Statistiques

La statistique est aujourd'hui un fait social total : elle règne sur la société, régente les institutions et domine la politique. Un vêtement de courbes, d'indices, de graphiques, de taux recouvre l'ensemble de la vie. L'éducation disparaît derrière les enquêtes PISA, l'université derrière le classement de Shanghai, les chômeurs derrière la courbe du chômage... La statistique devait refléter l'état du monde, le monde est devenu un reflet de la statistique.

OLIVIER REY
Mathématicien et philosophe contemporain

Que ce soit à l'issue d'une enquête, par observation directe ou à la suite d'une expérimentation, nous disposons à présent d'un ensemble de données qu'il convient maintenant de traiter. C'est-à-dire, les organiser, les représenter et en déterminer les éléments caractéristiques comme la moyenne, la variance ou l'écart-type. Tel est l'enjeu de la statistique descriptive.
Si les données ne sont relatives qu'à une seule variable (respectivement deux variables), on parle de statistique descriptive univariée (respectivement bivariée).

1 Rappels en statistiques univariées

Soit Ω un ensemble.

- Ω désigne **la population** qui fera l'objet de l'étude.
- Les éléments $\omega_i \in \Omega$ sont les **individus**.
- Une **variable statistique** (ou **caractère**) X sur la population Ω est une application $X : \Omega \to F$. Dans le cas où F est une partie de $\mathbb{R}$, on dit que le caractère est **quantitatif**. Sinon, on dit que le caractère est qualitatif.

  Remarque. Dans la suite, on ne considère que les caractères quantitatifs.

- Si Ω est un ensemble fini, le nombre d'éléments de Ω, noté Card(Ω) est **l'effectif de la population**.
- Quand il n'est pas possible d'étudier chaque individu de la population, on étudie seulement les individus d'une partie finie E de la population Ω. Dans ce cas, la partie E est appelée **échantillon**. Le nombre d'individus de l'échantillon, Card(E), est **la taille de l'échantillon**. Ainsi, si N désigne la taille, on peut écrire $E = \{e_1, e_2, \cdots, e_N\}$ où les e_i sont distincts deux à deux.

On distinguera deux types de variables statistiques.

- Si l'ensemble des valeurs prises par la variable, noté X(Ω), est un ensemble fini, on dit que la variable quantitative est **discrète**.
- Dans le cas contraire, on dit que la variable quantitative est **continue.**

1.1 Cas discret

Soit X une variable statistique discrète et $E = \{e_1, e_2, \cdots, e_N\}$ un échantillon.

- La donnée du N-uplet des observations

$$x = \big(\mathrm{X}(e_1), \mathrm{X}(e_2), \cdots, \mathrm{X}(e_\mathrm{N})\big)$$

définit une **série statistique**.

- L'échantillon étant un ensemble fini, l'ensemble des valeurs prises par la variable X sur E est aussi fini. On peut l'écrire

$$\mathrm{X}(\mathrm{E}) = \{m_1, m_2, \cdots, m_p\} \quad \text{où} \quad m_1 < m_2 < \cdots < m_p$$

où chaque m_i est une **valeur** ou **modalité.**

- **L'effectif d'une modalité** m_i est le nombre d'individu de E pour lequel le caractère prend la modalité m_i. C'est-à-dire

$$n_i = \mathrm{Card}\big\{e \in \mathrm{E} \mid \mathrm{X}(e) = m_i\big\}.$$

Si N est la taille de l'échantillon, $\mathrm{N} = \sum_{i=1}^{p} n_i$.

- **La fréquence d'une modalité** m_i est la quantité

$$f_i = \frac{\text{Effectif de la modalité}}{\text{Taille de l'échantillon}} = \frac{n_i}{\mathrm{N}}.$$

On a alors $\sum_{i=1}^{p} f_i = 1$.

- **La fréquence cumulée d'une modalité** m_i est la somme de toutes les fréquences des modalités qui lui sont inférieures. Autrement dit,

$$\mathrm{F}_i = \sum_{m_j \leqslant m_i} f_j.$$

Remarque. Donner une série statistique est équivalent à la donnée du couple (m, n) où

- $m = (m_1, m_2, \cdots, m_p)$ est le p-uplet constitué des modalités,
- $n = (n_1, n_2, \cdots, n_p)$ est le p-uplet constitué des effectifs.

Il est alors commode de représenter une série statistique à l'aide d'un tableau :

Modalités	m_1	m_2	…	m_p	Total
Effectifs	n_1	n_2	…	n_p	N

Exemples. Donnons en deux que l'on suivra tout au long de ce chapitre.

- *Exemple 1.* Voici les notes d'Inès cette année en mathématiques

$$x = (15, 12, 8, 14, 12, 15, 10, 12, 10, 12, 15, 15).$$

Série que l'on résume sous la forme :

Modalités	8	10	12	14	15	Total
Effectifs	1	2	4	1	4	12

- *Exemple 2.* Le site de L'I.N.S.E.E (Institut national de la statistique et des études économiques) contient de très nombreuses statistiques en libre accès. Voici le début de la liste du nombre d'habitants par département.

Estimations de population par sexe et âge au 1er janvier 2022 : comparaisons départementales

	Ensemble	Part des femmes (en %)	Part des hommes (en %)	Part des 0 à 24 ans (en %)	Part des 25 à 59 ans (en %)	Part des 60 ans ou plus (en %)	dont part des 75 ans ou plus (en %)
Ain	665 391	50,7	49,3	30,5	44,7	24,8	*8,6*
Aisne	524 403	51,2	48,8	29,8	41,6	28,6	*9,7*
Allier	331 757	52,0	48,0	24,9	38,9	36,1	*14,4*
Alpes-de-Haute-Provence	165 582	51,4	48,6	24,6	40,2	35,2	*13,4*
Hautes-Alpes	141 059	51,2	48,8	24,8	41,3	33,9	*12,3*
Alpes-Maritimes	1 103 555	52,8	47,2	26,3	42,4	31,2	*13,0*
Ardèche	330 865	51,2	48,8	25,8	40,8	33,5	*12,4*
Ardennes	265 285	51,1	48,9	27,8	42,0	30,3	*10,4*
Ariège	153 126	51,1	48,9	24,3	40,7	35,1	*13,4*
Aube	311 083	51,4	48,6	29,9	41,5	28,6	*10,4*

Ci-dessous, la seconde colonne exportée en Python dans une liste.

```
L=[665391,
524403,
331757,
165582,
141059,
1103555,
330865,
265285,

...

294436,
868846,
299348]
```

Par exemple, on peut y lire que le département Nord est le plus peuplé (au 1er janvier 2022) avec 2 606 873 habitants.

1.2 Cas continu

On découpe l'ensemble des valeurs possibles $X(\Omega)$ en un certain nombre d'intervalles. Notons $p \in \mathbb{N}^*$ le nombre d'intervalles choisis. Ces intervalles doivent être deux à deux disjoints, et leur réunion est égale à (ou contient) l'ensemble des valeurs possibles. Plus précisément, en notant $a_1, a_2, \dots, a_{p+1} \in \mathbb{R}$, les bornes de tous ces intervalles avec $a_1 < a_2 < \cdots < a_{p+1}$, on a

$$X(\Omega) \subset \bigcup_{i=1}^{p} [a_i ; a_{i+1}[.$$

L'intervalle $[a_i ; a_{i+1}[$ est une **classe**.

