Probabilités et Statistiques

Année 2010/2011

laurent.carraro@telecom-st-etienne.fr

olivier.roustant@emse.fr

telecom Saint-Etienne

Cours n°5

Statistique exploratoire

Plan

- Un problème : Peut-on reconnaître des variétés d'iris par les dimensions de leurs fleurs ?
- Données historiques (R. Fisher)
- Statistiques descriptives
  - Indicateurs chiffrés
  - Outils de visualisation : fonction de répartition empirique, histogramme, boxplot (boîtes à moustaches !), estimation non paramétrique d'une densité

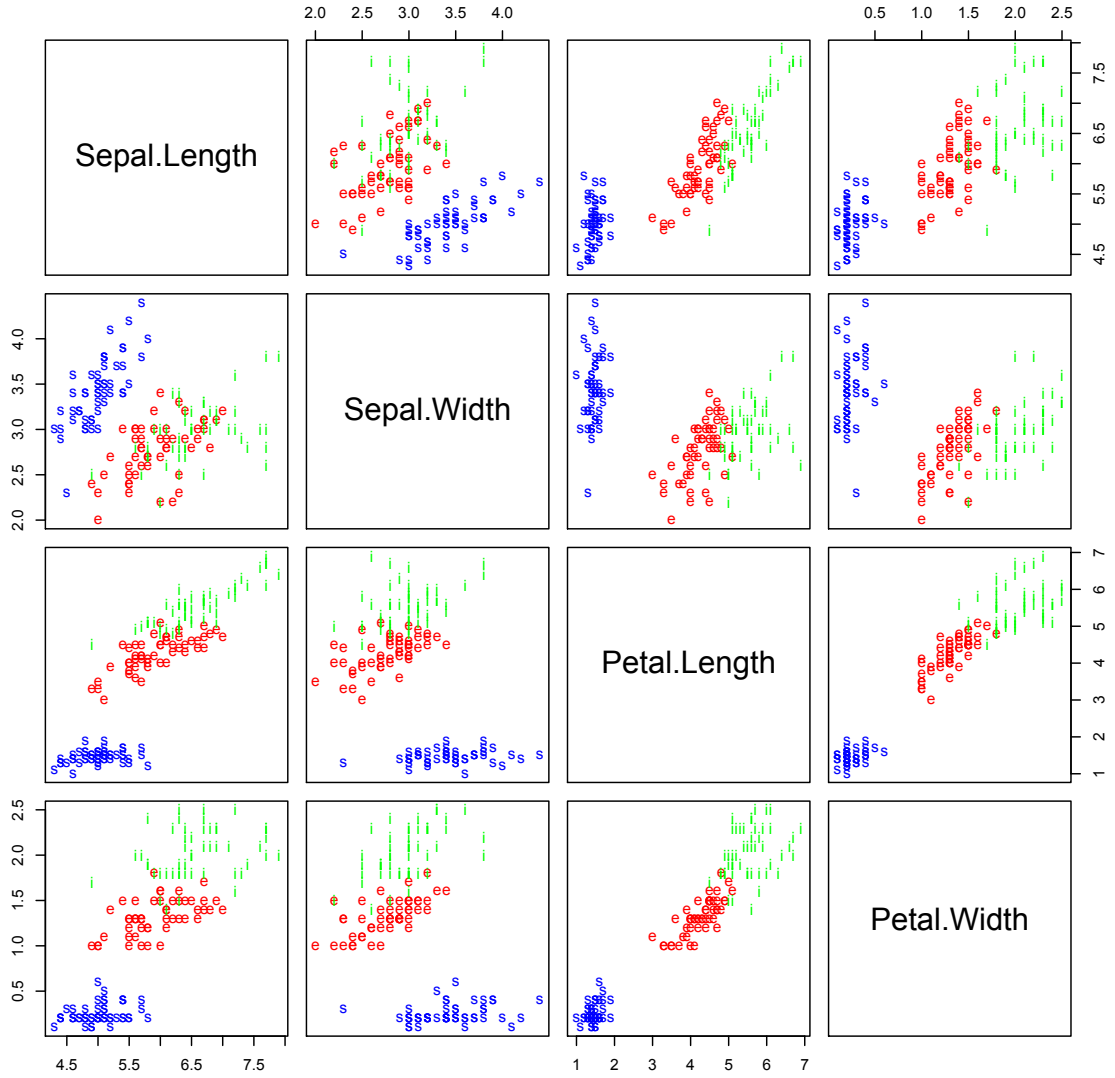
Les iris de Fisher

- **Question :**
  - Pour 3 variétés d'iris (setosa, versicolor, virginica), on mesure largeur et longueur du sépale et du pétale.
  - Les mesures permettent-elles de deviner la variété ?
- **Peut-on identifier des facteurs qui expliquent l'appartenance à un groupe ?**
  - Santé :
    - facteurs = résultats d'analyses
    - groupes = malades, sains
  - Etude financière :
    - facteurs = indicateurs macroéconomiques
    - groupes = ratings (cf. agences de notation)

Les données

numéro	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
...	...	...	...	...	...
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

Premier graphique



plot(iris[1:4], pch=c("s", "e", "i")[as.numeric(iris$Species)])

Premières observations

- Les dimensions du sépale semblent peu discriminantes
- On se concentre donc sur :
  - longueur pétale
  - largeur pétale

