M2 Data Sciences **Introduction é la statistique**
2018-2019 **non paramétrique**

TP 1 : Tests non paramétriques et estimation de densité

Préambule : fonctions utiles.

- Voici le nom des principales fonctions de test sous R dont vous aurez besoin : `ks.test` (test de Kolmogorov-Smirnov, pour un ou deux échantillons), `wilcox.test` (tests de Wilcoxon et Mann-Whitney),
  Il vous faut également charger le paquet `nortest`, qui vous donnera accés au test de Lilliefors `lillie.test` (qui correspond au test KS d'adéquation é la famille gaussienne).
- En plus du mémento de commandes usuelles R, vous pourriez avoir besoin de `qqplot` ou `qqnorm` et `rank`.
- Les jeux de données utilisés dans les exercices sont issus des packages `MASS` et `datasets`
- La fonction `density` implémente l'estimation de densité par noyau. La fonction `hist` avec l'option `freq=F` implémente l'estimation de densité par histogrammes réguliers. La fonction `histogram` du package `histogram`, avec l'option `type='irregular'` implénte les estimateurs de densité par histogrammes irréguliers.

1 Tests non paramétriques : cas pratiques

Note. Dans des situations pratiques, le niveau de test acceptable dépend du domaine et du probléme considérés. Dans l'ensemble des exercices, on considérera un niveau de 5%.

Exercice 1 Les données `anorexia` du package `MASS` comportent des mesures du poids de 72 patientes avant et aprés traitement. La premiére colonne donne le traitement reéu (3 valeurs), la deuxiéme et la troisiéme colonne le poids avant et aprés traitement.

1) On veut tout d'abord tester la différence de poids avant et aprés traitement.
   a) Quel test paramétrique pourrait-on envisager ? Sous quelles conditions ?
   b) Proposer une procédure de test non-paramétrique et conclure.

2) On veut maintenant tester si le changement de poids avant-aprés traitement différe selon le traitement reéu. Plus précisément, on veut tester la différence entre le traitement contréle ("Cont") et le traitement familial ("FT").
   a) A l'aide de la fonction `qqplot`, donner une réponse intuitive.
   b) Quel test paramétrique pourrait-on envisager ? Préciser les conditions d'application ?
   c) Appliquer le test de Mann Whitney. Un message d'avertissement mentionne l'existence d'ex-aequos. Qu'en pensez-vous ? Conclure en répondant é la question posée.

Exercice 2 Un fabricant garantit que la fiabilité des appareils qu'il vend est telle que leur durée de vie suit une loi exponentielle, de moyenne 1700 heures. Afin de tester cette affirmation, on mesure la durée de vie en heures de 10 de ces appareils pris au hasard. On obtient les valeurs suivantes $\{555, 653, 1801, 678, 3635, 502, 2044, 3359, 3546, 774\}$.

1) a) Tracez la fonction de répartition (fdr) empirique associée àé l'échantillon des $n = 10$ appareils.

b) Tracez sur le même graphique la fdr de la loi exponentielle $\mathcal{E}(1/1700)$.

2) a) Expliquez pourquoi la statistique de Kolmogorov Smirnov est donnée par

$$D_n = \max_{1 \leq i \leq n} \{|F_0(X_{(i)}) - i/n|, |F_0(X_{(i)}) - (i-1)/n|\},$$

où $(X_{(1)}, \ldots, X_{(n)})$ est la statistique ordonnée associée à l'échantillon.

b) Déterminez la valeur observée de la statistique D_n pour cet échantillon.

c) En vous reportant à la table statistique fournie à la fin de l'énoncé, déterminez le seuil de rejet du test au niveau $\alpha = 5\%$ et concluez sur le test avec ce niveau.

3) Retrouvez ces résultats avec la fonction *ks.test*.

Exercice 3 On considére le jeu de données `birthwt` du package `MASS` qui rassemble les données d'une étude de 1986, dans le Massachusetts (USA) analysant les facteurs de risque au cours de la grossesse d'accoucher d'un bébé de "faible poids", c'est-à-dire de poids inférieur à 2500 g. Proposer un étude de ce jeu de données permettant de répondre é la question.

