NOM, Prénom : ..... Date : ...../...../.....

Partiel de Traitement de données

Consignes

— Le partiel dure 1h45.
— Vous avez le droit à une feuille A4 manuscrite recto-verso, les autres documents sont interdits.
— Le sujet est à rendre avec la copie.

1 Questions de cours (8 pts)

Les questions sont indépendantes, elles peuvent être traitées dans l'ordre qui vous convient.

Question 1 – Rappelez la définition de la variance. Quelle interprétation faîtes-vous d'une variable avec une variance élevée par rapport à une variable avec une faible variance ?

Question 2 – Que représente l'histogramme d'une variable X, comment influe le nombre N d'observations de cette variable sur l'histogramme ?

Question 3 – Que signifie un coefficient de corrélation positif entre deux variables X et Y ? Donner deux exemples différents (représentations graphiques) de variables X et Y liées par un coefficient de corrélation positif.

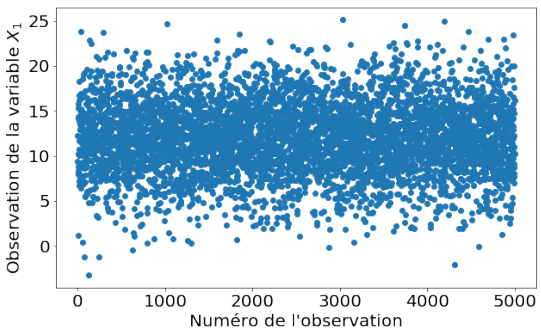
Question 4 – Dans quel(s) but(s) effectue-t-on une analyse en composantes principales (ACP) ?

Question 5 – Rappelez les étapes à réaliser pour appliquer une ACP sur un jeu de données multivariées.

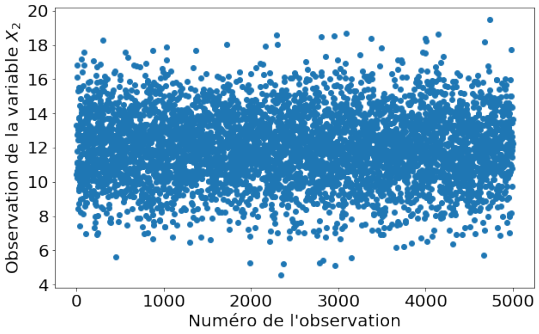
2 Exercices d'application (12 pts)

Exercice 1 – Mettre en correspondance différentes représentations graphiques.

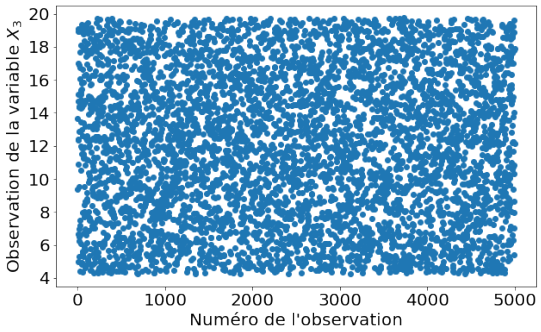
Pour les quatre variables $X1, X2, X3$ et $X4$ présentées sur la figure 1 :
— associer les boîtes à moustaches de la figure 2 avec la représentation par nuage de points correspondante. Chaque association doit être précisément justifiée, sinon elle ne sera pas prise en compte dans la notation,
— représenter l'allure générale de l'histogramme de chacune des variables (on prendra soin d'indiquer quelques valeurs remarquables sur l'axe des abcisses).



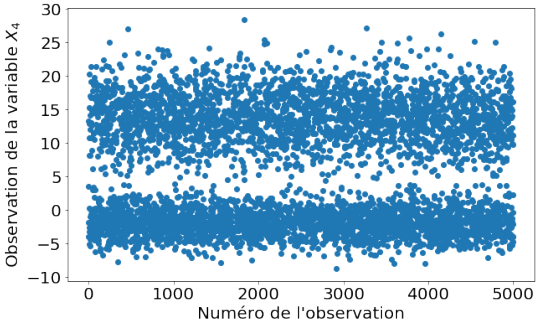
($X1$)



($X2$)

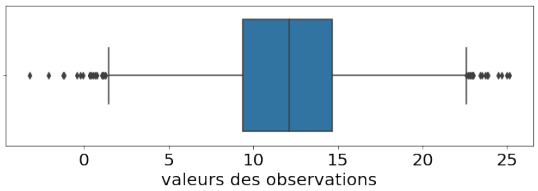


($X3$)

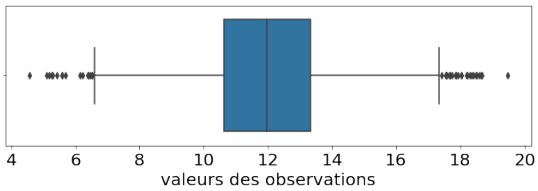


($X4$)

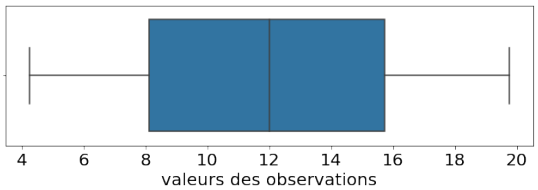
FIGURE 1 – Représentation graphique par nuage de points des variables $X1, X2, X3$ et $X4$.



