TP7 : Statistiques bivariées

Commencer par importer les bibliothèques suivantes dans chaque fichier **Python** utilisé :

```
import numpy as np
import matplotlib.pyplot as plt
```

Objectifs du TP

- Connaître les principaux indicateurs statistiques : moyenne empirique, variance empirique, écart-type empirique, covariance empirique, coefficient de corrélation linéaire.
- Savoir tracer et interpréter le nuage de points associé à une série statistique double.
- Etre capable de juger de la pertinence d'un ajustement linéaire via le coefficient de corrélation linéaire.
- Savoir calculer l'équation de la droite de régression linéaire.
- Etre capable de se ramener à un ajustement linéaire par un changement de variable.

I. Le cours

I.1. Moyenne empirique, variance empirique et écart-type empirique d'une série statistique (statistiques univariées)

Soit X une v.a.r. discrète définie sur un univers Ω (typiquement, Ω est une population et X est un attribut d'une personne, par exemple sa taille). On suppose que l'on a observé un n-échantillon d'individus (autrement dit, un échantillon de n individus, ou bien encore un échantillon de taille n). On obtient un n-échantillon observé que l'on note x :

$$x = (x_1, \dots, x_n) = (X(\omega_1), \dots, X(\omega_n))$$

x_1 est la valeur prise par X lors de l'observation du premier individu, x_2 la valeur prise par X lors de l'observation du deuxième individu, etc. Les x_i ne sont donc pas des v.a.r. mais bien des réels. On dit que x est la *série statistique* associée à X sur le n-échantillon observé.

Définition On définit trois indicateurs statistiques associés à la série statistique x :

- La moyenne empirique :

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

- La variance empirique :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

- L'écart-type empirique :

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{s_x^2}$$

Commentaire

Considérons une v.a.r. X d'ensemble image $X(\Omega) = \{z_1, \dots, z_n\}$. Rappelons que

$$\mathbb{E}(X) = \sum_{k=1}^{n} z_k \mathbb{P}([X = z_k])$$

Considérons maintenant le n-échantillon observé de X : $x = (x_1, \dots, x_n)$. Notons, pour tout $k \in [\![1, n]\!]$, N_k le nombre de fois où la valeur z_k a été observée dans ce n-échantillon. Avec cette notation, il vient :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{n} N_k z_k = \sum_{k=1}^{n} z_k \frac{N_k}{n}$$

Ainsi, la formule de la moyenne empirique apparaît être la formule de l'espérance où l'on a remplacé $\mathbb{P}([X = z_k])$ par la fréquence empirique observée $\frac{N_k}{n}$ de la valeur z_k.

Si x est déclaré sous forme d'un vecteur (via le module `numpy`), on calculera ces trois indicateurs statistiques à l'aide de commandes prédéfinies en **Python** :

- La moyenne empirique : `np.mean(x)`
- La variance empirique : `np.var(x)`
- L'écart-type empirique : `np.std(x)`

I.2. Série statistique double, covariance empirique et coefficient de corrélation linéaire (statistiques bivariées)

On considère maintenant un couple (X, Y) de v.a.r. discrètes définies sur un univers Ω (par exemple le couple (taille, poids)). On suppose que l'on a observé un n-échantillon d'individus. On obtient une *série statistique double* :

$$(x_i, y_i)_{i\in[\![1,n]\!]} = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) = ((X(\omega_1), Y(\omega_1)), (X(\omega_2), Y(\omega_2)), \dots, (X(\omega_n), Y(\omega_n)))$$

Définition On appelle *covariance empirique* de la série statistique double $(x_i, y_i)_{i\in[\![1,n]\!]}$ le réel :

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Théorème 1 (Formule de Koenig-Huygens)**.**

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

Définition On appelle *coefficient de corrélation linéaire* de la série statistique double $(x_i, y_i)_{i\in[\![1,n]\!]}$ le réel :

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

Théorème 2 (Inégalité de Cauchy-Schwarz)**.**

$$|s_{x,y}| \leqslant s_x s_y \qquad \textit{donc} \qquad |r_{x,y}| \leqslant 1$$

Si x et y sont déclarés sous forme de vecteurs (via le module `numpy`), on pourra calculer ces deux indicateurs statistiques en **Python** :

- La covariance empirique :

```
s = np.mean(x*y) - np.mean(x)*np.mean(y)
```

- Le coefficient de corrélation linéaire :

