
Probabilités et Statistiques

Année 2009/2010

laurent.carraro@telecom-st-etienne.fr

olivier.roustant@emse.fr

Cours n°5

Statistique exploratoire

Plan

- Un problème : un traitement est-il efficace ?
- Des données aux probabilités : modélisation
- Statistiques descriptives
 - Indicateurs chiffrés
 - Outils de visualisation : fonction de répartition empirique, histogramme, boxplot (boîtes à moustaches !), estimation non paramétrique d'une densité
 - Comparaison à une transformation affine près : qq-plot, droite de Henri

Les faiseurs de pluie

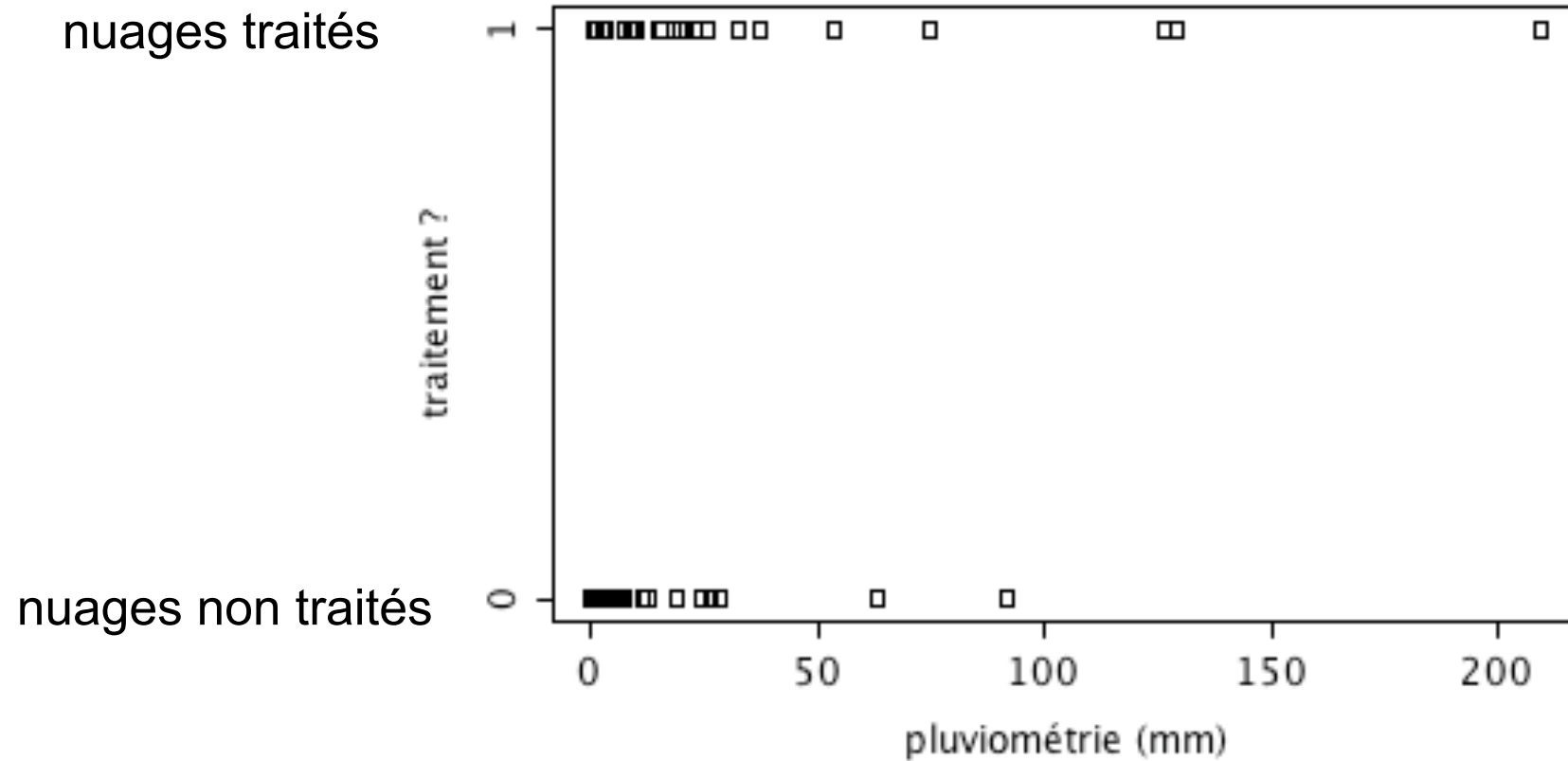
➤ Question :

- Une société propose un traitement des nuages à base de nitrate d'argent pour augmenter la pluviométrie.
- Quelle est l'efficacité de ce traitement ?

➤ Protocole expérimental

- Sur 26 nuages choisis au hasard, application du traitement et mesure de la pluviométrie
- Sur 26 autres nuages, choisis au hasard, sans rapport avec les nuages traités, mesure de la pluviométrie

Données



Notation et modélisation

- $x_1, \dots, x_n$: pluviométries des nuages non traités
- $y_1, \dots, y_n$: pluviométries des nuages traités

- *Hypothèses :*
 - $x_1, \dots, x_n$ sont des réalisations de v.a. $X_1, \dots, X_n$, indépendantes et de même loi μ_X
 - *Vocabulaire : on dit que $x_1, \dots, x_n$ est un **échantillon** de la loi μ_X*
 - $y_1, \dots, y_n$ sont des réalisations de v.a. $Y_1, \dots, Y_n$, indépendantes et de même loi μ_Y
 - $X_1, \dots, X_n, Y_1, \dots, Y_n$ sont indépendantes

- Reformulation du problème ?

