⋆ Statistiques descriptives bivariées ⋆

Compétences attendues

- ✓ Représenter un nuage de points associé à une série statistique double.
- ✓ Représenter la droite des moindres carrées.
- ✓ Calculer le coefficient de corrélation linéaire et interpréter sa valeur.

Liste des commandes Python exigibles aux concours

- Dans la librairie `numpy` : `sum`, `min`, `max`, `cumsum`, `mean`, `median`, `var`, `std` ;
- Dans la librairie `matplotlib.pyplot` : `hist`, `plot`, `show`

Objectifs. Les données statistiques ne vont pas toujours toutes seules, et pour un même individu, il est possible de s'intéresser à plusieurs caractères. Dans ce TP, nous nous limiterons à l'étude simultanée de deux caractères. Nous nous poserons alors la question suivante : peut-on exprimer l'un de ces caractères en fonction de l'autre ? De cette recherche de correspondances peuvent alors découler des analyses fines, explicatives voire prédictives, ou au contraire mettre en évidence une absence de corrélation entre ces caractères.

1 Série statistique double ou bivariée

1.1 Définition

Soit $\Omega = \{\omega_1, \dots, \omega_n\}$ une population d'effectif n, sur laquelle nous étudions deux caractères quantitatifs $X, Y : \Omega \to \mathbb{R}$ avec X supposé non constant. Pour tout $i \in [\![1, n]\!]$, on note :

- $x_i = X(\omega_i)$ la modalité de X prise par l'individu ω_i,
- $y_i = Y(\omega_i)$ la modalité de Y prise par l'individu ω_i.

Définition 1.1

On appelle *série statistique double* (ou *bivariée*) de la population Ω pour le couple de caractères (X, Y) la donnée du n-uplet $((x_i, y_i))_{1 \leq i \leq n}$ des modalités de (X, Y) sur Ω.

Exemple 1.1. 10 enfants de 6 ans d'une même classe sont mesurés et pesés. On note X le caractère désignant la taille de l'enfant (en centimètres) et Y celui désignant le poids de l'enfant (en kilogrammes). On obtient la série statistique double suivante :

Enfant	1	2	3	4	5	6	7	8	9	10
X	121	123	108	118	111	109	114	103	110	115
Y	25	22	19	24	19	18	20	15	20	21

Représentation informatique. On représentera une série statistique double sur Python par deux vecteurs `x` et `y` de taille n, où `(x(i),y(i))` est la modalité $(X, Y)(\omega_i)$.

On importera pour la suite du TP les librairies `numpy` (avec le raccourci `np`), `numpy.linalg` (avec le raccourci `al`), `numpy.random` (avec le raccourci `rd`) et `matplotlib.pyplot` (avec le raccourci `plt`).

1.2 Représentation graphique

On représente une série statistique double à l'aide d'un nuage de points. C'est l'ensemble des points M_i du plan de coordonnées (x_i, y_i) pour tout $1 \leq i \leq n$.

Pour tracer un nuage de points sur Python, on utilise l'instruction plt.plot.

Définition 1.2

Soient `x` et `y` deux vecteurs de même taille.

L'instruction `plt.plot(x,y, ".")` trace le nuage de points dont les abscisses sont données par `x` et les ordonnées par `y`.

Remarque 1.1. L'option `"."` a pour effet de ne pas relier les points et de les représenter par des .. On peut également utiliser `"+"` ou `"o"` pour changer la forme des points.

1.3 Modèle de régression

Lorsqu'on étudie une série statistique bivariée, on peut penser que l'un des caractères, par exemple X, est une cause de l'autre, par exemple Y. On dit alors que X est le caractère *explicatif* et Y le caractère *expliqué*. On cherche alors un modèle de régression, c'est-à-dire une expression de Y en fonction de X :

$$Y = f(X)$$

où la fonction f est appelée fonction de régression.

Pour envisager un modèle de régression satisfaisant, on trace le nuage des points (X, Y) afin de proposer une fonction de régression.

Exercice 1

1. Pour la série statistique double proposée en exemple, quel est le caractère explicatif et le caractère expliqué ?
2. Représenter le nuage de points associé à cette série statistique double. Quelle fonction de régression vous semble appropriée ?
3. Proposer une droite qui passe «très près » de tous ces points.