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

I.3. Variable explicative et variable à expliquer

Dans une population donnée, on peut souhaiter étudier simultanément deux caractères X et Y. On peut alors s'intéresser aux propriétés de chacun des 2 caractères pris séparément (statistiques univariées), mais aussi au lien entre ces 2 caractères (statistiques bivariées) ; on étudie alors le couple de caractères $Z = (X, Y)$. En particulier, on peut penser que l'une des variables, X par exemple, est une cause de l'autre, par exemple Y. On dit alors que X est la *variable explicative* et Y est la *variable à expliquer*. Dans ce cas, on tentera d'exprimer Y en fonction de X. On commencera toujours par tracer le nuage des points de Y en fonction de X pour deviner la relation entre ces données. Voir partie I.4.

Dans ce TP, nous allons décrire un critère, qui permet d'établir (ou non) si il y a une corrélation linéaire entre deux variables. Il faut cependant garder en tête le fait suivant :

Une causalité entraîne une corrélation mais la réciproque est fausse.

Il y a deux écueils à éviter lorsqu'on établit une corrélation linéaire entre deux variables :

- Le premier consiste à conclure qu'il y a un lien de causalité. Comme dit précédemment, ce n'est pas toujours le cas. En effet, il pourrait y avoir une variable cachée C (C comme cause) qui explique X et Y simultanément.

  Exemple : on observe une corrélation positive entre activité physique et cancer de la peau. Une conclusion simpliste consiste à affirmer que l'activité physique peut, d'une manière ou d'une autre, provoquer le cancer. Une explication plus fine consiste à dire qu'il existe une variable cachée : l'exposition au soleil. Plus les personnes sont exposées au soleil et plus il est probable qu'elles fassent de l'activité physique. De même, plus les personnes sont exposées au soleil et plus il est probable qu'elles développent un cancer de la peau.
- Quand bien même une causalité existerait entre X et Y, le deuxième écueil consiste à se tromper dans le sens de la causalité. En effet, même en cas de corrélation forte entre X et Y, celle-ci est symétrique en X et Y et ne permet pas de dire quelle variable explique l'autre.

  Exemple : un lien de corrélation est établi entre l'usage de cannabis et le fait de développer une psychose. Est-ce que fumer participe au développement d'une psychose ou est-ce que fumer est une conséquence visant à calmer l'angoisse générée par la maladie ?

C'est une partie du travail des chercheurs et chercheuses en sciences humaines et sociales que de déterminer si une variable peut en expliquer une autre ou non, et la seule analyse mathématique ne peut pas répondre définitivement à une telle question, puisque celle-ci ne démontre que des corrélations et jamais de causalité.

I.4. Nuage de points, point moyen et modèle de régression

Considérons une série statistique double $(x_i, y_i)_{i\in[\![1,n]\!]}$.

Définition On appelle *nuage de points* associé à cette série le tracé des points de coordonnées (x_i, y_i) dans un repère orthonormé du plan.

Définition On appelle *point moyen* de la série statistique double $(x_i, y_i)_{i\in[\![1,n]\!]}$ le point de coordonnées $(\overline{x}, \overline{y})$.

L'examen du nuage de points permet de faire des constatations qualitatives :

× est-il concentré ou dispersé ?

× relève-t-on une tendance ?
(variations dans le même sens (covariance positive) ? dans le sens contraire (covariance négative) ? suit une courbe particulière ? ...)

× y a-t-il des valeurs aberrantes ?

