Objectif Décrire efficacement d'importants jeux de données.
Rechercher l'existence d'une relation (corrélation) affine entre deux variables.

I - Série statistique à une variable

1) Un peu de vocabulaire...

Un **caractère**, ou **variable**, est une propriété commune aux **individus** d'une **population**.
Un **échantillon** est une partie de la population complète.
L'**effectif** d'une population ou d'un échantillon est le nombre d'individus qui la compose.
Un caractère peut-être **quantitatif**, s'il peut s'exprimer par un nombre, ou **quantitatif** dans le cas contraire.
On peut de plus distinguer les caractères quantitatifs **discrets**, qui ne prennent que des valeurs numériques isolées, des caractères quantitatifs **continus**, lorsque toutes les valeurs peuvent être prises dans un intervalle.

2) Caractéristiques de position

Les caractéristiques, ou indicateurs, de position d'une série les plus utilisés sont les suivantes.
Ces caractéristiques sont aussi appelées **indicateurs de tendance centrale** et permettent de situer le niveau global de la série.

Définition *On considère N valeurs d'un caractère x_1, x_2,...,x_N.*
La moyenne, notée $\overline{x}$, est :

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Si la valeurs x_1 est prise n_1 fois par le caractère, la valeur x_2 prise n_2 fois, ..., alors

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_Nx_N}{N} = \frac{1}{N}\sum_{i=1}^{n} n_ix_i \quad avec \quad N = \sum_{i=1}^{n} n_i$$

Définition *La médiane M_e d'une série statistique* **ordonnée** *est une valeur qui partage la population en deux groupes de même effectif.*

Si l'effectif total de la série est impair : $N = 2p+1$, la médiane est la $(p+1)^{ème}$ valeur.
Si l'effectif est pair : $N = 2p$, on prend en général pour médiane la moyenne de la $p^{ème}$ et de la $(p+1)^{ème}$ valeur.

Définition *Le* **mode** *d'une série statistique est la valeur du caractère la plus fréquente.*

Exemple : Soit la série statistique :

Notes x_i	6	8	10	12	15	18
Nombre d'élèves n_i	1	5	3	4	2	2

La moyenne de cette série est $\overline{x} \simeq 11,18$.

L'effectif total est $N = 17$. La médiane est donc la $9^{ème}$ valeur de la série ordonnée, soit $M_e = 10$. Il y a ainsi 8 valeurs qui lui sont inférieures et 8 supérieurs.

Son mode est 8.

La moyenne ou la médiane d'une série permet de situer le niveau global de celle-ci, mais ne donne pas d'information sur la répartion des valeurs.

Ainsi, les séries statistiques :

10, 10, 10, 10, 10, 10, 10 et 2, 2, 2, 10, 18, 18 18 ont le même effectif et les mêmes moyennes et médianes. On voit bien néanmoins que la description par uniquement une de ces deux caractéristiques est limitée et ne rend pas compte de la **dispersion** de l'ensemble des valeurs.

Pour décrire une série statistique, on doit donc fournir, en plus d'une caractéristique de position, une caractéristique de dispersion.

Définition *L'***étendue** *d'une série est l'écart entre les valeurs extrêmes de la série.*

Définition *La* **variance d'une série** *est la moyenne des carrés des écarts à la moyenne :*

$$V = \frac{1}{N}\sum_{i=1}^{n} n_i \left(x_i - \overline{x}\right)^2$$

*L'***écart type** σ *de la série est la racine carrée de la variance :* $\sigma = \sqrt{V}$.

Propriété *La variance est égale à la moyenne des carrés moins le carré de la moyenne :*

$$V = \overline{x^2} - \overline{x}^2$$

Définition *Les quartiles* Q_1, Q_2 *et* Q_3 *d'une série sont trois valeurs de la série ordonnée qui la partagent en quatre séries de même effectif (25% de l'effectif total).*
Le deuxième quartile est la médiane : $Q_2 = M_e$.
L'écart inter-quartile est le nombre $Q_3 - Q_1$.

On définit de la même façon les déciles D_1, D_2,...,D_9 *d'une série, en partageant la série en dix séries de même effectif (10% de l'effectif total).*
L'écart inter-décile est le nombre $D_9 - D_1$.