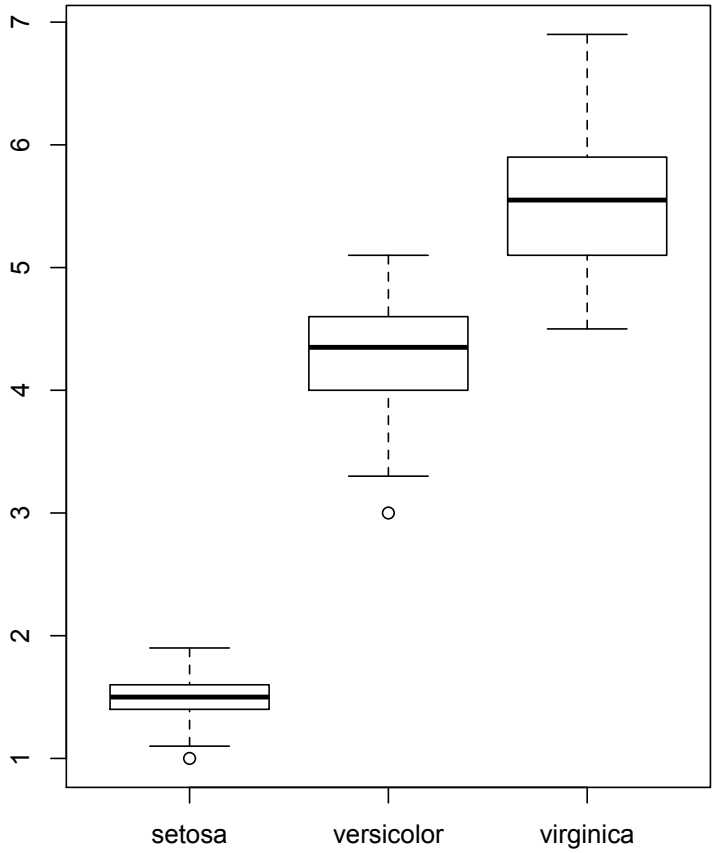
Résumés numériques

Longueur des pétales

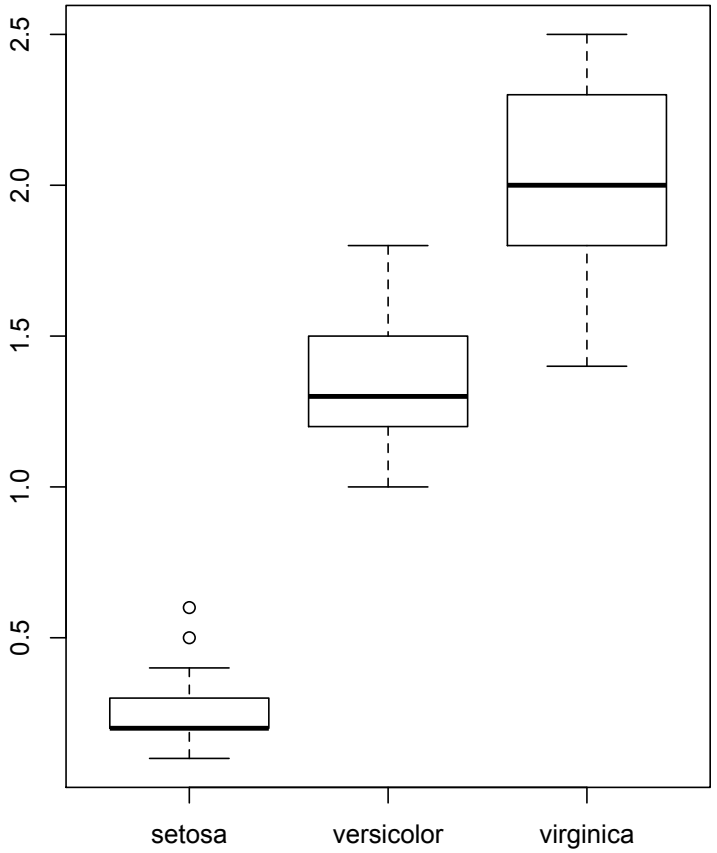
	Setosa	Versicolor	Virginica
moyenne	1.462	4.260	5.552
médiane	1.50	4.35	5.55
écart-type	0.174	0.470	0.552
interquartiles	0.175	0.600	0.775
quantile 5%	1.200	3.39	4.845
quantile 95%	1.700	4.90	6.655
quantile 25%	1.400	4.00	5.100
quantile 75%	1.575	4.60	5.875

```
Fonction utiles : mean, median, sd, quantile
```

Boxplot (boîte à moustaches)