Reformulation du problème

➤ Le traitement est efficace si :

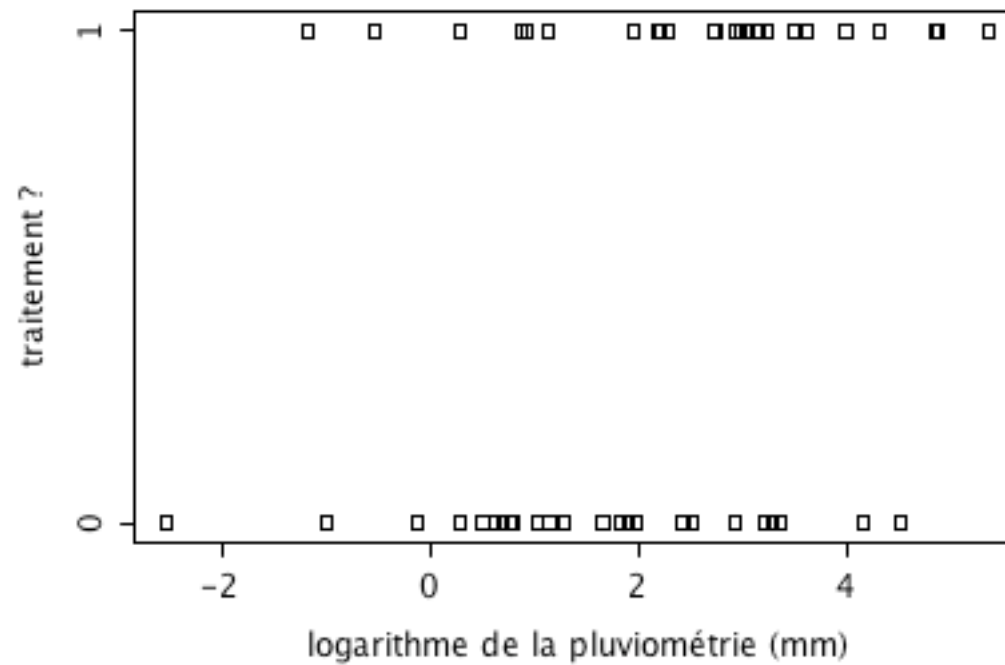
- Pour tout x , la probabilité pour que la pluviométrie dépasse x est plus grande pour les nuages traités que pour les nuages non traités :
- i.e. $P(Y \geq x) > P(X \geq x)$
- i.e. $F_Y(x) < F_X(x)$
- avec F_X fonction de répartition des X_i , et F_Y fonction de répartition des Y_j
- Si tel est le cas, quel lien peut-on donner entre $F_X(x)$ et $F_Y(x)$?

Quelques indicateurs statistiques

	Sans traitement	Avec traitement
POSITION		
Moyenne	12.5	33.7
Médiane	3.37	16.9
DISPERSION		
Ecart-type	21.2	49.6
q(75%) - q(25%)	10.2	23.5
q(5%)	0.37	0.78
q(95%)	54.5	128.6

