Université de Bordeaux

MASTER 1 - Ingénierie Mathématique

TP 1 - Estimation non-paramétrique d'une densité

Si l'on dispose de n observations $X_1, \cdots, X_n$ issues de la répétition d'une expérience, une méthode de modélisation revient à prétendre que ces mesures sont les réalisations de variables aléatoires indépendantes et suivant une même loi de densité f. Si on ne dispose d'aucune information a priori sur f , on se trouve dans le cadre d'estimation non paramétrique.

Pour ce TP nous prendrons l'exemple d'un fabricant textile qui aurait besoin de représenter la distributions des tailles de sa clientèle. Pour celà l'entreprise réalise un sondage auprès de 1000 clientes.

Question 0 Pour la suite de l'exercice on considérera la densité cible connue et comme étant une $\mathcal{N}(171, 6.8^2)$. A l'aide de la fonction **rnorm** créer un échantillon de taille 1000 qui représentera les résultats du sondage.

1ère méthode : l'histogramme On suppose que la densité f est à support compact $[a, b[$. Pour estimer f par la méthode de l'histogramme, nous devons découper $[a, b[$ en k intervalles disjoints $I_j = [a_j, a_{j+1}[$ de taille $2h$ avec $0 \leq j \leq k-1$ et où $a = a_0 \leq a_1 \leq \cdots \leq a_{k-1} \leq a_k = b$.
Donc $\forall x \in [a, b[$, il existe un j tel que $x \in I_j$. On a alors l'estimateur histogramme :

$$\hat{f}_n(x) = \frac{n_j}{n(a_{j+1} - a_j)} = \frac{n_j}{2nh}$$

où n_j est le nombre d'observations appartenant à I_j.

Question 1.1 Représenter graphiquement la distribution de Taille avec **hist**. Combien de sous-intervalles sont pris en compte.
Question 1.2 Superposer la densité d'une $\mathcal{N}(171, 6.8^2)$.
Question 1.3 Représenter un histogramme de 100 sous-intervalles. Superposer la densité d'une $\mathcal{N}(171, 6.8^2)$.
Question 1.4 Le nombre de sous-intervalles le mieux adapté est $\sqrt{n}$ où n est la taille de l'échantillon. Représenter l'histogramme correspondant. Superposer la densité d'une $\mathcal{N}(171, 6.8^2)$. Comparer les résultats.

2ème méthode : l'histogramme mobile Pour estimer f au point x avec cette méthode, on n'utilise que que les X_i d'un intervalle $[x-h, x+h[$ de longueur fixe $2h$ et centré en x. On obtient ainsi l'estimateur histogramme mobile :

$$\hat{f}_n(x) = \frac{1}{2nh} \sum_{k=1}^{n} \mathbf{1}_{(X_k \in]x-h, x+h])}$$

Question 2.1 Créer une fonction **indic_x** qui prend en argument x, a et b et renvoie 1 si $x \in]a, b]$ et 0 sinon.
Question 2.2 A l'aide de la fonction **indic_x** créer une fonction **indic_v** qui qui pour un vecteur V renvoie renvoie le nombre de terme compris entre a et b.
Question 2.3 A l'aide de la fonction **indic_v** créer une fonction **hist_mobile_x** qui pour un point x renvoie la valeur de $\hat{f}_n(x)$ calculée avec un histogramme mobile de fenêtre h.

On définit la fonction **hist_mobile** par :

```
> hist_mobile=function(X,h,min,max,N){
> x=seq(min,max,length.out = N)
> res=x
> for(i in 1 :N)res[i]=hist_mobile_x(x[i],X,h)
> return(list(x=x,y=res))}
```

Question 2.4 Tester la fonction **hist_mobile** pour différentes valeurs de h. Les représenter sur un même graphe et superposer la densité d'une $\mathcal{N}(171, 6.8^2)$. Comparer les résultats.

3ème méthode : l'estimateur à noyau On définit l'estimateur à noyau pour tout $x \in \mathbb{R}$ par :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - X_k}{h_k}\right)$$

Question 3.1 Définir une fonction **K** qui prend en argument un point x et renvoie la valeur du noyau gaussien en ce point. On rappelle que le noyau gaussien est défini par

$$K(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$$

Pour un point x donné on définit l'estimateur à noyau par :

```
> estim_noy_x=function(x,X,h){
> n=length(X)
> res=sum(K((x-X)/h))/(n*h)
> return(res)}
```

Question 3.2 A l'aide de la fonction **estim_noy_x** et en utilisant comme modèle la fonction **hist_mobile**, créer une fonction **estim_noy**=function(X, h,min, max, N) qui calcule $\hat{f}_n$ avec l'estimateur à noyau pour N points compris entre min et max
Question 3.3 Tester la fonction **estim_noy** pour différentes valeurs de h. Les représenter sur un même graphe et superposer la densité d'une $\mathcal{N}(171, 6.8^2)$. Comparer les résultats.. Com- parer les résultats.
Question 3.4 Représenter sur un même graphe des estimations de la densité par les trois méthodes et superposer la densité d'une $\mathcal{N}(171, 6.8^2)$. Comparer les résultats.