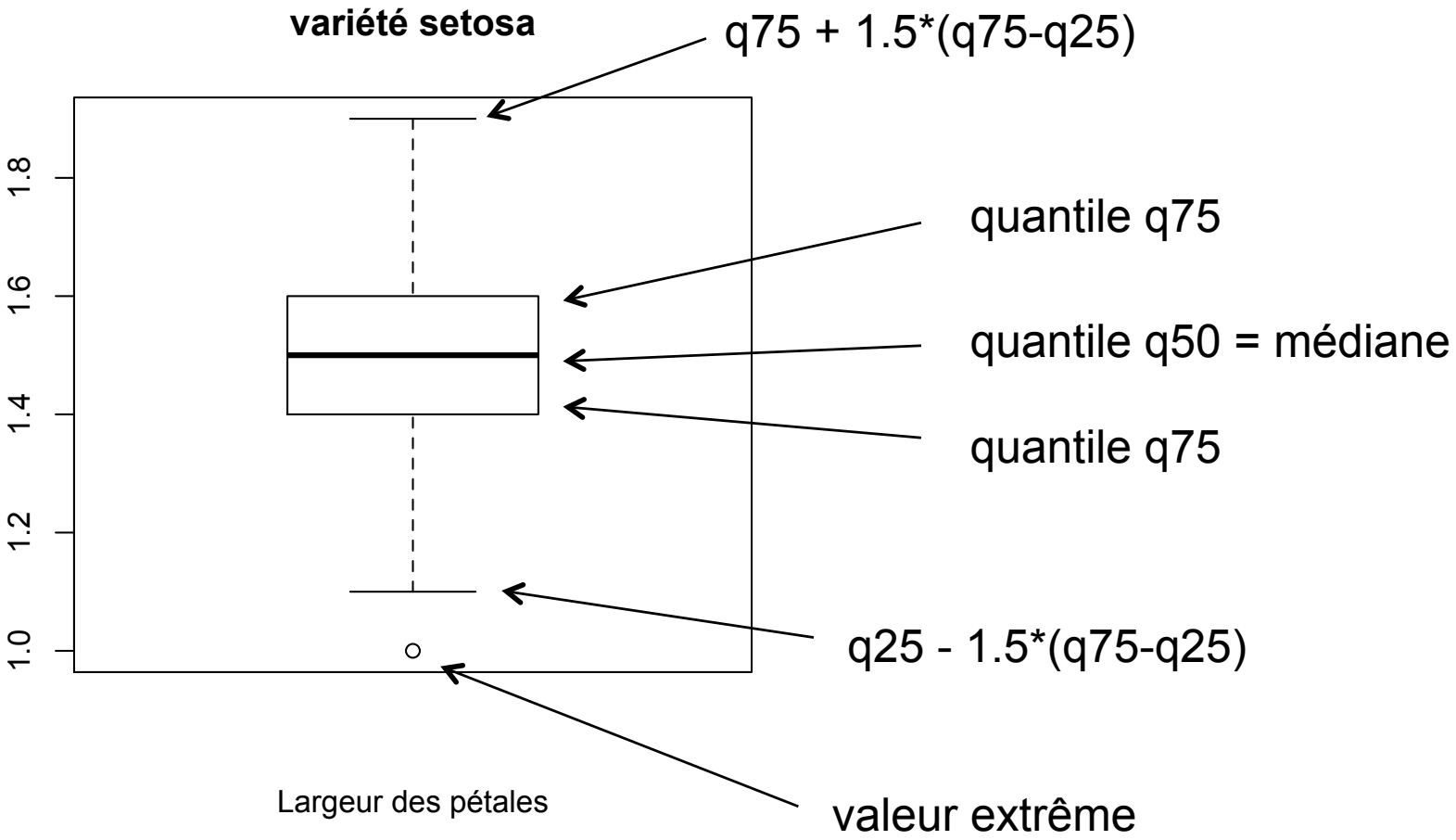
Longueur des pétales



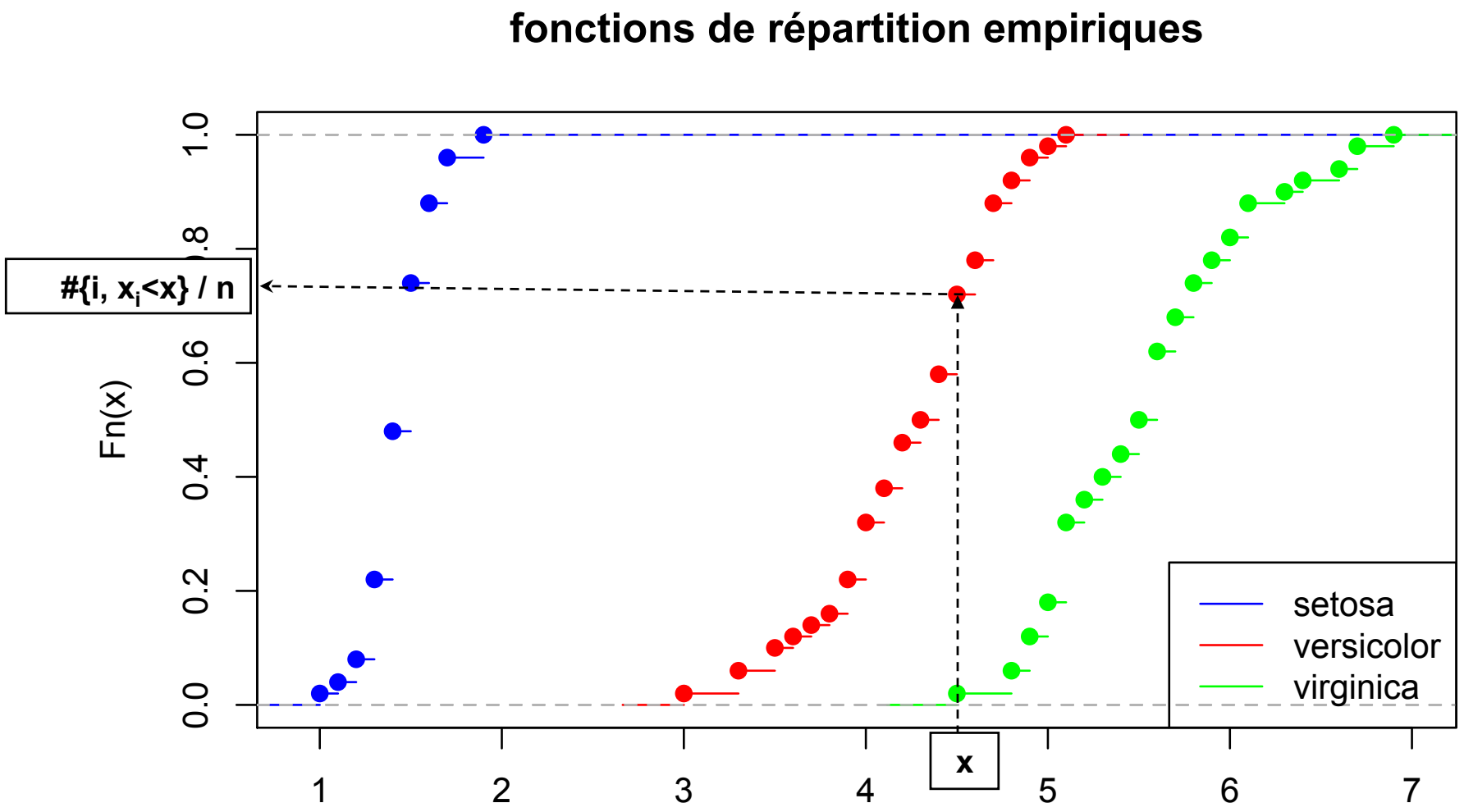
Largeur des pétales

```
boxplot(iris[,3]~iris$Species,xlab="Longueur des pétales")
boxplot(iris[,4]~iris$Species,xlab="Largeur des pétales")
```

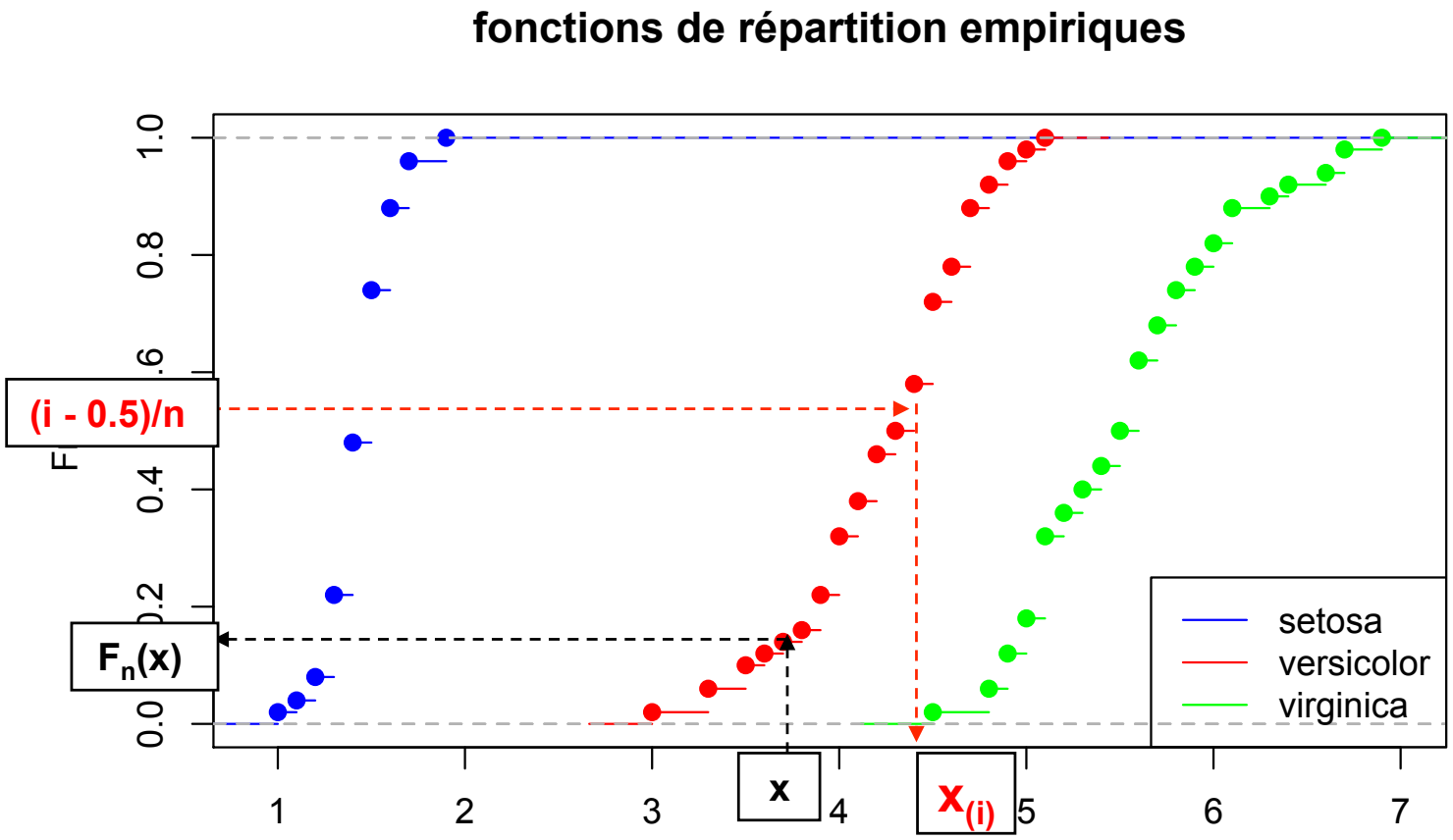
Comment est faite la boîte ?