Fonction utiles : mean, median, sd, quantile

Transformation des données

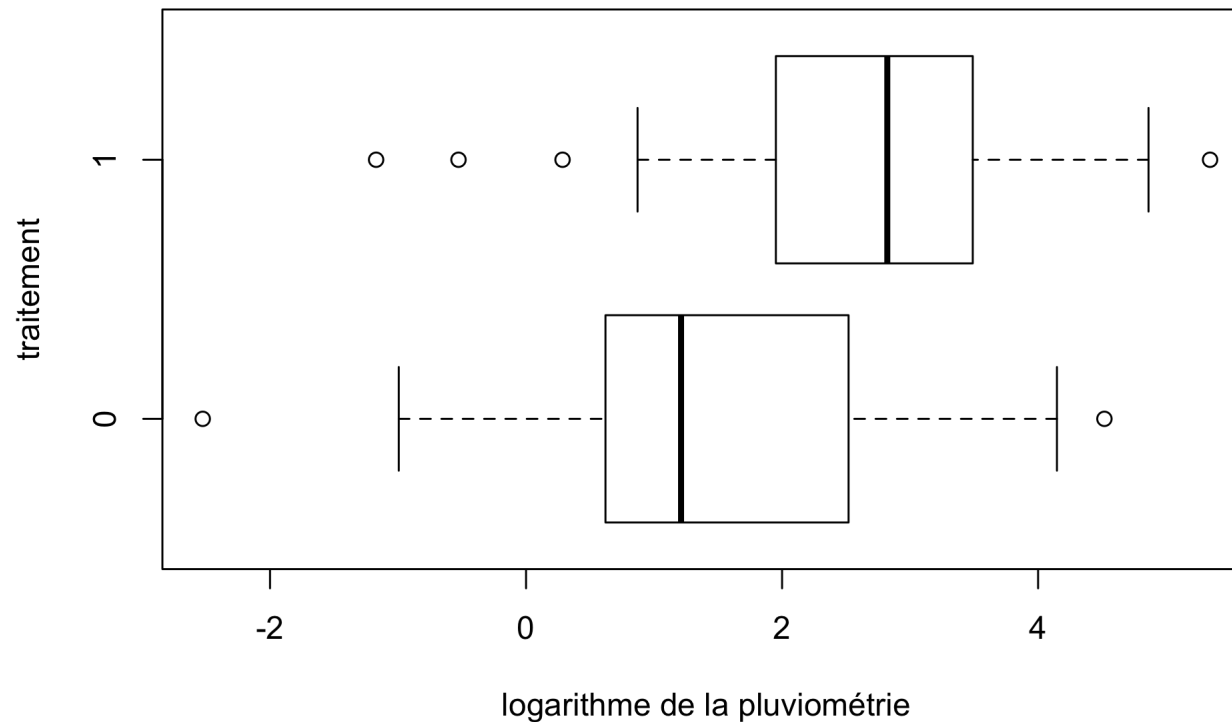


Indicateurs pour le log

	Sans traitement	Avec traitement
POSITION		
Moyenne	1.42	2.56
Médiane	1.21	2.82
DISPERSION		
Ecart-type	1.64	1.60
q(75%) - q(25%)	1.86	1.42
q(5%)	- 0.99	- 0.32
q(95%)	3.95	4.86

Fonction utiles : mean, median, sd, quantile

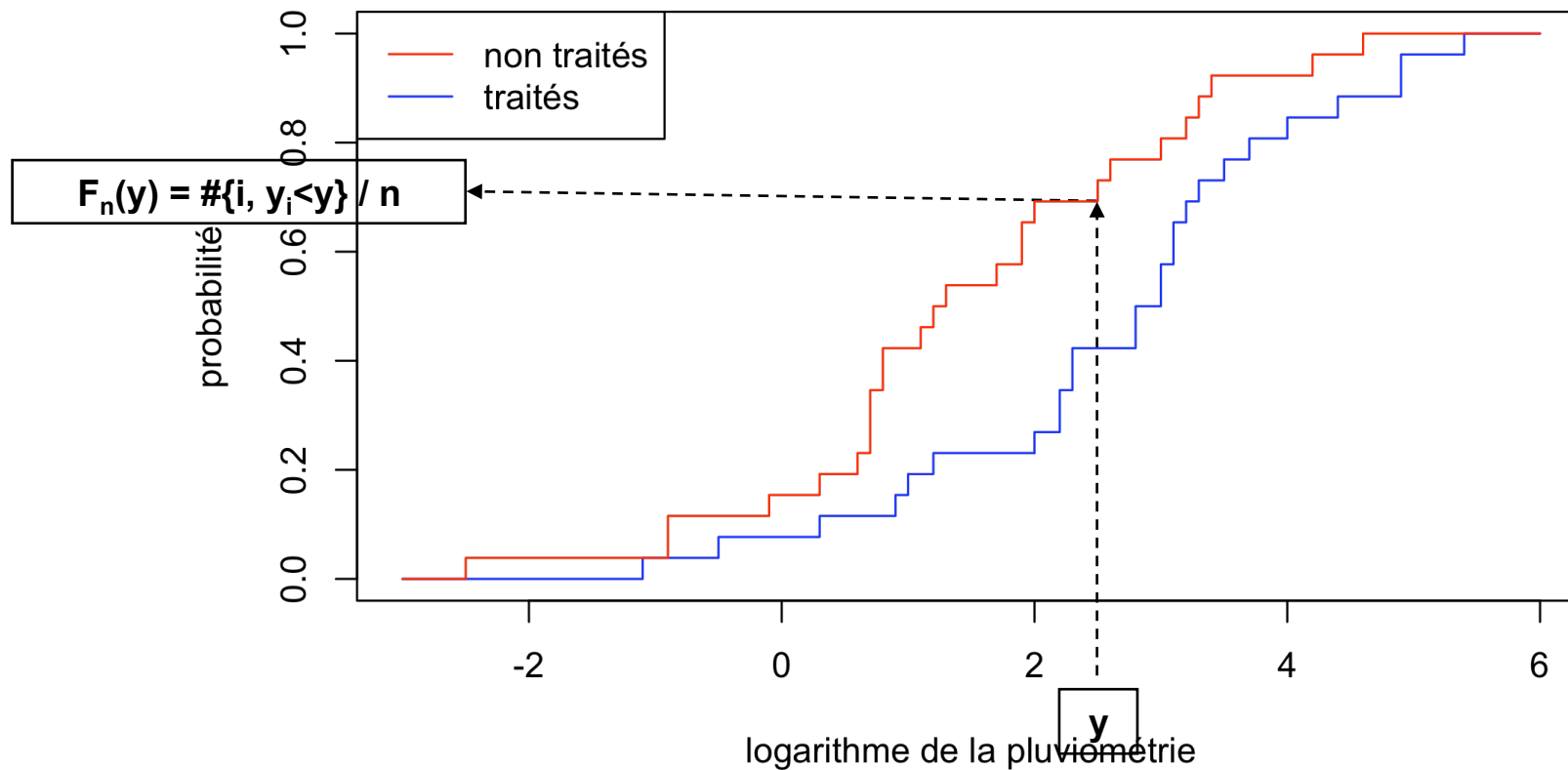
Boxplot (boîte à moustaches)



