Statistiques TP 5
Test de Kolmogorov Smirnov pour deux échantillons, régression linéaire

Rappel : taper `krdc vnc://nom_du_serveur` dans un terminal.
Les sujets et les corrigés des TPs sont mis le lendemain des séances sur ma page :
http://www.math.u-bordeaux1.fr/~chabanol/stat.html

1. Test de comparaison de Kolmogorov Smirnov

Le test de Kolmogorov-Smirnov peut aussi servir à déterminer si deux lois inconnues sont identiques. En effet, soit $(X_1, \dots, X_n)$ un n-échantillon de fonction de répartition F_X, et $(Y_1, \dots Y_m)$ un m-échantillon de fonction de répartition F_Y. On suppose que les deux échantillons sont indépendants et que F_X et F_Y sont continues. On veut tester $H_0 : F_X = F_Y$ contre $F_X \neq F_Y$. Or on peut montrer que sous H_0, $D_n = \sqrt{\frac{nm}{n+m}} \sup_{x \in R} |F_X^n(x) - F_Y^m(x)|$ (où F_X^n et F_Y^n désignent les fonctions de répartition empiriques des deux échantillons) converge en loi vers la même loi que pour le test de Kolmogorov Smirnov vu précédemment, de fonction de répartition F_{KS}

(1) *Écrire une fonction qui prend en entrée un échantillon X et un réel x et fournit la fonction de répartition empirique évaluée en ce point (il faut donc compter le nombre d'éléments de léchantillon inférieurs ou égaux à x; commencer par trier l'échantillon peut être une bonne idée...).*

(2) *Écrire une fonction qui prend en entrée deux échantillons X et Y, ainsi que α, calcule D_n et donne le résultat du test au niveau α, ainsi que le seuil critique.*

(3) Voici deux tableaux représentant le revenu net en 2002 de 20 groupes français et 24 groupes allemands, en milliards d'euros :

Groupes français :

0.2	3.8	7.6	4.0	4.1	-2.8	4.7	3.6	5.4	-0.2
1.6	5.6	-0.6	0.8	-5.0	0.1	2.9	3.7	3.9	1.1

Groupes allemands :

1.8	4.0	1.4	1.9	1.9	1.8	1.4	1.9	1.4	4.5	2.2	2.4
3.1	0.3	-1.4	0.4	2.3	0.2	1.5	4.8	0.6	1.0	1.5	5.5

Représenter les deux fonctions de répartition empiriques, puis tester l'homogénéité des deux groupes.

(4) On a mesuré la hauteur des arbres dans 2 forêts, et on a trouvé (en mètres)
Forêt 1 : 23.4 24.4 24.6 24.9 25 26.2 26.3 26.8 26.9 27 27.6 27.7
Forêt 2 : 22.5 22.9 23.7 24 24.4 24.5 25.3 26 26.2 26.4 26.7 26.9 27.4 28.5
Tester l'homogénéité des deux forêts.

2. Régression linéaire

On considère le modèle de régression linéaire simple $Y_i = a + bX_i + \epsilon_i, i = 1, \dots n$ où a et b sont des paramètres inconnus, et où les ϵ_i sont des variables aléatoires indépendantes de loi $N(0, \sigma^2)$, où σ est un troisième paramètre également inconnu.

(5) *Prendre $n = 50$, $X_i = \frac{i}{n}$, $a = 1$, $b = 2$, $\sigma = 0.2$; générer $(Y_1, \dots, Y_{50})$.*

(6) On estime a et b par la méthode des moindres carrés, qui consiste comme on l'a vu à projeter le vecteur des observations Y sur le plan engendré par les deux vecteurs X et $u = (1, \dots, 1)$ et à prendre comme estimateurs $\hat{a}$ et $\hat{b}$ les coordonnées du projeté $\hat{Y}$ dans la base (u, X). On rappelle que si on note M la matrice dont les vecteurs colonnes sont u et X, on trouve $\hat{a}$ et $\hat{b}$ en effectuant $({}^tMM)^{-1}\,{}^tMY$.