Exercice 4 Le data frame `UScrime` du package `MASS` rassemble des données sociologiques sur la criminalité dans 47 états des Etats Unis en 1960. Le fichier descriptif est disponible par la commande `?UScrime` et le début du tableau de données peut étre affiché par la commande `head(UScrime)`. La variable binaire `So` vaut 1 pour les Etats du sud, et 0 pour les autres. La variable `Prob` indique le taux de criminalité. On veut tester si le taux de criminalité différe entre les Etats du nord et du sud. Proposer une procédure de test et conclure.

Exercice 5 Le data frame `Animals` du package `MASS` donne le poids moyen du cerveau et le poids moyen du corps pour 28 espéces. Le fichier descriptif est disponible par la commande `?Animals` et le début du tableau de données peut étre affiché par la commande `head(Animals)`. On veut savoir si le poids du cerveau augmente avec le poids du corps.

a) On veut tout d'abord tenter de répondre é la question posée par un modéle linéaire. Qu'en pensez-vous ?

b) Proposer un test non paramétrique pour répondre é la question.

2 Estimation de densité : aspects théoriques

Exercice 6 Considérons $X_1, \cdots, X_n$ n variables aléatoires i.i.d. de densité f_X. On cherche à estimer f_X à partir des observations $X_1, \cdots, X_n$.

1. Donner la définition d'un estimateur à noyau de f_X, en précisant toutes les quantités qui interviennent.
2. Donner des exemples de noyaux
3. Montrer que le risque quadratique ponctuel de cet estimateur s'écrit comme la somme d'un biais et d'une variance que l'on définira.
4. Calculer la variance de cet estimateur à noyau. En déduire une majoration de cette variance, ainsi que les conditions requises pour le noyau et pour f_X, permettant d'établir cette majoration.

5. Calculer le biais de cet estimateur à noyau. En déduire une majoration de ce biais, ainsi que les conditions requises pour le noyau et pour f_X, permettant d'établir cette majoration.
6. En déduire une majoration du risque quadratique ponctuel de cet estimateur à noyau.
7. Commenter

Exercice 7 Considérons $X_1, \cdots, X_n$ n variables aléatoires i.i.d. de densité $f_X \in \mathbb{L}([0,1])$. On cherche à estimer f_X à partir des observations $X_1, \cdots, X_n$. Considérons la base d'histogrammes définie par

$$\varphi_k(x) = \sqrt{D}\mathbb{1}_{\left[\frac{(k-1)}{D};\frac{k}{D}\right[}(x), \quad D > 0, \quad k = 1 \cdots D.$$

Soit $S_D = \text{vec}\{\varphi_k, k = 1, \cdots, D\}$. On note $\| f \|_\infty = \sup x \in [0,1]|f(x)|$ et $\| f \|^2 = \int_0^1 f^2(x)dx$.

1. Montrer que pour toute $g \in S_D$, on a

$$\| g \|_\infty^2 \leq D \| g \|^2 \quad \text{et} \quad \| \sum_{k=1}^{D} \varphi_k^2 \|_\infty \leq D.$$

2. Donner l'expression de la projection orthogonale de f_X sur S_D notée $\Pi_{S_D}(f_X)$.
3. En déduire un estimateur sans biais de $\Pi_{S_D}(f_X)$, noté $\widehat{f}_D$.
4. Montrer que le risque quadratique intégré de $\widehat{f}_D$ s'écrit

$$\mathbb{E} \| \widehat{f}_D - f_X \|^2 = \| \Pi_{S_D}(f_X) - f_X \|^2 + \mathbb{E} \| \widehat{f}_D - \Pi_{S_D}(f_X) \|^2 .$$

5. Montrer que

$$\mathbb{E} \| \widehat{f}_D - \Pi_{S_D}(f_X) \|^2 \leq D/n$$

6. Montrer que si f_X est telle que $|f_X(x) - f_Y(y)| \leq C|x-y|^\alpha$ pour $\alpha \in]0,1[$, alors

$$\| \widehat{f}_D - f_X \|^2 \leq C^2 D^{-2\alpha}.$$

7. En déduire une majoration du risque quadratique intégré de $\widehat{f}_D$.
8. Trouver D_{opt} qui minimise cette majoration du risque quadratique intégré.
9. En déduire la majoration du risque quadratique intégré pour ce D_{opt}.
10. Commenter

3 Estimation de densité : implémentation

Nous allons ici implémenter les estimateurs de densité par histogrammes et par noyaux.