Remarques.

- On peut ainsi définir l'effectif d'une classe et sa fréquence.

- On peut utiliser ces définitions pour une variable discrète lorsque l'ensemble des valeurs prises est trop important. Par exemple, pour étudier le nombre d'habitants par département, on pourra faire des tranches de 200 000 habitants.

2 Paramètres

Statistics may be regarded as
(i) the study of populations,
(ii) as the study of variation,
(iii) as the study of methods of the reduction of data.

RONALD AYLMER FISHER
biologiste et statisticien britannique (1890-1962).

La collecte massive de données permise par la révolution numérique de ces dernières années a rendu l'emploi des statistiques de plus en plus nécessaire. Comme le souligne R.A Fisher, un des objectifs des statistiques est d'introduire des quantités pertinentes pour l'analyse des données afin d'en extraire les informations pertinentes. Les plus élémentaires étant la moyenne, la variance, l'écart-type...

2.1 Paramètres de position

DÉFINITION **Mode - (hors-programme)**

*Un **mode** d'une série statistique est une modalité pour laquelle l'effectif est maximal.*

⚠ **Attention.** Un mode n'est pas nécessairement unique.

- *Exemple 1.* Il y a deux modes dans la série des notes d'Inès : 12 et 15 d'effectif 4.

DÉFINITION **La moyenne empirique**

*Soit $x = (x_i)_{1\leqslant i\leqslant \mathrm{N}}$, une série statistique. On définit la **moyenne** de la série par*

$$\overline{x} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} x_i.$$

Si $(m_i)_{1\leqslant i\leqslant p}$, $(n_i)_{1\leqslant i\leqslant p}$ et $(f_i)_{1\leqslant i\leqslant p}$ désignent respectivement les modalités, les effectifs et les fréquences, on a aussi

$$\overline{x} = \frac{1}{\mathrm{N}} \sum_{i=1}^{p} n_i\, m_i = \sum_{i=1}^{p} f_i\, m_i$$

Exemples.

- *Exemple 1.* On a directement $\overline{x} = \frac{1}{12}\left(1\times 8 + 2\times 10 + 4\times 12 + 1\times 14 + 4\times 15\right) = 12{,}5$.

- *Exemple 2.* Avec Python, la moyenne est obtenue par la commande `mean()` qui prend en argument une liste.
Ci-contre, le calcul du nombre d'habitants moyen par département.

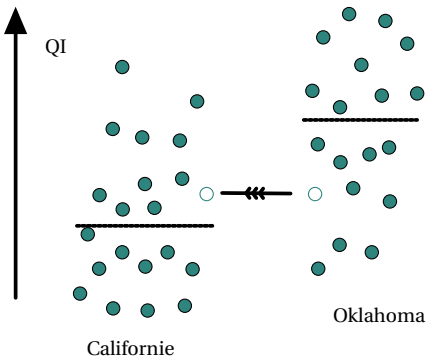
Console

```
>>> import numpy as np
>>> np.mean(L)
671419.7623762377
```

Remarque. Phénomène de Will Rogers

« Quand, pendant la Grande Dépression, les ouvriers pauvres ont quitté l'Oklahoma pour la Californie, le niveau intellectuel moyen des deux États a augmenté. »

Cette citation est attribué au comédien américain originaire d'Oklahoma : Will Rogers. C'est une façon déguisée de dire que les habitants de Californie sont en moyenne plus bêtes que ceux de l'Oklahoma. Le quotient intellectuel de la Californie augmente car il accueille des Oklahomans plus intelligents alors que celui de l'Oklahoma augmente car les émigrés faisaient baisser la moyenne.



Considérons deux populations Ω_1 et Ω_2 ainsi qu'un caractère X défini sur $\Omega_1 \cup \Omega_2$. Cet exemple illustre le fait qu'en déplaçant de la population à la moyenne la plus élevée un élément en dessous de la moyenne de son groupe mais au-dessus de la moyenne de l'autre groupe, on peut augmenter la moyenne des deux populations.

DÉFINITION — **La médiane d'une série statistique - (hors-programme)**

Soit $x = (x_i)_{1 \leqslant i \leqslant N}$ une série statistique ordonnées suivant l'ordre croissant. On définit la ***médiane*** *de la série statistique par*

- x_p *si* N *est impair avec* $p = (N+1)/2$,
- $\frac{x_p + x_{p+1}}{2}$ *si* N *est pair avec* $p = N/2$.

Autrement dit, la médiane un nombre réel m_e tel que le nombre d'individus pour lesquels X prend une valeur inférieure ou égale à m_e soit égal au nombre d'individus pour lesquels X prend une valeur supérieure ou égale à m_e.

Exemples.

• *Exemple 1.* Ordonnons les douze notes d'Inès.

$$x = (15, 12, 8, 14, 12, 15, 10, 12, 10, 12, 15, 15), \quad \tilde{x} = (8, 10, 10, 12, 12, 12, 12, 14, 15, 15, 15, 15).$$

La médiane est donnée par $$m_e = \frac{\tilde{x}_6 + \tilde{x}_7}{2} = 12.$$

• *Exemple 2.* Avec Python, la médiane est obtenue par la commande `median()` qui prend en argument une liste.

Console

```
>>> import numpy as np
>>> np.median(L)
524506.0
```

• Ci-dessous, une comparaison entre l'âge médian et l'âge moyen des françaises et français.

	Age moyen			Âge médian		
Année	Femmes	Hommes	Ensemble	Femmes	Hommes	Ensemble
2021	43,4	40,6	42,1	42,6	39,6	41,1

• Selon l'INSEE, la salaire mensuel moyen (équivalent temps plein en 2019) d'un Français est de 2 424 euros net. Le salaire médian, lui, s'élève à 1 940 euros par mois. La moyenne est relevée par les très hauts revenus qui ne concernent qu'une minorité de salariés.
Ce dernier exemple illustre un fait important : la moyenne est sensible aux valeurs extrêmes alors que la médiane ne l'est pas.

Exercice 1

✧ **Transformation affine d'une série statistique**

Soient a, b deux réels et x une série statistique. Exprimer en fonction de la moyenne de x, la moyenne de la série $x' = ax + b$. Faire de même avec la médiane. p. 15

Remarque. La série $x' = ax + b$ est la série obtenue à partir de x en replaçant chaque donnée x_i par le réel $ax_i + b$.

DÉFINITION — **Quartiles - (hors-programme)**

Soit x une série statistique.

• *Le* ***premier quartile*** *de x est la plus petite modalité de la série pour laquelle la fréquence cumulée est supérieure ou égale à* 1/4.

• *Le* ***troisième quartile*** *de x est la plus petite modalité de la série pour laquelle la fréquence cumulée est supérieure ou égale à* 3/4.

Remarque. De même, on définit le k-ième décile est la plus petite modalité de la série pour laquelle la fréquence cumulée est supérieure ou égale à $k/10$.

Exemple. Ci-dessous, la répartition du patrimoine net des français en 2018 suivant les déciles (source I.N.S.E.E).