Fonction de répartition empirique

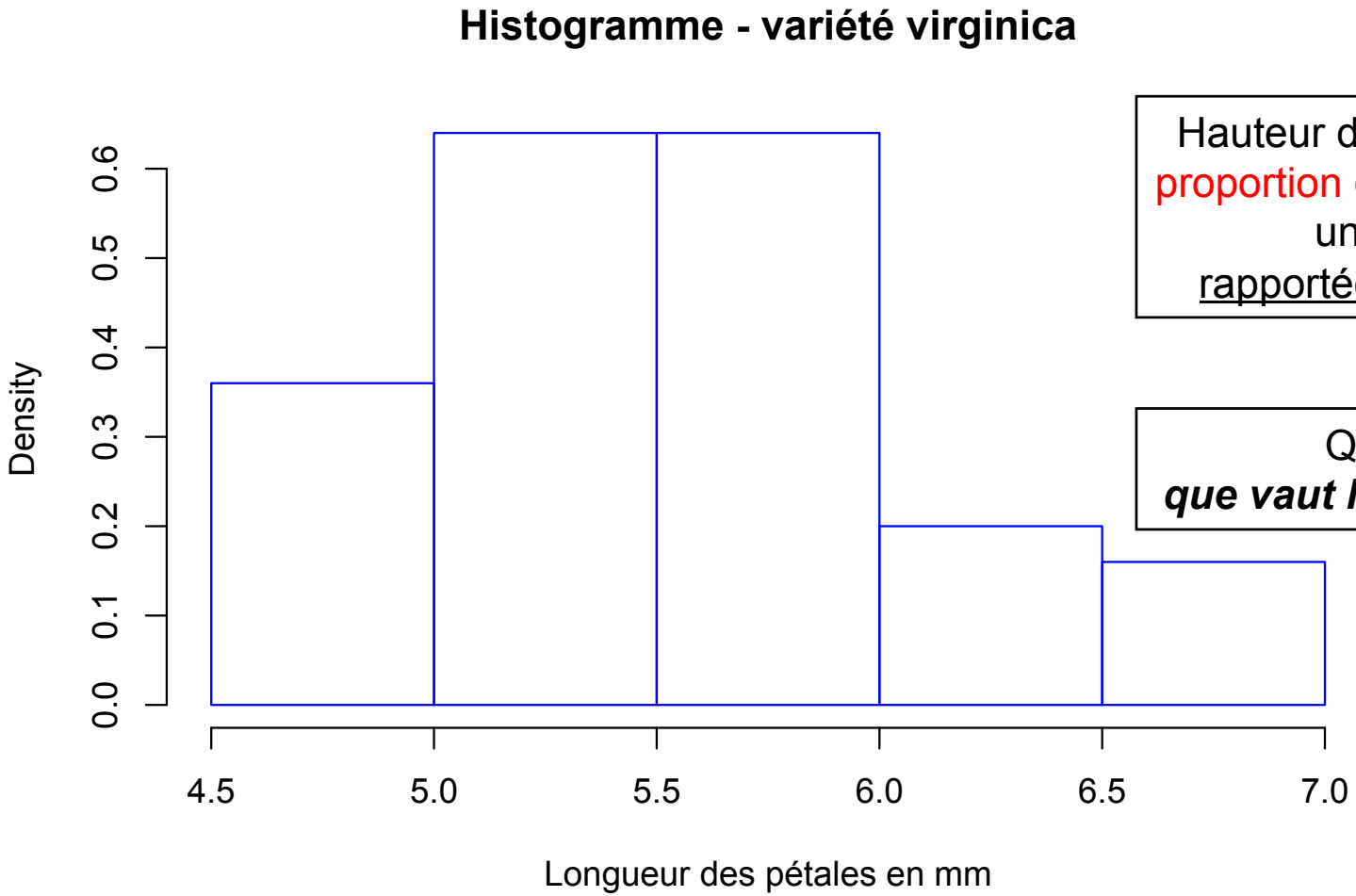


Quantiles empiriques



Si : $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ sont les données classées dans l'ordre croissant
: $x_{(i)} = q((i-0.5)/n)$ **quantile empirique** d'ordre (i-0.5)/n

Histogramme

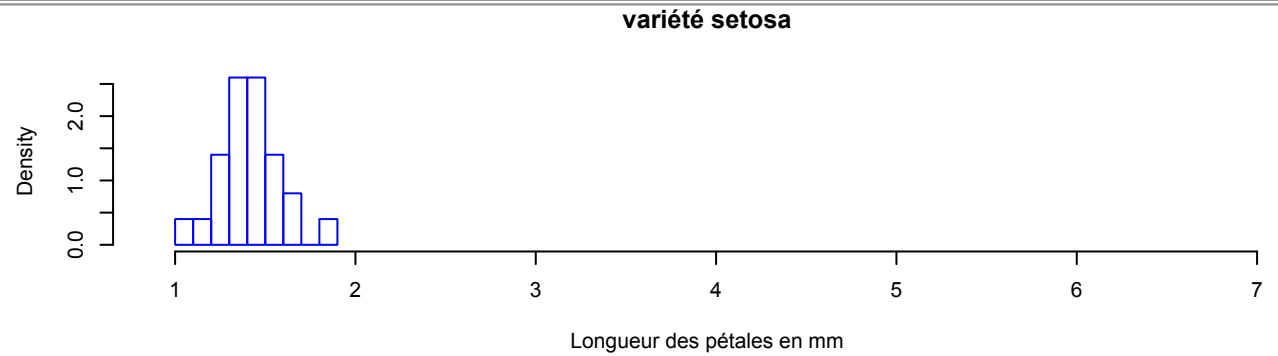


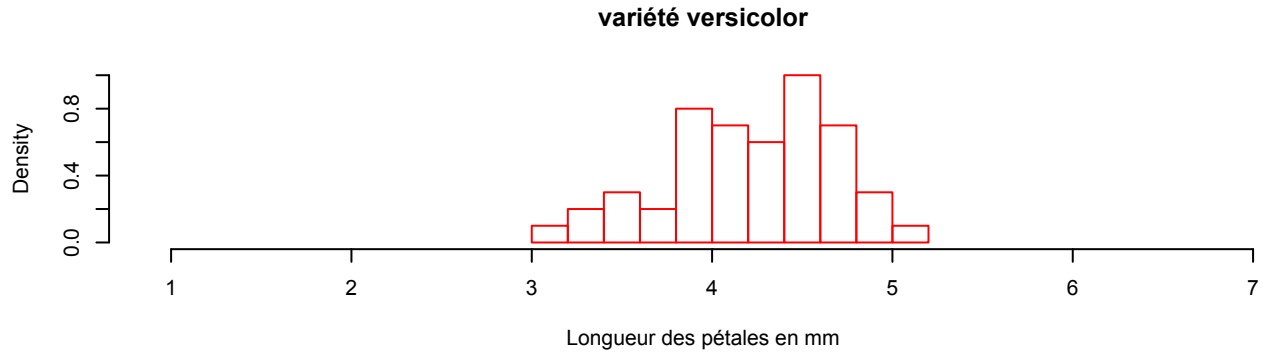
Hauteur de chaque barre :
proportion des données dans une classe,
rapportée à sa longueur

