \|$ sur un $\mathbb{R}-$espace vectoriel E est une application de $E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x} \in E, \quad \| \mathbf{x} \| = 0 \Leftrightarrow \mathbf{x} = 0$

(ii) $\forall \mathbf{x} \in E, \lambda \in \mathbb{R}, \quad \| \lambda \mathbf{x} \| = |\lambda| \, \| \mathbf{x} \|$

(iii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad \| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \|$

Equivalence Distance et Norme

- A une norme $\| \cdot \|$, on peut associer la distance définie par :

$$d(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \|$$

- A une distance d, on peut associer sous certaines conditions la norme $\| \cdot \|$ définie par

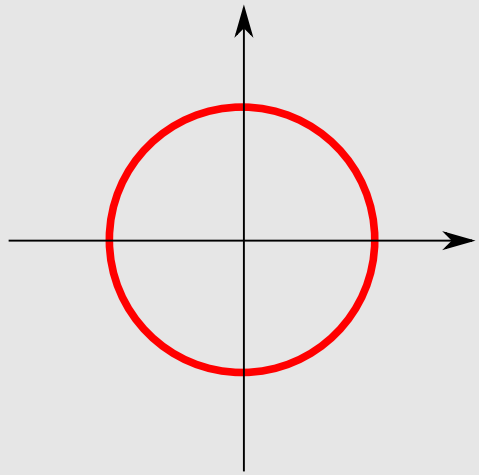
$$\| \mathbf{x} \| = d(0, \mathbf{x})$$

Distances usuelles pour variables quantitatives

Distance euclidienne ou distance L_2

$$\begin{aligned} d^2(\mathbf{x},\mathbf{y}) &= \sum_{j=1}^{p}(x_j - y_j)^2 \\ &= (\mathbf{x}-\mathbf{y})'\mathbf{I}(\mathbf{x}-\mathbf{y}) \end{aligned}$$

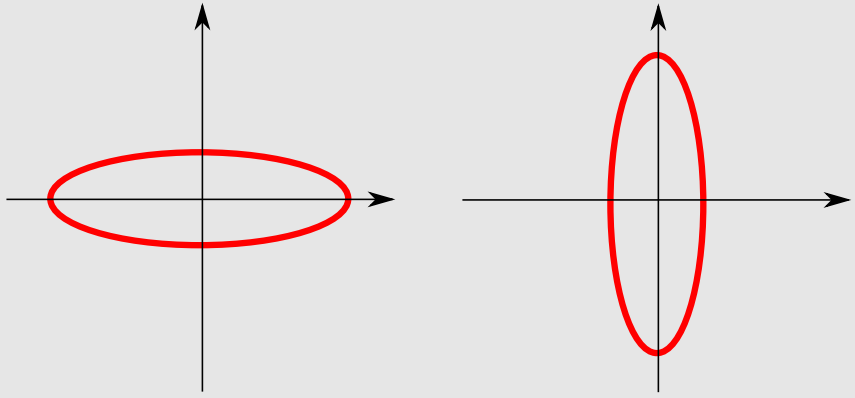
avec $\mathbf{I}$ = matrice identité



Distance euclidienne pondérée

$$\begin{aligned} d^2(\mathbf{x},\mathbf{y}) &= \sum_{j=1}^{p} w_j(x_j - y_j)^2 \\ &= (\mathbf{x}-\mathbf{y})'\mathbf{D}(\mathbf{x}-\mathbf{y}) \end{aligned}$$

avec $\mathbf{D} = diag(w_1, \ldots, w_p)$

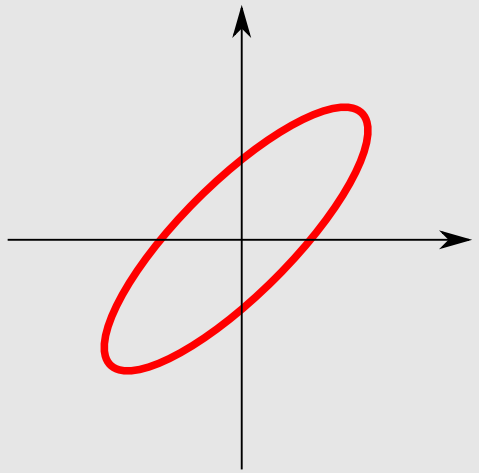


Distances usuelles pour variables quantitatives

Distance de Mahalanobis

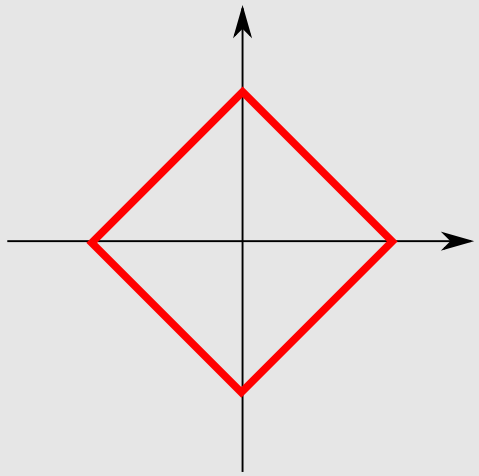
$$d^2(\mathbf{x},\mathbf{y}) \quad = \quad (\mathbf{x}-\mathbf{y})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})$$

avec $\mathbf{\Sigma}$ = matrice de covariance des données, symétrique définie positive



