Université de Strasbourg
M1 - Magistère
Statistique

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

Sujet 2

• **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).

$\mapsto$ **Estimateur ASH**[1]: on se donne K histogrammes, notés $\widehat{f}^{[1]}$, ..., $\widehat{f}^{[K]}$, tous calculés à partir du même échantillon de taille n et ayant tous la même taille de fenêtre h mais avec des origines décalées: a, $a + \dfrac{h}{K}$, $a + \dfrac{2h}{K}$, ...,$a + \dfrac{(K-1)h}{K}$ respectivement. On moyenne ces K histogrammes: on obtient l'estimateur ASH de la densité noté $\widehat{f}_{\text{ASH}}$ qui s'écrit donc:

$$\widehat{f}_{\text{ASH}}(x) = \frac{1}{K} \sum_{k=1}^{K} \widehat{f}^{[k]}(x).$$

L'estimateur ainsi obtenu est encore une fonction en escalier, constante sur de petits intervalles de longueur h/K. Par conséquent, on peut considèrer que les classes sont les petits intervalles $B_i = [a + (i-1)\dfrac{h}{K}, a + i\dfrac{h}{K}[$ pour $i = 1, .., JK$. Pour $x \in B_i$, on peut exprimer $\widehat{f}_{\text{ASH}}(x)$ sous la forme

$$\begin{aligned}\widehat{f}_{\text{ASH}}(x) &= \frac{1}{K} \sum_{k=1-K}^{K-1} (K - |k|) \frac{\#\{i \in \{1, ..., n\} : X_i \in B_{i+k}\}}{nh} \\ &= \frac{1}{nh} \sum_{k=1-K}^{K-1} \left(1 - \frac{|k|}{K}\right) \#\{i \in \{1, ..., n\} : X_i \in B_{i+k}\}.\end{aligned}$$

[1] Scott D.W. (1985) *Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions*, The Annals of Statistics, 13(3):1024-1040 et Scott D.W. (2009) *Averaged shifted histogram*, Wiley Interdisciplinary Reviews: Computational Statistics, 2(2):160-164.

$\mapsto$ **Estimateur FP** (FP= *frequency polygon* = polygone des fréquences): on se donne un histogramme $\widehat{f}$ déterminé avec la fenêtre h. L'estimateur FP de la densité noté $\widehat{f}_{\text{FP}}$ est obtenu en reliant par un segment le centre de chaque couple de classes adjacentes de l'histogramme. Ainsi, pour $j = 1, ..., (J-1)$, pour $x \in \Big[a + (j-1/2)h, a + (j+1/2)h\Big[$, l'expression de $\widehat{f}_{\text{FP}}(x)$ est celle du segment qui relie les points $\left(a + (j-1/2)h, \dfrac{N_j}{nh}\right)$ et $\left(a + (j+1/2)h, \dfrac{N_{j+1}}{nh}\right)$, ce qui fournit

$$\widehat{f}_{\text{FP}}(x) = \frac{N_{j+1} - N_j}{nh^2}\, x + \frac{N_j}{nh} - \frac{N_{j+1} - N_j}{nh^2}\, (a + (j-1/2)h)\,.$$

Aux bords de $[a, b]$ ie pour $x \in [a, a + h/2[$ et $x \in [a + (J-1/2)h, b]$, on définit $\widehat{f}_{\text{FP}}(x) = \widehat{f}(x)$. L'estimateur ainsi obtenu est cette fois une fonction continue.

• **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

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

• **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[2] est utilisé par défaut par la fonction `hist`:

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

Deux autres choix sont implémentés dans la fonction `hist`, à savoir la règle de Diaconis et Freedman[3]:

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

en notant IQR l'intervalle inter-quartile, ainsi que la règle de référence gaussienne de Scott[4]:

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

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[5]:

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

et la formule de Doane[6]:

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

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 ASH: le package `ash` contient la fonction `ash1` qui permet de calculer $\widehat{f}_{\mathrm{ASH}}$ dans le cas de données univariées.

Polygône des fréquences: La règle de référence à la loi gaussienne standard préconise

$$h = 2.15\,\frac{\widehat{\sigma}}{n^{1/5}}.$$

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 $\Gamma(1, 1/2)$, $\Gamma(2, 1/2)$ et $\Gamma(5, 1)$,

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

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

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

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

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

- la loi triangulaire de densité donnée pour $x \in \mathbb{R}$ par

$$g(x) = \max\left(1 - |x|, 0\right),$$

- la loi de Student à 2 degrés de liberté, dilatée, translatée: $X = 0.25Y - 0.5 = (Y-2)/4$ où $Y \sim T(2)$,

- on note $\widetilde{\varphi}_{(0,1)}$ la restriction de $\varphi_{(0,1)}$ à $[-1, 2]$ puis l'on considère la loi de densité définie comme suit:

$$f(x) = \frac{\widetilde{\varphi}_{(0,1)}(x)}{\int_{-1}^{2} \varphi_{(0,1)}(u)du}, \quad x \in \mathbb{R},$$

- la loi de densité donnée par la fonction en escalier suivante:

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

- la loi dont la densité est en dents de scie comme ci-dessous:

$$f(x) = g(x+9) + g(x+7) + ... + g(x-7) + g(x-9),$$

- le mélange de lois gaussiennes de densité:

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

1. Simuler M échantillons de taille n suivant les distributions exposées ci-dessus pour différentes valeurs de n.

2. Déterminer les estimateurs de la densité par la méthode de l'histogramme, de l'ASH et du polygône des fréquences en utilisant différentes tailles de fenêtre (ou de manière équivalente en faisant varier le nombre de classes).

3. Analyser le comportement des trois estimateurs (histogramme, ASH, FP) de la densité à l'aide des critères objectifs précédemment présentés.