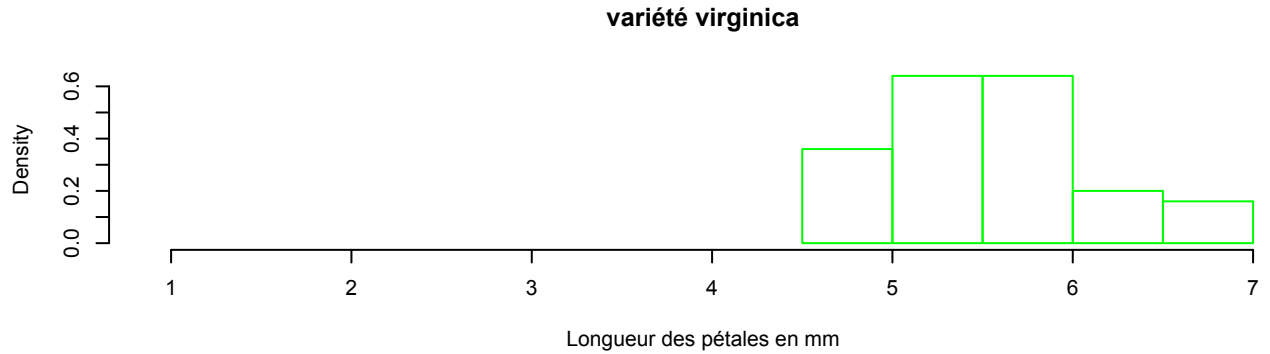
Question :
que vaut la surface totale?

```
hist(Petal.Length[Species=="virginica"], freq=FALSE, border="blue",
xlab="Longueur des pétales en mm",main="Histogramme - variété
virginica")
```

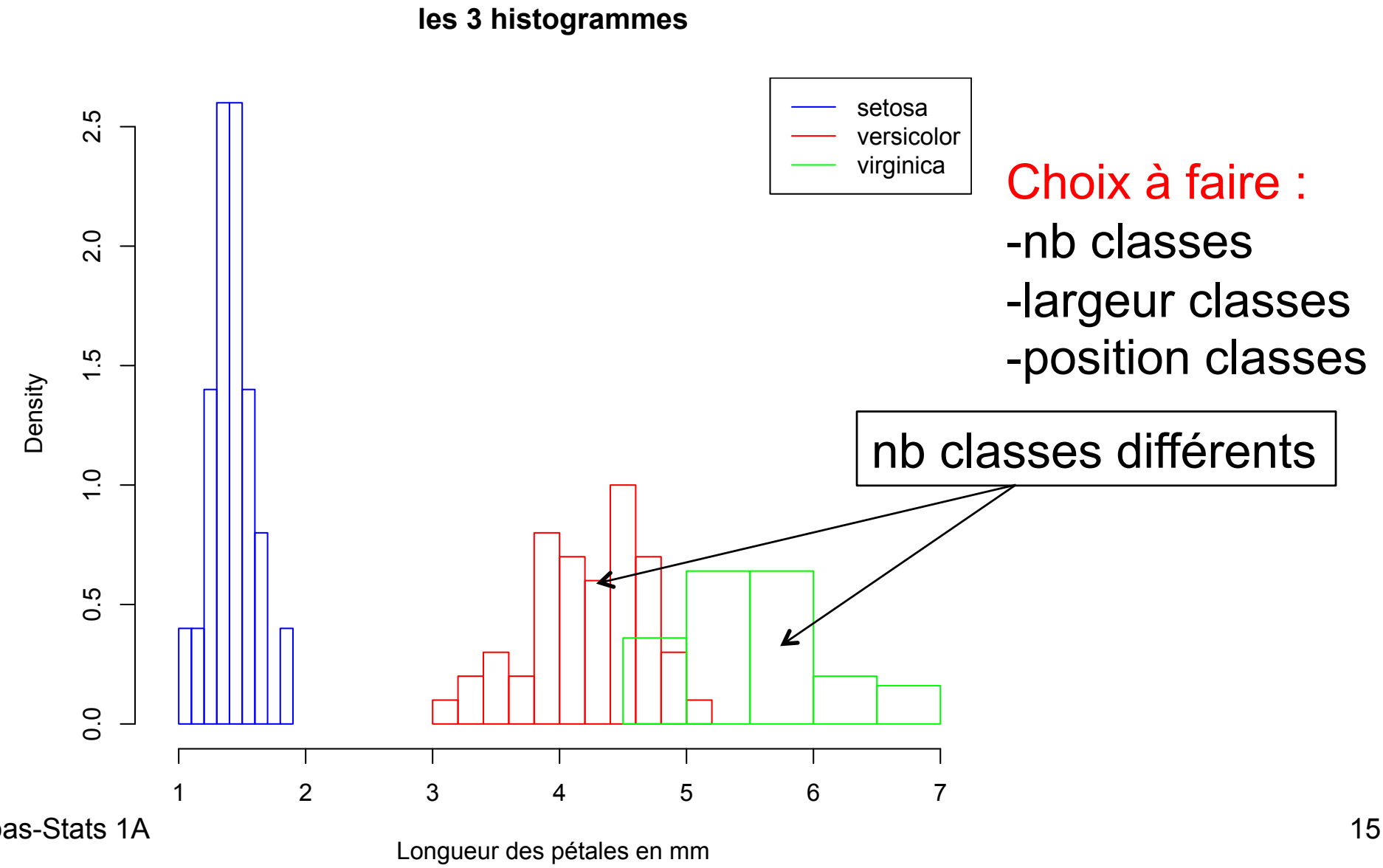
Les trois histogrammes







Histogrammes superposés



Choix à faire :