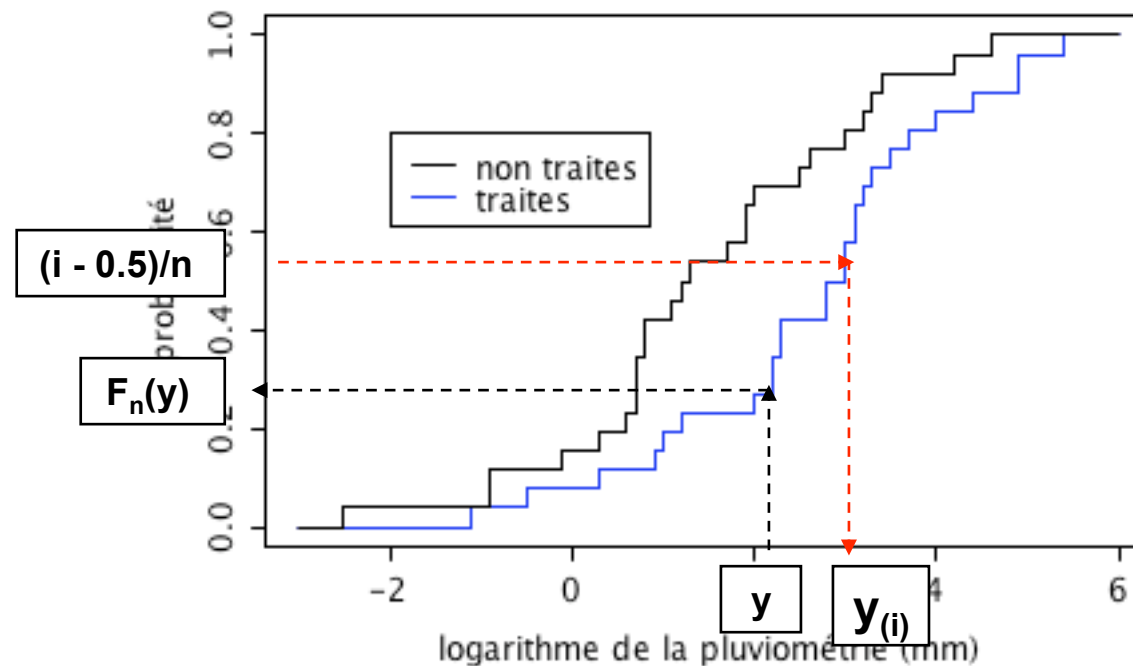
```
with(data=pluie, boxplot(log(hauteur)~traitement, horizontal=TRUE, range=1,
xlab="logarithme de la pluviométrie (mm)", ylab="traitement?"))
```

