Statistiques descriptives élémentaires

C. Ambroise

Laboratoire Statistique et Génome
UMR CNRS 8071

Automne 2007

Plan

Plan

Analyse statistique exploratoire élémentaire

- Avant d'étudier les méthodes d'AD :
  $\longrightarrow$ rappel des méthodes exploratoires élémentaires
- Utilisation de ces outils pouvant aller très loin :
  $\longrightarrow$ EDA : Exploratory data analysis
- Ici, on se limite aux données individus-variables quantitatives

Le tableau de données

- n individus mesurés par p variables quantitatives
- Matrice réelle $X = (x_i^j) = \begin{pmatrix} x_1^1 & x_1^j & x_1^p \\ & \cdot & \cdot \\ x_i^1 & x_i^j & x_i^p \\ & \cdot & \cdot \\ x_n^1 & x_n^j & x_n^p \end{pmatrix}$
- Chaque variable est représentée par le vecteur $\mathbf{x}^j = (x_1^j, \ldots, x_n^j)'$
- Chaque individu est représenté par le vecteur $\mathbf{x}_i = (x_i^1, \ldots, , x_i^p)'$
- X : réalisation d'un échantillon de taille n du vecteur aléatoire de dimension p

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

Plan

Vecteur aléatoire

- Loi de probabilité de $\mathbf{X}$: $P_{\mathbf{X}}(A) = P(\mathbf{X} \in A)$
- $\mathbf{X}$ discret : définie par la probabilité élémentaire $p : \mathbb{R}^p \to \mathbb{R}$

$$p(\mathbf{x}) = p(\mathbf{X} = \mathbf{x}) \quad \text{ou} \quad p(x^1, \ldots, x^p) = P(X^1 = x^1, \ldots, X^p = x^p)$$

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} p(\mathbf{x}) = \sum_{(x^1, \ldots, x^p) \in A} p(x^1, \ldots, x^p)$$

- $\mathbf{X}$ absolument continu : définie par la densité $f : \mathbb{R}^p \to \mathbb{R}$

$$P(\mathbf{X} \in A) = \int_A f(\mathbf{x}) d\mathbf{x} = \int_A f(x^1, \ldots, x^p) dx^1 \ldots dx^p$$

Fonction de répartition

- Définition

$$F(x^1, \ldots, x^p) = P(X^1 \leq x^1, \ldots, X^p \leq x^p)$$

- Continu :

$$F(x^1, \ldots, x^p) = \int_{-\infty}^{x^1} \ldots \int_{-\infty}^{x^p} f(u^1, \ldots, u^p) du^1 \ldots du^p$$

$$f(x^1, \ldots, x^p) = \frac{\partial^p F}{\partial x^1 \ldots \partial x^p}(x^1, \ldots, x^p)$$

Lois marginales

- Tout sous-vecteur de $\mathbf{X}$ est aussi un vecteur aléatoire
- Loi marginale : loi d'un tel sous-vecteur
- Notation : $p_{j_1,\ldots,j_q}$ ou $f_{j_1,\ldots,j_q}$
- Une seule variable :
  - Discret : $p_j(x^j) = \sum_{x^1,\ldots,x^{j-1},x^{j+1},\ldots,x^p} p(x^1,\ldots,x^p)$
  - Continu : $f_j(x^j) = \int_{\mathbb{R}^{p-1}} f(x^1,\ldots x^p)dx^1\ldots dx^{j-1}dx^{j+1}\ldots dx^p$

Espérance

- Définition : $\mathbb{E}(\mathbf{X}) = (E(X^1), \ldots, E(X^p))'$
- Linéarité : Si $\mathbf{X}$ et $\mathbf{Y}$ sont de dimension p et si A et B sont 2 matrices de dimensions (q, p), alors

$$\mathbb{E}(A\mathbf{X} + B\mathbf{Y}) = AE(\mathbf{X}) + BE(\mathbf{Y})$$

- Ex. : si $u'\mathbf{X}$ combinaison linéaire : $E(u'\mathbf{X}) = u'E(\mathbf{X})$
- Espérance d'une fonction réelle d'un vecteur aléatoire :
  - Continu : $\mathbb{E}(\varphi(\mathbf{X})) = \int_{\mathbb{R}^p} \varphi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$
  - Discret : $\mathbb{E}(\varphi(\mathbf{X})) = \sum_V \varphi(\mathbf{x}) p(\mathbf{x})$

Matrice de variance I

- Variance et covariance d'une variable aléatoire :

$$\sigma_j^2 = \text{var}(X^j) = \mathbb{E}[(X^j - \mathbb{E}(X^j))^2]$$

$$\sigma_{jj'} = \text{cov}(X^j, X^{j'}) = \mathbb{E}[(X^j - \mathbb{E}(X^j))(X^{j'} - \mathbb{E}(X^{j'}))]$$

- Variance et covariance d'un vecteur aléatoire

$$\Sigma = \text{var}(\mathbf{X}) = \mathbb{E}([\mathbf{X} - \mathbb{E}(\mathbf{X})][\mathbf{X} - \mathbb{E}(\mathbf{X})]')$$

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

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1j} & \sigma_{1p} \\ & \cdot & \cdot \\ \sigma_{j1} & \sigma_j^2 & \sigma_{jp} \\ & \cdot & \cdot \\ \sigma_{p1} & \sigma_{pj} & \sigma_p^2 \end{pmatrix}$$

