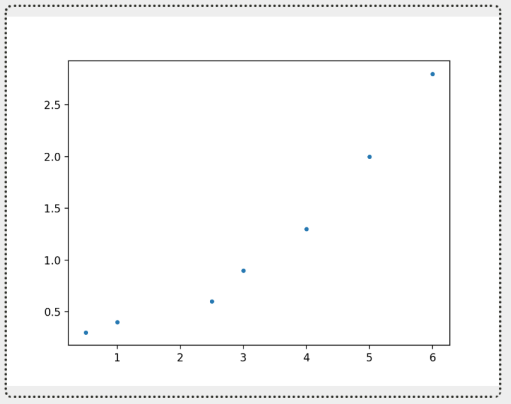
Statistiques bivariées - éléments de correction

Exercice 1 (Une première régression linéaire)

1. Si on devait faire le dessin à la main, on ferait attention à bien respecter une éventuelle échelle données par l'énoncé, et à réfléchir avant de tracer les axes à l'encombrement de la figure en fonction de l'échelle et des valeurs extrêmes des séries statistiques représentées. On pensera également à utiliser un crayon de papier bien taillé, à renseigner les légendes sur les axes.
Sinon, l'affichage python donne :

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 x=np.array([0.5,1,2.5,3,4,5,6])
5 y=np.array([0.3,0.4,0.6,0.9,1.3,2,2.8])
6 plt.plot(x,y,'.')
7 plt.show()
```



2. Le nuage de points semble avoir une direction privilégiée, il est donc judicieux de réaliser un ajustement linéaire.
Calculons le coefficient de corrélation linéaire $r_{X,Y}$ pour confirmer cette conjecture.
On commence par calculer la moyenne des deux séries, les variances des deux séries, et la covariance de la série double, grâce à la méthode de Koënig-Huygens. Les calculs sont donnés ci-après et les codes python également.

$\overline{X} = \frac{1}{7}(0,5+1+2,5+3+4+5+6) = \frac{22}{7}$

$\overline{Y} = \frac{1}{7}(0,3+0,4+0,6+0,9+1,3+2+2,8) = \frac{8,3}{7} = \frac{83}{70}$

$s_X^2 = \frac{1}{7}(0,5^2+1^2+2,5^2+3^2+4^2+5^2+6^2) - \overline{X}^2 = \frac{93,5}{7} - \frac{484}{49} = \frac{341}{98}$

$s_Y^2 = \frac{1}{7}(0,3^2+0,4^2+0,6^2+0,9^2+1,3^2+2^2+2,8^2) - \overline{Y}^2 = \frac{14,95}{7} - \frac{6889}{4900} = \frac{894}{1225}$

$s_{X,Y} = \frac{1}{7}(0,5\times 0,3+1\times 0,4+2,5\times 0,6+3\times 0,9+4\times 1,3+5\times 2+6\times 2,8) - \overline{X}\ \overline{Y} = \frac{36,75}{7} - \frac{22}{7}\times\frac{83}{70} = \frac{1493}{980}$

Enfin, $r_{X,Y} = \dfrac{s_{X,Y}}{s_X \times s_Y} = \dfrac{\frac{1493}{980}}{\sqrt{\frac{341}{98}}\times\sqrt{\frac{894}{1225}}} \simeq 0,956$

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

Python donne : `0.9560241549849806` Le coefficicent de corrélation linéaire étant très proche de 1, on en déduit, d'une part que la corrélation entre le nombre de d'inernautes en ligne et la durée du temps de chargement est très bonne, d'autre part, que ces deux variables ont le même sens de variation.

3. Le coefficient directeur de la droite de régression linéaire de Y en X par la méthode des moindres carrés est donc donné par :

$$a = \frac{s_{X,Y}}{s_X^2} = \frac{\frac{1493}{980}}{\frac{341}{98}} = \frac{1493}{3410} \simeq 0,438$$

Le point moyen de coordonnées $(\overline{X},\overline{Y})$ appartenant à la droite de régression linéaire $\mathscr{D}$, ses coordonnées vérifient son équation $y = ax+b$ d'où $\overline{Y} = a\times\overline{X}+b$ i.e. $\frac{83}{70} = \frac{1493}{3410}\times\frac{22}{7}+b$ au final $b = \frac{83}{70} - \frac{1493}{3410}\times\frac{22}{7} = -\frac{59}{310} \simeq -0,19$
En python :

