TP 1 : Correction complète

1 Estimation de densité

Partie B

1-b) La valeur $D = 15$ semble la plus appropriée. Si D est plus petit, les variations de la densité ne sont pas bien restituées. Si D est plus grand, l'estimateur présente de grandes irrégularités qui ne correspondent pas à de véritables irrégularités de la densité (il est trop instable).

2) La valeur $D = 50$ semble la plus appropriée.

3) On remarque que le D optimal est plus grand pour f_X qui est plus irrégulière que pour f_Y. En effet, d'un point de vue théorique, le D optimal diminue quand la régularité de la densité augmente.

Partie C : estimation par noyaux

1-b) La fenêtre par défaut vaut 0.2.

1-c) La valeur de la fenêtre varie de 0.09 à 0.23 selon les méthodes.

1-d) La fenêtre optimale parmi ces 4 valeurs est 0.22. Si la fenêtre est trop petite, l'estimateur présente d'importantes irrégularités qui ne correspondent pas à la vraie densité. Si la fenêtre est trop petite, l'estimateur est trop lissé.

2) La fenêtre optimale parmi ces 4 valeurs est 0.1.

Partie D : exemples sur des données réelles

1) On observe deux sous groupes : les Etats avec faible pluviométrie (0-20 cm par an) et les Etats avec forte pluviométrie (25-60 cm par an). Le deuxième groupe comporte plus d'Etats

2) La majorité des diagnostiques intervient entre 20 et 60 ans. L'âge auquel les patients sont le plus fréquemment diagnostiqués est environ 35 ans.

2 Tests non paramétriques : cas pratiques

Exercice 1.

Soit $Z = (Z_1, \ldots, Z_{72})$ la différence de poids (poids après - poids avant traitement)

pour les 72 patients.
1-a) On peut considérer le test de Student à condition que l'échantillon Z soit gaussien. Le test de Lilliefors au niveau 0.05, ainsi que les observations graphiques nous permettent ne nous permettent pas de rejetter l'hypothèse de normalité de Z. Nous pouvons donc appliquer le test de Student.

— Conditions d'application : $Z \sim \mathcal{N}(\mu, \sigma^2)$

— H0 : "$\mu = 0$"

Le test de Student fournit une p-value de 4.10^{-3}, donc au niveau 0.05, H0 est rejetée. Ainsi on conclut que le poids des patients change au cours du traitement.

1-b) On peut également tester la différence de poids avant/après traitement par un test de signe et rang de Wilcoxon. L'hypothèse H0 est "la distribution d'où est issu l'échantillon Z est symmétrique". Nous obtenons une p-value de 0.01, donc H0 est rejetée. Ainsi, comme de la même façon qu'avec le test de Student, on conclut à un changement de poids significatif au cours du traitement.

2) Soit $Z^1 = (Z_1^1, \ldots, Z_{26}^1)$ et $Z^2 = (Z_1^2, \ldots, Z_{29}^2)$ les changements de poids avant/après traitement pour les individus sous traitement "contrôle" et sous traitement "familial".

2-a) Avec qqplot, le changement de poids semble plus faible avec le traitement "contrôle" qu'avec le traitement "familial".

2-b) On peut appliquer un test de Student, à condition que les échantillons Z^1 et Z^2 soient gaussiens. D'après le test de Lilliefor (p-values 0.6 et 0.8), on ne rejette pas l'hypothèse de normalité de Z^1 et Z^2. On peut donc appliquer le test de Student.

2-c) En alternative, on peut également appliquer le test de Mann Whitney. L'hypothèse nulle H0 est "le changement de poids avant/après traitement est identiquement distribué pour les traitements contrôle et familial", c'est à dire "Z^1 et Z^2 sont issus de la même distribution".

Le test de Mann Whitney suppose que les distributions sont diffuses, c'est à dire qu'il n'y a pas d'ex-aequo ; en en cas d'ex-aequos, la fonction `wilcox.test` fournit une valeur approchée de la p-value. Néanmoins, on constate qu'il y a seulement un ex-aequo dans Z^1 et aucun ex-aequo dans Z^2, donc les conclusions du test de Mann Whitney demeurent fiables.

On obtient une p-value de de 0.004, donc on rejette H0 au niveau 0.05. Ainsi, il y a une différence significative sur le changement de poids entre les traitements contrôle et familial.

Exercice 2.

Soit $X = (X_1, \ldots, X_{16})$ et $Y = (Y_1, \ldots, Y_{31})$ le taux de criminalité dans les 16 Etats du sud et les 31 Etats du nord. La fonction qqplot semble indiquer une différence entre les distributions de X et Y, la médiane de X étant inférieure à celle de Y.

Nous allons tester si cette différence est significative avec un test de Mann Whitney. L'hypothèse nulle H0 est "la criminalité est identiquement distribuée dans les Etats du nord et du sud" c'est à dire "X et Y sont issus de la même distribution". Le test donne une p-value de 9.10^{-5}, H0 est donc rejetée au niveau 0.05 : la criminalité n'est pas la même dans les Etats du sud et du nord ; elle est plus élevée dans les Etats du sud.

Exercice 3.
Soit $X = (X_1, \ldots, X_{126})$ et $Y = (Y_1, \ldots, Y_{79})$ la durée de survie après diagnostique pour 126 hommes et 79 femmes. On veut tester si la durée de survie dépend du sexe par un test de Mann Whitney. L'hypothèse nulle H0 est "La durée de survie ne dépend pas du sexe" c'est à dire "X et Y sont issus de la même distribution". On obtient une p-value de 0.04, donc H0 est rejetée au niveau 0.05 : la durée de survie dépend du sexe. De plus, la fonction qqplot indique que la durée de survie est supérieure pour les hommes.

Exercice 4.
Soit $X = (X_1, \ldots, X_{28})$ et $Y = (Y_1, \ldots, Y_{28})$ le poids moyen du corps et du cerveau pour 28 espèces. Ces deux échantillons sont appariés.

a) Considérons un modèle linéaire où Y est la variable réponse et X la variable explicative : $Y_i = \alpha X_i + \beta + \varepsilon_i$ avec $(\varepsilon_i,)_{i=1,\ldots,28}$ i.i.d. de distribution $\mathcal{N}(0, \sigma^2)$ et indépendants de X. On peut vérifier a posteriori les hypothèse du modèle en regardant les résidus : l'hypothèse de normalité des résidus n'est pas cruciale, mais l'hypothèse d'indépendance par rapport à X doit être vérifiée. Or, on constate que les résidus dépendent fortement de X, donc le modèle linéaire est innaproprié.

b) On applique le test de corrélation de Spearman unilatéral. L'hypothèse nulle H0 est "le poids du corps et du cerveau ne sont pas corrélés" c'est à dire "X et Y ne sont pas corrélés" contre l'hypothèse alternative "X et Y sont en relation croissante". On obtient une p-value de 9.10^{-6}, donc H0 est rejetée au niveau 0.05 : le poids moyen du cerveau tend à augmenter avec le poids moyen du corps.