Par exemple, 10% des français ont un patrimoine net supérieur à 549 600 euros.

1er décile	2600
2e décile	9000
3e décile	23400
4e décile	60800
5e décile	117000
6e décile	176700
7e décile	246200
8e décile	348700
9e décile	549600
95e centile	794800
99e centile	1745800

2.2 Paramètres de dispersion

DÉFINITIONS — **La variance empirique et l'écart-type**

Soit x, une série statistique.

- ***La variance*** *de la série est le réel, noté $\sigma(x)^2$, défini par*

$$\sigma(x)^2 = \frac{1}{N}\left((x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_N-\overline{x})^2\right) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2.$$

- ***L'écart-type****, $\sigma(x)$ d'une série statistique est défini comme la racine carrée de la variance.*

Remarques.

- Si les modes $m_1, m_2, \cdots, m_p$ de la série sont donnés avec leurs effectifs $n_1, n_2, \cdots, n_p$ ou leurs fréquences $f_1, f_2, \ldots, f_p$, alors

$$\sigma(x)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i(m_i-\overline{x})^2 = \sum_{i=1}^{p} f_i(m_i-\overline{x})^2.$$

- Une variance est toujours un nombre positif. L'écart-type est donc bien défini.
- Une série a une variance nulle si et seulement si toutes les valeurs de la série sont identiques.

Exemples.

- *Exemple 1.*

$$\sigma(x)^2 = \frac{1}{12}\left(1\times(8-12,5)^2+2\times(10-12,5)^2+4\times(10-12)^2+1\times(10-14)^2+4\times(10-15)^2\right) \simeq 5,08$$

puis un écart type de 2,25.

- *Exemple 2.* Avec Python, la variance et l'écart-type sont obtenus par les commandes `var()` et `std()` qui prennent en argument une liste.

Console

```
>>> np.var(L)
267615024984.00293
>>> np.std(L)
517315.20853731234
```

PROPOSITION — **Transformation affine**

Soient x une série statistique et $a, b \in \mathbb{R}$.
Si x' est la série statistique obtenue en modifiant chaque donnée x_i en $ax_i + b$, alors

$$V' = a^2 V \quad et \quad \sigma' = |a|\sigma$$

où V, σ, (resp. V', σ') désignent la variance et l'écart-type de x (resp. de x').

Exercice 2

✧ Prouver la proposition précédente. p. 15

En pratique, on utilise la formule suivante pour calculer la variance.

THÉORÈME — **Formule de Koenig-Huygens**

Soit x une série statistique de modalités $(m_i)_{1\leqslant i\leqslant p}$, d'effectifs $(n_i)_{1\leqslant i\leqslant p}$ et de variance $\sigma(x)^2$, alors

$$\sigma(x)^2 = \left(\frac{1}{N}\sum_{i=1}^{N} {x_i}^2\right) - \overline{x}^2 = \left(\frac{1}{N}\sum_{i=1}^{p} n_i {m_i}^2\right) - \overline{x}^2.$$

Exercice 3

✧ Prouver ce théorème. p. 15

Remarque. Si x^2 désigne la série dont chaque donnée est ${x_i}^2$. On démontre que $\frac{1}{N}\sum_{i=1}^{p} n_i {m_i}^2$ est la moyenne $\overline{(x^2)}$. La formule de formule de Koenig-Huygens devient alors

$$\sigma(x)^2 = \overline{(x^2)} - \left(\overline{x}\right)^2.$$

Exercice 4

✦✦ **L'écart type, un indicateur de dispersion par rapport à la moyenne**
Soit $x = (x_i)_{1\leqslant i\leqslant N}$ une série statistique avec $\overline{x}$ et $\sigma(x)$, respectivement la moyenne et l'écart-type de la série.
Soit r le nombre d'éléments de la série statistique compris entre $\overline{x} - 2\sigma(x)$ et $\overline{x} + 2\sigma(x)$.

1. Montrer que
$$\sum_{k=1}^{N} \left(x_k - \overline{x}\right)^2 \geqslant 4\sigma(x)^2(N - r).$$
2. En déduire qu'au moins les trois quarts des éléments de la série statistique sont compris entre $\overline{x} - 2\sigma(x)$ et $\overline{x} + 2\sigma(x)$.

p. 15

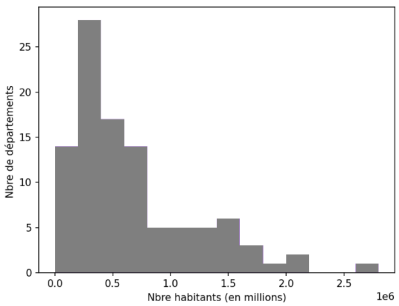
Histogrammes

Un histogramme donne une idée graphique de la répartition des valeurs du caractère sur l'échantillon. Ce graphique est obtenu en traçant, pour chaque $i \in [\![1, p]\!]$, le rectangle de base $[a_i, a_{i+1}]$ sur l'axe des abscisses et en ordonnées, l'effectif de la classe $[a_i, a_{i+1}]$.

Exemple 2.

Editeur

```
# Création d'un tableau avec les intervalles de
    longueurs 200 000 (habitants)
inter =np.linspace(0,2.8*10**6,15)
plt.hist(L, bins=inter)
plt.xlabel('Nbre habitants (en millions)')
plt.ylabel('Nbre de départements ')
plt.show()
```



3 Statistiques bivariées

Dans la suite, x, y désigne deux séries statistiques. La question est de savoir dans quelle mesure l'une des deux, dite **expliquée**, dépend de l'autre, dite **explicative**.

3.1 Premières définitions

DÉFINITION — **La covariance empirique**

*Soient $x = (x_i)_{1\leqslant i\leqslant N}$ et $y = (y_i)_{1\leqslant i\leqslant N}$, deux séries statistiques. On définit la **covariance** des deux séries par*

$$\mathrm{Cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right).$$

*Si les séries x et y ne sont pas constantes, on définit aussi le **coefficient de corrélation linéaire empirique** par*

$$\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)}.$$

Remarques.
→ $\rho(x,y)$ est un réel de $[-1,1]$.
→ On a l'égalité $\rho(x,y) = \pm 1$ si et seulement s'il existe a et b tels que pour tout indice i, $y_i = ax_i + b$. Dans ce cas, le signe de a est le même que celui de $\rho(x,y)$.