Dans la suite de ce TP (et le plus souvent en statistique) nous nous intéresserons plus particulièrement aux modèles de régression linéaire, c'est-à-dire les modèles pour lesquels la fonction de régression f est affine. Nous cherchons ainsi à savoir si le caractère expliqué Y peut s'exprimer comme fonction affine de X, ce qui revient à chercher $(a, b) \in \mathbb{R}^2$ tel que :

$$Y = aX + b.$$

2 Régression linéaire

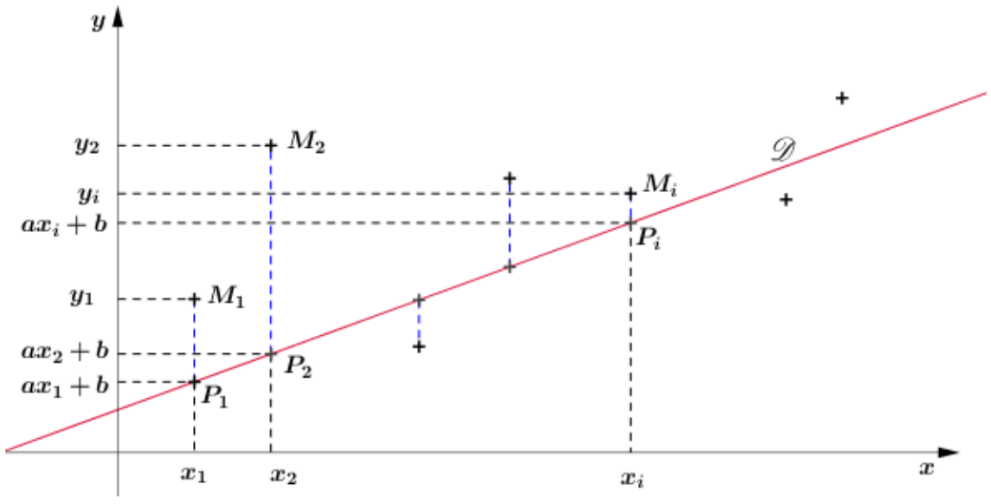
2.1 Méthode des moindres carrés

Étudions l'existence d'une relation linéaire pour la série statistique double $((x_i, y_i))$. On souhaite ainsi "placer" tous les points M_i de coordonnées (x_i, y_i) sur une même droite $\mathscr{D}$. On cherche donc $(a, b) \in \mathbb{R}^2$ tel que :

$$\forall i = 1, \dots, n, \quad y_i = ax_i + b.$$

Seulement, il y a très peu de chance qu'une telle droite existe, nos points n'étant très probablement pas alignés.

Pour obtenir un modèle de régression linéaire "le plus satisfaisant possible ", on va chercher la "meilleure" droite $\mathscr{D}$ approchant l'ensemble des points M_i au sens suivant : pour tout $1 \leq i \leq n$, on mesure la distance M_iP_i entre M_i et le point $P_i \in \mathscr{D}$ d'abscisse x_i.



On cherche $(a, b) \in \mathbb{R}^2$ rendant minimale la quantité [1] :

$$\Delta = \sum_{i=1}^{n} M_iP_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 .$$

1. Pourquoi vouloir minimiser cette quantité en particulier ? Pourquoi pas une autre, comme par exemple la somme des longueurs, ou encore la plus grande des longueurs ? Une des raisons pour lesquelles on s'intéresse à la somme des carrés est qu'on dispose alors d'un résultat garantissant l'existence et l'unicité d'une telle droite, comme on va le voir dans la section suivante.

Une telle droite, si elle existe, est appelée droite des moindres carrés.

2.2 Existence et unicité de la droite des moindres carrés

Réécrivons matriciellement ce problème. Notons $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ et $A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$.

Les points M_i sont tous sur la même droite si et seulement s'il existe $U = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathscr{M}_{2,1}(\mathbb{R})$ tel que :

$$Y = AU.$$

Si cette équation admet une solution, c'est-à-dire si Y appartient à $\mathrm{Im}(A)$, alors c'est bon : la droite $y = ax + b$ est solution du problème.