Fonction de répartition empirique

fonctions de répartition empiriques

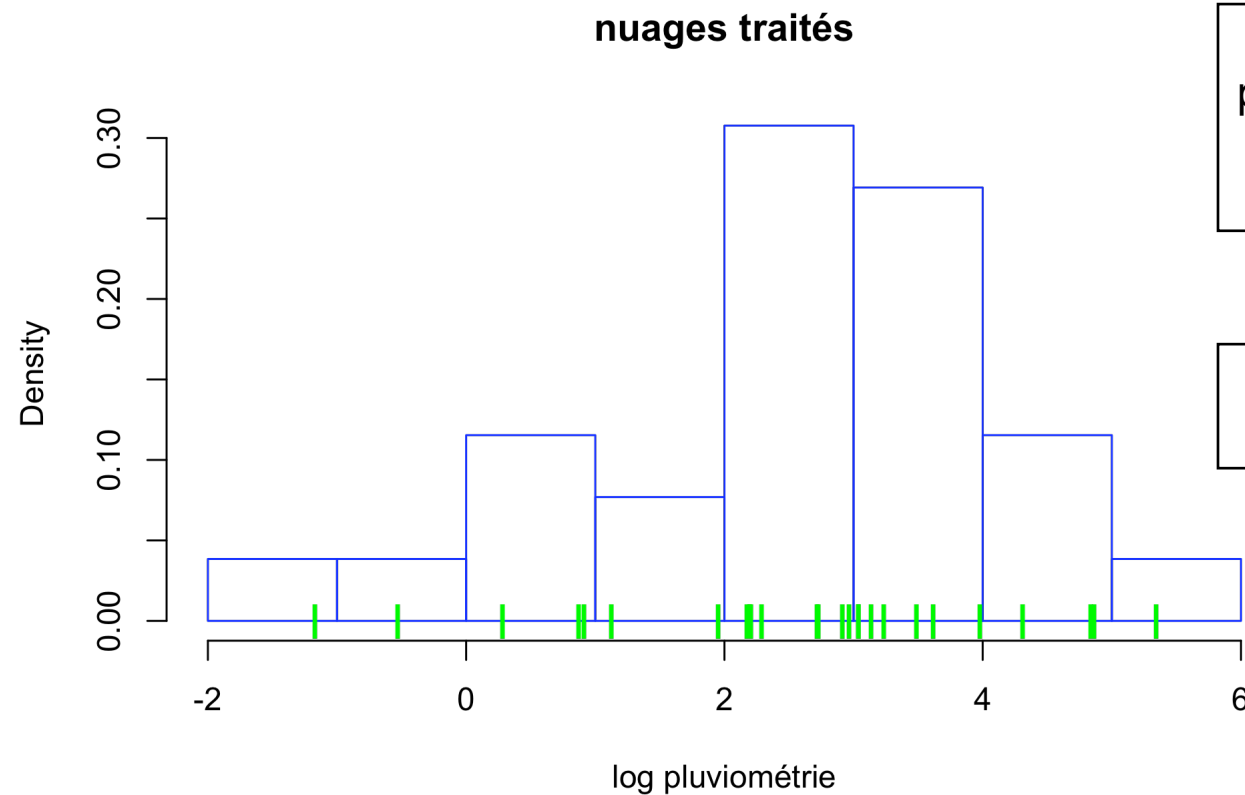


Quantiles empiriques



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

Histogramme

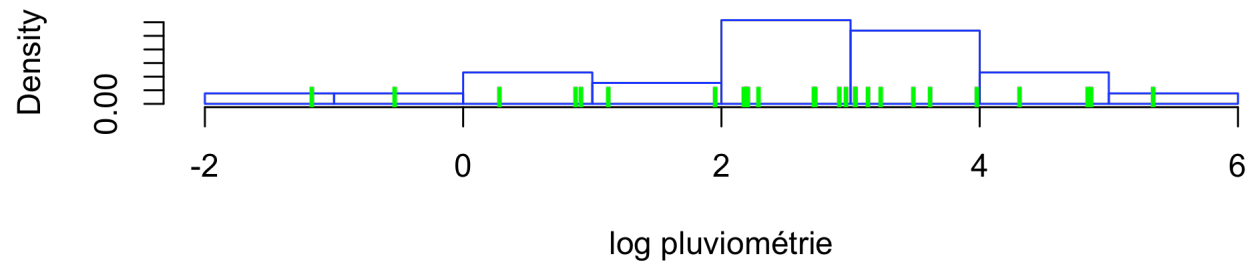


Hauteur de chaque barre :
proportion des données dans
une **classe**,
rapportée à sa longueur

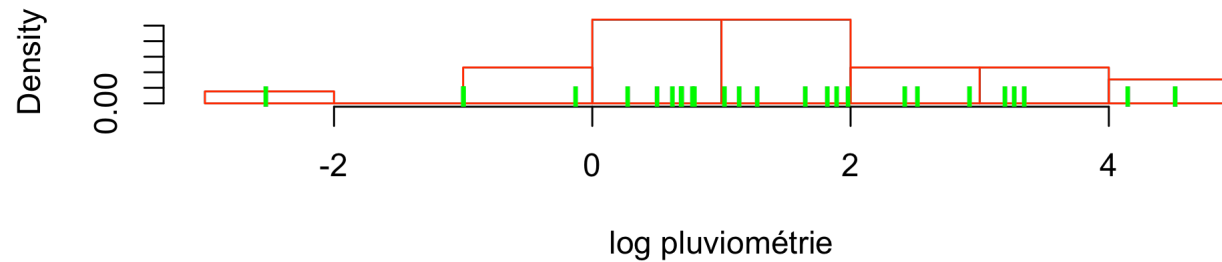
Question :
que vaut la surface totale?

