Université de Strasbourg
M1 - Magistère
Statistique

Ségolen Geffray
geffray@math.unistra.fr
Année 2018/2019

Sujet 10

• **Rappels**: Soit f une fonction de densité à estimer. Soit $(X_1, ..., X_n)$ un échantillon i.i.d. de variables aléatoires distribuées comme une variable aléatoire X dont la loi admet la densité $f(.)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$.

$\mapsto$ **Histogramme**: Soit $(I_j)_{j=1,...,J}$ une partition de $[a, b]$ où le segment $[a, b]$ est approprié à la distribution considérée. Supposons que les intervalles I_j sont tous de même longueur notée $|I_j| = h > 0$ de sorte que $I_j = [a + (j-1)h, a + jh]$ pour $j = 1, ..., J$ et que $h = (b-a)/J$. L'estimateur de $f(.)$ au moyen d'un histogramme est défini pour $x \in \mathbb{R}$ par:

$$\widehat{f}_{n,h}(x) = \frac{1}{nh} \sum_{j=1}^{J} N_j I(x \in I_j)$$

avec

$$N_j = \sum_{i=1}^{n} I(X_i \in I_j).$$

L'estimateur ainsi obtenu est une fonction en escalier, constante sur chaque intervalle I_j. En pratique, l'utilisateur doit choisir a, b et J (ou de manière équivalente h).

$\mapsto$ L'**estimateur à noyau** de la densité (obtenu par convolution avec un noyau) est défini pour $x \in \mathbb{R}$ par:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

où la fenêtre $h > 0$ est le paramètre de lissage et où $K(.)$ est un noyau positif ie $K : \mathbb{R} \to \mathbb{R}$ est une fonction intégrable telle que $\displaystyle\int_{\mathbb{R}} K(u)du = 1$ et $K(.) \geq 0$.

• **Implémentation au moyen du logiciel** `R`:

Histogramme: La fonction `hist` du logiciel `R` détermine l'estimateur de la densité par histogramme.
La fonction `histde` du package `ks` propose aussi une implémentation de l'estimateur de la densité par histogramme.
La relation entre le nombre de classes J et la longueur h des intervalles est donnée par:

$$J = \left\lceil \frac{\max(x_1, ..., x_n) - \min(x_1, ..., x_n)}{h} \right\rceil.$$

La règle de Sturges[1] est utilisé par défaut par la fonction `hist`:

$$J = 1 + \lceil \log_2 n \rceil.$$

[1]Sturges H A. (1926) *The choice of a class interval*, Journal of the American Statistical Association, 21 (153): 65-66.

Deux autres choix sont implémentés dans la fonction `hist`, à savoir la règle de Diaconis et Freedman[2]:

$$h = 2\,\frac{\text{IQR}}{n^{1/3}}$$

en notant IQR l'intervalle inter-quartile, ainsi que la règle de référence gaussienne de Scott[3]:

$$h = 3.5\,\frac{\widehat{\sigma}}{n^{1/3}}$$

en notant $\widehat{\sigma}$ un estimateur de l'écart-type. D'autres choix ont été proposés dans la littérature dont la règle de la racine carrée (choix du logiciel `Excel`):

$$J = \left\lceil \sqrt{n} \right\rceil,$$

la règle de Rice[4]:

$$J = \left\lceil 2n^{1/3} \right\rceil,$$

et la formule de Doane[5]:

$$J = 1 + \left\lfloor \log_2 n + \log_2 \left(1 + \frac{\widehat{\kappa}_1}{c}\right) \right\rfloor$$

où $\widehat{\kappa}_1$ est le coefficient d'asymétrie empirique (disponible dans le package `moments`) et où c est donné par:

$$c = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$

Estimateur à noyau: La fonction `density` du logiciel `R` détermine l'estimateur de la densité par méthode à noyau. Il est possible de choisir à la fois le noyau et la fenêtre. Voici quelques choix possibles de noyau:

noyau	expression
Epanenchnikov	$\frac{3}{4}(1-u^2)I(\lvert u\rvert \leq 1)$
gaussien	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$
triangulaire	$(1-\lvert u\rvert)I(\lvert u\rvert \leq 1)$
rectangulaire	$\frac{1}{2}I(\lvert u\rvert \leq 1)$

Vous pouvez fournir une valeur de votre choix pour la taille de fenêtre h dans l'argument `bw`. Sinon, des méthodes de choix automatiques de la fenêtre sont implémentées. L'argument `bw="nrd0"` correspond au choix

$$h = 0.9\frac{\widehat{\sigma}}{n^{1/5}}$$

[2]Freedman D., Diaconis P. (1981) *On the histogram as a density estimator: L2 theory*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57 (4): 453–476.