Partie A : simulation des données

On considère une taille déchantillon $n = 1000$, et un intervalle d'estimation $I = [0,5]$. Définir un vecteur x de 2000 points régulièrement espacés sur $[0,5]$.

1. Générer un échantillon i.i.d. `Y` de taille n selon une distribution normale $\mathcal{N}(2.5,1)$. Tracer la densité de la distribution $\mathcal{N}(2.5,1)$ sur I.

2. Soit U une variable discrète à valeur dans $\{1, 2, 3\}$ et de distribution :

$$(1) \quad \begin{cases} \mathbb{P}[U = 1] = p_1 \\ \mathbb{P}[U = 2] = p_2 \\ \mathbb{P}[U = 3] = p_3 \end{cases}$$

avec $p_1 + p_2 + p_3 = 1$. Soit X une variable dépendant de U telle que

$$(2) \quad X|U \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1) & \text{si} \quad U = 1 \\ \mathcal{N}(\mu_2, \sigma_2) & \text{si} \quad U = 2 \\ \mathcal{N}(\mu_3, \sigma_3) & \text{si} \quad U = 3 \end{cases}$$

Alors on peut montrer que X a pour densité :

$$f_X(x) = p_1\varphi_{\mu_1,\sigma_1}(x) + p_2\varphi_{\mu_2,\sigma_2}(x) + p_3\varphi_{\mu_3,\sigma_3}(x)$$

oé $\varphi_{\mu,\sigma}$ désigne la densité de la distribution gaussienne de moyenne μ et variance σ. Cette distribution est ce qu'on appelle un "mélange de gaussiennes".

(a) Tracer la densité de X sur l'intervalle I pour les valeurs suivantes des paramètres :

$$\begin{cases} \mu_1 = 1 \\ \mu_2 = 3 \\ \mu_3 = 4 \end{cases} \quad \begin{cases} \sigma_1 = 0.5 \\ \sigma_2 = 0.3 \\ \sigma_3 = 0.2 \end{cases} \quad \begin{cases} p_1 = 0.2 \\ p_2 = 0.5 \\ p_3 = 0.3 \end{cases}$$

(b) Générer un échantillon i.i.d. de taille n, de même loi que `X` selon la distribution f_X.

Indication. On pourra générer un échantillon i.i.d. `U` de taille n selon la distribution (1) en utilisant `U <- sample(x=c(1,2,3), size=n, replace=TRUE, prob=c(0.2,0.5,0.3))`

On pourra ensuite créer un vecteur `X` de longueur n, puis pour tout $i = 1, \ldots, n$, générer `X[i]` selon la distribution $\mathcal{N}(\mu_j, \sigma_j^2)$ où $j =$`P[i]` et `P` est le vecteur (p_1, p_2, p_3).

Partie B : estimation par histogrammes

1. (a) Représenter sur une multi-figure les estimateurs par histogrammes réguliers de la densité de `Y` ainsi que la vraie densité selon laquelle est tiré l'échantillon, pour les valeurs suivantes de D : 5,15,50,300

   (b) Quelle valeur de D vous parait la plus appropriée ? Décrivez ce qu'on observe si D est trop grand ou trop petit.

2. Mêmes questions pour l'échantillon `X`

3. La valeur de D optimale est-elle la même pour ces deux échantillons ? Pourquoi ?

4. Pour chacun des échantillons `X` et `Y`, tracer l'histogramme irrégulier à l'aide de la fonction `histogram` ainsi que la vraie densité.