Distance de Manhattan ou distance L_1

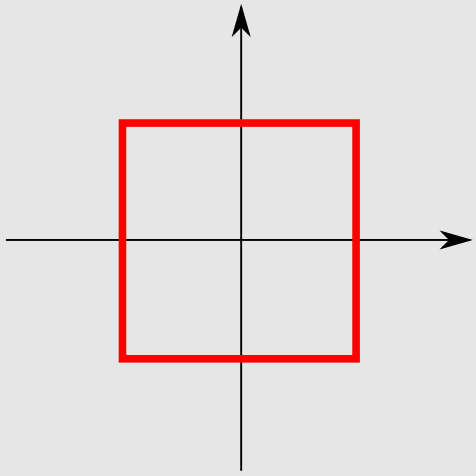
$$d(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|$$



Distances usuelles pour variables quantitatives

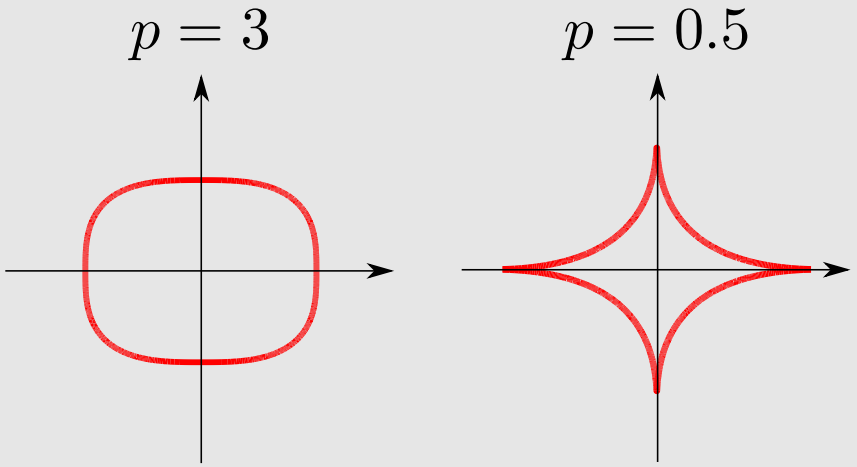
Distance de Tchebychev ou distance L_∞

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq j \leq p} |x_j - y_j|$$



Distance de Minkowski ou distance L_p (généralisant L_1, L_2)

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^{p} |x_j - y_j|^p \right)^{1/p}$$



Dissimilarité et similarité

Dissimilarité

Une mesure de dissimilarité est une application de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symétrie)

(ii) $\mathbf{x} = \mathbf{y} \quad \Longrightarrow \quad d(\mathbf{x}, \mathbf{y}) = 0$

Similarité

Une mesure de similarité est une application de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symétrie)

(ii) $s(\mathbf{x}, \mathbf{y}) = s_{max} \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{y}$

(iii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad s(\mathbf{x}, \mathbf{y}) \leq s_{max}$

Equivalence entre dissimilarité et similarité

- Si s est une mesure de similarité, alors l'application d définie par

$$d(\mathbf{x}, \mathbf{y}) = s_{max} - s(\mathbf{x}, \mathbf{y})$$

  est une mesure de dissimilarité

- Si d est une mesure de dissimilarité, alors l'application s définie par

$$s(\mathbf{x}, \mathbf{y}) = d_{max} - d(\mathbf{x}, \mathbf{y})$$

  est une mesure de similarité

Ultramétrique

Une ultramétrique δ sur un ensemble E est une fonction de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad \delta(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$

(ii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad \delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{y}, \mathbf{x})$

(iii) $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in E, \quad \delta(\mathbf{x}, \mathbf{z}) \leq \max(\delta(\mathbf{x}, \mathbf{y}), \delta(\mathbf{y}, \mathbf{z}))$
(inégalité ultramétrique)

Propriétés de l'ultramétrique

- L'inégalité (iii) entraine l'inégalité triangulaire
- On peut donc vérifier qu'une ultramétrique est une distance

L'ultramétrique joue un rôle fondamental en classification (on verra dans la suite qu'il y a un lien direct entre ultramétrique et hiérarchie)

Exemples

$$\mathbf{X} = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 4 & 3 \\ 5 & 4 \\ 5 & 1 \end{pmatrix}$$

Matrice de dissimilarités associée (distance euclidienne)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{x}_1$	0	2	3.16	4	5
$\mathbf{x}_2$	2	0	3.16	4.47	4.12
$\mathbf{x}_3$	3.16	3.16	0	1.41	2.24
$\mathbf{x}_4$	4	4.47	1.41	0	3
$\mathbf{x}_5$	5	4.12	2.24	3	0

Matrice de similarités

$d_{max} = 5$ et $s(\mathbf{x}_i, \mathbf{x}_{i'}) = d_{max} - d(\mathbf{x}, \mathbf{x}_{i'})$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{x}_1$	5	3	1.84	1	0
$\mathbf{x}_2$	3	5	1.84	0.53	0.88
$\mathbf{x}_3$	1.84	1.84	5	3.59	2.76
$\mathbf{x}_4$	1	0.53	3.59	5	2
$\mathbf{x}_5$	0	0.88	2.76	2	5

Similarité entre deux variables binaires

Jaccard

$$s(\mathbf{x},\mathbf{y}) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$$

Dice et Czekanowski

$$s(\mathbf{x},\mathbf{y}) = \frac{2n_{11}}{2n_{11}+n_{10}+n_{01}}$$

Ochiaï

$$s(\mathbf{x},\mathbf{y}) = \frac{n_{11}}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})}}$$

Russel et Rao

$$s(\mathbf{x},\mathbf{y}) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}+n_{00}}$$

Rogers et Tanimoto

$$s(\mathbf{x},\mathbf{y}) = \frac{n_{11}+n_{00}}{n_{11}+n_{00}+2(n_{10}+n_{01})}$$

où n_{11}, n_{10}, n_{01}, n_{00} représentent le nombre de fois où le couple $(\mathbf{x},\mathbf{y})$ vaut $(1,1)$, $(1,0)$, $(0,1)$, $(0,0)$.