Matrice de variance II

- Matrice carrée, symétrique et positive de dimension (p, p)
- Propriétés
  - Rappels
    - $\text{var}(X^j) = E[(X^j)^2] - (E[X^j])^2$
    - $\text{cov}(X^j, X^{j'}) = \mathbb{E}(X^j X^{j'}) - \mathbb{E}(X^j)\mathbb{E}(X^{j'})$
    - $\text{var}(X^j + X^{j'}) = \text{var}(X^j) + \text{var}(X^{j'}) + 2\text{cov}(X^j, X^{j'})$
    - $\text{var}(aX^j) = a^2\text{var}(^j)$
    - $\text{cov}(aX^j, bX^{j'}) = a.b.\text{cov}(X^j, X^{j'})$
  - $\text{var}(A\mathbf{X}) = A.\text{var}(\mathbf{X}).A' = A\Sigma A'$
  - $\text{var}(u'\mathbf{X}) = u'\Sigma u$
  - $\text{cov}(u'\mathbf{X}, v'\mathbf{X}) = u'\Sigma v$

Matrice de corrélation

- Matrice formée des coefficients de corrélation linéaire

$$\rho_{jj'} = \frac{\mathrm{cov}(X^j, X^{j'})}{\sqrt{\mathrm{var}(X^j)\mathrm{var}(X^{j'})}}$$

- Valeurs $\in [-1, +1]$
- Diagonale à 1
- Symétrique

$$\begin{pmatrix} 1 & \rho_{1j} & \rho_{1p} \\ & . & . \\ \rho_{j1} & 1 & \rho_{jp} \\ & . & . \\ \rho_{p1} & \rho_{pj} & 1 \end{pmatrix}$$

Indépendance

- Indépendance (mutuelle) et indépendance 2 à 2
- Discret : $p(x^1, \ldots, x^p) = p_1(x^1) \ldots p_p(x^p)$
- Continu : $f(x^1, \ldots, x^p) = f_1(x^1) \ldots f_p(x^p)$
- Propriétés

$$X^1, \ldots, X^p \text{ indépendantes } \Longrightarrow \mathbb{E}(X^1 \ldots X^p) = \mathbb{E}(X^1) \ldots \mathbb{E}(X^p)$$

$$X^1, \ldots, X^p \text{ indépendantes } \Longrightarrow \text{var}(\sum_{j=1}^{p} X^j) = \sum_{j=1}^{p} \text{var}(X^j).$$

- Indépendance 2 à 2 $\Rightarrow \Sigma$ diagonale
- Réciproque fausse

Exemple : la loi normale multidimensionnelle

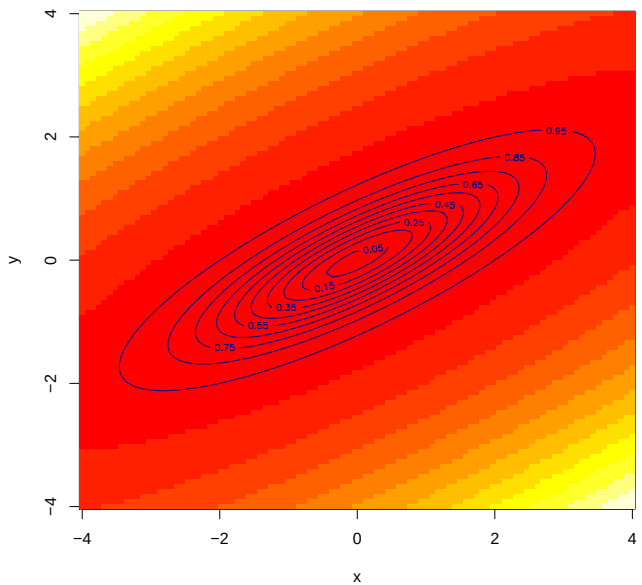
- Densité

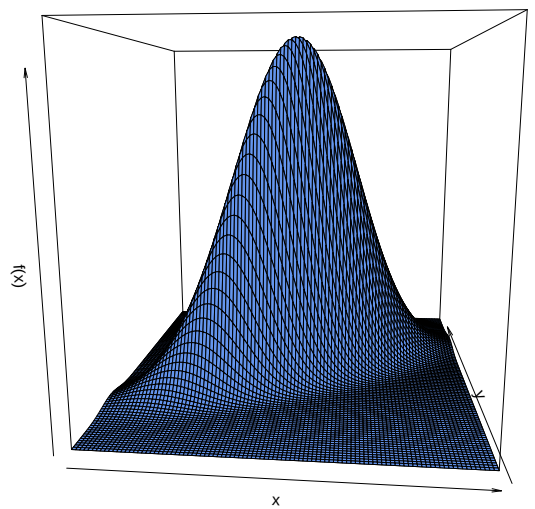
$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

  où $\mathbf{x} = (x^1, \ldots, x^p)'$.
- Généralise la distribution normale réelle
- Tout sous-vecteur est normal : donc $X^1, \ldots, X^p$ normales
- Matrice de variance diagonale = indépendance des variables

Loi normale

Loi normale bidimensionnelle et ellipses d'équiprobabilité





Plan

Statistiques associées à un vecteur aléatoire I

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

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

- Variance empirique

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

Statistiques associées à un vecteur aléatoire II

- Covariance empirique

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

- Coefficient de corrélation linéaire empirique

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

- Matrice de variance empirique

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

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

Statistiques associées à un vecteur aléatoire III

- Matrice de corrélation empirique

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

Plan

Statistiques élémentaires

- Minimum et maximum
- Valeurs centrales : moyenne et médiane
- Dispersions : étendue, variance, écart-type, étendue interquartile (Iqr)
- Sensibilité ou non aux valeurs extrêmes

Les iris

	Mean	Median	Etendue	IQR
Sepal.Length	5.84	5.80	3.60	1.30
Sepal.Width	3.06	3.00	2.40	0.50
Petal.Length	3.76	4.35	5.90	3.50
Petal.Width	1.20	1.30	2.40	1.50

