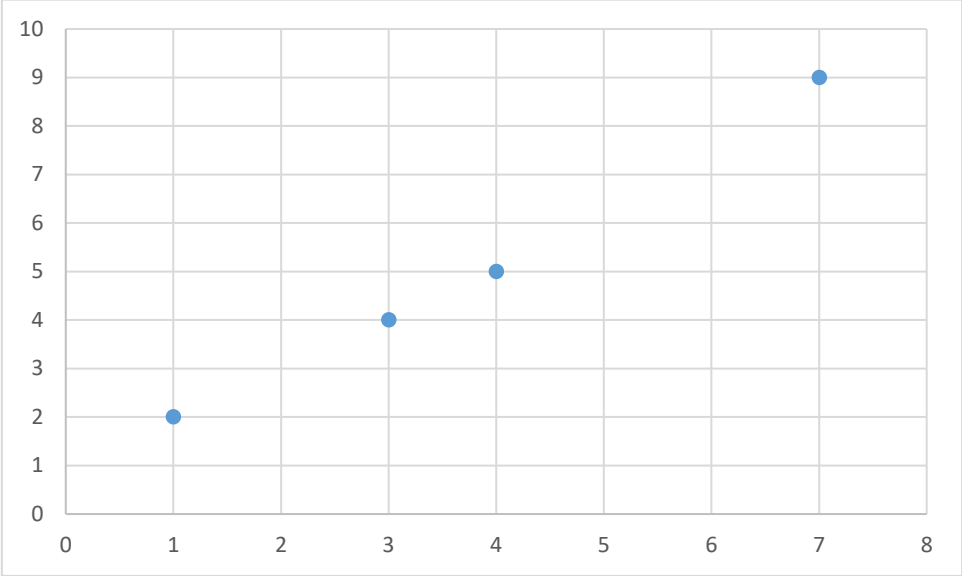
Analyse de la corrélation linéaire entre deux variables statistiques

Considérons une expérience pour laquelle on mesure deux grandeurs x et y. En faisant varier l'une des grandeurs, x par exemple, on obtient deux séries de mesure comme par exemple :

x	y
$x_1 = 1$	$y_1 = 2$
$x_2 = 3$	$y_2 = 4$
$x_3 = 4$	$y_3 = 5$
$x_4 = 7$	$y_4 = 9$

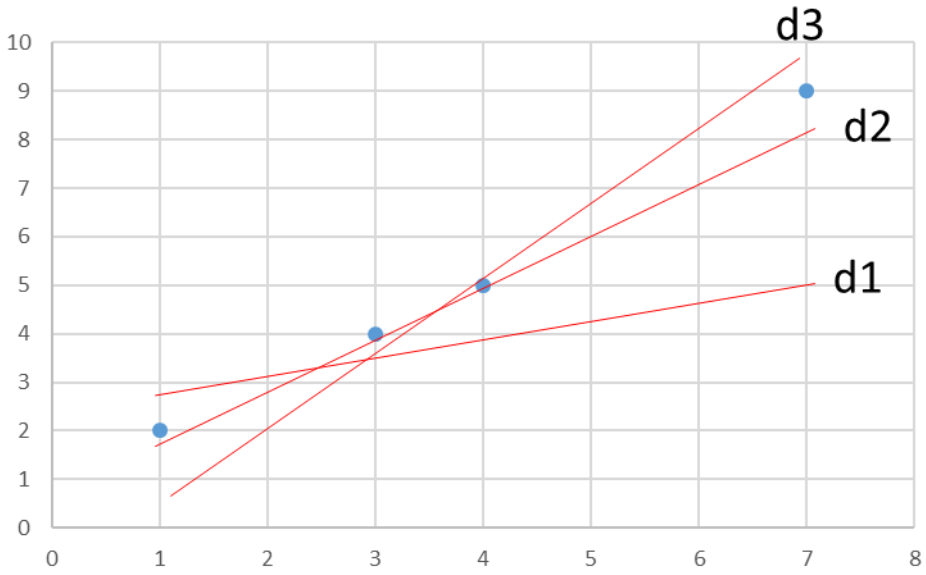
Formons avec ces résultats un nuage de 4 points :

$$A_1(x_1, y_1), \quad A_2(x_2, y_2), \quad A_3(x_3, y_3), \quad A_4(x_4, y_4)$$



Nous observons que les points semblent tous proches d'une même droite. La problématique est alors de déterminer cette droite.

