TP8 : Statistiques descriptives bivariées

▶ Dans votre dossier `Info_2a`, créez le dossier `TP_8`.

I. Avant-propos

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 en commençant par tracer le nuage des points de Y en fonction de X pour deviner la relation entre ces données.
Dans ce TP, on s'intéresse au problème suivant.

Un chercheur en sociologie veut analyser s'il existe une relation linéaire entre la densité de population dans les ville et le taux de criminalité correspondant dans ces villes. Que choisiriez-vous comme variable explicative X et comme variable à expliquer Y ?

- × 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.

On récupère alors les données un échantillon de villes. Pour la ville i, on note $M_i = (x_i, y_i)$ le couple de résultats obtenus.

Série statistique
Soient un échantillon $\{\omega_1, \omega_2, \dots, \omega_n\}$ de Ω. On appelle **série statistique** la donnée de la liste

$$x = [x_1, x_2, \dots, x_n]$$

Chaque x_i est associé à un seule réalisation ω_i : $x_i = X(\omega_i)$.

Série statistique double
Soient un échantillon $\{\omega_1, \omega_2, \dots, \omega_n\}$ de Ω et deux séries statistiques $x = [x_1, x_2, \dots, x_n]$ et $y = [y_1, y_2, \dots, y_n]$. On appelle **série statistique double** la donnée de la liste

$$[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$$

Chaque couple (x_i, y_i) est associé à un seule réalisation ω_i : $(x_i, y_i) = (X(\omega_i), Y(\omega_i))$.

II. Modèle de régression

II.1. Nuage de points et point moyen

▶ Taper les instructions suivantes :

```
import numpy as np
import matplotlib.pyplot as plt
x = np.linspace(1,5,5)
y = [2, 8, 1, 5, 4]
plt.scatter(x, y, marker = ''+'')
plt.xlim(0,6)
plt.ylim(0,10)
```

Que fait exactement la fonction `plt.plot` ? À quoi sert les fonctions `plt.xlim` et `plt.ylim` ?

- La fonction `plt.plot` trace
  - × les points $M_i = (x_i, y_i)$
  - × avec des points en croix (`marker = ''+''`)
- La fonction `plt.xlim` permet de délimiter l'axe des abscisses : il démarre à 0 et fini à 6.
- La fonction `plt.ylim` permet de délimiter l'axe des ordonnées : il démarre à 0 et fini à 10.

- On appelle **nuage de points** associé à la série des n couples (x_i, y_i) l'ensemble des points M_i de coordonnées (x_i, y_i) tracés dans un repère orthonormé du plan.
  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) ? suit une courbe particulière ? ...)
  - × y a-t-il des valeurs aberrantes ?

  ▶ Reprenons l'exemple sur la criminalité. On note Y le taux de criminalité en nombre de crimes par 10 000 habitants, et X la densité de population mesurée en milliers d'habitants par km^2.

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

Si la ville 3 a une superficie de 20 km^2, quel est le nombre de crimes dans cette ville ?

D'après le tableau, il y a 11.5 crimes pour 10 000 habitants. De plus, il y a 15 000 habitants par km^2.
Donc il y a $11.5 \times \dfrac{15000}{10000} = 17.25$ crimes par km^2. Ainsi il y a $17.25 \times 20 = 345$ crimes dans cette ville.

▶ Tracer le nuage de points de ces observations.

```
x = [12, 9, 15, 4, 4, 2, 10, 3, 5, 11, 10, 11]
y = [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.scatter(x, y, marker = ''+'')
```

On rappelle les définitions suivantes :

× On appelle **moyenne empirique** de la série statistique $x = [x_1, \ldots, x_n]$ le réel :

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

× On appelle **variance empirique** de la série statistique $x = [x_1, \ldots, x_n]$ le réel :

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

▶ La moyenne empirique de la série statistique x s'effectue avec la commande `np.mean(x)`. Proposer une méthode « à la main » pour effectuer la même opération.