Estimation de densité

- une observation de loi $f()$ tombe dans une région $\mathcal{R}$ avec une probabilité

$$P = \int_{\mathcal{R}} f(\mathbf{y}) d\mathbf{y}.$$

- $(\mathbf{x}_1, ..., \mathbf{x}_n)$ de loi parente $f()$, alors la probabilité que r de ces vecteurs tombent dans $\mathcal{R}$ suit une loi binomiale $\mathcal{B}(P, n)$ et

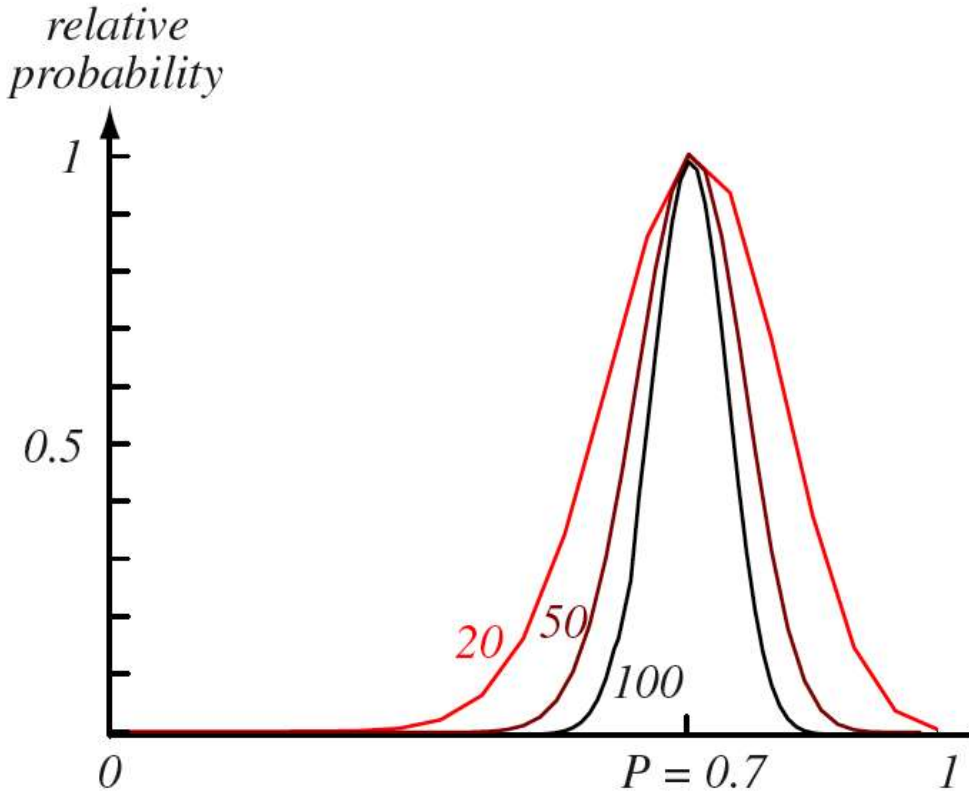
$$P(R = r) = C_r^n P^r (1 - P)^{n-r}.$$

Estimation de densité

- L'espérance de la variable aléatoire R est

$$\mathbb{E}[R] = n \cdot P,$$

- P est classiquement estimé par $\hat{P} = r/n$



Un estimateur

- Si le volume V de la région $\mathcal{R}$ est suffisamment petit,

$$P = \int_{\mathcal{R}} f(\mathbf{y}) d\mathbf{y} \approx V \cdot f(\mathbf{x}),$$

  avec $\mathbf{x}$, un vecteur de $\mathcal{R}$.
- En remplacant P par son estimation, on obtient :

$$\hat{f}(\mathbf{x}) = \frac{r/n}{V}.$$

- Convergence sous certaines hypothèses concernant le choix de V et de r, en fonction de n.

Histogramme I

- Estimateur de la fonction de densité

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

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

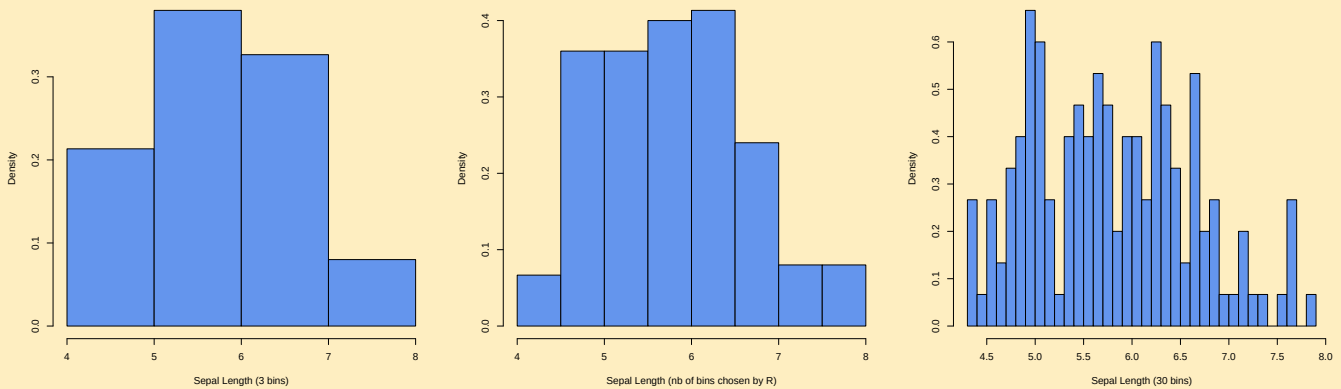
$$\sum_i h_i(a_{i+1} - a_i) = 1 \text{ et } h_i(a_{i+1} - a_i) = \hat{P}_F(X \in [a_i, a_{i+1}[)$$