5. On souhaite maintenant étudier le comportement du MISE en fonction de D dans un cas particulier. On considère $Z_1, \cdots, Z_n$ i.i.d. de la beta de paramètres 1.9 et 1.9. On note f la densité des Z_i.

   Considérons la base d'histogrammes définie par

$$\varphi_k(x) = \sqrt{D}\mathbb{1}_{\left[\frac{(k-1)}{D};\frac{k}{D}\right[}(x), \quad D > 0, \quad k = 1 \cdots D.$$

Soit $S_D = \text{vec}\{\varphi_k, k = 1, \cdots, D\}$. On note $\| f \|_\infty = \sup x \in [0,1] |f(x)|$ et $\| f \|^2 = \int_0^1 f^2(x)dx$. On note f_D la projection orthogonale de f sur S_D et θ_j tel

$$\theta_j = < f, \varphi_j > .$$

(a) Rappeler la définition du MISE

(b) Montrer qu'il s'écrit

$$\mathbb{E} \| \hat{f}_D - f_D \|^2 + \| f_D - f_D \|^2 .$$

(c) Dans le cas spécifique où f est une densité appartenant à la famille des loi *beta* de paramètres 1.9 et 1.9.

(d) Tracer la densité ainsi que les histogrammes pour plusieurs valeurs de D

(e) Quelles sont les valeurs de D qui semblent raisonnables pour estimer f ?

(f) Etude du terme de variance

i. Montrer que

$$\mathbb{E} \| \hat{f}_D - f_D \|^2 = \sum_{j=1}^{D} \mathbb{E}(\hat{\theta}_j - \theta_j)^2,$$

où $\hat{\theta}_j$ est un estimateur de θ_j que l'on explicitera.

ii. Montrer que

$$\mathbb{E}(\hat{\theta}_j - \theta_j)^2 = DP_j(1 - Pj)/n,$$

où

$$P_j = \mathbb{P}\left[Z \in \mathbb{1}_{\left[\frac{(j-1)}{D}; \frac{j}{D}\right[} \right].$$

(g) Etude du terme de biais : montrer que

$$\| f_D - f_D \|^2 = -\sum_{j=1}^{D} \theta_j^2 + \int f^2(x)dx.$$

6. Implémentation : tracer le MISE en fonction de D pour D allant de 3 à 25.

Partie C : estimation de densité par noyaux

1. (a) Considérons l'échantillon `Y`. Tracer sur un même graphe l'estimateur de densité par noyaux fournit par la fonction `density` avec la fenêtre par défaut, et la vraie densité.

(b) En regardant le fichier d'aide de la fonction `density`, déterminer la valeur de la fenêtre par défaut ("bandwidth" en anglais) pour l'estimateur ci-dessus.

(c) Déterminer la valeur de la fenêtre pour les autres méthodes de sélection de fenêtre.

(d) Représenter sur une multi-figure les estimateurs par noyaux de la densité de `Y` ainsi que la vraie densité selon laquelle est tiré l'échantillon, pour les valeurs suivantes de la fenêtre : 0.01,0.1, 0.22, 2.

(e) Quelle valeur de la fenêtre vous parait la plus appropriée ? Décrivez ce qu'on observe si la fenêtre est trop grande ou trop petite.

2. Mêmes questions avec l'échantillon `X`.

3. L'espérance f_h de l'estimateur par noyau d'une densité par noyau admet l'expression suivante :

$$f_h(x_0) = \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\right] = \frac{1}{nh}\int K\left(\frac{x - x_0}{h}\right) f(x)dx$$

où f_h correspond a un lissage de f, et plus h est grand, plus l'effet de lissage est important. On veut observer ce phénomène sur les données en traçant f_h pour plusieurs valeurs de h.

D'après la loi des grands nombres, si N est très grand

$$\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right) \simeq f_h(x_0). \tag{1}$$

Ainsi, en calculant l'estimateur de densité par noyau à partir d'un échantillon de taille N pour N très grand, on obtient une approximation numérique de f_h.