On utilise la définition de la moyenne empirique.

```
1 xmean = (1/n) * sum(x)
```

▶ Reprenons l'exemple de la criminalité. Quel est le taux de criminalité moyen sur les 12 villes ?

```
1 ymean = np.mean(y)
```

▶ La variance empirique de la série statistique x s'effectue avec la commande `np.var(x)` et l'écart-type avec la commande `np.std(x)`. Proposer une méthode « à la main » pour le calcul de l'écart-type.

```
1 ecart = np.sqrt(np.var(x))
```

▶ Reprenons l'exemple de la criminalité. Quel est la variance empirique des densités de population sur les 12 villes ?

```
1 yvar = np.var(y)
```

- Si l'on note $\overline{x}$ la moyenne des x_i et $\overline{y}$ la moyenne des y_i, alors le point de coordonnées $(\overline{x}, \overline{y})$ est le **point moyen du nuage**.

▶ Quel est le point moyen du nuage de la série statistique double précédente ?

```
1 xm = np.mean(x)
2 ym = np.mean(y)
```

Le point moyen de ce nuage est donc $(8, 6.33)$.

- On rappelle que X est la variable explicative et Y la variable à expliquer.
  Chercher un **modèle de régression** consiste à savoir si Y est une fonction de X à un bruit près. Plus formellement, on cherche à déterminer une fonction f telle que

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

  où la fonction f est appelée **fonction de régression** et ε est une variable aléatoire appelée **erreur d'ajustement** (ou résidu).
- Nous verrons dans la suite le modèle de régression linéaire, c'est-à-dire le cas où f est une fonction affine : $f(X) = aX + b$. Il existe cependant bien d'autres choix pour la fonction f. Par exemple,
  - × d'autres modèles paramétriques, c'est-à-dire déterminés par un nombre fini de paramètres : $f(X) = \exp(aX) + b$, etc.
  - × des modèles non paramétriques, c'est-à-dire déterminés par un nombre infini de paramètres : forêts aléatoires (on approche f par des fonctions constantes par morceaux), estimation par noyau ($f(X) = \sum_{k\in\mathbb{Z}} \beta_k K(X - x_k)$, où K est une fonction positive d'intégrale 1 appelée noyau), etc.

II.2. Covariance

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

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

Il ne faut pas confondre les notations X, Y avec x, y. En effet, X et Y sont deux **variables aléatoires**, alors que x et y sont deux réalisations de ces variables ($x = X(\omega)$ et $y = Y(\omega)$), ce sont donc **des réels**.

▶ Que fait l'instruction `np.cov` ? *On pourra se reporter à la rubrique d'aide.*

La commande `np.cov(x,y)` renvoie la matrice de covariance **non biaisée** des séries x et y. Pour obtenir la matrice de covariance empirique classique des série x et y, on utilise l'instruction `np.cov(x, y, bias = True)`. Elle renvoie plus précisément le tableau :

$$\begin{bmatrix} s^2(x) & \text{cov}(x, y) \\ \text{cov}(x, y) & s^2(y) \end{bmatrix}$$

▶ Reprenons l'exemple sur la criminalité. Quelle est la covariance empirique de la série statistique double $(x_i, y_i)_{i\in[\![1,n]\!]}$?

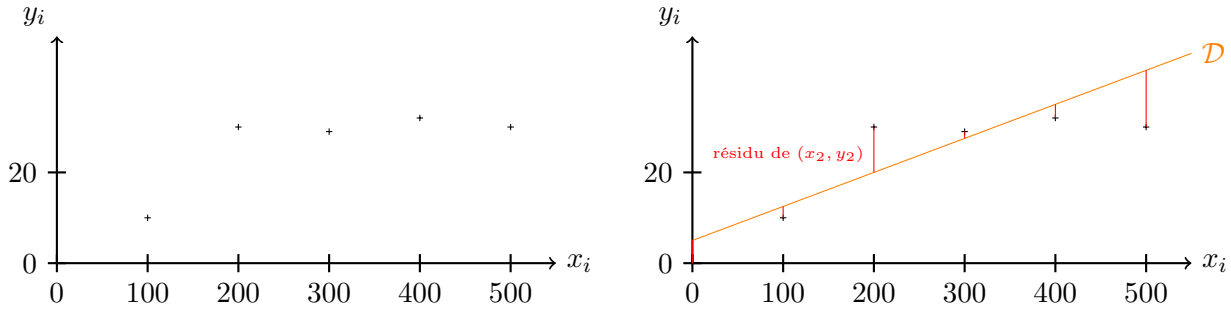
```
1  cov = np.cov(x, y, bias = True)[0,1]
```

La covariance empirique est donc $10,38$.