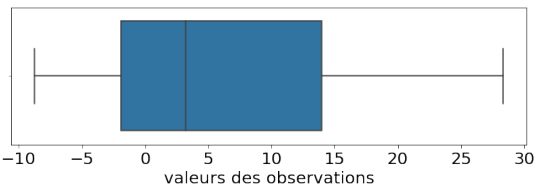
(a)



(b)



(c)



(d)

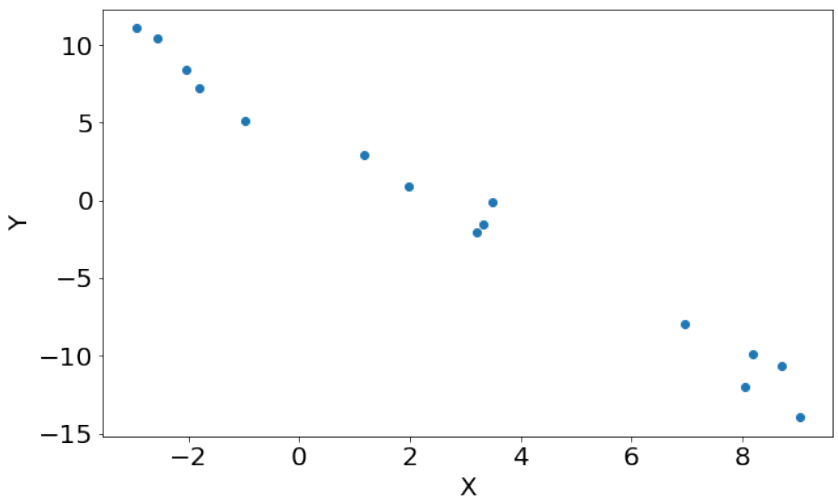
FIGURE 2 – Boites à moustaches (diagramme de Tukey) associées aux 4 variables $X1, X2, X3$ et $X4$.

Exercice 2 – Analyser les relations entre deux variables X et Y.

Dans les deux cas proposés sur les figures 3 et 4 :

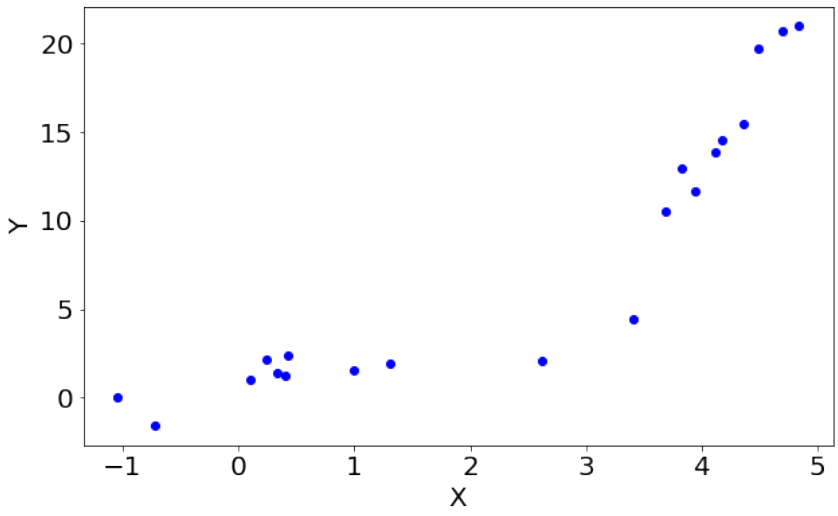
- décrivez la relation qui lie X et Y (le modèle exact n'est pas demandé, seulement une description générale de la relation)
- que peut-on dire du coefficient de corrélation $cor(X, Y)$?

Les coordonnées des différents points (i.e. les valeurs prises par le couple de variables (X, Y)) sont données à titre indicatif.



X	Y
9.0	-13.9
3.2	-2.1
3.3	-1.5
8.7	-10.7
3.5	-0.1
-2.0	8.4
8.2	-9.9
1.2	2.9
-2.9	11.1
8.1	-12.0
2.0	0.9
7.0	-8.0
-2.6	10.4
-1.0	5.1
-1.8	7.2

FIGURE 3 – Cas 1 : représentation graphique de 15 réalisations du couple de variables (X, Y). Les coordonnées des différents points sont données dans le tableau de droite (les points ne sont pas ordonnés).



X	Y
0.4	2.4
-1.0	0.1
-0.7	-1.6
2.6	2.1
1.0	1.6
1.3	2.0
0.3	1.4
0.2	2.2
0.4	1.2
0.1	1.0
4.7	20.7
4.8	21.0
3.7	10.6
4.4	15.5
4.2	14.6
4.1	13.9
3.4	4.4
3.9	11.7
3.8	13.0
4.5	19.7

FIGURE 4 – Cas 2 : représentation graphique de 20 réalisations du couple de variables (X, Y). Les coordonnées des différents points sont données dans le tableau de droite (les points ne sont pas ordonnés).

Exercice 3 – Histogramme et probabilité.

On appelle Z la variable "année de réalisation" de 29 films sélectionnés au hasard dans la base de données IMdB.

$$Z = [1994, 2008, 2010, 1999, 1994, 1994, 2001, 1999, 2003, 1972, 2012, 2002, 1995, 2012,$$
$$2000, 2005, 2012, 2014, 1997, 1991, 2009, 2009, 1998, 2006, 1993, 2000, 2006, 1980, 1999]$$

— Construire l'histogramme de cette variable, on choisira soigneusement le nombre et la largeur des classes.
— Donner la valeur des premier et troisième quartiles ainsi que la médiane.
— Peut-on conclure que dans la base de données IMdB on a 25% de chance de trouver un film produit après 2008 ? Justifier la réponse.