Histogramme II

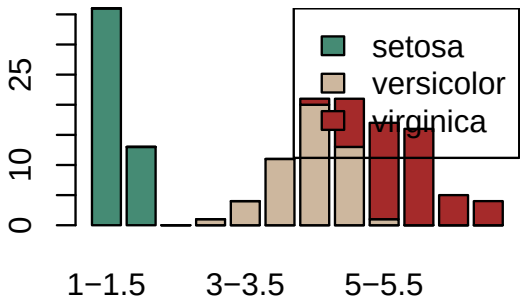
- Attention : hauteur proportionnelle à la fréquence si et seulement si les intervalles ont tous la même largeur
- Nombre d'intervalles :
  - Important
  - Réglage difficile
  - Règle empirique : règle de Sturges $1 + 10/3 * \log_{10}(n)$

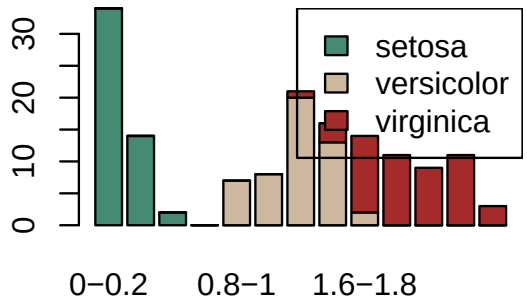
Histogramme

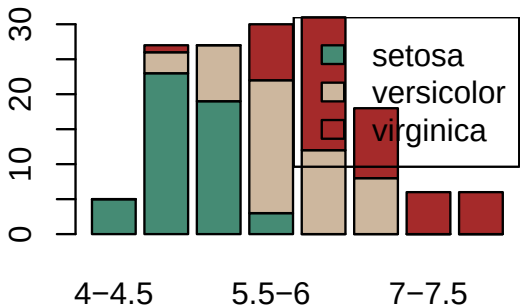
Histogrammes des longueurs de sépales

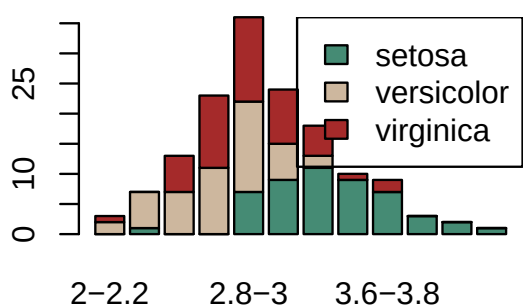


Histogrammes et variable qualitative









Méthode des noyaux (Parzen) I

L'approche de Parzen a été proposée par Rosenblatt (1956) dans le cas unidimensionnel puis par Parzen (1962).

- Supposons que la région $\mathcal{R}_n$ est un hypercube de côté h_n et de dimension d ;
- le volume est $V_n = h_n^d$;

Méthode des noyaux (Parzen) II

- Soit la fonction

$$K_1(u) = \begin{cases} 1 & \text{si } |u^j| < \frac{1}{2},\ \forall j \in \{1, \ldots, d\} \\ 0 & \text{sinon.} \end{cases}$$

- $K_1(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}) = 1$ si l'observation $\mathbf{x}_i$ tombe dans l'hypercube $\mathcal{R}_n$ centré autour de $\mathbf{x}$.
- le nombre d'observations de l'échantillon tombant dans l'hypercube $\mathcal{R}_n$ est

$$r_n = \sum_{i=1}^{n} K_1(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}).$$

Méthode des noyaux (Parzen) III

- l'estimateur de la densité au point $\mathbf{x}$ devient

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}K_1(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}).$$

- il est possible d'utiliser d'autres fonction $K()$ Pour garantir que $\hat{f}()$ définit bien une densité, il faut que

$$K(\mathbf{y}) > 0$$

et

$$\int \frac{1}{V_n}K(\mathbf{y})d\mathbf{y} = 1.$$

En effet, dans ce cas, ces deux conditions seront aussi satisfaites par $\hat{f}()$.

Exemples de noyaux

Noyau triangle

$$K_2(x) = \begin{cases} x+1 & \text{si } -1 < x \leq 0 \\ 1-x & \text{si } 0 < x \leq 1 \\ 0 & \text{sinon} \end{cases}$$

Noyau gaussien

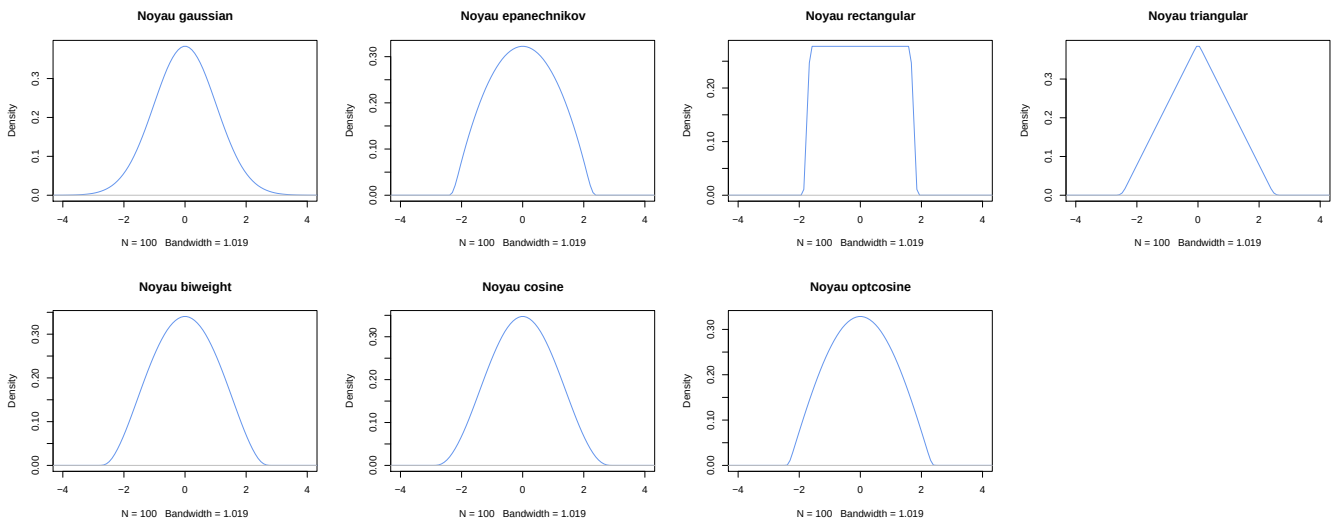
$$K_3(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}x^2)$$