Observons d'abord qu'en traçant trois droites au hasard sur ce graphique, il nous semble que si nous devions définir la droite la plus proche du nuage de points ce serait la droite d_2



Mais nous avons procédé là qu'à un tracé hasardeux. Comment pourrait-on alors définir précisément la droite la plus proche du nuage de points et en quel sens ?

La réponse est la méthode des moindres carrés. Pour cela on considère une droite d'équation générale : $y = a\,x + b$. Puis on calcule les carrés des écarts entre les points du nuage et les points de cette droite de même abscisse. La somme de ces carrés est la somme des aires figurées en vert sur le graphique ci-dessous. Elle s'exprime mathématiquement par la formule :

$$S(a,b) = (a\,x_1 + b - y_1)^2 + (a\,x_2 + b - y_2)^2 + (a\,x_3 + b - y_3)^2 + (a\,x_4 + b - y_4)^2$$

On détermine alors les valeurs des paramètres a et b qui donnent à la fonction $S(a,b)$ la valeur minimale. Pour cela on écrit que la dérivée de $S(a,b)$ à b fixé est nulle et la dérivée de $S(a,b)$ à a fixé est nulle. Cela conduit aux formules, généralisables à un nombre de points du nuage quelconque :

$$\boldsymbol{a = \frac{s_{xy}}{s_{xx}}}\ , \quad \boldsymbol{b = m_y - a\ m_x}$$

Où :

$$\boldsymbol{m_x = \frac{x_1 + x_2 + x_3 + x_4}{4}}, \quad \boldsymbol{m_y = \frac{y_1 + y_2 + y_3 + y_4}{4}}$$

$$\boldsymbol{s_{xx} = \frac{{x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2}{4} - {m_x}^2}$$

$$\boldsymbol{s_{xy} = \frac{x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4\ y_4}{4} - m_x\ m_y}$$

m_x est la moyenne de la série des x, m_y la moyenne de la série des y

s_{xx} est appelée variance de la série des x, et correspond à la moyenne des carrés moins le carré de la moyenne des valeurs de la série.

s_{xy} covariance des séries x et y correspond à la moyenne du produit moins le produit des moyennes des valeurs des deux séries.

La droite d'équation $y = a\ x + b$ est appelée droite d'ajustement linéaire.

Preuve :

Pour une valeur fixée de b, on dérive par rapport à a, ce qui donne :

$$\frac{\partial S}{\partial a}(a,b) =$$

$$2\ x_1\ (a\ x_1 + b - y_1) + 2\ x_2\ (a\ x_2 + b - y_2) + 2\ x_3\ (a\ x_3 + b - y_3) + 2\ x_4\ (a\ x_4 + b - y_4)$$

Pour une valeur fixée de a, on dérive par rapport à b, ce qui donne :

$$\frac{\partial S}{\partial b}(a,b) =$$

$$2\ (a\ x_1 + b - y_1) + 2\ (a\ x_2 + b - y_2) + 2\ (a\ x_3 + b - y_3) + 2\ (a\ x_4 + b - y_4)$$

Le minimum de la fonction $S(a,b)$ s'obtient pour le couple (a,b) rendant nulles ces deux expressions, c'est-à-dire vérifiant le système :

$$\begin{cases} 2\ x_1\ (a\ x_1 + b - y_1) + 2\ x_2\ (a\ x_2 + b - y_2) + 2\ x_3\ (a\ x_3 + b - y_3) + 2\ x_4\ (a\ x_4 + b - y_4) = 0 \\ 2\ (a\ x_1 + b - y_1) + 2\ (a\ x_2 + b - y_2) + 2\ (a\ x_3 + b - y_3) + 2\ (a\ x_4 + b - y_4) = 0 \end{cases}$$

Soit, en simplifiant par 2 et en développant:

$$\begin{cases} a\ {x_1}^2 + b\ x_1 - x_1 y_1 + a\ {x_2}^2 + b\ x_2 - x_2 y_2 + a\ {x_3}^2 + b\ x_3 - x_3 y_3 + a\ {x_4}^2 + b\ x_4 - x_4 y_4 = 0 \\ a\ x_1 + b - y_1 + a\ x_2 + b - y_2 + a\ x_3 + b - y_3 + a\ x_4 + b - y_4 = 0 \end{cases}$$

