

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 corrélation 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'` implémente 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 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.

- 1) On veut tout d'abord tester la différence de poids avant et après 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-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 ? 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 à 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 à 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 $(X_{(1)}, \dots, X_{(n)})$ est la statistique ordonnée associée à l'échantillon.

- b) Déterminez la valeur 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 considère 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 à 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 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 5 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.

2 Estimation de densité : aspects théoriques

Exercice 6 Considérons X_1, \dots, X_n n variables aléatoires i.i.d. de densité f_X . On cherche à estimer f_X à partir des observations X_1, \dots, 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, \dots, 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, \dots, X_n . Considérons la base d'histogrammes définie par

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

Soit $S_D = \text{vec}\{\varphi_k, k = 1, \dots, D\}$. 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

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

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 \hat{f}_D .
4. Montrer que le risque quadratique intégré de \hat{f}_D s'écrit

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

5. Montrer que

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

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

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

7. En déduire une majoration du risque quadratique intégré de \hat{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 d'échantillon $n = 1000$, 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

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

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 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 $i = 1, \dots, n$, 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 paraît 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 fourni 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 paraît 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 à 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

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

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)).

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. Estimer et commenter la distribution des données (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 diagnostiqués ? Vers quel âge les patients sont-ils le plus fréquemment diagnostiqués ?

La table ci-dessous donne la valeur critique t_ε telle que $P(D_n > t_\varepsilon) = \varepsilon$ pour le test de Kolmogorov Smirnov, avec $D_n := \|F_n - F\|_\infty$, F_n étant la fonction de répartition empirique.

$n \backslash \varepsilon$	0,2	0,1	0,05	0,02	0,01
4	0,492 6	0,565 2	0,623 9	0,688 9	0,734 2
5	0,447 0	0,509 4	0,563 3	0,627 2	0,668 5
6	0,410 4	0,468 0	0,519 3	0,577 4	0,616 6
7	0,381 5	0,436 1	0,483 4	0,538 4	0,575 8
8	0,358 3	0,409 6	0,454 3	0,506 5	0,541 8
9	0,339 1	0,387 5	0,430 0	0,479 6	0,513 3
10	0,322 6	0,368 7	0,409 3	0,456 6	0,488 9
11	0,308 3	0,352 4	0,391 2	0,436 7	0,467 7
12	0,295 8	0,338 2	0,375 4	0,419 2	0,449 1
13	0,284 7	0,325 5	0,361 4	0,403 6	0,432 5
14	0,274 8	0,314 2	0,348 9	0,389 7	0,417 6
15	0,265 9	0,304 0	0,337 6	0,377 1	0,404 2
16	0,257 8	0,294 7	0,327 3	0,365 7	0,392 0
17	0,250 4	0,286 3	0,318 0	0,355 3	0,380 9
18	0,243 6	0,278 5	0,309 4	0,345 7	0,370 6
19	0,237 4	0,271 4	0,301 4	0,336 8	0,361 2
20	0,231 6	0,264 7	0,294 1	0,328 7	0,352 4
21	0,226 2	0,258 6	0,287 2	0,321 0	0,344 3
22	0,221 2	0,252 8	0,280 9	0,313 9	0,336 7
23	0,216 5	0,247 5	0,274 9	0,307 3	0,329 5
24	0,212 0	0,242 4	0,269 3	0,301 0	0,322 9
25	0,207 9	0,237 7	0,264 0	0,295 2	0,316 6
26	0,204 0	0,233 2	0,259 1	0,289 6	0,310 6
27	0,200 3	0,229 0	0,254 4	0,284 4	0,305 0
28	0,196 8	0,225 0	0,249 9	0,279 4	0,299 7
29	0,193 5	0,221 2	0,245 7	0,274 7	0,294 7
30	0,190 3	0,217 6	0,241 7	0,270 2	0,289 9
31	0,187 3	0,214 1	0,237 9	0,266 0	0,285 3
32	0,184 5	0,210 9	0,234 2	0,261 9	0,280 9
33	0,181 7	0,207 7	0,230 8	0,258 0	0,276 8
34	0,179 1	0,204 7	0,227 4	0,254 3	0,272 8
35	0,176 6	0,201 9	0,224 3	0,250 7	0,269 0
36	0,174 2	0,199 1	0,221 2	0,247 3	0,265 3
37	0,171 9	0,196 5	0,218 3	0,244 0	0,261 8
38	0,169 7	0,193 9	0,215 4	0,240 9	0,258 4
39	0,167 5	0,191 5	0,212 7	0,237 9	0,255 2

FIGURE 1 – Table de Kolmogorov-Smirnov