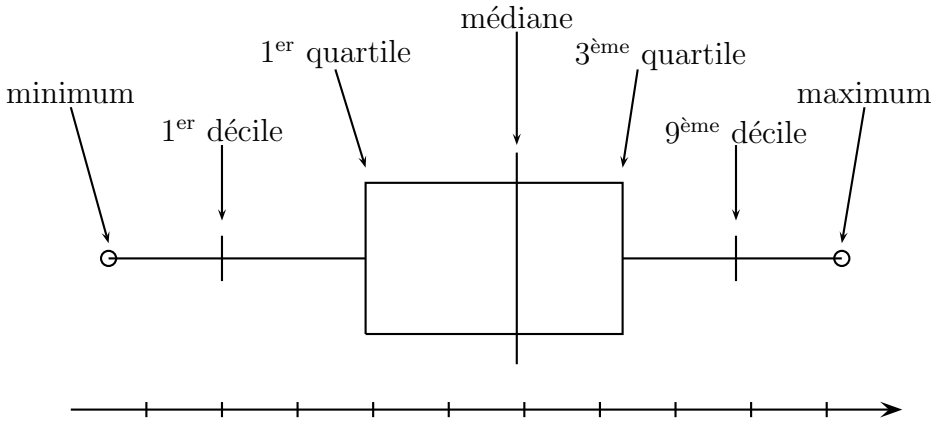
Remarque : Dans le cas d'une série statistique continue, on regroupe les valeurs en classes (ou intervalles). Les indicateurs de position et de dispersion sont alors calculés en utilisant le centre des classes.

4) Diagramme en boîte

La représentation d'une série à l'aide d'un diagramme en boîte (ou diagramme à pattes, ou boîte à moustaches, ou Whiskers plots) repose sur la description de la série par ses quantiles.

Cette représentation a été introduite en 1977 par John Tukey[1].

1. John Wilder Tukey (16 juin 1915 - 26 juillet 2000) est un important statisticien américains. Il a créé et développé de nombreuses méthodes statistiques.
Il est notamment connu pour son développement en 1965, avec James Cooley, de l'algorithme de la transformée de Fourier rapide (*fft*).



Exercice 1 Soit la série statistique :

Longueur x_i (mm)	4.7	4.8	4.9	5.0	5.1	5.2	5.3
Effectifs n_i	1	4	23	30	27	9	6

Déterminer le mode, la moyenne, l'écart type, la médiane, l'étendue, et les écarts inter-quartiles et inter-deciles de cette série.

Exercice 2 On mesure, en millimètres, le diamètre de 100 pièces prises au hasard dans la production d'une machine. On obtient les résultats suivants :

Diamètre x_i (mm)	Effectifs n_i
80,36	8
80,37	19
80,38	55
80,39	36
80,40	10
80,41	11
80,42	5

Soit σ l'écart type de cette série statistique. Un réglage de la machine est nécessaire lorsque $\sigma > 0,013$.

Faut-il régler la machine ?

II - Série statistique à deux variables - Ajustement affine

On s'intéresse à l'étude, sur une population donnée, du lien qui peut exister entre deux caractères.

On peut présenter l'étude générale sous la forme suivante :

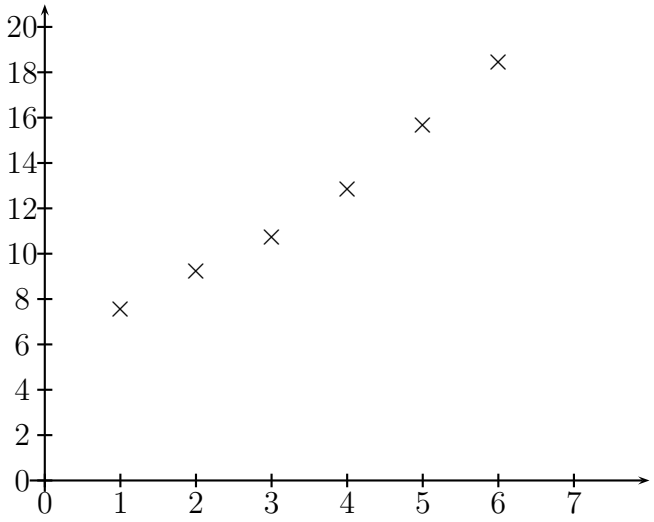
Valeurs du 1er caractère x_i	x_1	x_2	x_3	...	x_k
Valeurs du 2ème caractère y_i	y_1	y_2	y_3	...	y_k

Exemple : L'étude du coût de maintenance annuel d'une installation de chauffage dans un immeuble de bureaux, en fonction de l'âge de l'installation, a donné les résultats suivants :

Age x_i (années)	1	2	3	4	5	6
Coût y_i (k€)	7,55	9,24	10,74	12,84	15,66	18,45

Y'a-t-il un lien entre l'âge de l'installation et le coût de maintenance ? Si oui, peut-on le quantifier, et peut-on, par exemple, prévoir le coût de maintenance d'une installation de 7 ans ? 8 ans ? 10 ans ?

On appelle nuage de points, l'ensemble des points A_i de coordonnées $(x_i ; y_i)$.