Oui en regroupant les termes de même nature :

$$\begin{cases} a\ ({x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2) + b\ (x_1 + x_2 + x_3 + x_4) - (x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4) = 0 \\ a\ (x_1 + x_2 + x_3 + x_4) + 4\ b - (y_1 + y_2 + y_3 + y_4) = 0 \end{cases}$$

Puis en faisant apparaitre les variances et la covariance :

$$\begin{cases} a\left(4\left(s_{xx}+{m_x}^2\right)\right)+b\left(4\,m_x\right)-\left(s_{xy}+m_x\,m_y\right)=0 \\ a\left(4\,m_x\right)+4\,b-\left(4\,m_y\right)=0 \end{cases}$$

En simplifiant par 4 et en développant :

$$\begin{cases} a\,s_{xx}+a\,{m_x}^2+b\,m_x-s_{xy}-m_x\,m_y=0 \\ b=m_y-a\,m_x \end{cases}$$

A noter que la seconde relation fait apparaitre que la droite d'ajustement passe par le point moyen de coordonnées (m_x, m_y). En la substituant dans la première, on aboutit à :

$$\begin{cases} a\,s_{xx}+a\,{m_x}^2+\left(m_y-a\,m_x\right)m_x-s_{xy}-m_x\,m_y=0 \\ b=m_y-a\,m_x \end{cases}$$

Ce qui après simplification donne :

$$\begin{cases} a\,s_{xx}-s_{xy}=0 \\ b=m_y-a\,m_x \end{cases}$$

Et donc aux formules à démontrer.

Pour savoir si un nuage de points est proche d'un alignement, on calcule le **coefficient de corrélation linéaire :**

$$\boldsymbol{r=\frac{s_{xy}}{s_x\,s_y}}$$

Où :

$$\boldsymbol{s_x=\sqrt{s_{xx}}} \quad , \quad \boldsymbol{s_y=\sqrt{s_{yy}}}$$

s_x est l'écart type de la série des x, s_y est l'écart type de la série des y

Une propriété de r est qu'il est compris entre -1 et 1. Un nuage proche de l'alignement a un coefficient de corrélation linéaire voisin de 1 ou de -1.

La valeur de 1 (respectivement -1) correspond à une droite d'ajustement de coefficient directeur $a>0$ (respectivement $a<0$)

Programme Python :

Voici un programme Python qui affiche le coefficient de corrélation de deux séries et sort un graphique avec le nuage de points et la droite d'ajustement :

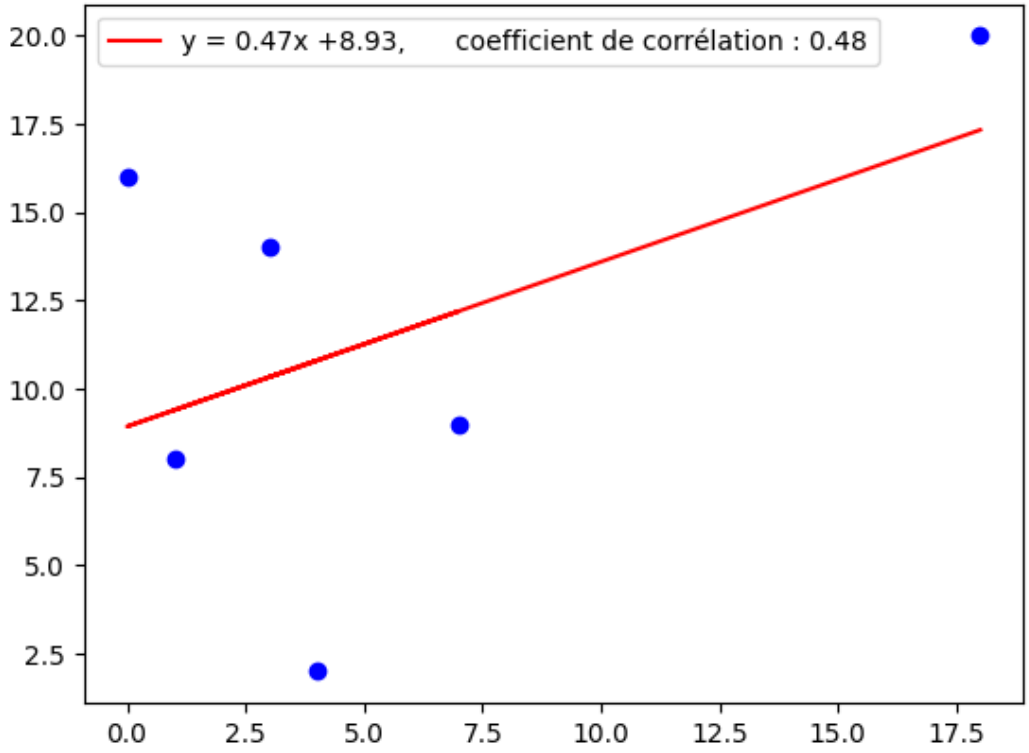
```
import matplotlib.pyplot as plt

import numpy as np

from math import *
```

```
x=np.array([1,7,3,4,0,18])
y=np.array([8,9,14,2,16,20])
#y=-0.5*x+14
x2=x**2
y2=y**2
xy=x*y
n=len(x)
moyx=sum(x)/n
moyy=sum(y)/n
moyx2=sum(x2)/n
moyy2=sum(y2)/n
moyxy=sum(xy)/n
varx=moyx2-moyx**2
vary=moyy2-moyy**2
covxy=moyxy-moyx*moyy
sx=sqrt(varx)
sy=sqrt(vary)
r=covxy/(sx*sy)
#droite d'ajustement y = a x + b
a=covxy/varx
b=moyy-a*moyx
y2=[a*t+b for t in x]
lab="y = "+str(round(a,2))+"x +"+str(round(b,2))+",     coefficient de corrélation : "+str(round(r,2))
#nuage de points
plt.plot(x,y,'ob')
plt.plot(x,y2,'r-',label=lab)
plt.legend()
plt.show()
```