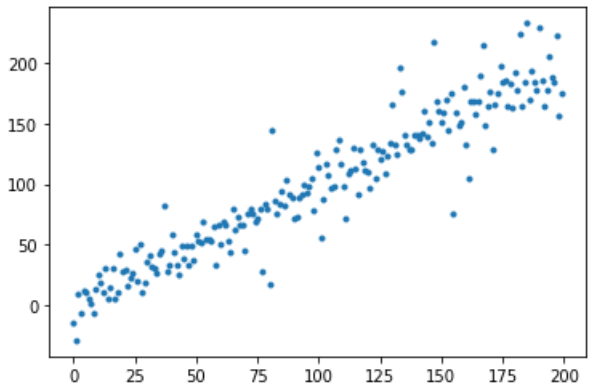


Fig. 1

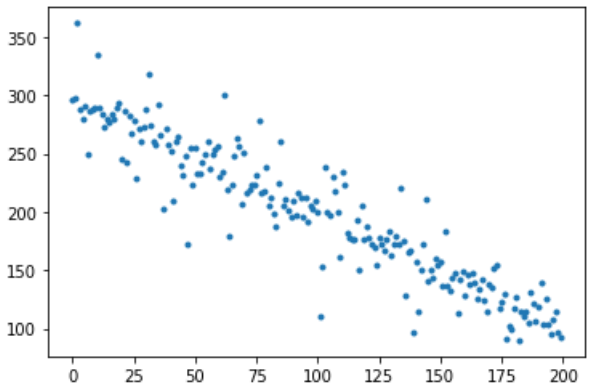


Fig. 2

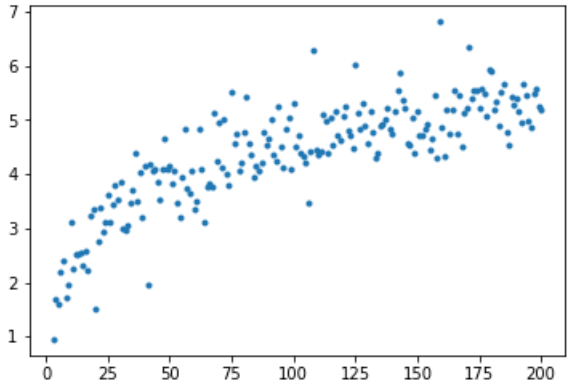


Fig. 3

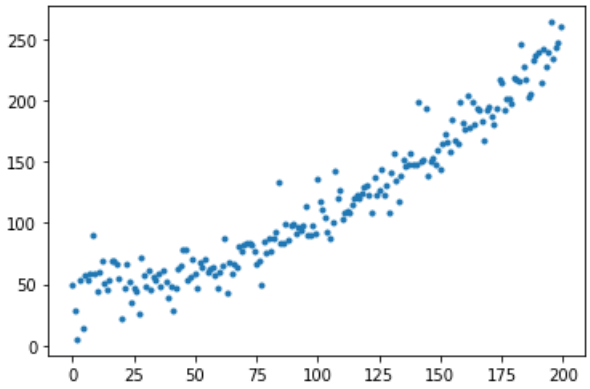


Fig. 4

Chercher à expliquer la variable Y par la variable X, c'est tenter de mettre en place un certain *modèle de régression* qui explicite le lien de causalité entre X et Y. Plus précisément, un modèle de régression consiste à postuler l'existence d'une fonction f telle que :

$$Y = f(X) + \varepsilon$$

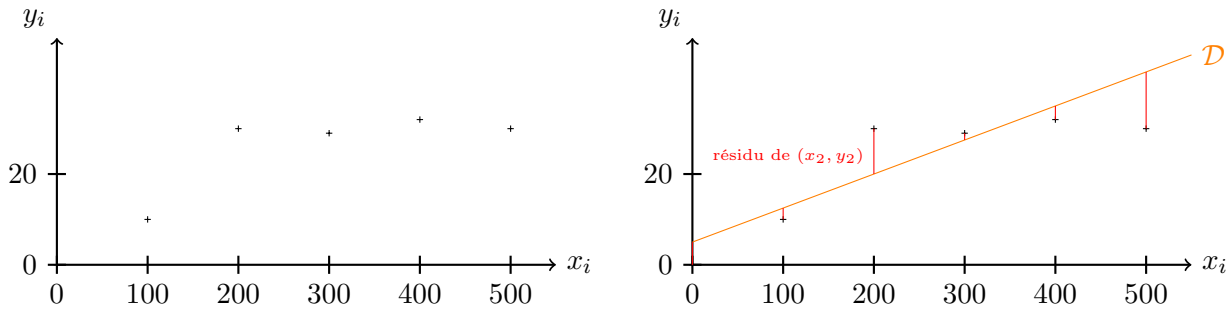
où ε est une variable aléatoire, appelée *erreur d'ajustement*.

► Proposer une fonction f pour les quatre nuages de points représentés ci-dessus.