```
a=cov/np.var(x)
b=np.mean(y)-a*np.mean(x)
```

On peut considérer que $\mathscr{D}$ a pour équation $y = 0,438x - 0,19$

4. Pour la suite, on prendra $y = 0,44x - 0,19$ pour équation de la droite $\mathscr{D}$.

a. Pour tracer la droite $\mathscr{D}$: $y = 0,44x - 0,19$, il suffit de connaître deux de ses points. Le premier point est obtenu grâce à l'ordonnée à l'origine, ses coordonnées, sont $(0\ ;\ -0,19)$.
Pour obtenir un deuxième point et avoir un tracé précis, on choisit une autre abscisse pas trop proche de la première. En

choisissant $x=6$, on trouve $y=0,44\times 6-0,19=2,45$. Les coordonnées de ce point sont donc $(6\ ;\ 2,45)$. Le droite $\mathscr{D}$ passe par ces deux points.
Remarque : On aurait aussi pu placer le point moyen dont on a déjà calculé les coordonnées dans la question précédente. Il appartient également à la droite de régression linéaire.

Explications du code Python :
Pour tracer une droite (en réalité python dessinera un segment), il suffit de mettre dans un vecteur les abscisses des points pour lesquels on veut tracer la droite, ici on peut se servir du vecteur x déjà créé (mais sinon, on peut prendre deux valeurs dans le vecteur : l'abscisse minimum et l'abscisse maximum puisque deux points suffisent pour tracer une droite) et dans un autre vecteur les images des composantes du vecteur x par la fonction affine ad-hoc. (pas besoin de créer et nommer ce vecteur, on peut le mettre directement de manière formelle dans le plot) :

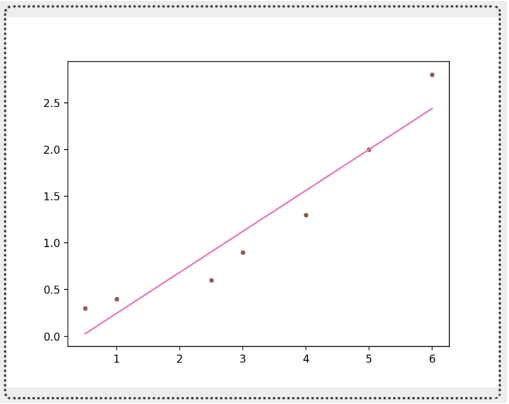
```
import numpy as np
import matplotlib.pyplot as plt

x=np.array([0.5,1,2.5,3,4,5,6])
y=np.array([0.3,0.4,0.6,0.9,1.3,2,2.8])
plt.plot(x,y,'.')
plt.show()

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

a=cov/np.var(x)
b=np.mean(y)-a*np.mean(x)