Noyau d'Epanechnikov

$$K_4(x) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - x^2/5) & \text{si } |x| \leq \sqrt{5} \\ 0 & \text{sinon,} \end{cases}$$

Noyau de Lejeune

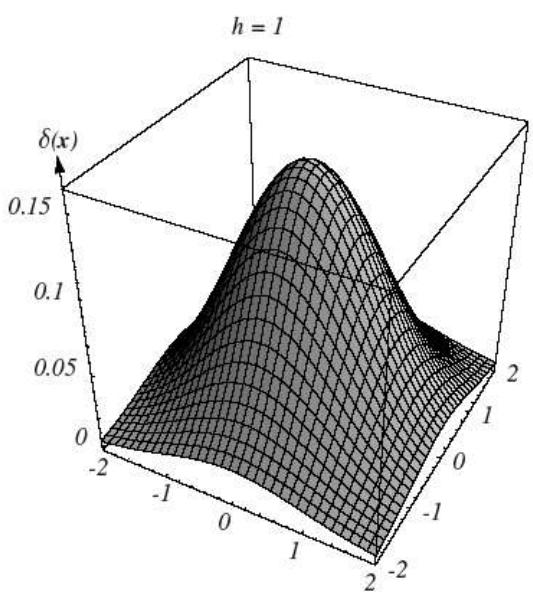
$$K_5(x) = \begin{cases} \frac{105}{64}(1-x^2)^2(1-3x^2) & \text{si } |x| \leq 1 \\ 0 & \text{sinon,} \end{cases}$$

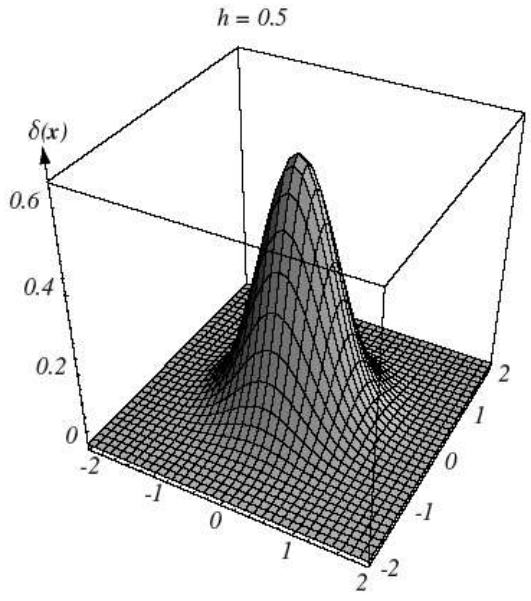


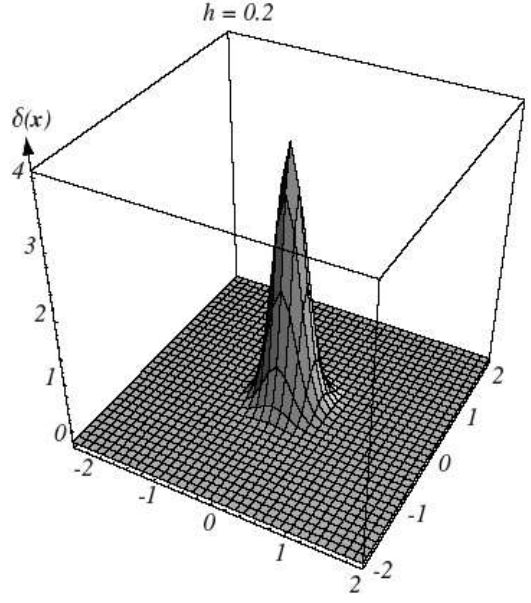
Largeur de bande I

- K pondère la contribution du vecteur $\mathbf{x}_i$
- La fonction fenêtre hypercubique et la gaussienne multivariée constituent des exemples de fonction K.
- nécessité de choisir un paramètre h, qui détermine "la zone d'influence" associée à chaque observation $\mathbf{x}_i$.