Sinon, et c'est ce cas qui est intéressant puisque les points ne sont pas alignés en général, on cherche à minimiser la quantité Δ qui se réécrit :

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 = \|Y - AU\|^2,$$

où $\|\cdot\|$ désigne la norme associée au produit scalaire canonique $\langle \cdot, \cdot \rangle$ sur $\mathscr{M}_{n,1}(\mathbb{R})$. On se retrouve dans la situation de votre cours de mathématiques sur la recherche de pseudo-solutions d'un système linéaire : on sait (si A est de rang 2, ce qui est effectivement le cas car les x_i ne sont pas tous égaux) que ce problème admet une unique solution $U = \begin{pmatrix} a \\ b \end{pmatrix}$. On obtient donc le théorème suivant.

Théorème 2.1 [Problème des moindres carrés : régression linéaire]

Considérons une série statistique double $((x_i, y_i))_{1 \leq i \leq n}$.

Il existe une et une seule droite minimisant la quantité Δ. On l'appelle *la droite des moindres carrés associée à la série statistique double* $((x_i, y_i))_{1 \leq i \leq n}$.

Remarque 2.1. La droite des moindres carrés est la droite qui passe "le plus près" de tous les points du nuage de points au sens des moindres carrés (c'est-à-dire au sens où elle minimise la quantité Δ). Elle fournit donc "le meilleur" modèle de régression linéaire. Cela ne dit cependant pas si ce modèle est pertinent ou non...

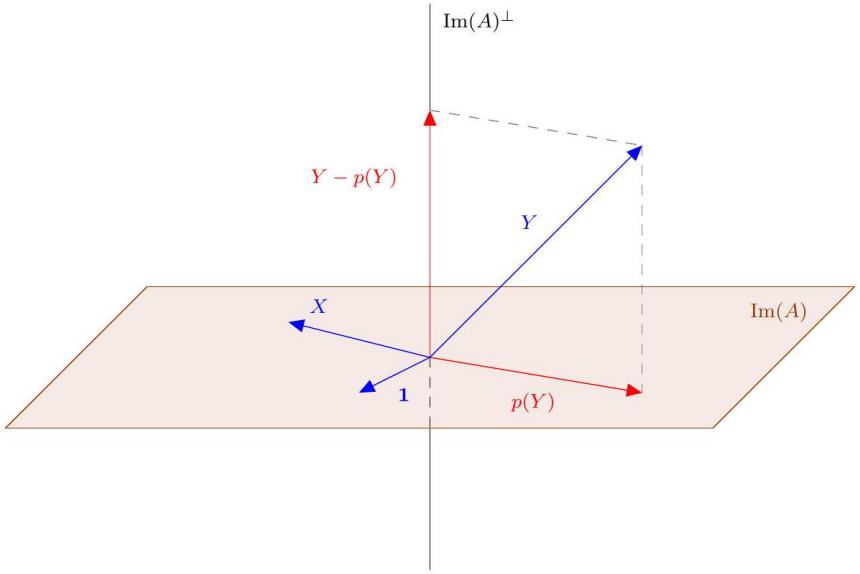
3 Équation de la droite des moindres carrés

Précisons le théorème précédent. On sait que ce minimum est atteint en un unique point $p(Y)$:

$$\min_{V \in \mathscr{M}_{2,1}(\mathbb{R})} \|Y - AV\| = \|Y - p(Y)\|$$

où p est la projection orthogonale sur le sous-espace F de $\mathscr{M}_{n,1}(\mathbb{R})$ suivant :

$$F = \{AV, V \in \mathscr{M}_{2,1}(\mathbb{R})\} = \mathrm{Im}(A) = \mathrm{Vect}\left(A\begin{pmatrix} 1 \\ 0 \end{pmatrix}, A\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \mathrm{Vect}\left(\underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}_{=:X}, \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{=:1}\right).$$