Définition *Le* **point moyen** *du nuage de points est le point de coordonnées* $(\overline{x} ; \overline{y})$.

Exemple : Dans l'exemple précédent, le point moyen G a pour coordonnées $(3,5 ; 12,41)$.

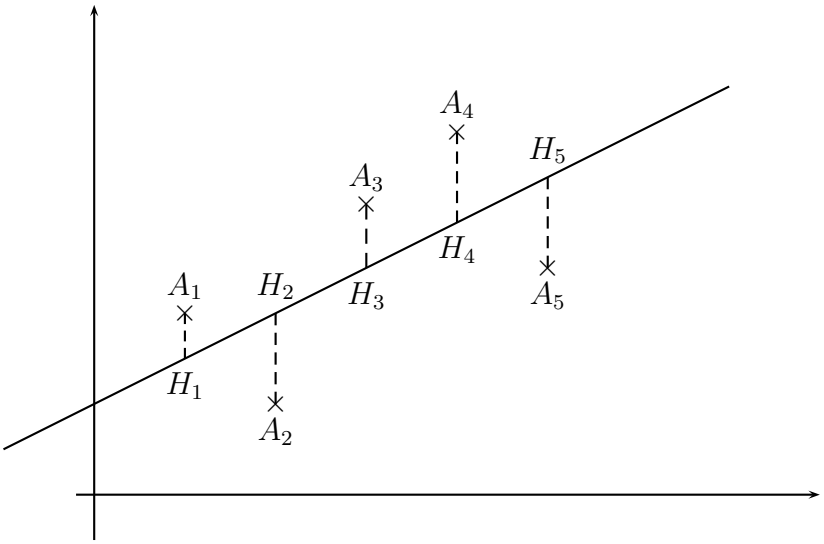
2) Ajustement affine par la méthode des moindres carrés

Les points de l'exemple précédents ne sont pas alignés. Néanmoins, ces points semblent se distribuer approximativement autour d'une droite.

La méthode des moindres carrés permet de déterminer l'équation de la "meilleure" droite passant dans le nuage de points, ainsi que de quantifier la "qualité de l'alignement des points" du nuage.

On considère un nuage de points $A_k(x_k ; y_k)$.

Pour une droite quelconque, on peut définir la "distance" de la droite au nuage de points par la somme des distances $A_k H_k$. Ainsi, la "meilleure" droite passant dans le nuage de points est celle dans la distance au nuage de points est la plus petite.



$$\sum_{k=1}^{n} A_k H_k^2$$

soit minimum.

Cette droite est appelée **droite de régression de** y **en** x, *ou encore* **droite des moindres carrés**.

Cette droite de régression passe par le point moyen $G(\overline{x};\overline{y})$ et a pour coefficient directeur $m = \dfrac{\text{cov(x ;y)}}{V(x)}$, où $V(x)$ est la variance de x :

$$V(x) = \frac{1}{n}\sum_{k=1}^{n}(x-\overline{x})^2$$

et cov(x ;y) est la covariance de x et de y :

$$\text{cov}(x,y) = \frac{1}{n}\sum_{k=1}^{n}(x-\overline{x})(y-\overline{y})$$

Remarque : $V(x) = \text{cov}(x,x)$.
On note en général $\sigma_{xy} = \text{cov(x,y)}$ la covriance de x et de y, et $\sigma_x = \sigma_{xx} = \sqrt{V(x)}$.

Propriété *L'équation de la droite de régression de* y *en* x, *ou droite des moindres carrés, est :*

$$y - \overline{y} = \frac{cov(x,y)}{V(x)}(x-\overline{x})$$

soit aussi

$$y - \overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x-\overline{x})$$

Remarque : La droite de régression de x en x est $y = x$!

Exemple : La droite de régression de l'exemple précédent a pour équation $y = 2,17x + 4,83$.

3) Coefficient de corrélation

La droite de régression est la droite la plus proche de tous les points du nuage. Néanmoins, l'idée d'approcher tous les points du nuage par une droite peut-être plus ou moins pertinent.
Le coefficient de corrélation est un nombre qui quantifie justement ce degré de pertinence.

Définition *Le coefficient de corrélation linéaire entre* x *et* y *est le nombre*

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Propriété

- $-1 \leqslant r \leqslant 1$
- r *a le même signe que le coefficient directeur de la droite de régression.*
- *La corrélation est d'autant meilleure que* $|r|$ *est proche de* 1 *(si* $r = 1$ *ou* $r = -1$, *les points sont alignés et la corrélation est parfaite).*

Une erreur (malheureusement) assez répandue consiste à confondre corrélation avec causalité.
Observer que deux variables sont corrélées entre elles ne signifie pas que l'une soit la conséquence de l'autre, c'est-à-dire qu'il ait un lien de cause à effet.