Exercice 5

✦ Prouver les deux remarques précédentes. p. 15

Par un calcul, on montre :

PROPOSITION — **Calcul de la covariance**

Soient $x = (x_i)_{1\leqslant i\leqslant N}$ et $y = (y_i)_{1\leqslant i\leqslant N}$, deux séries statistiques. On a

$$\mathrm{Cov}(x,y) = \overline{x*y} - \overline{x}\cdot\overline{y}$$

où $\overline{x*y}$ désigne la moyenne de la série $x*y = (x_i\cdot y_i)_{1\leqslant i\leqslant N}$.
Dès lors, la commande Python pour obtenir la covariance empirique est

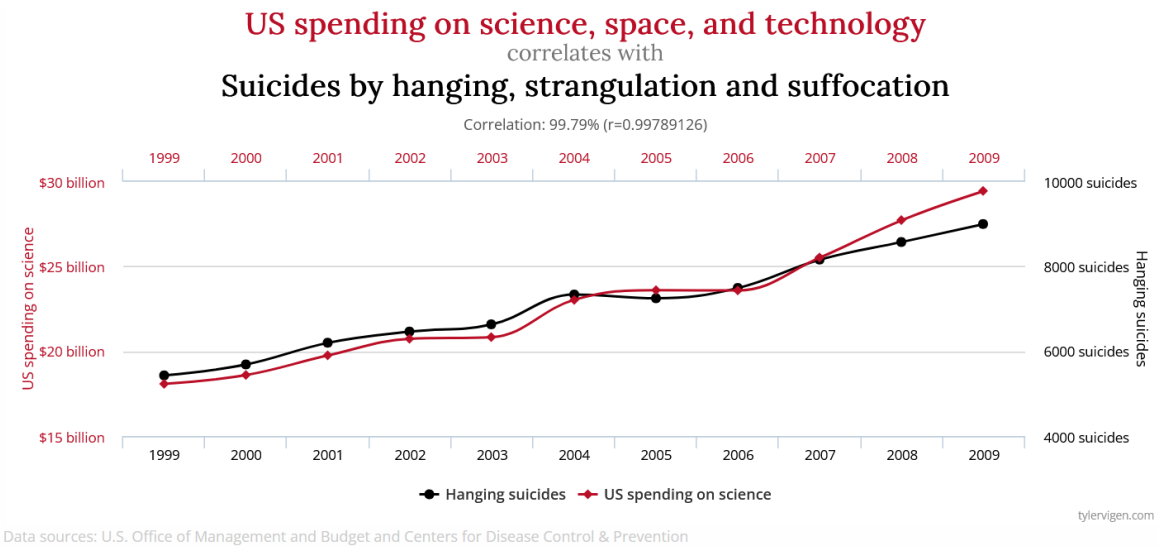
```
np.mean(x*y)-np.mean(x)*np.mean(y)
```

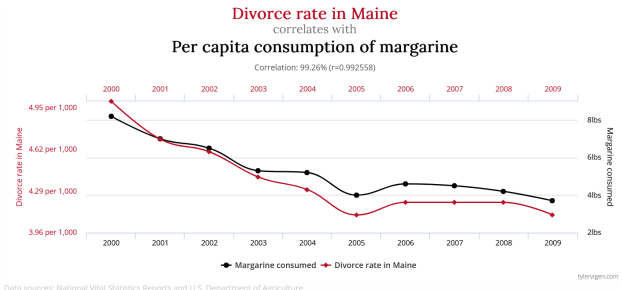
et pour le coefficient de corrélation

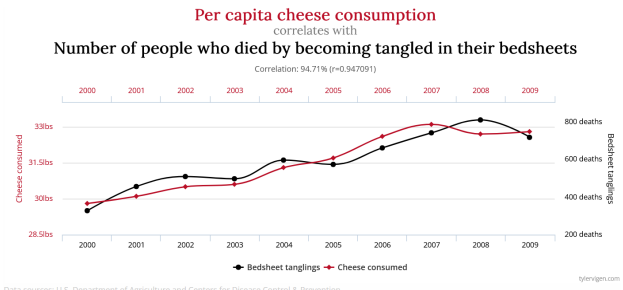
```
(np.mean(x*y)-np.mean(x)*np.mean(y))/(np.std(x)*np.std(y))
```

Corrélation n'est pas causalité

Quelques exemples issus du site : *http ://tylervigen.com/spurious-correlations.*







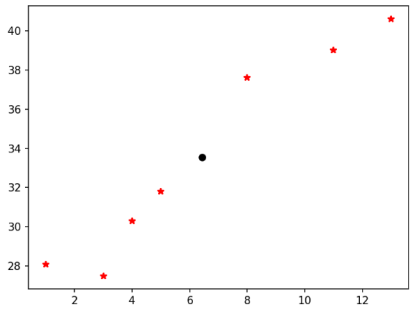
3.2 Nuage de points

DÉFINITION — **Nuage de points**

Soient $x = (x_i)_{1\leqslant i\leqslant N}$ et $y = (y_i)_{1\leqslant i\leqslant N}$, deux séries statistiques.
*Le **nuage de points** associé à ces deux séries est l'ensemble des points M_i du plan de coordonnées (x_i, y_i) où $i \in [\![1, N]\!]$.*

Remarque. Le point $(\overline{x}, \overline{y})$ est le **point moyen** du nuage.

• Voici le code pour tracer un nuage de points associé à deux séries statistiques de même effectif, ainsi que son point moyen.



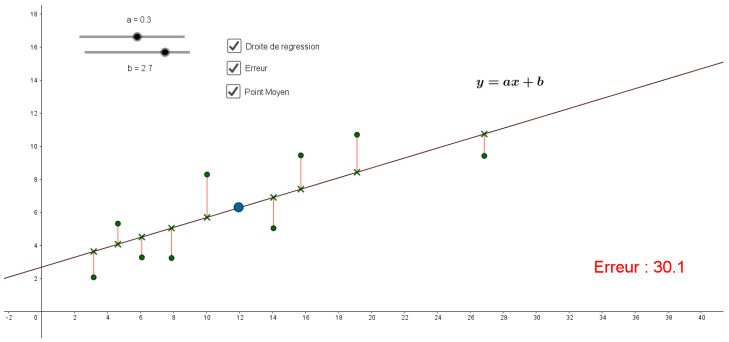
Editeur

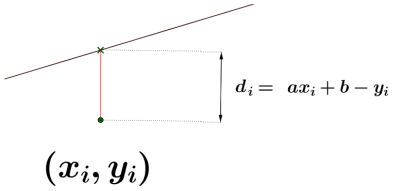
```
import matplotlib.pyplot as plt
import numpy as np

# Exemples avec deux séries
x=np.array([1,3,4,5,8,11,13])
y=np.array([28.1,27.5,30.3,31.8,37.6,39,40.6])


# Tracé du nuage de points et du point moyen

plt.plot(x,y,'r*')
plt.plot(np.mean(x),np.mean(y),'ko')
```

3.3 Problème des moindres carrés et droite de régression

Soit n un entier supérieur à 2.
Considérons n points de $\mathbb{R}^2$, $(x_1, y_1), \ldots, (x_n, y_n)$ non alignés verticalement. On cherche la droite qui « approxime » au mieux ces n points. Si on note $y = ax + b$, l'équation d'une droite, on cherche à minimiser l'erreur

$$\mathrm{E}_r = \sum_{i=1}^{n} {d_i}^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2.$$





Point de vue algèbre linéaire

Traduisons matriciellement le problème. Posons :

$$\mathrm{X} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathrm{A} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \quad \text{et} \quad \mathrm{B} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \text{de sorte que} \quad \mathrm{AX} - \mathrm{B} = \begin{bmatrix} ax_1 + b - y_1 \\ ax_2 + b - y_2 \\ \vdots \\ ax_n + b - y_n \end{bmatrix}.$$