On cherche donc $U \in \mathscr{M}_{2,1}(\mathbb{R})$ tel que $p(Y) = AU$. On a :

$$
\begin{aligned}
Y - p(Y) \in \mathrm{Im}(A)^\perp = \mathrm{Vect}\left(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right)^\perp &\Leftrightarrow \begin{cases} \langle X, Y - p(Y)\rangle = 0 \\ \langle \mathbf{1}, Y - p(Y)\rangle = 0 \end{cases} \\
&\Leftrightarrow \begin{cases} X^\top \times (Y - p(Y)) = 0 \\ \mathbf{1}^\top \times (Y - p(Y)) = 0 \end{cases} \\
&\Leftrightarrow \begin{cases} \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \times (Y - p(Y)) = 0 \\ \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} \times (Y - p(Y)) = 0 \end{cases} \\
&\Leftrightarrow \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix} \times (Y - p(Y)) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
&\Leftrightarrow A^\top \times (Y - p(Y)) = 0_{2,1} \\
&\Leftrightarrow A^\top \times p(Y) = A^\top Y \\
&\Leftrightarrow A^\top \times A \times U = A^\top Y
\end{aligned}
$$

D'où finalement si $A^\top \times A \in \mathscr{M}_2(\mathbb{R})$ est inversible, la solution au sens des moindres carrés est :

$$U = \left(A^\top \times A\right)^{-1} A^\top Y.$$

Cette formule n'est pas au programme en ECG.

Remarque 3.1. La matrice $A^\top \times A$ est bien inversible car :

$$A^\top A = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}$$

et $\det\left(A^\top \times A\right) = n\sum x_i^2 - \left(\sum x_i\right)^2 > 0$. En effet par l'inégalité de Cauchy-Schwarz, on a :

$$\left(\sum x_i \times 1\right)^2 \leq \left(\sum 1\right) \times \left(\sum x_i^2\right) = n\sum x_i^2.$$

De plus, on a égalité si et seulement si X et 1 sont colinéaires, c'est-à-dire si les x_i sont égaux, ce qui n'est pas le cas par hypothèse.

Exercice 2

Reprenons la série statistique de l'exemple de départ.

1. Appliquer la méthode qui précède afin d'obtenir les valeurs de a et b.
2. Représenter la droite des moindres carrés sur le même graphique que le nuage de points. Le résultat est-il conforme à vos attentes ?
3. Quel est le signe du coefficient directeur de cette droite ? Comment l'interprétez-vous ?

On peut également obtenir l'équation de la droite des moindres carrés à l'aide de la commande Python suivante (hors programme).

Définition 3.1

Soient `x`, `y` des vecteurs de même taille, `x` ayant au moins deux coefficients distincts.

La commande `a,b=np.polyfit(x,y,1)` renvoie deux réels `a`, `b` tels que $y = ax+b$ est l'équation de la droite des moindres carrés pour la série statistique double $((x_i, y_i))$.

Remarque 3.2. Plus généralement, `np.polyfit(x, y, k)` donne les coefficients de la courbe polynomiale de degré k qui approche "le mieux" le nuage de points.

Point Culture. Cérès est la plus petite planète naine du système solaire. Avec un diamètre d'environ 950 kilomètres, il s'agit de l'objet le plus grand et le plus massif de la ceinture d'astéroïdes située entre les orbites des planètes Mars et Jupiter. Elle fut découverte le 1er janvier 1801 par Giuseppe Piazzi, astronome italien, qui pu suivre sa trajectoire jusqu'au 14 février 1801, date à laquelle l'astéroïde s'approcha trop près du Soleil pour continuer à être observé. Difficile alors pour les astronomes de l'époque de prédire la position exacte de Cérès sur la base des seules observations de Piazzi afin de confirmer sa découverte.

Afin de retrouver l'astéroïde, Carl Friedrich Gauss (1777 - 1855) développa la méthode des moindres carrés permettant de comparer les données expérimentales de Piazzi au modèle mathématique censé décrire ces données. Il obtint ainsi un résultat suffisamment précis pour permettre à Zach, un astronome allemand, de localiser à nouveau Cérès à la fin de l'année 1801.

La méthode des moindres carrés de Gauss, dont la régression linéaire est un exemple d'application, ne fut publiée qu'en 1809. Elle fut indépendamment élaborée par le mathématicien français Adrien-Marie Legendre (1752 - 1833) en 1805.

3.1 Covariance, corrélation linéaire

Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ deux séries statistiques. On rappelle que la *moyenne* de la série statistique x est donnée par :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

On appelle *point moyen* de la série statistique double $((x_i, y_i))$ le point de coordonnées $(\bar{x}, \bar{y})$.

Exercice 3 Représenter le point moyen dans notre exemple (on pourra utiliser la commande `plt.plot(.,.,"o")` pour le différencier des autres points). Que constate-t-on ?