Par exemple, dans le monde, au 20ème siècle par exemple, le nombre de mariages a augmenté ainsi que le nombre de décés. Ces deux variables sont sûrement corrélées, ce qui ne montre en aucun cas l'existence d'un lien de cause à effet d'un phénomène à l'autre (en fait ces deux augmentations peuvent être directement reliées à une augmentation commune : l'augmentation de la démographie mondiale).

5) Exercices

Exercice 3 Droite de Mayer

Le tableau suivant donne la durée moyenne d'intervention, en minutes, sur les postes de télévision en panne dans un atelier de dépannage, de 1992 à 2000.

Rang x_i	1	2	3	4	5	6	7	8	9
Année	1992	1993	1994	1995	1996	1997	1998	1999	2000
Durré moyenne d_i	83	82	80	75	73	74	71	71	70

Partie A. Ajustement à l'aide de la droite de régression

1. Calculer le coefficient de corrélation linéaire entre x et d (à 10^{-3} près).
   Semble-t-il y avoir une dépendance affine entre le rang de l'année et la durée moyenne des interventions ?
2. Par la méthode des moindres carrés, donner la droite de régression de d en x (valeurs arrondies à 10^{-2} près).
3. En supposant que l'évolution se poursuit ainsi pendant les 5 années futures, estimer la durée moyenne d'intervention dans cet atelier en 2002.

Partie B. Ajustement à l'aide de la droite de Mayer

La méthode de Mayer consiste à partager la série en 2.
Soit S_1 la série correspondant aux années 1992-1996, et S_2 la série correspondant aux années 1997-2000.

1. Déterminer les coordonnées du point moyen G_1 de la série S_1, et du point moyen G_2 de la série S_2.
2. Déterminer l'équation de la droite (G_1G_2) appelée droite de Mayer.
3. Estimer la durée moyenne d'intervention dans cet atelier en 2002 avec la droite de Mayer et comparer avec la droite de régression.

Après un accident nucléaire, on procède à intervalles de temps réguliers à des mesures de radioactivité sur un site donné. Le tableau suivant donne les résultats de ces mesures.

Rang x_i de la mesure	1	2	3	4	5	6
Valeur y_i mesurée	100	61	37	22	14	7

Pour chaque mesure on pose $z_i = \ln y_i$ et on étudie alors la série statistique $(x_i; z_i)$.

1. Compléter le tableau :

Rang x_i de la mesure	1	2	3	4	5	6
$z_i = \ln y_i$						

2. Calculer le coefficient de corrélation de cette série à 0,001 près. Commenter le résultat.
3. Donner une équation de la droite D de régression de z en x (on arrondira les coefficients à 0,01 près).
4. En déduire une relation entre x et y du type $y = \alpha e^{\beta x}$, où α et β sont deux constantes à déterminer.
5. En supposant que le modèle reste valable, en déduire pour la prochaine mesure ($x_i = 7$) une estimation de y.
6. En supposant toujours que le modèle reste valable, déteminer à partir de quelle mesure la valeur y mesurée sera inférieure à $0,01$.

Exercice 5 **Ajustement caré**

Au cours d'une séance d'essai, un pilote automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule. Au moment du top sonore, on mesure la vitesse de l'automobile puis la distance nécessaire pour arrêter le véhicule.

Pour six expériences, on a obtenu les résultats suivants :

v_i (km/h)	27	43	62	80	98	115
distance y_i d'arrêt (m)	6,8	20,5	35,9	67,8	101,2	135,8

On pose $x_i = v_i^2$ et on considère la série $(x_i; y_i)$.

1. Compléter le tableau

x_i						
y_i	6,8	20,5	35,9	67,8	101,2	135,8

2. Dans un repère orthogonal représenter le nuage de points associé à cette nouvelle série (unités : 1cm pour 1000 en abscisse, et 1 cm pour 10 en ordonnée).
3. a. Déterminer, à l'aide de la calculatrice, l'équation de la droite de régression de y en x sous la forme $y = mx + p$. Tracer cette droite dans le repère précédent.
   b. A l'aide de cette équation, déterminer la valeur estimée de x correspondant à une distance d'arrêt de 180 m, puis la vitesse correspondante du véhicule.
   c. Quelle est la vitesse d'arrêt estimée correspondant à une vitesse de 150 km/h.
   d. Le manuel du code de la route donne, pour calculer la distance d'arrêt, en mètres, la méthode suivante : "Prendre le carré de la vitesse exprimé en dizaines de kilomètres par heure."
   Comparer le résultat obtenu au c. à celui que l'on obtiendrait par cette méthode.