Si on considère le produit scalaire canonique sur $\mathscr{M}_{n,1}(\mathbb{R})$ et la norme associée

$$\mathrm{E}_r = \|\mathrm{AX} - \mathrm{B}\|^2.$$

Les deux colonnes de la matrice A forment une famille libre (les points ne sont pas alignés verticalement). La matrice A est de rang 2. D'après le théorème précédent, il existe un seul vecteur minimisant $\|\mathrm{AX} - \mathrm{B}\|$. Calculons ce vecteur X_0 à partir des résultats sur les pseudo-solutions. On a

$${}^t\mathrm{AA} = \begin{bmatrix} \sum_{i=1}^{n} {x_i}^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} \in \mathscr{M}_2(\mathbb{R}).$$

Justifions que ${}^t\mathrm{AA}$ est inversible.

$$\det\left({}^t\mathrm{AA}\right) = n\sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2.$$

En appliquant l'inégalité de Cauchy-Schwarz (pour le produit scalaire canonique sur $\mathbb{R}^n$ aux vecteurs $x = (x_1, \ldots, x_n)$ et $u_0 = (1, \ldots, 1)$, il vient :

$$\langle u_0, x\rangle^2 = \left(\sum_{i=1}^{n} x_i\right)^2 \leqslant n\sum_{i=1}^{n} {x_i}^2.$$

Cette inégalité est en fait stricte car les deux vecteurs ne sont pas colinéaires. Le déterminant est donc non nul et la matrice ${}^t\mathrm{AA}$ est inversible. L'inverse est donné par

$$\left({}^t\mathrm{AA}\right)^{-1} = \frac{1}{n\sum\limits_{i=1}^{n} {x_i}^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} \begin{bmatrix} n & -\sum\limits_{i=1}^{n} x_i \\ -\sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} {x_i}^2 \end{bmatrix}.$$

De plus, on calcule

$${}^t\mathrm{AB} = \begin{bmatrix} \sum\limits_{i=1}^{n} x_i y_i \\ \sum\limits_{i=1}^{n} y_i \end{bmatrix}.$$

Ceci permet d'expliciter les vecteur X, puis ses composantes a et b.

$$a = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n\sum\limits_{i=1}^{n} {x_i}^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum\limits_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum\limits_{i=1}^{n} y_i\right)}{\frac{1}{n}\sum\limits_{i=1}^{n} {x_i}^2 - \left(\frac{1}{n}\sum\limits_{i=1}^{n} x_i\right)^2}.$$

Si on introduit les variances et covariances empiriques, on obtient

$$a = \rho(x,y)\sqrt{\frac{\mathbf{V}(y)}{\mathbf{V}(x)}} \quad \text{et} \quad b = \overline{y} - a\overline{x}.$$

L'équation de la droite est alors :

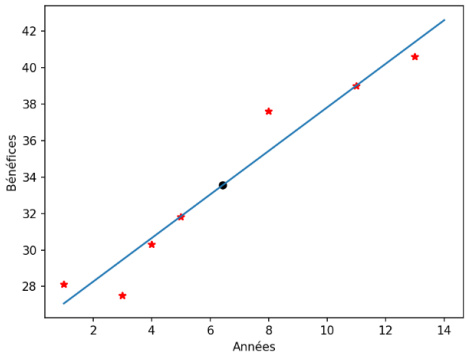
$$y - \overline{y} = \rho(x,y)\frac{\sigma(y)}{\sigma(x)}(x - \overline{x}).$$

Remarque. Avec cette dernière expression, on constate que le point moyen appartient à la droite de régression.

Exemple. On peut compléter l'exemple Python précédent en affichant la droite de régression. On comprend aussi sur cet exemple que la droite de régression peut être utilisée pour faire de la prédiction en prolongeant la droite. Précisons aussi qu'il existe des commandes Python plus directes pour afficher la droite de régression.

Editeur

```
# Tracé de la droite de régression

plt.xlabel('Années')
plt.ylabel('Bénéfices')
COVxy=np.mean(x*y)-np.mean(x)*np.mean(y)
a=COVxy/np.std(x)**2
b=np.mean(y)-a*np.mean(x)
t=np.linspace(1,14,2)
plt.plot(t,a*t+b)
```



Remarque. Un des inconvénients de la droite de régression est qu'elle est très sensible aux valeurs extrêmes : une seule valeur très éloignée de la droite modifie beaucoup le coefficient de corrélation linéaire et donc la pente de la droite. En particulier, la droite est très sensible aux erreurs de mesure.

Point de vue fonctions de plusieurs variables

Exercice 6

✦✦ *D'après HEC 2008 E*

Soient x, y deux séries statistiques avec $\sigma(x) \neq 0$. Soit f la fonction définie sur $\mathbb{R}^2$ par :

$$\forall\,(a,b)\in\mathbb{R}^2,\quad f(a,b)=\sum_{k=1}^{n}\left(ax_k+b-y_k\right)^2.$$

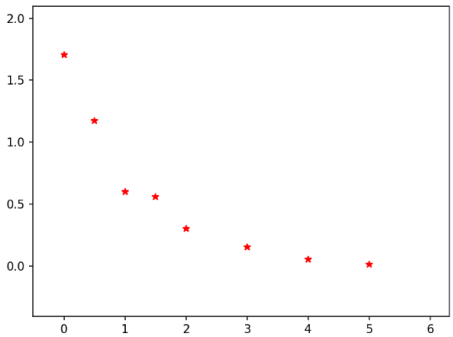
1. Justifier que f est de classe $\mathscr{C}^2$ sur $\mathbb{R}^2$.
2. (a) Écrire le système d'équations $\mathscr{S}$ permettant de déterminer les points critiques de f.
   (b) Résoudre le système $\mathscr{S}$.
   En déduire que f admet un unique point critique $(\hat{a},\hat{b})$ que l'on exprimera en fonction de $\overline{x}$, $\overline{y}$, $\sigma(x)$ et $\mathrm{cov}(x,y)$.
   (c) Montrer que ce point critique correspond à un minimum local de f.
   (d) Établir la formule suivante : $f(\hat{a},\hat{b})=n\sigma(y)^2\left(1-\rho(x,y)^2\right)$.
3. (a) Montrer que l'on a : $|\rho(x,y)|\leqslant 1$.
   (b) Que peut-on dire du nuage de points lorsque $|\rho(x,y)|=1$?

p. 16

Remarque. Retenons que lorsque le coefficient de corrélation linéaire est proche de ± 1, la droite de régression approche bien le nuage de points. Dans ce cas, on peut modéliser la dépendance et faire des prédictions.
De plus, si le coefficient de corrélation est positif, les données x_i « auront tendance » à augmenter lorsque y_i augmente et inversement si le coefficients est négatif.

3.4 Un exemple d'ajustement linéaire

Il arrive parfois que le nuage ne s'accorde pas à une droite mais à une courbe d'une fonction classique.