Définition 3.2

On définit :

- l'*écart-type* de la série statistique x par : $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$;
- la *covariance* (empirique) de la série statistique double $((x_i, y_i))$ par :

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Exercice 4 Montrer que :

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \quad \text{et} \quad \mathrm{Cov}(x, y) = \overline{xy} - \bar{x}\cdot\bar{y}.$$

Proposition 3.1 [Équation de la droite des moindres carrés - Hors Programme]

Soit $((x_i, y_i))$ une série statistique double. La droite des moindres carrés a pour équation :

$$y - \bar{y} = \frac{\text{Cov}(x, y)}{\sigma_x^2}(x - \bar{x}).$$

En particulier, elle passe par le point moyen $(\bar{x}, \bar{y})$.

Exercice 5 Montrer ce résultat en effectuant le calcul $U = (A^\top \times A)^{-1} \times A^\top Y$.

La droite des moindres carrés fournit "le meilleur" modèle de régression linéaire. Mais rien n'assure que ce modèle soit pertinent ou non, puisque le caractère expliqué Y n'est pas nécessairement corrélé linéairement avec le caractère explicatif X. Pour évaluer la corrélation linéaire entre X et Y, et donc la pertinence de notre modèle de régression linéaire, nous aurons besoin de la définition suivante.

Définition 3.3

On appelle *coefficient de corrélation linéaire* de x et y le réel définit par :

$$\rho_{x,y} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}.$$

Proposition 3.2

Le coefficient de corrélation linéaire vérifie les propriétés suivantes :

- $|\rho_{x,y}| \leq 1$;
- $\rho_{x,y} = \pm 1$ si et seulement s'il existe a et b tels que $Y = aX + b$. Dans ce cas, le signe de a est le même que celui de $\rho_{x,y}$.

Exercice 6 [Bonus] À l'aide de l'inégalité de Cauchy-Schwarz dans $\mathbb{R}^n$, démontrer la propriété précédente.

Proposition 3.3 [À retenir. Interprétation selon la valeur du coefficient de corrélation linéaire]

- Un coefficient de corrélation linéaire proche de ± 1 indique que la droite des moindres carrés approche bien le nuage. Si de plus il est positif, c'est que x_i "a tendance" à augmenter lorsque y_i augmente, alors que s'il est négatif, x_i diminue lorsque y_i augmente.
- Une corrélation linéaire proche de 0 indique une absence de relation de dépendance **linéaire** entre x et y.

En général, on estime que la corrélation linéaire entre les séries X et Y est forte quand $|\rho_{X,Y}| \leq 0,9$. Lorsque c'est le cas, notre modèle de régression linéaire va permettre de faire des prédictions correctes.

On peut utiliser les commandes `Python` suivantes pour calculer le coefficient de corrélation linéaire.

Définition 3.4

Soient `x`, `y` des vecteurs de même taille n.

- Les commandes `np.var(x)` et `np.std(x)` renvoient respectivement la variance et l'écart-type de la série statistique `x`.
- La commande `np.mean(x*y) - np.mean(x)*np.mean(y)` renvoie la covariance de la série statistique double $((x_i, y_i))$.

Remarque 3.3. Une commande hors programme mais utile.

- `np.corrcoef(x,y)` renvoie la matrice $\begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix}$;
- `np.corrcoef(x,y)[0,1]` renvoie $\rho_{x,y}$.

En se limitant aux commandes du programme, on obtient $\rho_{x,y}$ ainsi :

```
(np.mean(x*y)-np.mean(x)*np.mean(y))/(np.std(x)*np.std(y))
```

Exercice 7

1. Calculer le coefficient de corrélation linéaire pour notre série statistique double. Cela correspond-il à ce que vous vous attendiez ?
2. Estimer graphiquement le poids d'un enfant de 6 ans qui mesure 1m20.

Exercice 8 [Sensibilités aux valeurs extrêmes] Le petit Constantin arrive en retard en classe ce matin-là. Il prétend mesurer 105cm et peser 27 kg.

1. Calculer le coefficient de corrélation linéaire pour la série statistique double associée aux 11 enfants. Comparer cette valeur avec celle obtenue dans l'exercice précédent.
2. Représenter le nouveau nuage de points, en distinguant les points correspondants aux 10 enfants avec celui représentant Constantin.
3. Tracer la droite des moindres carrés correspondant à la nouvelle série statistique. Que constatez-vous ?