Les deux histogrammes

nuages traités

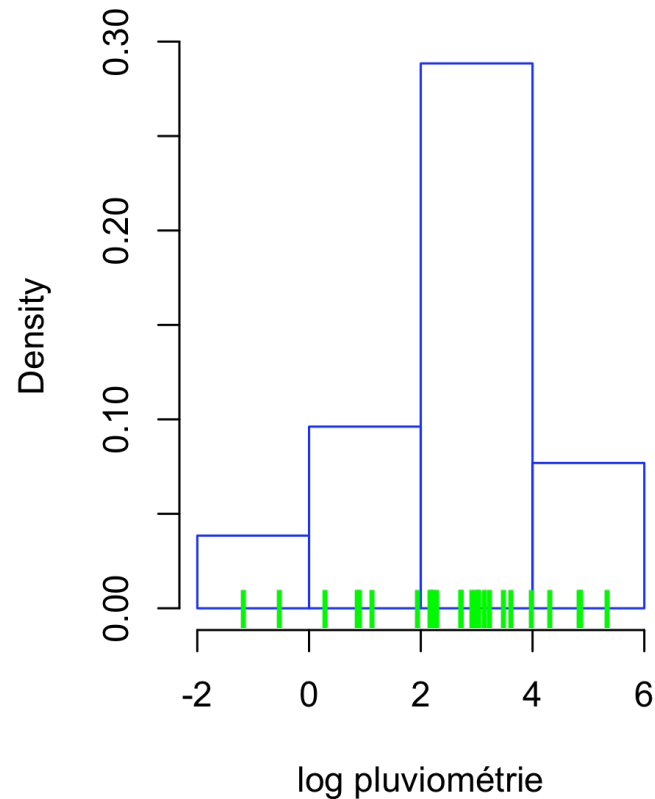


nuages non traités

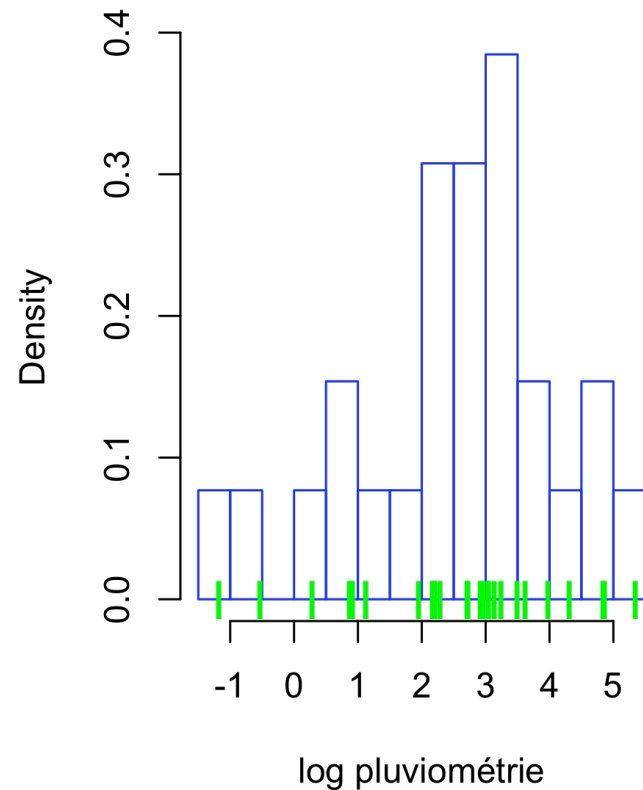


Influence du nombre de classes

nuages traités - 4 classes



nuages traités - 12 classes

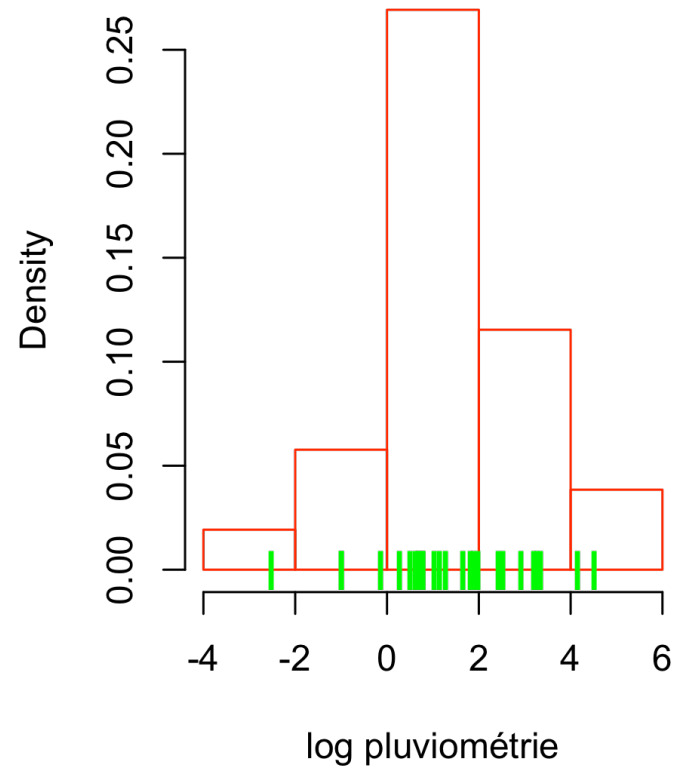


Choix à faire :

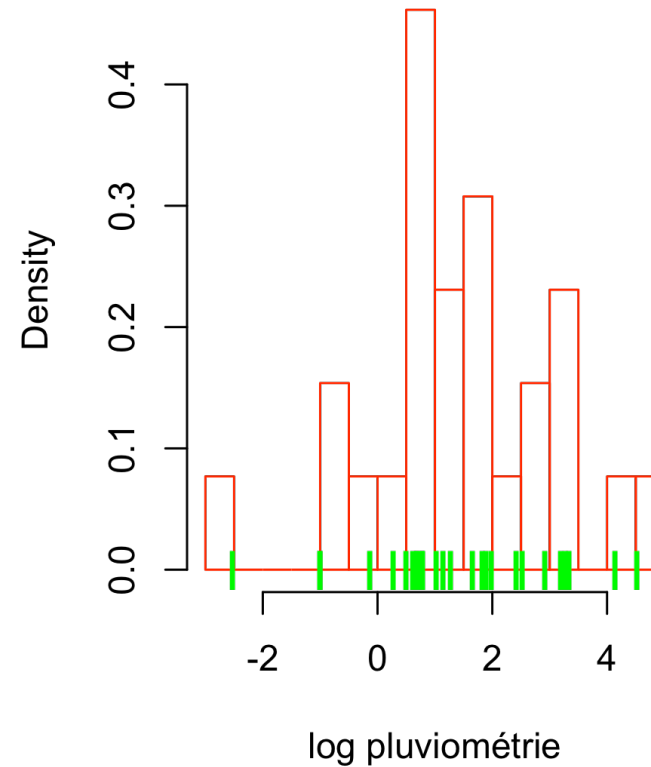
- nb classes
- largeur classes
- position classes