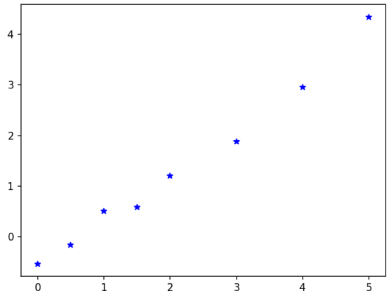


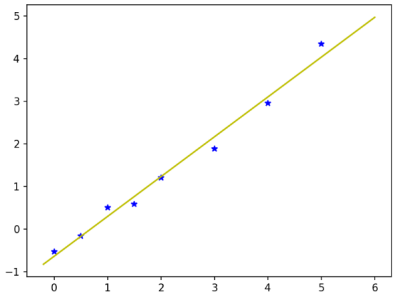
On peut alors supposer que le nuage s'organise autour de la courbe d'équation $y=f(ax+b)$ pour une certaine fonction f. Si f est bijective, on introduit la série $z=f^{-1}(x)$. Prenons l'exemple des séries :

Editeur

```
x=np.array([0,0.5,1,1.5,2,3,4,5])
y=np.array([1.707, 1.173, 0.601, 0.558,
    0.299,0.152, 0.052, 0.013])

plt.plot(x,y,'r*')
```

Dans l'exemple, on va tester avec la fonction f définie sur $\mathbb{R}$ par $f(t)=\exp(-t)$.
On procède alors à une régression linéaire sur les séries x, z pour estimer les réels a et b.





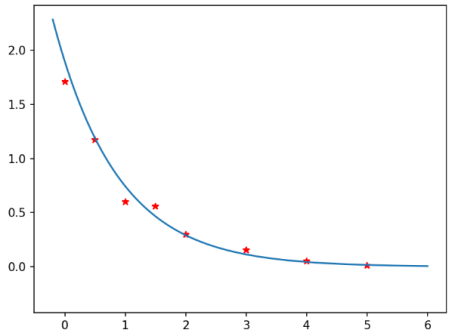
Console

```
>>> a
0.9349559793558717
>>> b
-0.6381866319796354
```

Testons le résultat, en comparant la courbe obtenue avec le nuage de points.

Editeur

```
z=-np.log(y)

COVxz=np.mean(x*z)-np.mean(x)*np.mean(z)
a=COVxz/np.std(x)**2
b=np.mean(z)-a*np.mean(x)

plt.plot(x,y,'r*')
t=np.linspace(-0.2,6,100)
plt.plot(t,np.exp(-(a*t+b)))
plt.show()
```

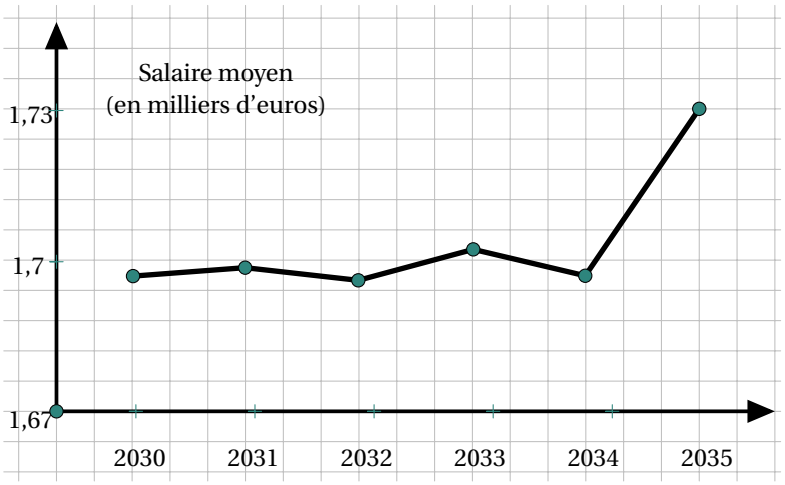


4 Compléments

Quelques biais statistiques

Exercice 7. ✦✦ Chercher l'erreur ou le biais dans les raisonnements suivants.

1. Dans la classe d'ECG, on compte qu'en moyenne les élèves ont 1.7 frères et sœurs. Ce qui donne 2.7 enfants par femme alors que le nombre d'enfants par femme n'est que de 1.8 dans la population française. Donc les enfants de familles nombreuses font plus facilement des études.
2. L'espérance de vie des professeurs de mathématiques est de 82 ans, soit 3 ans de plus que l'espérance de la population française. Donc les mathématiques sont bonnes pour la santé!
3. Lors de la seconde Guerre Mondiale, la Royal Air Force souhaite améliorer le taux de retour de ses bombardiers partis frapper les positions Allemandes. Les ingénieurs de la R.A.F décident d'étudier la localisation des impacts de balles sur les avions à leur retour de mission, puis de renforcer le blindage sur les zones les plus atteintes.
4. En 1936, le démocrate Franklin D. Roosevelt et le Républicain Alfred M. Landon concoururent pour la présidence des États-Unis. Juste avant l'élection, le journal *Literary Digest* réalisa un sondage de grande ampleur : 10 millions de bulletins de sondage sont distribués aux abonnés du magazine et à des gens figurant au bottin téléphonique. Après plus de 2 millions de réponses, Alfred Landon est annoncé président des États-Unis!
5. Comme le montre le graphe suivant, la nouvelle politique économique démarrée en 2034 a eu des effets considérables sur le salaire moyen des français.



6. Deux traitements contre les calculs rénaux présentent les résultats en fonction de la taille des calculs

petits calculs		gros calculs	
Traitement 1	Traitement 2	Traitement 1	Traitement 2
81 succès sur 87	234 succès sur 270	192 succès sur 263	55 succès sur 80

En regroupant les résultats, on obtient

Traitement 1	Traitement 2
78% (273 succès sur 350)	83% (289 succès sur 350)

Le second traitement est donc le plus efficace car il a le plus gros pourcentage de réussite.

7. Dans une université, à l'examen de la licence de biologie, les filles ont mieux réussi que les garçons et à l'examen de la licence de physique, les filles ont, là encore, mieux réussi que les garçons. Pourtant, en regroupant les résultats des deux licences, on découvre que les garçons ont mieux réussi que les filles. Il y a donc une erreur dans les résultats de la licence en faveur des garçons.

Les faits sont têtus, il est plus facile de s'arranger avec les statistiques.

MARK TWAIN

Indications et solutions

Exercice 1 — p. 5

Par définition de la moyenne

$$\overline{x'} = \frac{1}{N}\sum_{i=1}^{N} x'_i = \frac{1}{N}\sum_{i=1}^{N} (ax_i + b).$$

Par linéarité de la somme

$$\overline{x} = a\cdot\frac{1}{N}\sum_{i=1}^{N} x_i + b\frac{1}{N}\sum_{i=1}^{N} 1 = a\overline{x} + b.$$

Notons m_e, m'_e les médianes de x et x'.
Pour $a \geqslant 0$, l'ordre des termes de la série x' n'est pas changé. Donc

$$m'_e = am_e + b.$$

Si $a < 0$, l'ordre est inversé mais le terme « milieu » est inchangé et l'égalité précédente demeure.