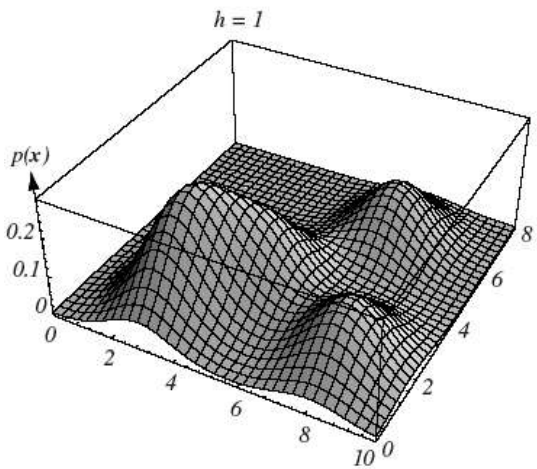
Largeur de bande II

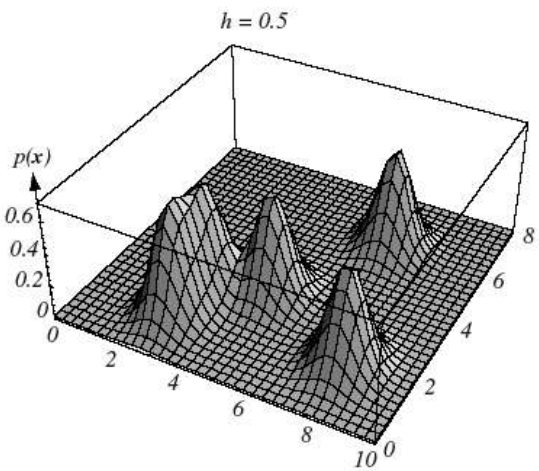


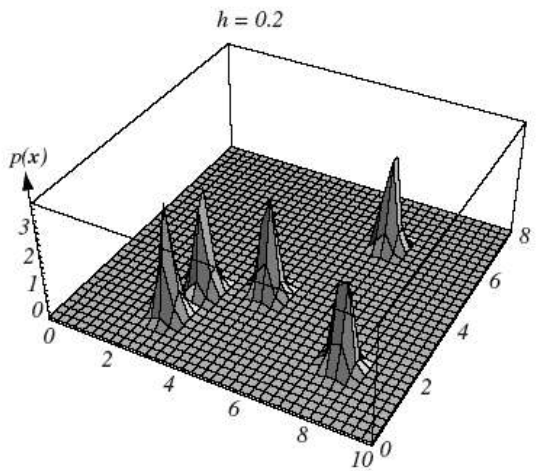




Largeur de bande III



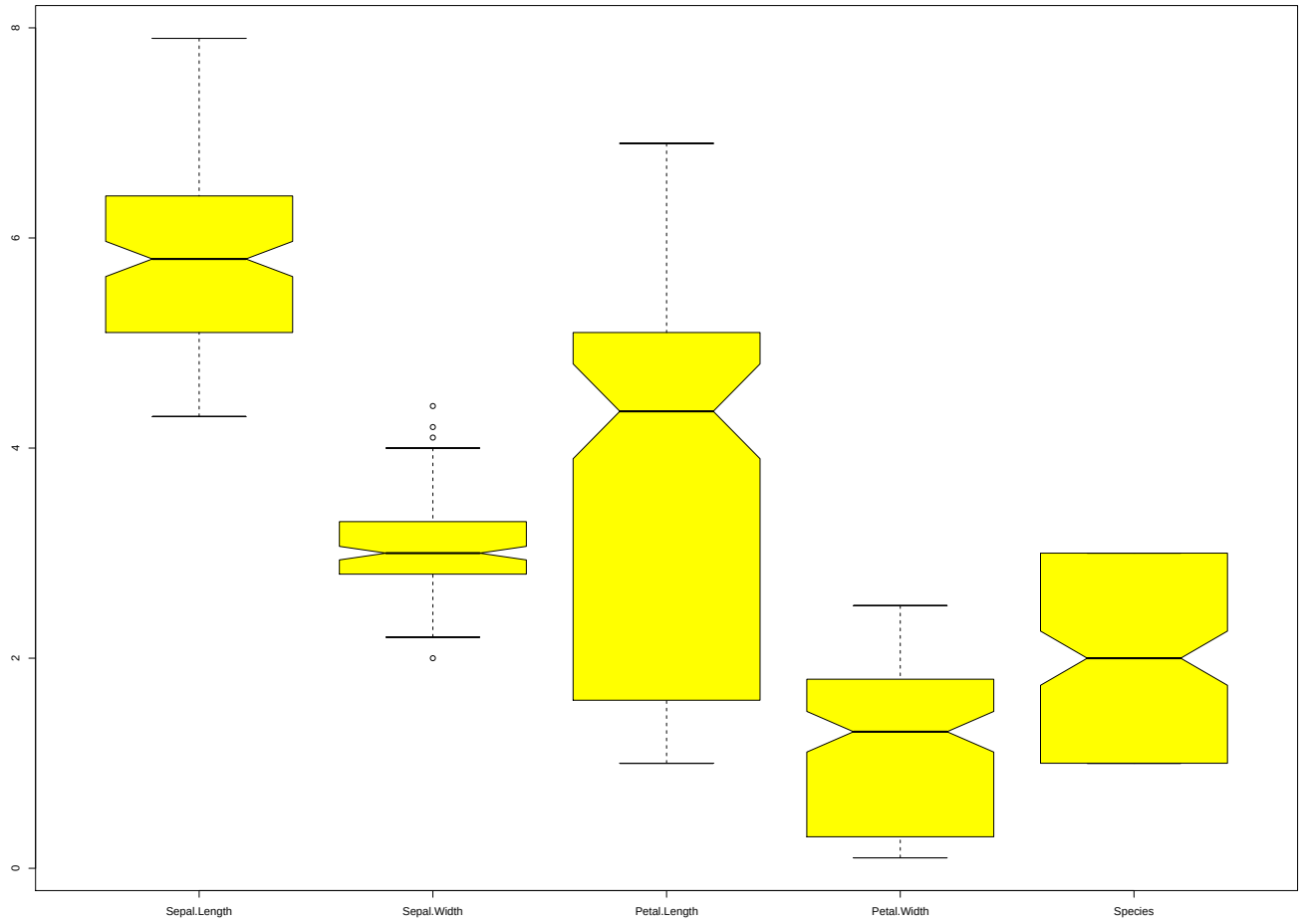