plt.plot(x,a*x+b)
plt.show()
```



b. 8000 personnes connectées correspond au point d'abscisse 8 sur la droite de régression linéaire.
Pour $x=8$, $y=0,44\times 8-0,19=3,33$.
Le temps de chargement pour 8000 personnes peut donc être estimé 3 seconde et 33 centièmes.

c. Un temps de chargement de 3, 5 secondes correspond à $y=3,5$ sur la droite de régression linéaire.
On résout donc $3,5=0,44\times x-0,19$. On obtient $x=\frac{3,5+0,19}{0,44}=\frac{369}{44}\simeq 8,386$
On peut donc estimer le nombre de visiteurs à 8386 lorsque le temps de chargement est de 3,5 secondes.
Puisque 60% d'entre eux cessent de charger la page au delà de ce délai, le site perd alors $\frac{369}{44}\times 0,6\simeq 5032$ visiteurs.

Exercice 2 (Choix du meilleur coefficient de corrélation)

Il est intéressant de choisir la variable X_3 puisqu'en valeur absolue, c'est celle qui a le coefficient de corrélation linéaire le plus proche de 1, c'est donc avec cette variable que la corrélation sera la meilleure.

Exercice 3 (Choix du meilleur coefficient de corrélation - autre paradigme)

Il est plus intéressant d'investir dans l'action P_2. En effet, on recherche le placement le moins bien corrélé à l'action EEE afin que les fluctuations des valeurs de P_2 absorbent au mieux celles de EEE.

Exercice 4 (Autour de la covariance et la corrélation linéaire)

Soient x et y deux séries statistiques. On pose $x'=2x$, $y'=2y$ et $x''=\frac{1}{2}x$

1. $s_{x',y'}=\frac{1}{n}\sum_{i=1}^{n}x'_iy'_i-\overline{X'}\,\overline{Y'}$

$s_{x',y'}=\frac{1}{n}\sum_{i=1}^{n}2x_i\times 2y_i-\overline{2X}\,\overline{2Y}$

$s_{x',y'}=4\times\frac{1}{n}\sum_{i=1}^{n}x_i\times y_i-4\overline{X}\,\overline{Y}$ en factorisant la somme par 2×2, et par linéarité de la moyenne

Au final, $s_{x',y'}=4\times s_{x,y}$.

2. $s_{x'',y'}=\frac{1}{n}\sum_{i=1}^{n}x''_iy'_i-\overline{X''}\,\overline{Y'}$

$s_{x'',y'}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}x_i\times 2y_i-\overline{\frac{1}{2}X}\,\overline{2Y}$

$s_{x',y'}=1\times\frac{1}{n}\sum_{i=1}^{n}x_i\times y_i-1\times\overline{X}\,\overline{Y}$ en factorisant la somme par $\frac{1}{2}\times 2$, et par linéarité de la moyenne

Au final, $s_{x',y'}=s_{x,y}$.

3. $r_{x',y'} = \dfrac{s_{x',y'}}{s^2_{x'}s^2_{y'}} = \dfrac{4s_{x,y}}{s^2_{2x}s^2_{2y}} \underset{Var(aX+b)=|a|Var(X)}{=} \dfrac{4s_{x,y}}{2s^2_x \times 2s^2_y} = r_{x,y}$

On rappelle que $Var(X)$ et s^2_X sont deux notations pour une seule et même notion qu'est la variance d'une série statistique X.

$r_{x'',y'} = \dfrac{s_{x'',y'}}{s^2_{x''}s^2_{y'}} = \dfrac{s_{x,y}}{s^2_{\frac{1}{2}x}s^2_{2y}} = \dfrac{s_{x,y}}{\frac{1}{2}s^2_x \times 2s^2_y} = r_{x,y}$

La corrélation entre x et y est la même que celle entre x' et y' et celle entre x'' et y'.
Par conséquent, si x et y sont très bien corrélées, il en sera de même pour x' et y', et également pour x'' et y'.
En revanche, si x et y sont mal corrélées, il en sera de même pour x' et y', et également pour x'' et y'.

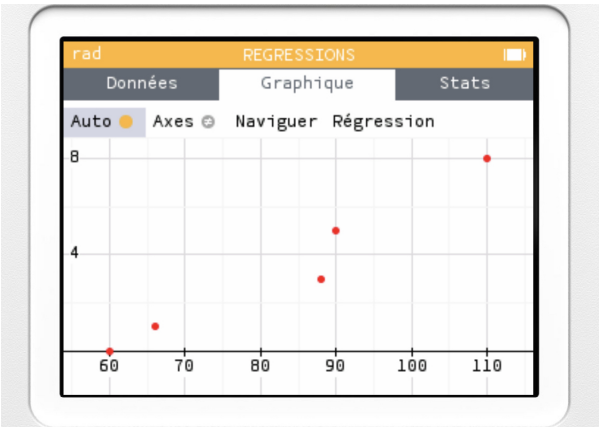
Exercice 5 (Entre corrélation et explication)

1.

Pour le tracer à la main :

- Utiliser un crayon de papier bien taillé
- Réfléchir **avant de tracer les axes** à l'encombrement de la figure en fonction de l'échelle et des valeurs extrêmes des séries statistiques représentées.
- Bien respecter l'échelle (faire un tableau au brouillon avec les conversion si besoin)
- Faire un dessin propre
- Renseigner les légendes sur les axes.

Ci contre, le nuage de point donné par une calculatrice Numworks :



2. Le nuage de points semble avoir une direction privilégiée, il est donc judicieux de réaliser un ajustement linéaire.
Calculons le coefficient de corrélation linéaire $r_{X,Y}$ pour confirmer cette conjecture.
On commence par calculer la moyenne des deux séries, les variances des deux séries, et la covariance de la série double, grâce à la méthode de Koënig-Huygens :

$\overline{X} = \frac{414}{5} = 82,4$

$\overline{Y} = \frac{17}{5} = 3,4$

$s^2_X = \dfrac{35\ 900}{5} - \overline{X}^2 = \frac{35\ 900}{5} - 82,8^2 = 324,16$

$s^2_Y = \dfrac{99}{5} - \overline{X}^2 = \frac{99}{5} - 3,4^2 = 8,24$

$s_{X,Y} = \dfrac{1660}{5} - \overline{X}\ \overline{Y} = 332 - 82,4 \times 3,4 = 50,48$

Enfin, $r_{X,Y} = \dfrac{s_{X,Y}}{s_X \times s_Y} = \dfrac{50,48}{\sqrt{324,16} \times \sqrt{8,24}} \simeq 0,977$

Le coefficicent de corrélation linéaire étant très proche de 1, on en déduit, d'une part que la corrélation entre le nombre de climatiseurs vendus et la nombre de noyades est très bonne, d'autre part, que ces deux variables ont le même sens de variation.

3. Le coefficient directeur de la droite de régression linéaire de Y en X par la méthode des moindres carrés est donc donné par :
$a = \dfrac{s_{X,Y}}{s^2_X} = \dfrac{50,48}{324,16} = \frac{631}{4051} \simeq 0,156$

Le point moyen de coordonnées $(\overline{X}, \overline{Y})$ appartenant à la droite de régression linéaire $\mathscr{D}$, ses coordonnées vérifient son équation $y = ax + b$ d'où $\overline{Y} = a \times \overline{X} + b$ i.e. $\frac{17}{5} = \frac{631}{4051} \times \frac{414}{5} + b$ au final $b = \frac{17}{5} - \frac{631}{4051} \times \frac{414}{5} = -\frac{192\ 367}{20\ 255} \simeq -9,5$

On peut considérer que $\mathscr{D}$ a pour équation $y = 0,156x - 9,5$

4. Si $x = 88$, on peut considérer que $y = 0,156 \times 88 - 9,5 \simeq 4,2$
Si 88 climatiseurs sont vendus, en s'appuyant sur cette régression linéaire, on peut craindre 4 noyades, MAIS...

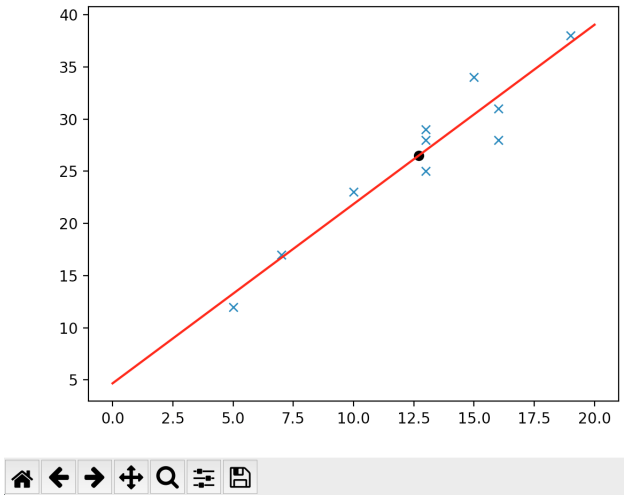
5. Corrélation n'est pas causalité. Il peut très bien y avoir une cause commune aux deux variables X et Y. Ici l'explication la plus vraisemblable est la météo. Dans un secteur littoral, on peut imaginer que lorsqu'il fait particulièrement chaud, on se baigne d'avantage, mais aussi on s'achète davantage de climatiseurs. La corrélation n'est donc pas très étonnante, mais ne signifie pas forcément une relation de cause à effet. Des études plus approfondies relativement à la température de l'atmosphère ainsi que d'autres explications alternatives doivent être faites. Dire que X explique Y n'est qu'un abus de langage, puisqu'il n'y a pas forcément de causalité, même lorsque la corrélation est très bonne.

Exercice 6 (Ajustement linéaire avec Python)