- × Figure 1 : $f(x) = ax + b$ avec $a > 0$
- × Figure 2 : $f(x) = ax + b$ avec $a < 0$
- × Figure 3 : $f(x) = \ln(x)$
- × Figure 4 : $f(x) = cx^2$ ou $f(x) = ce^x$ avec $c > 0$

I.5. Le modèle de régression linéaire (méthode des moindres carrés)

Si le nuage de points associé à une série statistique double paraît allongé et semble suivre une direction privilégiée, on peut avoir l'idée de chercher quelle droite approcherait au mieux les points de ce nuage. Le problème consiste donc à identifier une droite d'équation $y = ax + b$ qui ajuste bien le nuage de points. L'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i est $y_i - (ax_i + b)$.



Pour déterminer la valeur des coefficients a et b, on utilise le principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés de ces erreurs :

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

Méthode des moindres carrés

Il existe une unique droite rendant minimale la somme $\sum_{i=1}^{n} (y_i - (ax_i + b))^2$. Il s'agit de la droite d'équation :

$$y = \frac{s_{x,y}}{s_x^2}(x - \overline{x}) + \overline{y}$$

Cette droite est appelée *droite de régression linéaire de Y en X* et elle passe toujours par le point moyen $(\overline{x}, \overline{y})$.

L'équation de la droite de régression linéaire peut se réécrire :

$$y = \frac{s_{x,y}}{s_x^2}x + \left(\overline{y} - \frac{s_{x,y}}{s_x^2}\overline{x}\right)$$

Comment calculer en **Python** l'équation de la droite de régression linéaire puis la tracer ?

```
1  s = np.mean(x*y)-np.mean(x)*np.mean(y)
2  a = s/np.var(x)
3  b = np.mean(y) - a*np.mean(x)
4  plt.plot(x,a*x+b)
```

Il y a un lien très fort entre le modèle de régression linéaire et le coefficient de corrélation linéaire.

- Plus le coefficient de corrélation linéaire est proche de 1 en valeur absolue et plus les points du nuage sont proches de l'alignement.
- $|r_{x,y}| = 1$ ssi les points du nuage sont alignés.
- Si $r_{x,y} > 0$, alors la droite de régression linéaire est de pente positive : X et Y varient dans le même sens (lorsque l'une croît, l'autre croît, lorsque l'une décroît, l'autre décroît aussi).
- Si $r_{x,y} < 0$, alors la droite de régression linéaire est de pente négative : X et Y varient dans des sens opposés (lorsque l'une croît, l'autre décroît).

La droite de régression linéaire existe toujours. Pour autant, il n'est pas toujours pertinent de faire un ajustement linéaire. Comment justifier l'utilisation du modèle de régression linéaire ?

On retiendra qu'un ajustement linéaire est justifié lorsque $|r_{x,y}| \geqslant 0,9$.

II. Un exemple : densité de population et criminalité

Dans cette partie, et afin de pouvoir illustrer les différentes notions au programme, on se place dans la peau d'un·e doctorant·e en sociologie qui veut analyser si il existe une relation linéaire entre la densité de population dans les villes et le taux de criminalité correspondant dans ces villes.
On note Y le taux de criminalité en nombre de crimes par 10 000 habitants (sur une unité de temps), et X la densité de population mesurée en milliers d'habitants par km^2. On donne la série statistique double sous forme de tableau :

Région	1	2	3	4	5	6	7	8	9	10	11	12
x_i	12	9	15	4	4	2	10	3	5	11	10	11
y_i	7.7	5.8	11.5	2.1	3.7	3.6	7.5	4.2	3.8	10.3	8.6	7.2

► Que choisiriez-vous comme variable explicative et comme variable à expliquer ?

× X : la densité de population
× Y : le taux de criminalité

On cherche donc à expliquer des variations dans le taux de criminalité par des variations dans la densité de population.

II.1. Calcul des indicateurs statistiques

► Calculer à l'aide de **Python** la moyenne empirique $\overline{x}$ et l'écart-type empirique s_x.

$$\overline{x} = 8 \qquad \text{et} \qquad s_x = 4.02077936060494$$

► Calculer à l'aide de **Python** la moyenne empirique $\overline{y}$ et l'écart-type empirique s_y.

$$\overline{y} = 6.333333333333333 \qquad \text{et} \qquad s_y = 2.815532315961268$$

► Calculer à l'aide de **Python** la covariance empirique $s_{x,y}$ et le coefficient de corrélation linéaire $r_{x,y}$.

$$s_{s,y} = 10.383333333333347 \qquad \text{et} \qquad r_{x,y} = 0.9172042066584032$$

► Est-il pertinent de faire un ajustement linéaire de Y en X ?