[3]Scott D.W. (1979) *On optimal and data-based histograms*, Biometrika, 66 (3): 605–610.

[4]`http://onlinestatbook.com/2/graphing_distributions/histograms.html`

[5]Doane D.P. (1976) *Aesthetic frequency classification*, American Statistician, 30: 181–183.

où n est la taille de l'échantillon et où $\widehat{\sigma} = \min(S_n, R_n/1.349)$ en notant S_n l'écart-type empirique de l'échantillon et R_n l'étendue interquartile de l'échantillon. L'instruction `bw="nrd"` correspond au choix

$$h = 1.06\frac{\widehat{\sigma}}{n^{1/5}}.$$

L'instruction `bw="ucv"` correspond à un choix issu de la méthode de la validation croisée. L'idée est de déterminer un choix "optimal" de la fenêtre h noté h_{opt} de la façon suivante:

$$\begin{aligned}
h_{\text{opt}} &= \arg\min_{h>0} \texttt{ISE}(\widehat{f}_h) \\
&= \arg\min_{h>0} \int (\widehat{f}_h(x) - f(x))^2 dx \\
&= \arg\min_{h>0} \left\{ \int (\widehat{f}_h(x))^2 dx - 2\int \widehat{f}_h(x) f(x) dx + \int f(x)^2 dx \right\} \\
&= \arg\min_{h>0} \left\{ \int (\widehat{f}_h(x))^2 dx - 2\int \widehat{f}_h(x) f(x) dx \right\}.
\end{aligned}$$

Notons

$$J(h) = \int (\widehat{f}_h(x))^2 dx - 2\int \widehat{f}_h(x) f(x) dx$$

et remarquons que le premier terme est entièrement connu. Ensuite, remarquons que

$$\int \widehat{f}_h(x) f(x) dx = \mathbb{E}\left[\widehat{f}_h(X) | X_1, ..., X_n\right]$$

pour une variable aléatoire X de densité f, indépendante de $(X_1, ..., X_n)$. On estime alors $\mathbb{E}\left[\widehat{f}_h(X) | X_1, ..., X_n\right]$ sans biais par la méthode dite du "leave-one-out". Cela fournit le critère suivant à optimiser le critère en h:

$$\widehat{J(h)} = \int (\widehat{f}_h(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i,h}(X_i)$$

où $\widehat{f}_{-i,h}$ est l'estimateur de f calculé avec la fenêtre h mais en ôtant l'observation i, ce qui donne:

$$\widehat{f}_{-i,h}(x) = \frac{1}{(n-1)h}\sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h}\right).$$

Notons que la fonction $\widehat{J(h)}$ est un estimateur sans biais de $\texttt{MISE}(\widehat{f}_h)$ à une constante indépendante de h près puisque

$$\mathbb{E}\left[\widehat{J(h)}\right] = \texttt{MISE}(\widehat{f}_h) - \int f(x)^2 dx.$$

Soient les conditions suivantes:
(A_1): f est de carré intégrable et deux fois différentiable et sa dérivée seconde est continue bornée et de carré intégrable.
(A_2): le noyau $K : \mathbb{R} \to R$ est tel que $\int_{\mathbb{R}} K(x)^2 dx < \infty$, K est symétrique et admet un moment d'ordre 2, donc satisfait $\int xK(x)dx = 0$ et $\int x^2 K(x)dx < \infty$.

(A_3): Les fenêtres $h = h_n > 0$ forment une suite telle que $h \overset{n\to\infty}{\to} 0$ et $nh \overset{n\to\infty}{\to} \infty$. Sous ces conditions, un équivalent asymptotique du MISE est:

$$AMISE(\widehat{f}_h) = \frac{\int K(x)^2 dx}{nh} + \frac{h^4}{4}\left(\int x^2 K(x)dx\right)^2 \int f''(x)^2 dx.$$

La fenêtre optimale est alors celle qui minimise le AMISE. Cela fournit:

$$h_{\text{opt}} = \left(\frac{\int K(x)^2 dx}{\left(\int x^2 K(x)dx\right)^2 \int f''(x)^2 dx}\right)^{1/5} \frac{1}{n^{1/5}}.$$

Comme f'' est inconnue, il faut l'estimer: c'est le principe du *plug-in* (injection). On estime f'' par un estimateur à noyau avec une fenêtre que l'on qualifie de "pilote" et dont le choix est loin d'être une question triviale. Notons que les problèmes d'estimation de f et de f'' ne sont pas équivalents... Le AMISE correspondant à ce choix de fenêtre est l'ordre de $1/^{4/5}$.
L'argument `bw.SJ` de la fonction `density` permet d'implémenter ce choix de fenêtre.