```
import numpy as np
```

```
import matplotlib.pyplot as plt
X=np.array([15,13,5,10,16,7,13,13,16,19])    # saisie de la variable  X
Y=np.array([34,25,12,23,31,17,28,29,28,38]) # saisie de la variable  Y
plt.plot(X,Y,'x')                  # génération du nuage de points avec des croix
moyX=np.mean(X)                    # calcul de la moyenne de la série X
moyY=np.mean(Y)                    # calcul de la moyenne de la série Y
plt.plot(moyX,moyY,'o', color='black')# génération du point moyen matérialisé avec un rond noir
int=np.array([0,20])                    # saisie des abscisses des 2 points extrêmes de la droite
a=(np.mean(X*Y)-np.mean(X)*np.mean(Y))/np.var(X)# calcul du coef dir de la droite de rég.
b=np.mean(Y)-a*np.mean(X)          # calcul de son ordonnée à l'origine
ord=a*int+b                        # saisie des ordonnées des points extrêmes de la droite
plt.plot(int,ord,'r')              # tracé de la droite d'équation y=ax+b en rouge
plt.show()                         # affichage du dessin
```

Le rendu aura cette allure :



Exercice 7 (Compléments en Python - initiation aux bases de données)

```
import pandas as pd
```

```
import numpy as np
import matplotlib.pyplot as plt
donnees=pd.read_csv('appartements.csv') # récupération de la base de données dans python
X=donnees['Surface']          # affectation de la colonne dont l'entête est "Surface" à X
Y=donnees['Prix']             # affectation de la colonne dont l'entête est "Prix" à Y
moyX=np.mean(X)               # calcul de la moyenne de la série X
moyY=np.mean(Y)               # calcul de la moyenne de la série Y
plt.plot(X,Y,'+')             # génération du nuage de points
plt.plot(moyX,moyY,'go')      # génération du point moyen matérialisé avec un rond vert
plt.show()                    # affichage du dessin
```

Le rendu aura cette allure :