Idem pour nuages non traités

nuages non traités - 4 classes



nuages non traités - 12 classes



Estimation de densité

➤ Rappel :

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

➤ Histogramme :

Pour x dans la classe $[a, b]$

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

➤ Estimation de densité :

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

Interprétation (filtrage)

- Soit P_n la probabilité empirique :

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

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

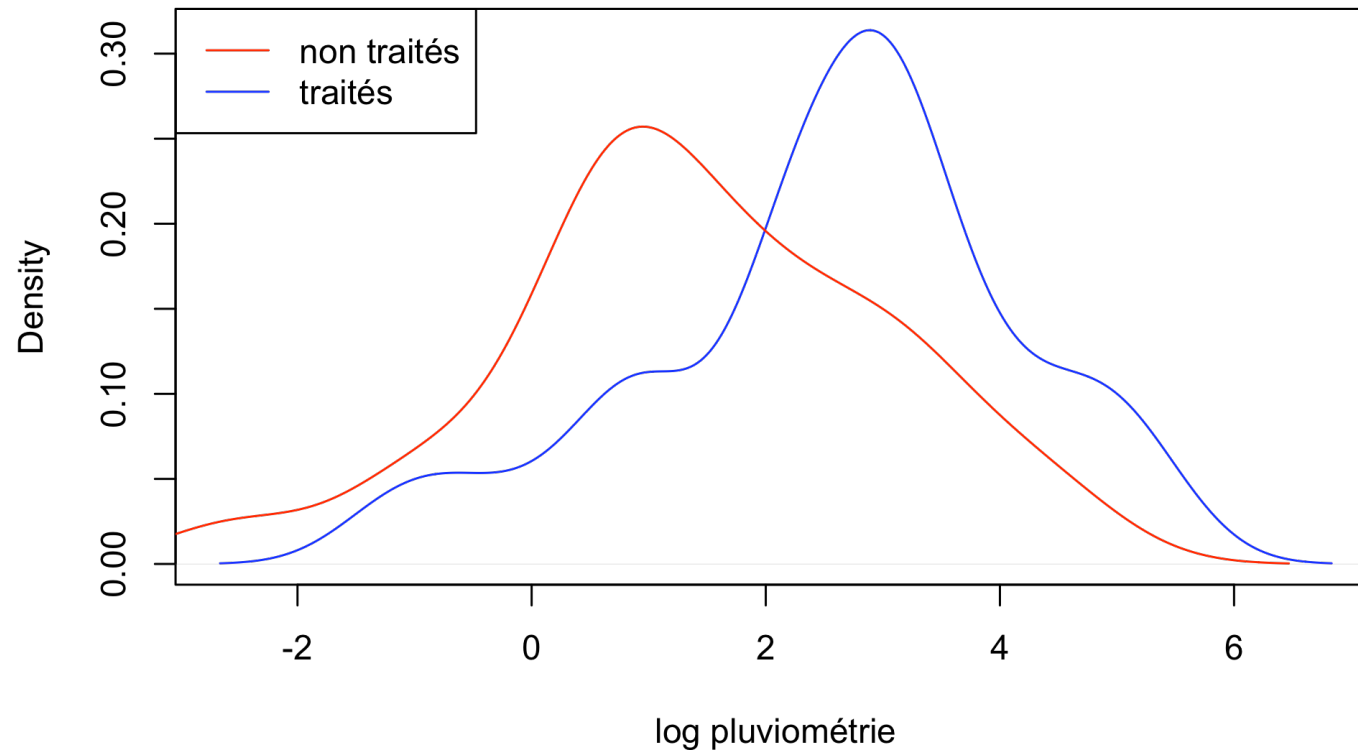
$$K_h(x) = 1/h K(x/h), \text{ où } K(u) = 1/2 \mathbf{1}_{[-1,1]}(u)$$