• **Critères objectifs**:
Critères ponctuels: pour un estimateur $\widehat{f}$ de f, le biais est ponctuellement donné par

$$b_f\left(\widehat{f}(x)\right) = \mathbb{E}\left[\widehat{f}(x)\right] - f(x),$$

la variance est ponctuellement donnée par

$$\text{Var}\left(\widehat{f}(x)\right) = \mathbb{E}\left[\left(\widehat{f}(x) - \mathbb{E}\left[\widehat{f}(x)\right]\right)^2\right],$$

et l'écart quadratique moyen est ponctuellement donné par

$$R_f\left(\widehat{f}(x)\right) = \mathbb{E}\left[\left(\widehat{f}(x) - f(x)\right)^2\right].$$

Critères globaux: le carré du biais intégré est donné par

$$\int \left(\mathbb{E}\left[\widehat{f}(x)\right] - f(x)\right)^2 dx,$$

la variance intégrée est donnée par

$$\int \text{Var}\left(\widehat{f}(x)\right) dx,$$

et l'écart quadratique moyen intégré (MISE = Mean Integrated Squared Error) est donné par

$$\int \mathbb{E}\left[\left(\widehat{f}(x) - f(x)\right)^2\right] dx.$$

Exercice 1.
On note $\varphi_{(m,\sigma^2)}$ la densité de la loi gaussienne de paramètres $m \in \mathbb{R}$ et $\sigma^2 > 0$. Considérons les distributions suivantes:

- la loi gaussienne standard $\mathcal{N}(0,1)$ de densité $\varphi_{(0,1)}$,
- les lois de Student $T(2)$ à 2 degrés de liberté et à 3 degrés de liberté $T(3)$,
- les lois $\Gamma(1,1/2)$, $\Gamma(2,1/2)$ et $\Gamma(5,1)$,
- la loi de densité donnée par la fonction en escalier suivante:

$$f(x) = \begin{cases} 1/2 & \text{si } 0 \leq x < 1/2, \\ 2 & \text{si } 1/2 \leq x < 2/3, \\ 5/4 & \text{si } 2/3 \leq x \leq 1, \\ 0 & \text{sinon,} \end{cases}$$

- les lois $B(1/2,1/2)$, $B(2,2)$ et $B(2,5)$ où $B(a,b)$ désigne la loi Beta de 1$^{\text{ère}}$ espèce de paramètre (a,b) pour $a,b>0$,
- la loi dont la densité est la fonction linéaire par morceaux suivante:

$$f(x) = \begin{cases} 10x & \text{si } 0 \leq x < 1/4, \\ 4-6x & \text{si } 1/4 \leq x < 1/2, \\ 2-2x & \text{si } 1/2 \leq x \leq 1, \\ 0 \quad \text{sinon,} \end{cases}$$

- la loi de densité:

$$f(x) = \begin{cases} \dfrac{1}{c}\,\varphi_{(1,0.4)}(x) & \text{si } 0 \leq x < 1/2, \\ \dfrac{1}{c}\,\varphi_{(0,0.4)}(x) & \text{si } 1/2 \leq x \leq 1, \end{cases}$$

où $c = \displaystyle\int_0^{1/2} \varphi_{(1,0.4)}(u)du + \int_{1/2}^{1} \varphi_{(1,0.4)}(u)du$,

- le mélange de lois gaussiennes de densité:

$$\frac{1}{2}\,\varphi_{(0,0.1)}(x) + \frac{1}{2}\,\varphi_{(1,0.4)}(x),$$

- la loi de densité:

$$0.5\varphi_{(0,1)}(x)+0.1\varphi_{(-1,0.1)}(x)+0.1\varphi_{(-0.5,0.1)}(x)+0.1\varphi_{(0,0.1)}(x)+0.1\varphi_{(0.5,0.1)}(x)+0.1\varphi_{(1,0.1)}(x).$$

1. Simuler M échantillons de taille n suivant les distributions exposées ci-dessus pour différentes valeurs de n.
2. Déterminer l'histogramme et l'estimateur à noyau de la densité en utilisant différents noyaux et différentes tailles de fenêtre implémentés dans le logiciel `R`.
3. Comparer le comportement de l'histogramme et de l'estimateur à noyau à l'aide des critères objectifs précédemment présentés.