Master Ingénierie mathématique,. Nantes Université

Statistique Inférentielle.

Anne Philippe
Université de Nantes, LMJL

Adresses email :
Anne.Philippe@univ-nantes.fr

Pages web :
http://www.math.sciences.univ-nantes.fr/~philippe
Notes de cours sur sur le logiciel R
https://www.math.sciences.univ-nantes.fr/~philippe/pdf/Anne-Philippe-cours-R.pdf

Fiche 1. Estimation non paramétrique

Exercice 1. Estimation d'une loi discrète finie

1) Récupérer les données dans le fichier http://www.math.sciences.univ-nantes.fr/~philippe/data/SampleDiscret.txt Les données ont été simulées suivant la loi binomiale de paramètre $n = 5$ et $p = 0.3$

   Soit X_1 une variable aléatoire discrète prenant les $\{0, ..., 5\}$. On veut estimer la loi de X_1 à partir d'un échantillon $X_1, ..., X_n$

   On définit les deux estimateurs suivants :

   -o- les fréquences empiriques

   $$\hat{p}_n = (\hat{p}_n(0), ..., \hat{p}_n(5)) = \frac{1}{n}(N_n(0), ..., N_n(5))$$

   où

   $$N_n(j) = \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(X_i)$$

   La fonction `table` calcule les valeurs des $N_n(j)$

   -o- la fonction de répartition empirique

   $$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{]-\infty,x]}(X_i)$$

   qui estime la fonction de répartition de X_1 :

   $$F(x) = P(X_1 \leq x)$$

   La fonction `ecdf` calcule la fonction F_n :

   > Fn = ecdf(x)

   > plot(Fn)

2) Justifier que $\hat{p}_n$ est un estimateur de la densité c'est à dire du vecteur de probabilités

$$(P(X_1 = 0), ..., P(X_1 = 5))$$

3) Pour $n = 50,\ 500,\ 1000,\ 10000$,

   a) Tracer l'estimation de la loi par l'estimateur des fréquences empiriques $\hat{p}_n$ calculées sur les n premières observations

   b) Ajouter les valeurs de $\{(i, P(X_1 = i)),\ i = 0...5\}$

   `Indication : Représenter les 4 graphiques sur une même fenêtre en utilisant la commande par(mfrow=c(2,2)) qui partage la fenêtre en 2 lignes et 2 colonnes`

4) Pour $n = 50,\ 500,\ 1000,\ 10000$,

   a) Tracer la fonction de répartition empirique F_n calculée sur les n premières observations

   b) Superposer la fonction de répartition théorique

      Indication : utiliser `pbinom` pour calculer les valeurs de la fonction de répartition, puis construire avec `stepfun` une fonction constante par morceaux et continue à droite.
      La syntaxe est
      `F=stepfun(x, z)` où
      — le vecteur `x` contient les points de discontinuité de la fonction,
      — le vecteur `z` est de la forme `c(a, y)` où `a` est la valeur prise par F avant le premier point de discontinuité et y les valeurs de la fonction aux points de discontinuités `x`.
      Pour représenter graphiquement la fonction F on utilise `plot` ou `lines` (pour superposer) : avec l'argument `vertical=FALSE`

      > plot(F, vertical=FALSE) ou lines(F, vertical=FALSE)

5) Commenter les résultats obtenus

Exercice 2. Le processus empirique

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant une loi continue. On rappelle que le processus empirique est défini par

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{]-\infty,t]}(X_i).$$

1) Illustrer la convergence presque sûre de $F_n(x)$ vers $F(x) = P(X_1 \leq x)$ et le théorème de Glivenko Cantelli sur les exemples suivants
   — X_1 suit une loi uniforme sur $[0, 1]$
   — X_1 suit une loi gaussienne standard
   — X_1 suit une loi beta de paramètre $(1/2, 1/4)$,
   — X_1 suit une loi de student à 1 degré de liberté.

2) Illustrer la propriété suivante :

   la loi de la variable aléatoire $\sup_t |F_n(t) - F(t)|$ ne dépend pas de la loi de X_1

   Reprendre les exemples de la question 1).

3) Par la simulation, trouver une suite (u_n) telle que

$$u_n \sup_t |F_n(t) - F(t)|$$

   converge en loi vers une variable aléatoire non dégénérée.