Boîte à moustaches ou boxplot

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

Exemple des iris



Conclusion

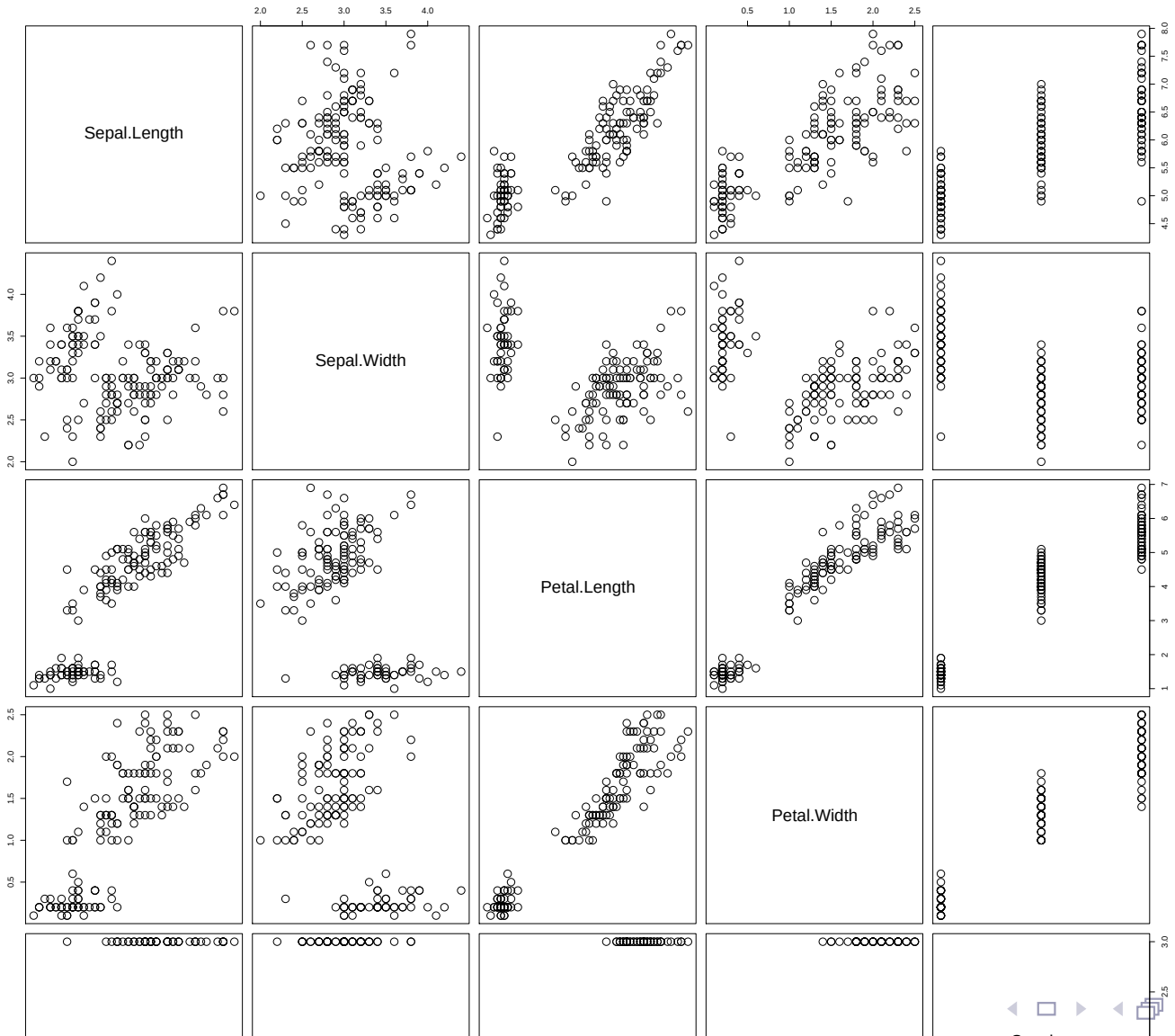
Mise en évidence de certaines caractéristiques :