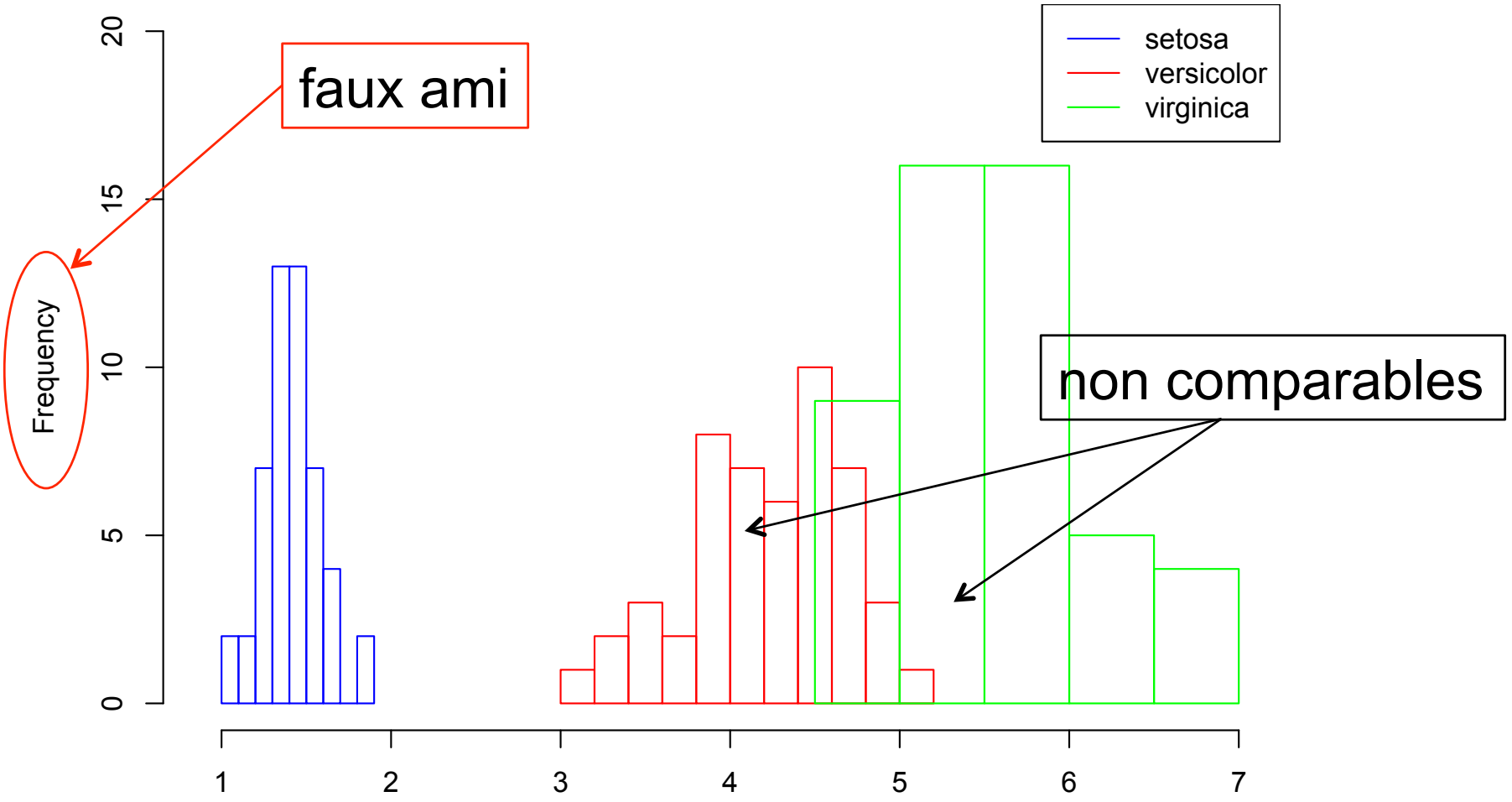
-nb classes

-largeur classes

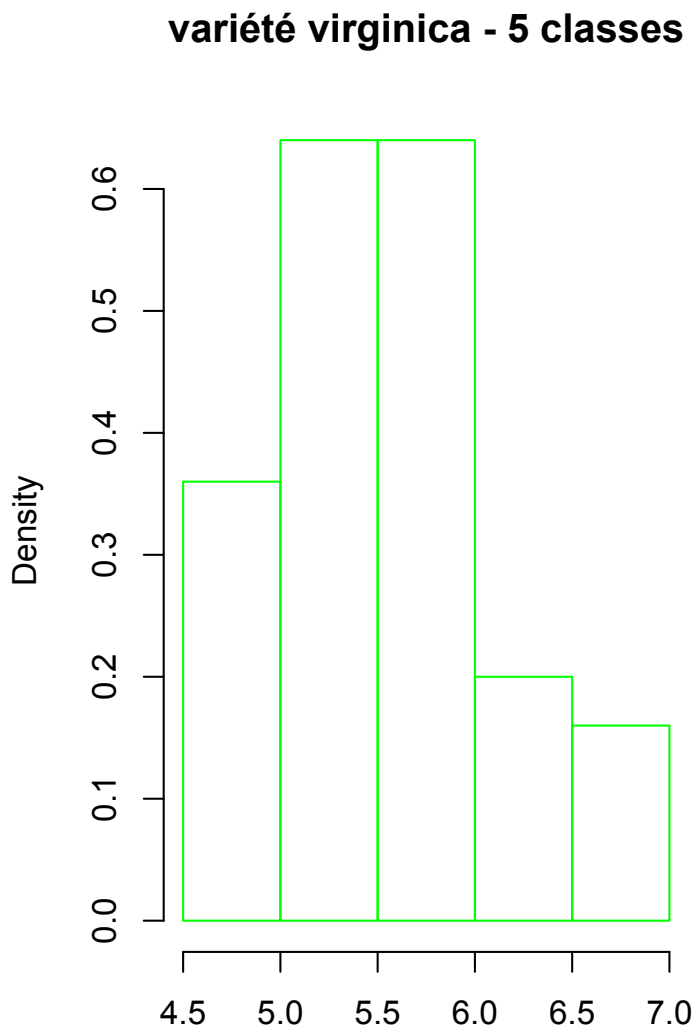
-position classes

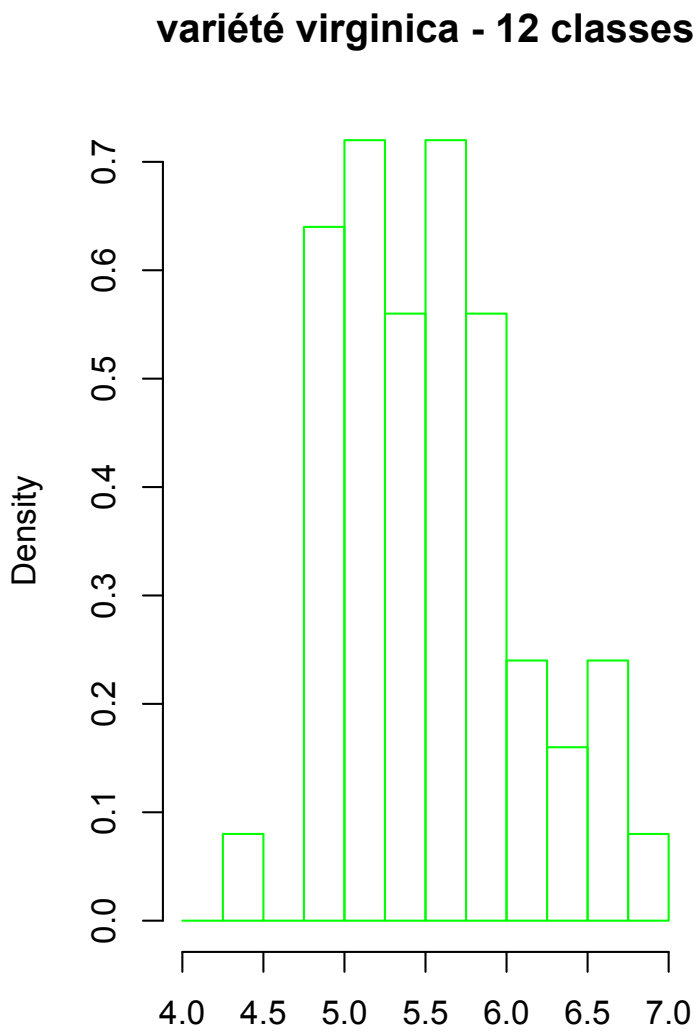
nb classes différents

Histogrammes superposés



Influence du nombre de classes





Estimation de densité

- Rappel : $$f_X(x) = \frac{P(X \in [x, x+dx])}{dx}$$
- Histogramme :