III. Régression linéaire

III.1. La méthode des moindres carrés

Si le nuage de points associé à une série statistique double possède une forme étirée, 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 $y = ax+b$ 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 :

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

$$y = \frac{\text{cov}(x,y)}{s^2(x)}(x - \overline{x}) + \overline{y}$$

Cette droite passe toujours par le point moyen, d'où l'équation $\overline{y} = a\,\overline{x} + b$ si et seulement si $b = \overline{y} - a\,\overline{x}$.

▶ Reprenons l'exemple sur la criminalité. Calculer les coefficients a et b de la droite de régression. Tracer cette droite sur le nuage de points précédents.

```
n = len(x)
a = np.cov(x, y, bias = True) / np.var(x)
b = np.mean(y) - a * np.mean(x)
z = [a * x[i] + b for i in range(n)]
plt.plot(x, z, ''r'')
```

Donc la droite de régression est $y = 0.59 \cdot x + 1.62$.

▶ 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 $y = ax + b$, c'est-à-dire

```
y0 = a * 7.5 + b
```

Une estimation est donc un taux de criminalité à 6.04.

III.2. Coefficient de corrélation linéaire

- On appelle **coefficient de corrélation linéaire empirique** de la série statistique double $(x_i, y_i)_{i\in[\![1,n]\!]}$ le nombre ρ défini par
$$\rho_{x,y} = \frac{\text{cov}(x,y)}{\sqrt{s^2(x)s^2(y)}}$$
- $\rho_{x,y} \in [-1,1]$.

> **Lien entre la régression linéaire et le coefficient de corrélation linéaire**
> Plus $|\rho_{x,y}|$ est proche de 1, plus les points sont proches de l'alignement et plus les prévisions données par les droites de régression sont pertinentes. $|\rho_{x,y}|$ ne valant 1 que lorsque les points du nuage sont alignés.

- Si $\rho_{x,y} > 0$ (resp. $\rho_{x,y} < 0$), alors les droites sont de pente positive (resp. négative) : X et Y varient dans le même sens (resp. en sens opposé).
  - ▶ Reprenons l'exemple sur la criminalité. Le taux de criminalité et la densité de population sont-ils corrélés ? positivement ?

```
1  cor = np.cov(x, y, bias = True) / (np.std(x) * np.std(y))
```

Le coefficient de corrélation linéaire vaut 0.84, donc est proche de 1. Le taux de criminalité et la densité de population sont donc corrélés positivement.

IV. Les statistiques bivariées au concours (HEC 2016)

Soit a un réel vérifiant $0 < a < 1$ et b un réel. Soit $n \geqslant 1$.
Pour tout $i \in [\![1,n]\!]$, on a $T_i = a\,u_i + b + R_i$, où R_1, R_2, …, R_n sont des variables aléatoires supposées indépendantes et de même loi normale centrée de variance $\sigma^2 > 0$. Le réel t_i est une réalisation de la variable aléatoire T_i.
On rappelle les définitions et résultats suivants :