A partir de cette observation, représenter sur une multi-figure :

$$f_{X,h}(x_0) = \frac{1}{nh}\int K\left(\frac{x - x_0}{h}\right) f_X(x)dx$$

ainsi que ainsi que la vraie densité f_X pour $h = 0.05, 0.2, 0.4, 0.8$ (prendre $N = 10^5$ dans (1)).

Exercice 8 La base de données *geyser* de la librairie *MASS* contient des données d'éruption (temps d'attente et durée) de l'Old Faithful geyser du Yellowstone National Park. Chargez les données et familiarisez vous avec.

1. Utilisez la fonction *density* pour estimer la densité des observations. Stockez le résultat dans une variable, et observez sa structure.
2. Tracez la courbe de l'estimateur à noyau à l'aide de la fonction *plot*.
3. Faites varier le noyau utilisé avec la variable *kernel*. Que constatez-vous ? Dans la suite, fixez un noyau de votre choix.
4. Faites à présent varier le paramètre de fenêtre *bw*. Que constatez-vous ?

Choix de la fenêtre optimale par validation croisée pour l'estimateur à noyau L'objectif est de sélectionner, par validation croisée *meilleur estimateur à noyau gaussien* (*i.e.* la meilleure fenêtre h), au sens du risque MISE.

1. Rappelez l'expression du risque MISE d'un estimateur à noyau $\hat{f}_{n,h}$ en tant que fonction de la fenêtre inconnue h. Donnez l'expression d'un estimateur $\hat{J}_{n,h}$ (à constante près) de ce risque obtenu par validation croisée.
2. Montrer que

$$\frac{2}{|C_v|(|C_v|-1)h}\sum_{\substack{i,j\in C_v \\ j\neq i}} K\left(\frac{X_i - X_j}{h}\right) \qquad = \frac{2}{(|C_v|-1)}\sum_{i\in C_v}\left\{\hat{f}_{n,h}^{C_v}(X_i) - \frac{K(0)}{|C_v|h}\right\},$$

où $\hat{f}_{n,h}^{C_v}(X_i)$ est l'estimateur à noyau de fenêtre h, construit avec les observations du paquet C_v, pris en la valeur X_i et $K(0) = (2\pi)^{-1/2}$.

3. On veut utiliser une grille sur les valeurs de $h \in [0, 8]$ et considérer l'échantillon comme composé de 5 paquets. Pour chaque valeur de h et chaque paquet d'observations, on veut calculer l'estimateur à noyau gaussien obtenu pour cette valeur de h et en utilisant les observations en-dehors du paquet considéré. Puis, on veut estimer l'erreur (via l'expression $\hat{J}_{n,h}$) commise par cet estimateur sur les observations du paquet considéré. Ensuite, on veut faire la moyenne de ces erreurs sur les 5 paquets. Enfin on veut sélectionner la valeur de h qui donne l'erreur moyenne la plus faible.

   (a) Commencez par écrire la structure d'un programme qui intègre une boucle sur h et une boucle sur les 5 paquets d'observations.

   (b) Réfléchissez au calcul de l'estimateur du risque MISE. ***Indications*** *:*

      i) Une intégrale s'approche par une somme de Riemann. On approcher donc la quantité $\int (\hat{f}^v_{n,h}(x))^2 dx$ par une somme finie (somme de Riemann) de la forme

$$\frac{1}{M} \sum_{i=1}^{M-1} (x_{i+1} - x_i)[\hat{f}^v_{n,h}(x_{i+1}))]^2,$$

      où les $\{x_i\}_{1 \leq i \leq M}$ forment une grille sur l'axe des abscisses.

      ii) Le second terme dans l'estimation du risque est la somme des valeurs prises par un estimateur à noyau en les observations.

   (c) Complétez votre programme pour sélectionner le meilleur estimateur. ***Attention*** *: si votre grille sur h est trop grossière, vous aurez peu de précision. Mais si elle est trop fine, votre calcul ne tournera pas dans le temps imparti ! Testez des sous parties de votre programme au fur et à mesure de son avancée.*