  Pour x dans la classe [a,b]

$$f_X(x) \approx \frac{Card\{x_i \in [a,b]\}/n}{b-a}$$

- Estimation de densité :

$$\hat{f}_X(x) = \frac{Card\{x_i \in [x-h, x+h]\}/n}{2h}$$

Interprétation (filtrage)

- Soit P_n la probabilité empirique :

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

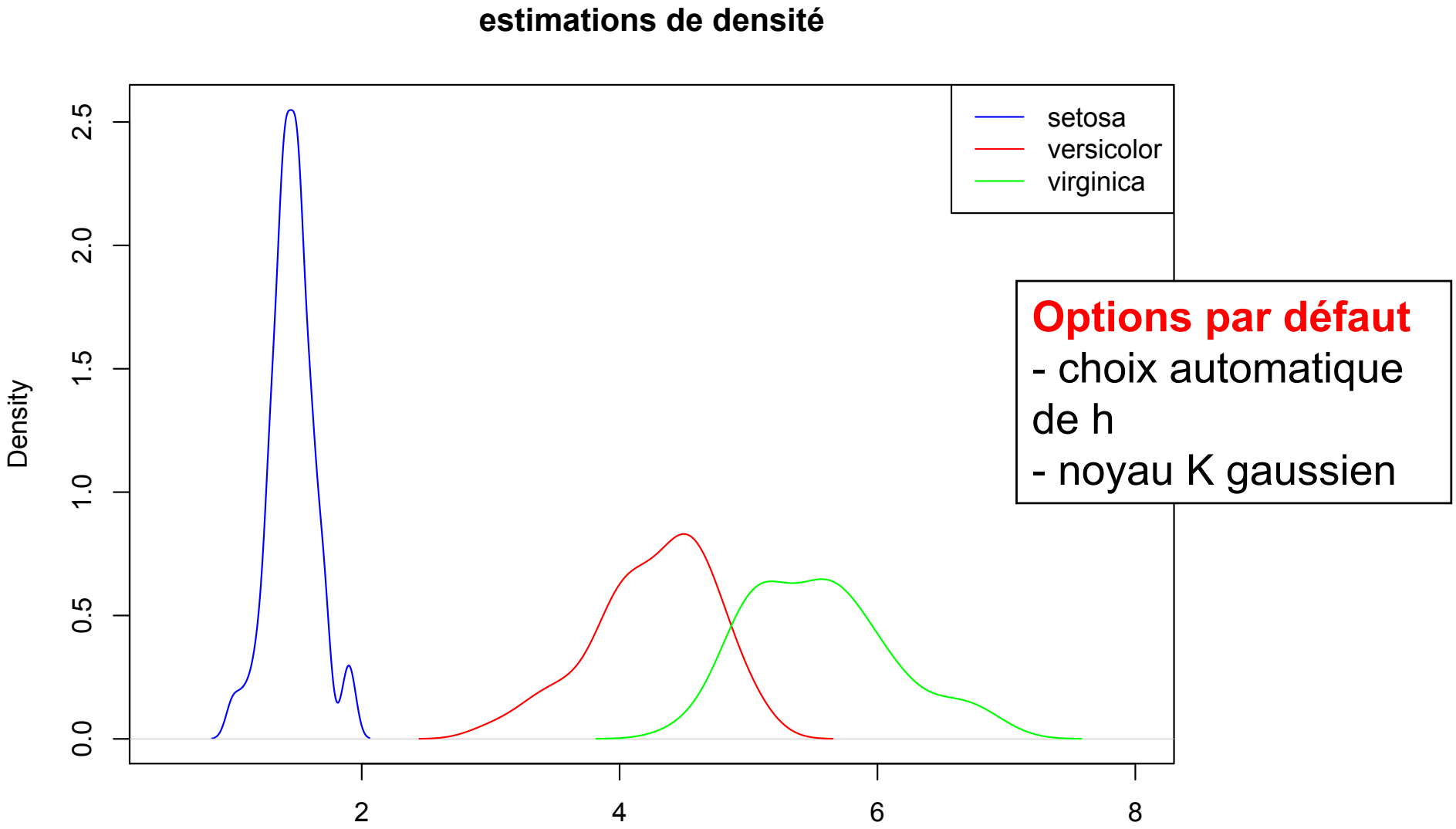
- Alors : $\hat{f}_X = K_h * P_n$

$$K_h(x) = 1/h\, K(x/h), \text{ où } K(u) = 1/2\ 1_{[-1,1]}(u)$$

- Pour K quelconque (densité de probabilité) :

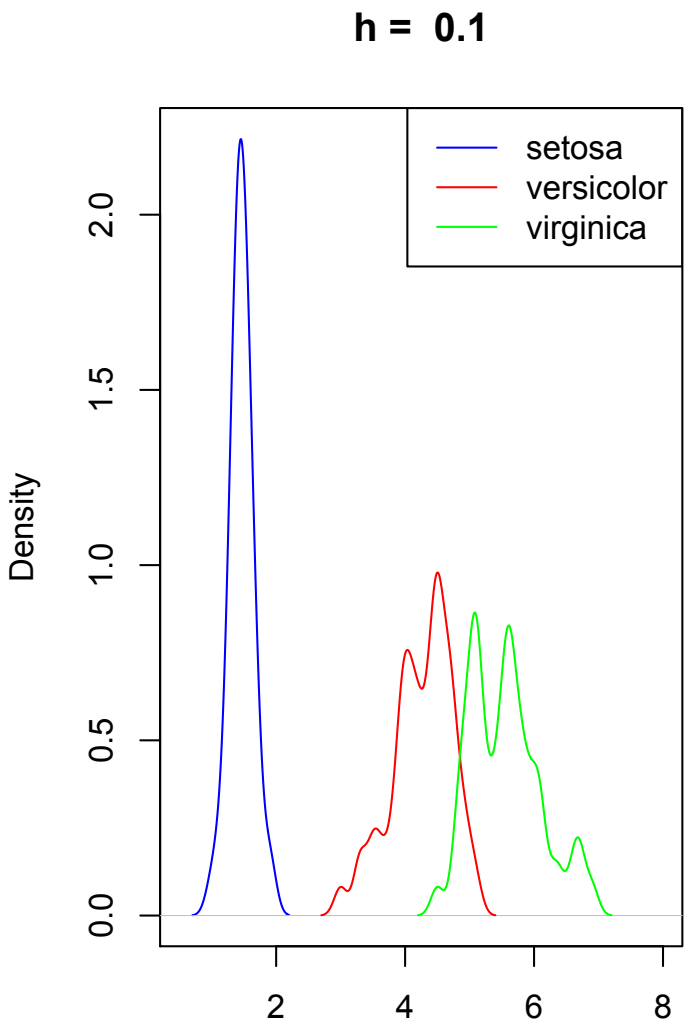
$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