- Si $(v_i)_{1\leqslant i\leqslant n}$ est une série statistique, la moyenne et la variance empiriques, notées respectivement $\overline{v}$ et s_v^2, sont données par : $\overline{v} = \dfrac{1}{n}\sum_{i=1}^{n} v_i$ et $s_v^2 = \dfrac{1}{n}\sum_{i=1}^{n}(v_i-\overline{v})^2 = \dfrac{1}{n}\sum_{i=1}^{n} v_i^2 - \overline{v}^2$.
- Si $(v_i)_{1\leqslant i\leqslant n}$ et $(w_i)_{1\leqslant i\leqslant n}$ sont deux séries statistiques, la covariance empirique de la série double $(v_i, w_i)_{1\leqslant i\leqslant n}$, notée $\text{cov}(v,w)$, est donnée par :
$$\text{Cov}(v,w) \;=\; \frac{1}{n}\sum_{i=1}^{n}(v_i-\overline{v})(w_i-\overline{w}) \;=\; \frac{1}{n}\sum_{i=1}^{n} v_i\,w_i - \overline{v}\,\overline{w} \;=\; \frac{1}{n}\sum_{i=1}^{n}(v_i-\overline{v})\,w_i$$

Pour tout $i \in [\![1,n]\!]$, soit φ_i la densité de T_i :
$$\forall d \in \mathbb{R},\quad \varphi_i(d) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\,\exp\left(-\frac{1}{2\sigma^2}\,\big(d-(a\,u_i+b)\big)^2\right)$$

Soit $\mathcal{F}$ l'ouvert défini par $\mathcal{F} =]0,1[\,\times\mathbb{R}$ et M la fonction de $\mathcal{F}$ dans $\mathbb{R}$ définie par :
$$M(a,b) \;=\; \ln\left(\prod_{i=1}^{n}\varphi_i(t_i)\right)$$

On suppose que : $0 < \text{cov}(u,t) < s_u^2$

- Montrer que M admet sur $\mathcal{F}$ un unique point critique, noté $(\hat{a},\hat{b})$. et l'exprimer en fonction de $\mathrm{cov}(u,t)$, s_u^2, $\overline{t}$ et $\overline{u}$. L'énoncé faisait ensuite démontrer que $(\hat{a},\hat{b})$ est un maximum global.
($\hat{a}$ *et* $\hat{b}$ *sont les estimations de* a *et* b *par la méthode dite du maximum de vraisemblance*)

Soit $(a,b)\in\mathcal{F}$.

$$\begin{aligned} M(a,b) &= \ln\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\,\exp\left(-\frac{1}{2\sigma^2}\,\big(d-(a\,u_i+b)\big)^2\right)\right)\\ &= \ln\left(\frac{1}{(\sqrt{2\pi})^n}\,\exp\left(-\frac{1}{2\sigma^2}\,\sum_{i=1}^{n}\big(d-(a\,u_i+b)\big)^2\right)\right)\\ &= -\frac{n}{2}\,\ln(2\pi)-\frac{1}{2\sigma^2}\,\sum_{i=1}^{n}\big(d-(a\,u_i+b)\big)^2 \end{aligned}$$

On remarque que M est de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$ en tant que fonction polynomiale.

$$\partial_1(M)(a,b) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}u_i\,t_i-a\,\sum_{i=1}^{n}u_i^2-b\,\sum_{i=1}^{n}u_i\right) = \frac{n}{\sigma^2}\left(\mathrm{cov}(u,t)+\overline{u}\,\overline{t}-a\,(s_u^2+\overline{u}^2)-b\,\overline{u}\right)$$

$$\partial_2(M)(a,b) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}t_i-a\,\sum_{i=1}^{n}-n\,b\right) = \frac{n}{\sigma^2}\left(\overline{t}-a\,\overline{u}-b\right)$$

Le couple (a,b) est un point critique de M si et seulement si $\left\{\begin{array}{rcl}\partial_1(M)(a,b) &=& 0\\ \partial_2(M)(a,b) &=& 0\end{array}\right.$.

$$\left\{\begin{array}{rcl}\partial_1(M)(a,b) &=& 0\\ \partial_2(M)(a,b) &=& 0\end{array}\right. \Leftrightarrow \left\{\begin{array}{rcl}\mathrm{cov}(u,t)+\overline{u}\,\overline{t}-a\,(s_u^2+\overline{u}^2)-(\overline{t}-a\,\overline{u})\,\overline{u} &=& 0\\ b &=& \overline{t}-a\,\overline{u}\end{array}\right.$$