Exercice 2 — p. 7

Par définition $\quad V' = \frac{1}{N}\sum_{i=1}^{N} (ax_i - b - \overline{x}')^2.$

$$\text{Or}\quad \overline{x'} = \frac{1}{N}\sum_{i=1}^{N}(ax_i+b) = a\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} b$$
$$= a\overline{x} + b.$$
$$\text{D'où}\quad V' = \frac{1}{N}\sum_{i=1}^{N}(ax_i - a\overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N} a^2(x_i-\overline{x})^2$$
$$= \frac{a^2}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 = a^2 V.$$

puis, par application de la racine carrée

$$\sigma' = \sqrt{V'} = \sqrt{a^2 V} = |a|\sqrt{V} = |a|\sigma.$$

Exercice 3 — p. 7

On a par linéarité de la somme

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i (m_i - \overline{x})^2$$
$$= \frac{1}{N}\sum_{i=1}^{p} n_i\left(m_i^2 - 2m_i\overline{x} + \overline{x}^2\right)$$
$$= \frac{1}{N}\sum_{i=1}^{p} n_i m_i^2 - \frac{2\overline{x}}{N}\sum_{i=1}^{p} n_i m_i + \frac{\overline{x}^2}{N}\sum_{i=1}^{p} n_i.$$

Or par définition des effectifs et de la moyenne

$$\frac{1}{N}\sum_{i=1}^{p} n_i m_i = \overline{x} \quad \text{et} \quad \frac{1}{N}\sum_{i=1}^{p} n_i = 1.$$

D'où

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i m_i^2 - 2\overline{x}\cdot\overline{x} + \overline{x}^2 = \frac{1}{N}\sum_{i=1}^{p} n_i m_i^2 - \overline{x}^2.$$

Exercice 4 — p. 7

1. Notons I, l'ensemble des indices k tels que la donnée x_k soit comprise entre $\overline{x} - 2\sigma(x)$ et $\overline{x} + 2\sigma(x)$. Autrement dit

$$k \in I \iff |x_k - \overline{x}| \leqslant 2\sigma(x).$$

On a alors

$$\sum_{k=1}^{N}(x_k-\overline{x})^2 = \sum_{k\in I}(x_k-\overline{x})^2 + \sum_{k\notin I}(x_k-\overline{x})^2.$$

Or pour $k \in [\![1;N]\!]$

$$|x_k - \overline{x}| > 2\sigma(x) \quad\Rightarrow\quad |x_k-\overline{x}|^2 \geqslant 4\sigma(x)^2.$$

Par somme

$$\sum_{k\notin I}(x_k-\overline{x})^2 \geqslant \sum_{k\notin I} 4\sigma(x)^2 = 4\sigma(x)^2\,\text{Card}(\bar{I})$$

Or, par définition de r,

$$\text{Card}(I) = r, \quad \text{Card}(\bar{I}) = N - r.$$

D'où le résultat puisque

$$\sum_{k=1}^{N}(x_k-\overline{x})^2 \geqslant \sum_{k\notin I}(x_k-\overline{x})^2 \geqslant 4\sigma(x)^2(N-r).$$

2. De plus, par définition de la variance

$$N\sigma(x)^2 = \sum_{k=1}^{N}(x_k-\overline{x})^2.$$

D'où $\quad N\sigma(x)^2 \geqslant 4\sigma(x)^2(N-r) = 4\sigma(x)^2 N - 4\sigma(x)^2 r.$

$$\Rightarrow\quad 4\sigma(x)^2 r \geqslant 3\sigma(x)^2 N$$
$$\Rightarrow\quad r \geqslant \frac{3}{4}N.$$

Noter que le résultat reste bien valable pour $\sigma(x) = 0$, puisque dans ce cas, toutes les données de la série sont identiques.
Finalement, au moins trois quarts de la série est compris entre $\overline{x} - 2\sigma(x)$ et $\overline{x} + 2\sigma(x)$.

Exercice 5 — p. 8

1. Rappelons l'inégalité de Cauchy-Schwarz dans $\mathbb{R}^N$. Pour $(x_i), (y_i) \in \mathbb{R}^N$

$$\left(\sum_{i=1}^{N} x_i y_i\right)^2 \leqslant \sum_{i=1}^{N} x_i{}^2 \cdot \sum_{i=1}^{N} y_i{}^2.$$

Dans notre cas, on obtient :

$$\left(\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})\right)^2 \leqslant \sum_{i=1}^{N}(x_i-\overline{x})^2\sum_{i=1}^{N}(y_i-\overline{y})^2.$$

En divisant par N^2 et en revenant aux définitions

$$\mathrm{Cov}(x,y)^2 \leqslant \sigma(x)^2\cdot\sigma(y)^2.$$

D'où l'encadrement de $\rho(x,y)$.

2. Il y a égalité $\rho(x,y)=\pm1$ si et seulement si il y a égalité dans l'inégalité de Cauchy-Schwarz. C'est-à-dire, si et seulement si $(x_i-\overline{x})_i$ et $(y_i-\overline{y})_i$ sont colinéaires. Comme $\sigma(x)\neq 0$, $(x_i-\overline{x})_i \neq 0_{\mathbb{R}^N}$ et on peut affirmer l'existence de $a\in\mathbb{R}$ tel

$$\forall i\in[\![1;N]\!],\quad y_i-\overline{y}=a(x_i-\overline{x})$$
$$y_i=ax_i+b$$

où $b=\overline{y}-a\overline{x}$.
Précisons que l'équivalence est bien conservée car s'il existe $a,b\in\mathbb{R}$ tels que

$$y_i=ax_i+b.$$

On a par somme

$$\overline{y}_i=a\overline{x}+b \quad\text{puis}\quad b=\overline{y}-a\overline{x}.$$

Dans ce cas

$$\begin{aligned}\mathrm{Cov}(x,y)&=\mathrm{Cov}(x,ax+b)\\&=a\underbrace{\mathrm{Cov}(x,x)}_{\geqslant 0}.\end{aligned}$$

Le réel a et $\rho(x,y)$ ont même signe.

Exercice 6 p. 12

1. La fonction f est de classe $\mathscr{C}^2$ sur $\mathbb{R}^2$ car polynomiale.

2.(a) On a

$$\partial_1 f(a,b)=\sum_{k=1}^{n}x_k\cdot 2(ax_k+b-y_k)$$
$$\partial_2 f(a,b)=\sum_{k=1}^{n}2(ax_k+b-y_k).$$

Le système d'équations donnant le point critique est :

$$\mathscr{S}:\ \begin{cases}\displaystyle\sum_{k=1}^{n}x_k(ax_k+b-y_k) &= 0\\ \displaystyle\sum_{k=1}^{n}(ax_k+b-y_k) &= 0\end{cases}$$
$$\iff \begin{cases}\displaystyle a\sum_{k=1}^{n}{x_k}^2+b\sum_{k=1}^{n}x_k &= \displaystyle\sum_{k=1}^{n}x_ky_k\\ \displaystyle a\sum_{k=1}^{n}x_k+bn &= \displaystyle\sum_{k=1}^{n}y_k.\end{cases}$$

En divisant par n, il vient :

$$\mathscr{S}\iff \begin{cases}a\overline{x^2}+b\overline{x} &= \overline{x*y}\\ a\overline{x}+b &= \overline{y}\end{cases}\qquad L_2$$

Matriciellement :

$$\begin{bmatrix}\overline{x^2}&\overline{x}\\ \overline{x}&1\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix}=\begin{bmatrix}\overline{x*y}\\ \overline{y}\end{bmatrix}.$$