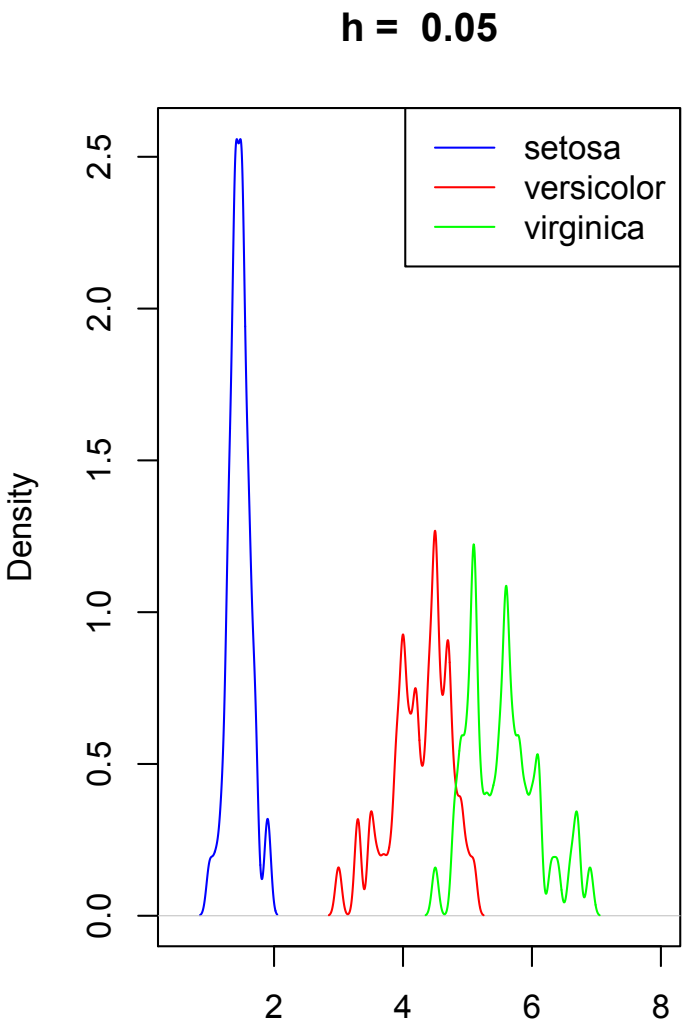
Estimation de densité



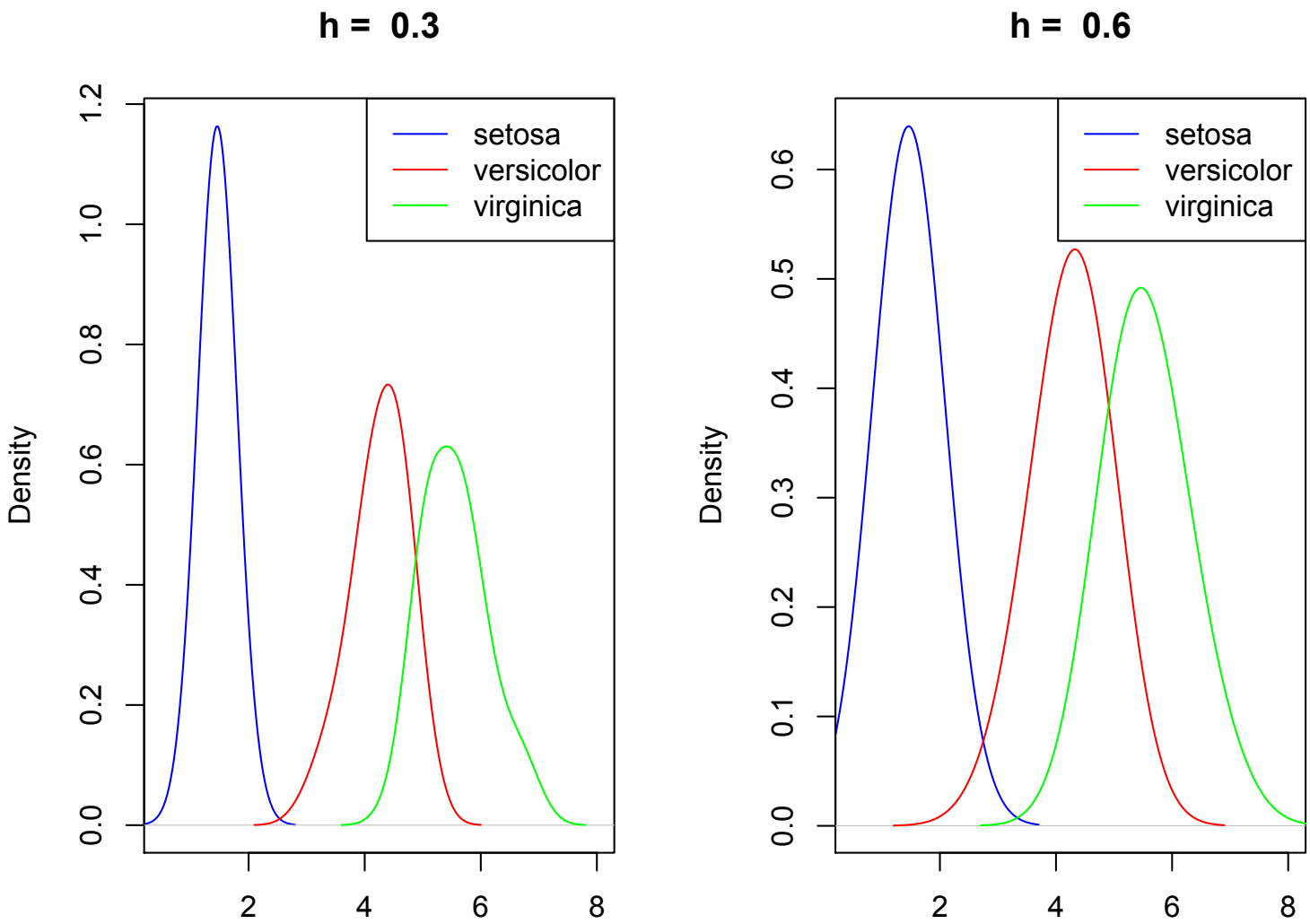
Options par défaut
- choix automatique de h
- noyau K gaussien

Influence de h (bandwidth)





Influence de h (bandwidth)



Conclusion ?

- Séparation des variétés :
  - Si Petal.Length < 2 : setosa
  - Si 2 < Petal.Length < 4.5 : versicolor
  - Si Petal.Length > 5.1 : virginica
  - Si 4.5 < Petal.Length < 5.1 : ???
- Pour aller plus loin :
  - règles valables hors de l'échantillon observé ?
  - donner une probabilité d'appartenance à la variété
  - raisonner en multidimensionnel (c'est l'**analyse discriminante**)