Exercice 3. Estimation non paramétrique sur des lois continues

Récupérer les fichiers de données suivantes :

-•- le fichier `SampleGauss01.txt` contient une réalisation $(x_1, \ldots, x_n)$ de l'échantillon $X = (X_1, \ldots, X_n)$ iid suivant la loi normale $\mathcal{N}(0, 1)$.
   `http://www.math.sciences.univ-nantes.fr/~philippe/data/SampleGauss01.txt`

-●●- le fichier `SampleGauss03.txt` contient une réalisation $(y_1, \ldots, y_n)$ de l'échantillon $Y = (Y_1, ..., Y_n)$ iid suivant la loi normale $\mathcal{N}(0, 3^2)$.
`http://www.math.sciences.univ-nantes.fr/~philippe/data/SampleGauss03.txt`

Estimation de la densité par un histogramme

Quelques fonctions `R` : La fonction `hist` trace l'histogramme de l'échantillon x.
La syntaxe est

- `hist(x, proba = TRUE)`, le choix du nombre de classes est optimisé pour minimiser l'erreur quadratique moyenne.
  On peut fixer les classes en ajoutant l'argument `breaks`.
- `hist(x, proba = TRUE , breaks = p)` avec `p` un entier, le nombre de classes est approximativement `p` et les classes sont de même longueur.
- `hist(x, proba = TRUE , breaks = a)` avec `a` un vecteur, les coordonnées de `a` définissent les classes de l'histogramme. Il y a donc `length(a) - 1` classes.

1) Dans cette question, on estime la densité par un histogramme en utilisant le nombre de classes optimal de la fonction hist.
   Pour $n = 50,\ 500,\ 1000\ 10000$:
   a) Tracer l'histogramme calculé sur les n premières observations $X_1, ..., X_n$.
   b) Superposer la densité de la loi théorique c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
   c) Commenter les résultats.
   `Tracer les 4 graphiques sur une même fenêtre en utilisant par(mfrow=c(2,2)).`
2) Refaire la question 1 sur l'échantillon Y et comparer les résultats.

3) Sur l'échantillon X, nous allons illustrer les effets du nombre de classes sur la qualité de l'estimateur.
   On fixe le nombre d'observations $n = 500$ et on construit l'histogramme avec k classes de même longueur pour différentes valeurs de $k = 3,\ 5,\ 10,\ 15,\ 20,\ 25,\ 50,\ 100,\ 150$.
   a) Construire une suite arithmétique a de longueur $k+1$ allant $min(X_1, ..., X_{500})$ à $max(X_1, ..., X_{500})$. Ces points définissent les classes de l'histogramme.
   b) Tracer l'histogramme de $X_1, ..., X_{500}$ ayant k classes définies par le vecteur a.
   c) Superposer la densité de la loi théorique c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
   d) Commenter les résultats obtenus. Comparer avec le résultat obtenu à la question 1.
   `Tracer les 9 histogrammes sur une même page : par(mfrow=c(3,3)).`

Estimation de la densité par l'estimateur à Noyau

Quelques fonctions `R` :

- La fonction `fn = density(x)` calcule sur l'échantillon x l'histogramme à noyau avec un choix automatique de h, optimisé pour minimiser l'erreur quadratique moyenne et par défaut le noyau gaussien
- Pour changer la valeur de h, on ajoute l'option `bw`
- Pour changer le noyau, on ajoute l'option `kernel`

— La commande `plot(fn)` permet ensuite de représenter graphiquement l'estimateur.

1) Dans cette question, on utilise l'estimateur avec la valeur de h optimisée par la fonction `density` et le gaussien (par défaut).
   Pour $n = 5,\ 10,\ 50,\ 100,\ 1000$ 10000, (`tracer les 6 graphiques sur une même fenêtre.`)
   a) Tracer l'estimateur à noyau calculé sur les n premières observations $X_1, ..., X_n$.
   b) Superposer la densité de la loi théorique c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
   c) Commenter les résultats.
2) On fixe $n = 1000$ et on fait varier le paramètre h.
   Pour $h = 0.01,\ 0.02,\ 0.1,\ 0.25,\ 0.5,\ 1\ ,\ 2$. On conserve le noyau gaussien.
   a) Tracer l'estimateur à noyau calculé sur les 1000 premières observations de X.
   b) Superposer la densité de la loi théorique, c'est à dire la densité de la loi $\mathcal{N}(0, 1)$.
   c) Commenter les résultats.
3) Refaire les questions 1 et 2 avec le noyau uniforme.