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

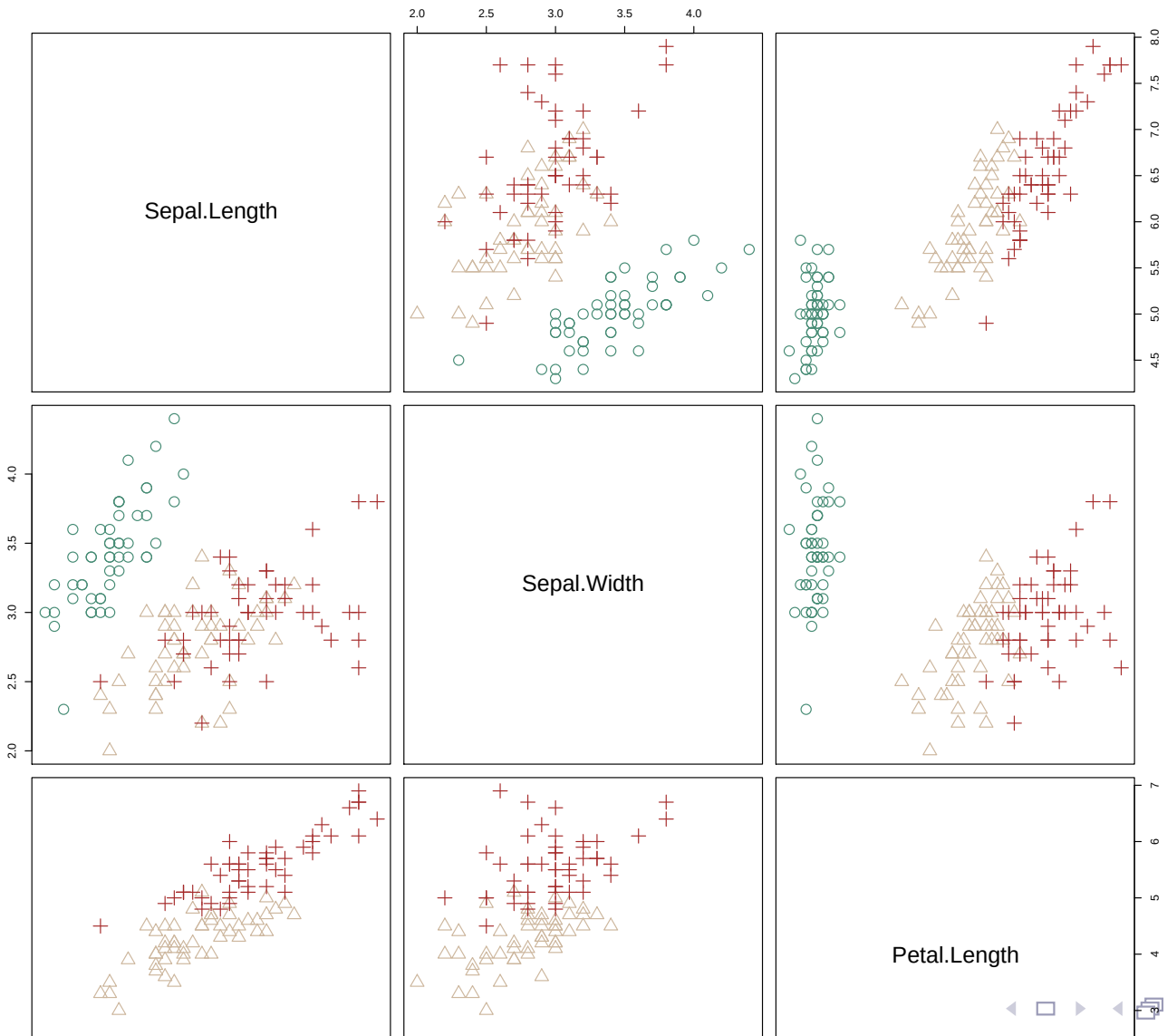
Graphique de dispersion

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

Les 5 variables des iris



Les 5 variables des iris en couleurs



Covariance et corrélation

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

Les iris

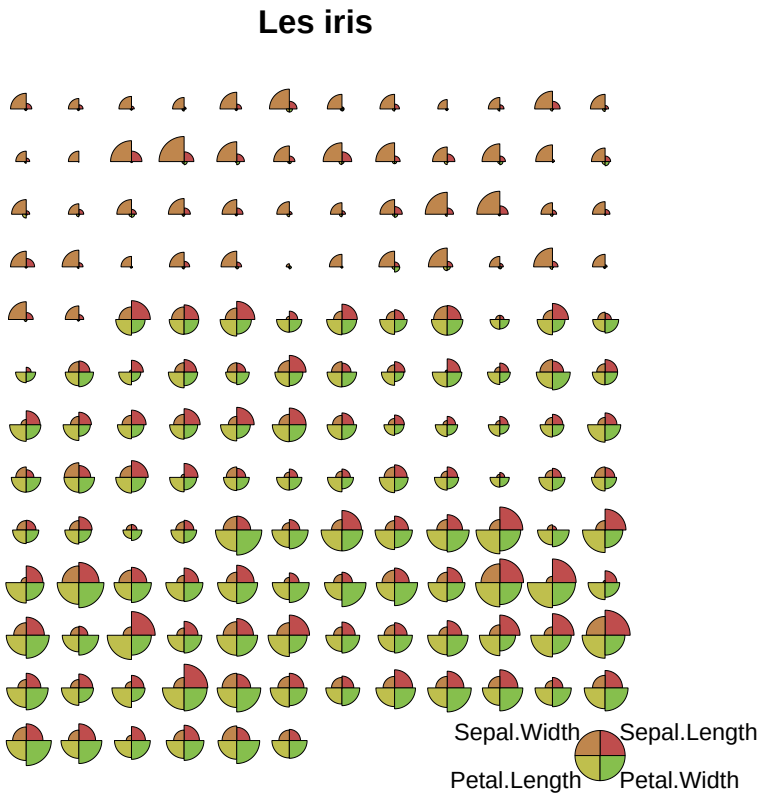
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.69	−0.04	1.27	0.52
Sepal.Width	−0.04	0.19	−0.33	−0.12
Petal.Length	1.27	−0.33	3.12	1.30
Petal.Width	0.52	−0.12	1.30	0.58

Tab.: Matrice de covariance

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.69	−0.04	1.27	0.52
Sepal.Width	−0.04	0.19	−0.33	−0.12
Petal.Length	1.27	−0.33	3.12	1.30
Petal.Width	0.52	−0.12	1.30	0.58

Tab.: Matrice de corrélation

Les diagrammes fleurs



Fléau de la dimension (curse of dimensionality) I

- Espace de grande dimension
- Calculs similaires à ceux du plan
- Mais difficile de généraliser
- Exemple 1 :
  - Dans $\mathbb{R}$
    - Pts uniformément répartis dans $[-1, +1]$
    - % de points situées à 1 distance ≤ 0.75 de l'origine : 75%
  - Dans $\mathbb{R}^{10}$
    - Pts uniformément répartis dans $[-1, +1]^{10}$
    - % de points situées à 1 distance ≤ 0.75 de l'origine : 5%
- Exemple 2 : on veut construire un histogramme en s'appuyant sur au moins une moyenne de 10 points par intervalle et 10 classes par variable
  - $\mathbb{R}$: 10 classes $n = 100$

Fléau de la dimension (curse of dimensionality) II

  - $\mathbb{R}^2$: 100 classes $n = 1000$
  - $\mathbb{R}^{10}$: 10^{10} classes $n = 10^{11} = 100 milliards$
- Si p assez grand, l'espace $\mathbb{R}^p$ est pratiquement vide et sauf si les données se situent au voisinage d'une variété de faible dimension, l'analyse des données n'apportera aucune information intéressante.
- Les points voisins d'un point donné sont tous très loin : difficultés dans l'emploi de méthodes du type k-plus proches voisins