- 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

estimations de densité

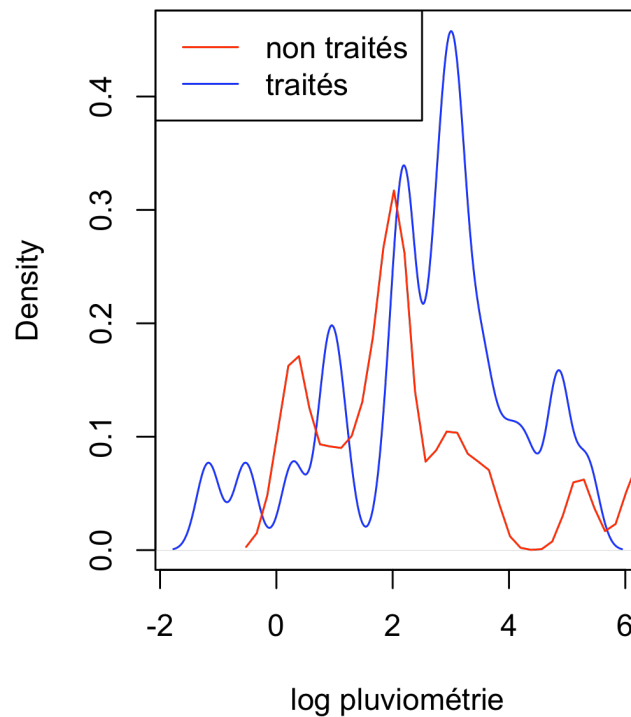


Options par défaut

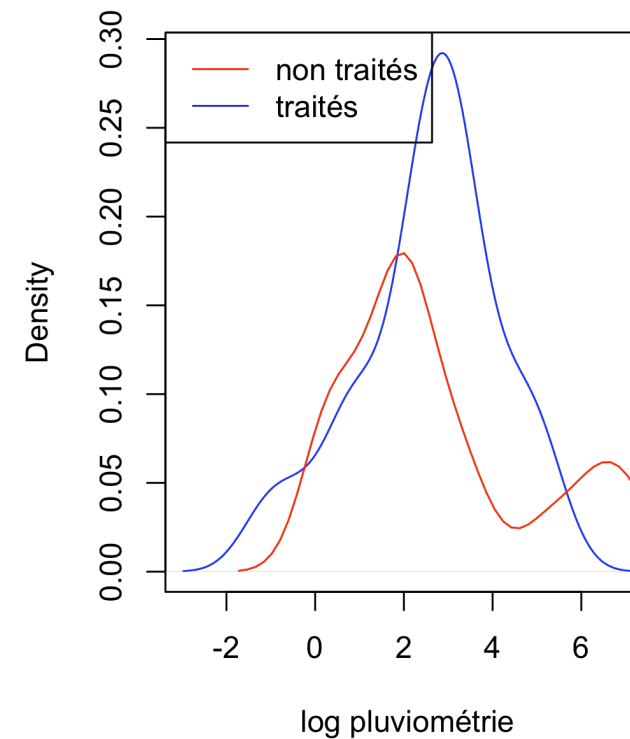
- choix automatique de h
- noyau K gaussien

Influence de h (bandwidth)

estimations de densité pour $h=0.2$

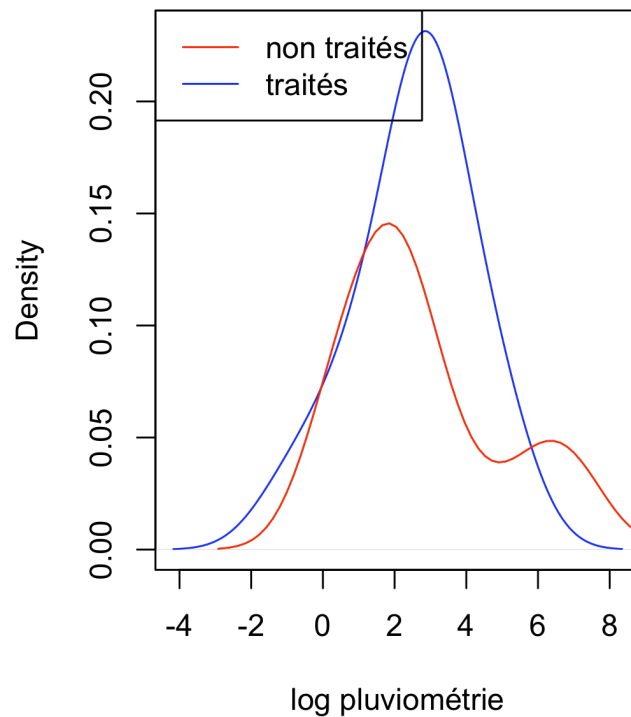


estimations de densité pour $h=0.6$

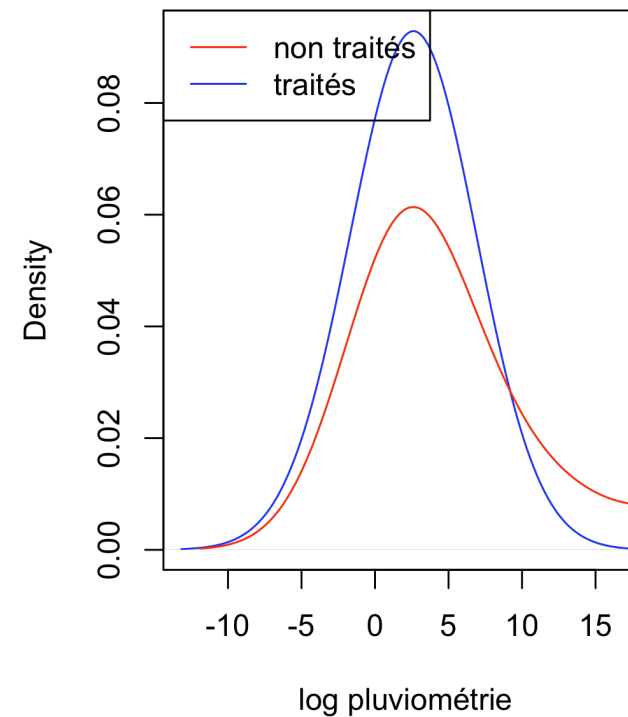


Influence de h (bandwidth)

estimations de densité pour $h=1$



estimations de densité pour $h=4$



Pour terminer ?

- Il semble, grosso modo, que

$$F_{\log(Y)}(u) = F_{\log(X)}(u-a)$$

autrement dit : $\log(Y)$ a même loi que $\log(X)+a$

- Peut-on préciser ? Comment savoir si des lois sont égales, à une transformation affine près
 - qq-plot (voir TD)