Le déterminant de la matrice de gauche est

$$\overline{x^2}-\overline{x}^2=\sigma(x)^2\neq 0$$

et la solution est donnée par :

$$\begin{bmatrix}a\\b\end{bmatrix}=\frac{1}{\sigma(x)^2}\begin{bmatrix}1&-\overline{x}\\ -\overline{x}&\overline{x^2}\end{bmatrix}\begin{bmatrix}\overline{x*y}\\ \overline{y}\end{bmatrix}.$$

D'où

$$a=\frac{1}{\sigma(x)^2}(\overline{x*y}-\overline{x}\cdot\overline{y})=\frac{\mathrm{Cov}(x,y)}{\sigma(x)^2}$$

et en reprenant la ligne L_2

$$b=\overline{y}-a\overline{x}=\overline{y}-\frac{\mathrm{Cov}(x,y)}{\sigma(x)^2}\overline{x}.$$

2.(c) La matrice Hessienne au point critique est

$$\begin{aligned}H_{\hat{a},\hat{b}}&=\begin{bmatrix}\partial^2_{1,1}f(\hat{a},\hat{b})&\partial^2_{2,1}f(\hat{a},\hat{b})\\ \partial^2_{1,2}f(\hat{a},\hat{b})&\partial^2_{2,2}f(\hat{a},\hat{b})\end{bmatrix}\\&=\begin{bmatrix}2\displaystyle\sum_{k=1}^{n}{x_k}^2&2\displaystyle\sum_{k=1}^{n}x_k\\ 2\displaystyle\sum_{k=1}^{n}x_k&2n\end{bmatrix}\\&=2\begin{bmatrix}\sigma(x)^2+\overline{x}^2&\overline{x}\\ \overline{x}&1\end{bmatrix}.\end{aligned}$$

$H_{\hat{a},\hat{b}}$ est symétrique donc diagonalisable. À l'aide du déterminant, on trouve que les valeurs propres vérifient

$$\lambda_1\lambda_2=\sigma(x)^2>0\quad\text{et}\quad\lambda_1+\lambda_2=\sigma(x)^2+\overline{x}^2+1>0.$$

Les valeurs propres sont positives et f admet un minimum local en $(\hat{a},\hat{b})$.

2.(d)

$$\begin{aligned}f(\hat{a},\hat{b})&=\sum_{k=1}^{n}(\hat{a}x_k-y_k+b)^2\\&=\sum_{k=1}^{n}(\hat{a}x_k-y_k+\overline{y}-\hat{a}\overline{x})^2\\&=\sum_{k=1}^{n}(\hat{a}(x_k-\overline{x})+\overline{y}-y_k)^2\\&=\sum_{k=1}^{n}\hat{a}^2(x_k-\overline{x})^2+2\hat{a}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})+\sum_{k=1}^{n}(y_k-\overline{y})^2\\&=\hat{a}^2n\sigma(x)^2+n\sigma(y)^2+\hat{a}(-2)n\,\mathrm{Cov}(x,y)\\&=\frac{\mathrm{Cov}(x,y)^2}{\sigma(x)^2}\cdot n+n\sigma(y)^2-2n\frac{\mathrm{Cov}(x,y)^2}{\sigma(x)^2}\\&=n\sigma(y)^2-n\frac{\mathrm{Cov}(x,y)^2}{\sigma(x)^2}\\f(\hat{a},\hat{b})&=n\sigma(y)^2\left(1-\rho(x,y)^2\right).\end{aligned}$$

3.(a) On a toujours $f(\hat{a},\hat{b})\geqslant 0$, d'où

$$1-\rho(x,y)^2\geqslant 0,$$

puis

$$|\rho(x,y)|\leqslant 1.$$

3.(b) Si $|\rho(x,y)|=1$, alors $f(\hat{a},\hat{b})=0$. D'où

$$\sum_{k=1}^{n} \underbrace{\left(\hat{a}x_k+\hat{b}-y_k\right)^2}_{\geqslant 0}=0.$$

Puis, pour tout $k\in[\![1;n]\!]$

$$\hat{a}x_k+\hat{b}-y_k=0.$$

Autrement dit, tous les points du nuage appartiennent à la même droite.

Remarque. Cette droite commune est la droite de régression.

Exercice 7 p. 13

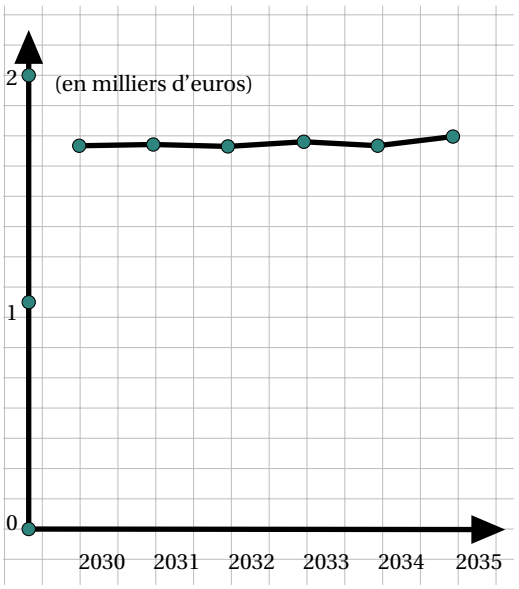
1. Il y a un biais dans le calcul précédent puisqu'en demandant uniquement aux enfants, on ne comptabilise que les femmes avec au moins un enfant en oubliant celles sans enfant.

2. De nouveau, le raisonnement précédent comporte plusieurs biais. Par exemple, il faut avoir plus d'une vingtaine d'années pour exercer ce métier. Dans un premier temps, il faudrait comparer l'espérance de 82 ans à l'espérance de vie à 25 ans.

3. On illustre ici ce qu'on appelle le biais du survivant. Il aurait fallu plutôt blinder les zones où les impacts de balles sont les moins nombreux. Ces zones sont probablement vitales pour l'avion car les avions touchés à ces endroits ne sont pas revenus.

4. L'échantillon de la population du journal et des abonnées du téléphone (en 1936) ne constitue pas un échantillon représentatif de la population américaine. On peut imaginer que l'on sonde plutôt la partie la plus aisée de la population.
Au même moment, le sondeur américain George Gallup prédisait avec justesse la réélection de Roosevelt avec seulement un échantillon plus représentatif de 5 000 personnes. Cet exemple marque les débuts du sondage d'opinion.

5. Il faut noter que l'axe des ordonnées ne s'étend que de 1670 à 1730. L'augmentation paraît plus importante qu'elle ne l'est en réalité (en moyenne 30 euros).



6. Détaillons le calcul en distinguant bien les traitements sur les petits et gros calculs.
⇀ Sur les petits calculs, le traitement A a un taux de réussite de $81/87\simeq 93\%$ alors que le B affiche un taux légèrement inférieur de $234/270\simeq 87\%$.
⇀ Sur les gros calculs, les traitements A et B ont respectivement des taux de 73% et 69%.
Dans les deux cas, le traitement A est plus efficace. Il est donc à privilégier pour traiter les petits et gros calculs.

Remarque. Ce résultat très contre-intuitif est connu sous le nom de paradoxe de Simpson.

7. C'est une nouvelle illustration du paradoxe de Simpson. En regroupant les résultats, les garçons ont une meilleure moyenne (comme pour le traitement B).