Proposition 3.4 [À retenir. Sensibilité de la méthode des moindres carrés aux valeurs extrêmes.]

La méthode des moindres carrés est très sensible aux valeurs extrêmes : une seule valeur très éloignée de la droite des moindres carrés a une "grosse" influence sur le coefficient de corrélation linéaire et également sur la position de la droite des moindres carrés. Pour y remédier, il peut être avantageux d'exclure au préalable les valeurs aberrantes des séries statistiques avant d'en faire l'étude.

Exercice 9 [Indépendance et corrélation linéaire]

1. (a) Créer deux vecteurs `x` et `y` contenant chacun 1000 nombres tirés au hasard selon une loi uniforme sur $[0;1]$.
   (b) Représenter le nuage de points ainsi que le point moyen et la droite des moindre carrés (à l'aide de la commande `np.polyfit`). Qu'en pensez-vous ?
   (c) Calculer le coefficient de corrélation linéaire. Comment expliquer le résultat obtenu ?
2. On pose à présent `x = 2*rd.random(100) - 1` et `y = x**2`.
   (a) Représenter le nuage de points associés à ces séries statistiques ainsi que la droite des moindres carrés (à l'aide de la commande `np.polyfit`). Qu'en pensez vous ?
   (b) Calculer le coefficient de corrélation linéaire. Les variables x et y sont-elles indépendantes ?

Proposition 3.5 [À retenir. Lien entre indépendance et corrélation linéaire.]

- Si deux séries statistiques proviennent de caractères indépendants, alors le coefficient de corrélation linéaire est proche de zéro.
- À l'inverse, un coefficient de corrélation linéaire proche de zéro n'assure en rien l'indépendance des deux caractères étudiés : il indique seulement une indépendance **linéaire** entre eux.

Exercice 10

Considérons les séries statistiques `x` et `y` suivantes :

```
x = np.arange(1,51)
y = np.log(x) + rd.normal(0,1/2,50)
```

1. Représenter le nuage de points associé.
2. (a) Calculer le coefficient de corrélation linéaire. Vous semble-t-il bon ?
   (b) Déterminer l'équation de la droite de régression de `y` par rapport à `x`.
   (c) Superposer la droite de régression linéaire au nuage de points.
3. Vérifier que le nuage de points se superpose bien avec courbe de la fonction $f : x \mapsto \ln(x)$.
4. Étudier ma corrélation linéaire de `y` par rapport à `log(x)`.

Proposition 3.6 [À retenir. Corrélations non linéaires.]

Le modèle de régression entre les caractères X et Y n'est pas nécessairement linéaire, il peut être logarithmique, exponentiel, ... L'étude faite dans ce TP peut cependant nous aider pour des régressions non linéaires, en étudiant la corrélation linéaire entre Y et $\ln(X)$, $\exp(X)$, ...

Exercice 11 [Corrélation et causalité]

Une bonne corrélation entre deux séries de données ne signifie pas pour autant qu'il existe un lien de cause à effet entre les deux. À titre d'exemple, considérons la série statistique suivante :

Année	1996	1997	1998	1999	2000
Morts	15.85	15.7	15.39	15.32	14.85
Importations de citrons	230	280	360	410	525

Ce tableau donne le nombre de morts (pour un million d'habitants) sur les autoroutes américaines, ainsi que le nombre de tonnes de citrons mexicains importés aux États-Unis de 1996 à 2000.
Calculer le coefficient de corrélation linéaire pour cette série double. En déduisez vous une information pertinente ?

Mise en Garde Attention donc à l'erreur courante, notamment dans les médias, qui est de croire qu'un coefficient de corrélation linéaire élevé (en valeur absolue) induit une relation de causalité entre les deux phénomènes mesurés. Voir à ce sujet cette page des Décodeurs du `monde.fr` présentant un outil de corrélation géographique sur la base de données sans rapport, de manière à générer « vos propres cartes pour ne rien démontrer du tout ». Vous y apprendrez par exemple que la consommation de fromage est fortement corrélée au nombre de licences de football. Mais ce n'est pas pour autant que les footballeurs mangent plus de fromage.