Et voici les résultats de la simulation :

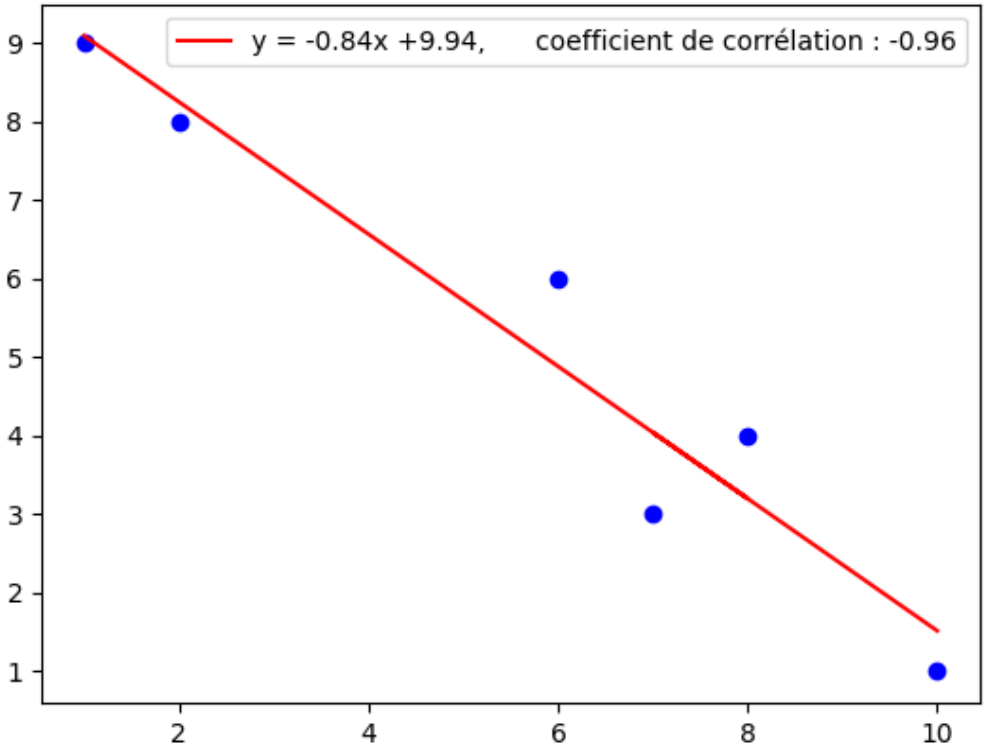


Ici le coefficient de corrélation est positif (la droite d'ajustement a une pente positive) et il est faible car les points sont très mal alignés.

Une autre simulation en remplaçant dans le programme deux lignes par :

x=np.array([10,7,8,6,2,1])

y=np.array([1,3,4,6,8,9])



Ici, l'alignement est meilleur, ce que traduit un coefficient de corrélation proche de -1.