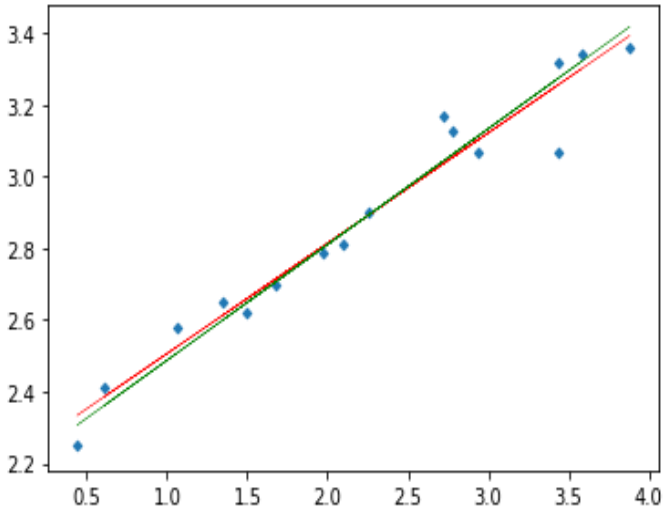
$$\Leftrightarrow \left\{\begin{array}{rcl}a\,s_u^2 &=& \mathrm{cov}(u,t)\\ b &=& \overline{t}-a\,\overline{u}\end{array}\right.$$

Finalement $(\hat{a},\hat{b})=\left(\dfrac{\mathrm{cov}(u,t)}{s_u^2},\overline{t}-\dfrac{\mathrm{cov}(u,t)}{s_u^2}\,\overline{u}\right)$.

- On rappelle qu'en `Python`, les commandes `np.var` et `np.cov` permettent de calculer respectivement la variance d'une série statistique et la covariance d'une série statistique double.
Si $(v_i)_{1\leqslant i\leqslant n}$ et $(w_i)_{1\leqslant i\leqslant n}$ sont deux séries statistiques, alors la variance de $(v_i)_{1\leqslant i\leqslant n}$ est calculable par `np.var(v)` et la covariance de $(v_i,w_i)_{1\leqslant i\leqslant n}$ est calculable par `np.cov(v, w, bias = True)`.
On a relevé pour $n=16$ entreprises qui produisent le bien considéré à l'époque donnée, les deux séries statistiques $(u_i)_{1\leqslant i\leqslant n}$ et $(t_i)_{1\leqslant i\leqslant n}$ reproduites dans les lignes 1 et 2 du code `Python` suivant dont la ligne 10 est incomplète :

```
1  u = [1.06,0.44,2.25,3.88,0.61,1.97,3.43,2.10,1.50,1.68,2.72,1.35,2.94,2.78,3.43,3.58]
2  t = [2.58,2.25,2.90,3.36,2.41,2.79,3.32,2.81,2.62,2.70,3.17,2.65,3.07,3.13,3.07,3.34]
3  a0 = np.cov(u, t, bias = True)[0,1] / np.var(u)
4  b0 = np.mean(t) - a0 * np.mean(u)
5  t0 = [a0 * u[i] + b0 for i in range(len(u))]
6  plt.scatter(u, t, marker = ''D'')
7      # ''D'' signifie que les points sont représentés par des losanges
8  plt.plot(u, t0, ''r'')
9      # équation de la droite de régression de t en u
10 plot2d(u,......................................................................, ''g'')
11     # équation de la droite de régression de u en t
```

Le code précédent complété par la ligne 10 donne alors la figure suivante :



▶ Compléter la ligne 10 du code permettant d'obtenir la figure précédente (*on reportera sur sa copie, uniquement la ligne 10 complétée*).