Ecrire une fonction qui prend en paramètre un échantillon Y (rentré comme vecteur colonne) et un vecteur colonne X, fournit les valeurs estimées $\hat{a}$ et $\hat{b}$, et trace les points correspondant aux valeurs observées ainsi que la droite de régression obtenue.
Indications : si u et X sont deux vecteurs colonnes, `M=[u,X]` fournit la matrice dont ce sont les deux vecteurs colonnes. On peut aussi faire `M(:,1)=u; M(:,2)=X`
Pour tracer des points avec `plot`, il faut lui donner en paramètre un nom de symbole, par exemple `'o'`.
Tester votre fonction avec votre échantillon.

La tester également avec un échantillon fourni par matlab, que vous pouvez obtenir grâce à getdata(5) : prendre X pour la première colonne, Y pour la deuxième. Recommencer en enlevant les valeurs particulières correspondant aux dinosaures (3 valeurs)
(si I est un sous-ensemble d'indices et X un vecteur, `X(I)=[]` fait disparaître les indices I)

(7) Quelle est la matrice de covariance du vecteur $\frac{1}{\sigma}\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix}$? Pour obtenir un intervalle de confiance pour a et b on a besoin d'estimer σ. Pour cela on utilise $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2$ (remarque : $\hat{Y} = \hat{a} + \hat{b}X_i = M({}^tMM)^{-1}\,{}^tMY$ est le projeté de Y sur le plan, et $\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2$ est la distance de Y au plan sur lequel on projette).
Modifier votre fonction pour qu'elle fournisse également un estimateur de σ.
Quelle est la loi de $(n-2)\frac{\hat{\sigma}^2}{\sigma^2}$? Pourquoi est-il indépendant du vecteur $(\hat{a}, \hat{b})$? Quelle est la loi de $\frac{\hat{a}-a}{\sqrt{var(\hat{a})}}\frac{\sigma}{\hat{\sigma}}$?
Modifier votre fonction pour qu'elle fournisse un intervalle de confiance pour a et pour b au risque 0.05 (faire help stixbox pour savoir comment obtenir les quantiles d'une loi de Student), et représenter les droites correspondant aux valeurs extrémales de vos intervalles. Tester avec vos deux jeux de valeurs.

(8) La méthode précédente peut en fait s'appliquer dans un cadre plus général où $Y_i = a + b_1X_i^{(1)} + b_2X_i^{(2)} + \ldots + b_kX_i^{(k)} + \epsilon_i$ où on a k variables "explicatives" $X^{(1)}, \ldots, X^{(k)}$. La méthode consiste toujours à estimer $a, b_1, \ldots, b_k$ en projetant cette fois sur un espace de dimension $(k+1)$, engendré par u et les $X^{(j)}$.
Certaines lois changent dans les calculs précédents. Lesquelles ?
Construire une fonction qui prend en entrée un échantillon (fourni en colonne), une matrice contenant en colonne les $X^{(j)}$, et fournit les estimations $\hat{a}, \hat{b_1} \ldots, \hat{b_k}, \hat{\sigma}$. Tester avec les données fournies par getdata(4) (en prenant pour X les 3 premières colonnes et pour Y la quatrième ou la cinquième) ainsi que par getdata(11) (en prenant pour Y la dernière colonne).

(9) On revient au cas $k = 1$. En fait $\hat{a}$ et $\hat{b}$ ne sont pas indépendants... Il peut être donc plus pertinent de chercher une région de confiance dans $\mathbb{R}^2$ pour le couple (a, b) que des intervalles de confiance. Pour cela, on utilise le fait que $\frac{||(\hat{Y}-(au+bX))||^2}{\sigma^2}$ suit une loi du chi2 à 2 degrés de liberté et est indépendant de $\hat{\sigma}^2$ (pourquoi ?) On peut alors en déduire que $\frac{||(\hat{a}-a)u+(\hat{b}-b)X||^2}{2\hat{\sigma}^2}$ suit une loi appelée loi de Fischer à $(2, n-2)$ degrés de liberté. Ses quantiles peuvent également être fournis par matlab. On peut ainsi écrire une région de confiance pour (a, b), qui est en fait un ellipsoïde de confiance.
Le dessiner pour votre premier jeu de valeur (indication : si une ellipse est donnée par une équation de la forme $A(x-x_0)^2 + B(y-y_0)^2 + C(x-x_0)(y-y_0) = D$, on peut la paramétrer par $x = x_0 + r(t)\cos(t), y = y_0 + r(t)\sin(t)$, où r est une fonction de t qui s'exprime facilement en fonction de A, B, C et D.)