ENSIIE **Introduction à la statistique**
2016-2017 **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), `cor.test( . , method='spearman')` (test de correlation de Spearman). Il vous faut également charger le paquet `nortest`, qui vous donnera accès au test de Lilliefors `lillie.test` (qui correspond au test KS d'adéquation à 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 Estimation de densité

Dans cette section, nous allons 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 $[0, 5]$ (fonction `seq`).

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. (fonctions `rnorm` et

`dnorm`)

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 :

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

où $\varphi_{\mu,\sigma}$ désigne la densité de la distribution gaussienne de moyenne μ et variance σ. Cette distribution est appelée "mélange de gaussiennes".

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

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

2-b) Générer un échantillon i.i.d. `X` de taille n selon la densité f_X.

Indication. On pourra générer un échantillon i.i.d. `U` de taille n selon la distribution (1) par la commande suivante.

```
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

(*La commande* `par(mfrow=c(k,l))` *permet d'afficher* $k \times l$ *graphiques sur une même figure. Elle doit être entrée avant les plots. La commande* `lines()` *permet*

d'ajouter une courbe sur un graphe existant.).

1-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.

Partie C : estimation 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.

1-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.

1-c) Déterminer la valeur de la fenêtre pour les autres méthodes de sélection de fenêtre disponibles.

1-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.

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 :

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

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 $h = 0.05, 0.2, 0.4, 0.8$ (on utilisera $N = 10^5$ dans (1)).

Partie D : exemples sur des données réelles

1) On considère le jeu de données `precip` du package `datasets`, qui fournit les précipitations annuelles pour les 70 états des Etats Unis. Regarder la distribution des données et décrivez ce que vous observez (existence de sous groupes, etc)

2) On considère le jeu de données `Aids2` du package `MASS`, qui comporte des données sur des patients australiens diagnostiqués séropositifs. Regarder la distribution de l'âge des patients. Dans quelle tranche d'âge se situe la majorité des disgnostiques ? Vers quel âge les patients sont-ils le plus fréquemment diagnostiqués ?

2 Tests non paramétriques : cas pratiques

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

Exercice 1. Les données `anorexia` du package `MASS` comportent des mesures du poids de 72 patientes avant et après traitement. La première colonne donne le traitement reçu (3 valeurs), la deuxième et la troisième colonne le poids avant et après

traitement. Le fichier descriptif est disponible par la commande `?anorexia` et le début du tableau de données peut être affiché par la commande `head(anorexia)`

1) On veut tout d'abord tester la différence de poids avant et après traitement.
   a) Quel test paramétrique pourrait-on envisager ? Les conditions d'application sont-elles remplies ?
   b) Proposer une procédure de test non-paramétrique et conclure.

2) On veut maintenant tester si le changement de poids avant-après traitement diffère selon le traitement reçu. Plus précisément, on veut tester la différence entre le traitement contrôle ("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 ? Les conditions d'application sont-elles remplies ?
   c) Appliquer le test de Mann Whitney. Un message d'avertissement mentionne l'existence d'ex-aequos. Qu'en pensez-vous ? Conclure en répondant à la question posée.

Exercice 2. 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 être 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 diffère entre les Etats du nord et du sud. Proposer une procédure de test et conclure.

Exercice 3. Le data frame `Melanoma` du package `MASS` rassemble des données épidémiologiques sur le mélanome collectées sur 205 patients danois. Le fichier descriptif est disponible par la commande `?Melanoma` et le début du tableau de données peut être affiché par la commande `head(Melanoma)`. La colonne `time` indique la durée de survie après le diagnostique, et la colonne `sex` le sexe du patient. La durée de survie après le diagnostique dépend-elle du sexe ?

Exercice 4. Le data frame `Animals` du package `MASS` donne le poids moyen du cerveau et le poids moyen du corps pour 28 espèces. Le fichier descriptif est disponible par la commande `?Animals` et le début du tableau de données peut être 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 à la question posée par un modèle linéaire. Qu'en pensez-vous ?
b) Proposer un test non paramétrique pour répondre à la question.