- En intervertissant le rôle de u et de t dans le raisonnement précédent, on obtient $u = \tilde{a}\,t + \tilde{b}$ où :

$$\tilde{a} \;=\; \frac{\mathrm{cov}(u,t)}{s_v^2} \quad \text{et} \quad \tilde{b} \;=\; \overline{u} - \frac{\mathrm{cov}(u,t)}{s_v^2}\,\overline{v}$$

Cependant, l'énoncé souhaite tracer la fonction f telle que $t = f(u)$ (et non $u = f(t)$). Or :

$$u \;=\; \tilde{a}\,t + \tilde{b} \quad \Leftrightarrow \quad t \;=\; dfrac1\tilde{a}\;(u - \tilde{b})$$

On peut donc compléter le script de la façon suivante :

```
1  a1 = np.cov(u, t, bias = True)[0,1] / np.var(t)
2  b1 = np.mean(u) - a1 * np.mean(t)
3  z = [ (u[i] - b1) / a1 for i in range(len(u))]
4  plt.plot(u, z, ‘‘g‘‘)
```

▶ Interpréter le point d'intersection des deux droites de régression.

Toutes les droites de régresion passent par le point moyen du nuage. Le point d'intersection de ces deux droites de régression est donc ce point moyen $(\overline{u}, \overline{t})$.

▶ Estimer graphiquement les moyennes empiriques $\overline{u}$ et $\overline{t}$.

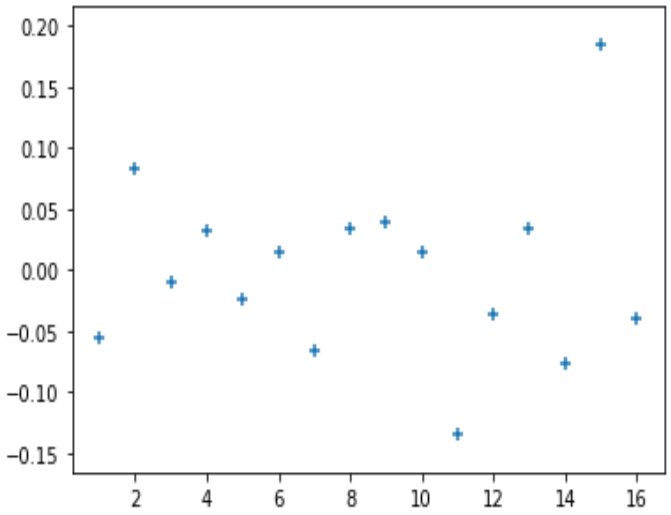
On sait que le point d'intersection des deux droites de régression est $(\overline{u}, \overline{t})$ donc on peut lire $\overline{u}$ comme l'abscisse du point d'intersection et $\overline{t}$ comme l'ordonnée de ce point. On lit $\overline{u} \approx 2.3$ et $\overline{t} \approx 2.9$.

- Le coefficient de corrélation empirique de la série statistique double $(u_i, t_i)_{1 \leqslant i \leqslant 16}$ est-il plus proche de -1, de 1 ou de 0 ?

  × $\rho_{u,t}$ n'est pas proche de 0, car les points sont proches de l'alignement.
  × $\rho_{u,t}$ n'est pas proche de -1 car les droites sont de pente positive.
  On en déduit que $\rho_{u,t}$ est plus proche de 1.

- On reprend les lignes 1 à 5 du code précédent que l'on complète par les instructions 6 à 8 qui suivent et on obtient le graphique ci-dessous :

```
6  e = [ t0[i] - t[i] for i in range(len(t)) ]
7  p = np.linspace(1, 16, 16)
8  plt.scatter(p, e, marker = ''+'')
9      # ''+'' signifie que les points sont représentés par des symboles d'addition
```



Que représente ce graphique ? Quelle valeur peut-on conjecturer pour la moyenne des ordonnées des 16 points obtenus sur le graphique ? Déterminer mathématiquement la valeur de cette moyenne.

Ce graphique représente les erreurs d'ajustement de chaque observation.
On peut conjecturer une moyenne des ordonnées de ces 16 points à 0.
On sait : $T = a\,u + b + R$. Donc : $T - (a\,u + b) = R$. D'où : $\mathbb{E}(T - (a\,u + b)) = \mathbb{E}(R) = 0$ car R suit une loi normale centrée. On retrouve bien mathématiquement une moyenne nulle.