$|r_{x,y}| \geqslant 0,9$ donc il est pertinent de faire un ajustement linéaire de Y en X.

II.2. Tracé du nuage de points et ajustement linéaire

- On écrit le script suivant pour tracer le nuage de points associé à la série statistique double $(x_i, y_i)_{i\in[\![1,12]\!]}$. L'option `'.'` permet de tracer les points du nuage sous forme de points ronds.

```
x = np.array([12,9,15,4,4,2,10,3,5,11,10,11])
y = np.array([7.7,5.8,11.5,2.1,3.7,3.6,7.5,4.2,3.8,10.3,8.6,7.2])
plt.plot(x,y,'.')
plt.show()
```

- Modifier le script précédent pour que le point moyen apparaisse sur le nuage de points.

```
x = np.array([12,9,15,4,4,2,10,3,5,11,10,11])
y = np.array([7.7,5.8,11.5,2.1,3.7,3.6,7.5,4.2,3.8,10.3,8.6,7.2])
plt.plot(x,y,'.')
plt.plot(np.mean(x),np.mean(y),'.')
plt.show()
```

- Modifier le script précédent pour que la droite de régression linéaire apparaisse également sur le nuage de points.

```
x = np.array([12,9,15,4,4,2,10,3,5,11,10,11])
y = np.array([7.7,5.8,11.5,2.1,3.7,3.6,7.5,4.2,3.8,10.3,8.6,7.2])
s = np.mean(x*y)-np.mean(x)*np.mean(y)
a = s/np.var(x)
b = np.mean(y) - a*np.mean(x)
plt.plot(x,a*x+b)
plt.plot(x,y,'.')
plt.plot(np.mean(x),np.mean(y),'.')
plt.show()
```

- Estimer le taux de criminalité le plus plausible pour une densité de population de 7500 habitants par km^2.

Ici, le taux de criminalité le plus plausible est donné par la droite de régression linéaire d'équation $y = 0,59 \cdot x + 1,62$ (on a calculé $a = 0,59$ et $b = 1,62$).
Une estimation est donc un taux de criminalité à $0,59 \times 7,5 + 1,62 = 6,04$.

III. Un exemple d'ajustement non linéaire

L'évolution du chiffre d'affaire (en millions d'euros) d'une entreprise depuis sa création en 2002 est donnée par le tableau suivant :

Année	2002	2003	2004	2005	2006	2007	2008	2009
Chiffre d'affaire	0,7	1,6	2	2,4	2,5	2,8	3	3

On note X la variable donnant l'année et Y celle donnant le chiffre d'affaire.

► Déclarez le vecteur x en utilisant la commande `np.arange`.

```
x = np.arange(2002,2010,1)
```

Dans la suite, on suppose que y est également déclaré via la commande

```
y = np.array([0.7,1.6,2,2.4,2.5,2.8,3,3])
```

► Etudier la pertinence d'un ajustement linéaire.

```
s = np.mean(x*y) - np.mean(x)*np.mean(y)
r = s/(np.std(x)*np.std(y))
print(r)
```

A l'aide du script précédent, on calcule $r_{x,y} = 0.9415838181386137$. Puisque $|r_{x,y}| \geqslant 0,9$, un ajustement linéaire de Y en X est justifié.

► Calculer le coefficient de corrélation linéaire associé à la série statistique double (x_i, e^{y_i}). Que peut-on en conclure ?

```
z = np.exp(y)
s = np.mean(x*z) - np.mean(x)*np.mean(z)
r = s/(np.std(x)*np.std(z))
print(r)
```

A l'aide du script précédent, on calcule $r_{x,\mathrm{e}^y} = 0.9914193917107901$. Ce coefficient de corrélation linéaire est plus proche de 1 en valeur absolue que le précédent. Ainsi, il paraît plus pertinent de procéder à un ajustement linéaire de e^Y en X que de Y en X.

► En déduire un ajustement logarithmique de Y en X et tracer la courbe de régression.

A la suite du script précédent, on écrit

```
a = s/np.var(x)
print(f'a = {a}')
b = np.mean(z) - a*np.mean(x)
print(f'b = {b}')
plt.plot(x,y,'.')
plt.plot(x,np.log(a*x+b))
plt.show()
```

Python affiche les valeurs $a = 2.7439416857197925$ et $b = -5491.202896687342$