TP5

Statistiques descriptives univariées et bi-variées

5.1 Statistiques univariées

Compétences attendues

- Regrouper une série statistique par modalités ou par classes.
- Connaître les indicateurs de position (moyenne, médiane, quartiles) et les commandes associées.
- Connaître les indicateurs de dispersion (écart-type, étendue, distance inter-quartiles) et mes commandes associées.
- Représenter graphiquement une série statistique.

5.1.1 Présentation des données

On considère un ensemble Ω appelé **population** en statistique descriptive. On appèlera ses éléments $\omega \in \Omega$ des **individus**.

DÉFINITION 5.1 (Caractère)

Un **caractère** ou **variable** sur la population Ω est une application $X : \Omega \longrightarrow E$ où E désigne un ensemble quelconque.
Si E est un ensemble de nombre on dit que X est un caractère **quantitatif**. Dans le cas contraire on par le de caractère **qualitatif**.

Exemple : Un caractère possible sur la population française est la taille (caractère quantitatif) ou alors la couleur des yeux (caractère qualitatif).
La connaissance complète d'un caractère X peut être rendue difficile, voire impossible, de part la taille de la population Ω. Afin de pouvoir l'étudier, on peut considérer ce caractère seulement pour une partie finie de la population Ω.

DÉFINITION 5.2

• On appelle **échantillon** de la population Ω toute partie finie $\{\omega_1, ..., \omega_n\}$ de Ω. Son cardinal n est alors la **taille** ou l'**effectif** de l'échantillon.

• On appelle **série statistique** d'un échantillon $\{\omega_1, ..., \omega_n\} \subset \Omega$ pour le caractère X, la donnée de la liste $(x_1, ..., x_n)$ des valeurs prises par X sur l'échantillon.

• Les valeurs prises par X sont appelées **modalités**.

• L'**effectif d'une modalité** m est le nombre n_m de fois où m apparaît dans la série statistique $(x_1, ..., x_n)$.

• La **fréquence d'une modalité** m est le réel $f_m = \dfrac{\text{effectif de } m}{\text{effectif total}} = \dfrac{n_m}{n}$.

• La **fréquence cumulée d'une modalité** m est le réel $p_m = \displaystyle\sum_{m' \leqslant m} f_{m'}$.

On notera la correspondance entre probabilités et statistiques:

Variable aléatoire X	$\leftrightarrow$	Caractère X
Support de X	$\leftrightarrow$	Ensemble des modalités de X
Probabilités $P(X = i)$	$\leftrightarrow$	fréquence f_i
Fonction de répartition $F_X : x \mapsto P(X \leqslant x)$	$\leftrightarrow$	Fréquences cumulées $p_k = \displaystyle\sum_{i \leqslant k} f_i$.

PROPRIÉTÉ 5.3

Si $(x_1,...,x_n)$ est une série statistique, $(m_1,...,m_p)$ ses modalités d'effectifs $(n_1,...,n_p)$ et de fréquences $f_1,...,f_p)$ alors on a :

$$\sum_{i=1}^{p} n_i = n \quad \text{et} \quad \sum_{i=1}^{p} f_i = 1.$$

Une série statistique brute ne permettant pas une lecture efficace des données, on souhaite la présenter de manière plus synthétique. Pour cela, on procède de deux manières distinctes selon le nombre de ses modalités.

Regroupement par modalités

Dans le cas où le nombre de modalités de la série est raisonnable, on regroupe la série par **modalités-effectifs**, c'est à dire qu'on donne :

- La liste (m_i) des modalités de X,
- Les effectifs (n_i) correspondants.

On peut aussi choisir de présenter la série par **modalités-fréquences**.

Regroupement par classe

Dans le cas où le nombre de modalités est trop grand, plutôt que de conserver toutes les valeurs, il est préférable de les regrouper par des intervalles appelés **classes** :

- $I_1 = [c_0, c_1[, I_2 = [c_1, c_2[, ..., I_k = [c_{k-1}, c_k[$, l'amplitude de la classe I_i étant $c_i - c_{i-1}$;
- on note n_i le nombre d'éléments de X appartenant à l'intervalle I_i.

On se ramène ainsi à une série statistique de taille k dont les modalités sont les milieux $y_i = \dfrac{c_{i-1}+c_i}{2}$ des classes et d'effectifs correspondants les n_i.

Python Commandes utiles

- `x = np.array([x1,...,xn])` représente la série statistique $(x_1,...,x_n)$.
- `len(x)` renvoie l'effectif de la série.
- `np.sum(x==xi)` renvoie l'effectif de la modalité x_i.
- `np.mean(x==xi)` renvoie la fréquence de la modalité x_i.
- `np.mean(x<=xi)` renvoie la fréquence cumulée de la modalité x_i.

EXERCICE 5.1.

1. Compléter la fonction suivante qui prend en argument une série statistique x et renvoie la liste des modalités de x.

```
def mod(x):
    L= []
    for k in .......:
      if k not in
    ......... :
    ...............
    return L
```

2. Compléter la fonction suivante, qui prend en argument une série statistique et renvoie le tri de x par modalités-effectifs.

```
def mod_eff(x):
    L=mod(x) #liste des modalites de x
    n= ..... #nombre de modalites de x
    T=np.zeros((2,n)) # on cree un tableau a double entree
    i=0
    for k in L:
      T[0,i]= ..... #  modalite
      T[1:i]=np.sum(x==k) # effectif
   i=i+1
    return T
```

3. En vous inspirant de ce fonction créer une fonction d'en-tête `def mod_freq(x):` qui prend en argument une série statistique et revoie la présentation modalités-fréquences.
4. Appliquer vos deux fonctions à la série $x = (2,11,7,2,15,4,5,5,5,13,5,15,7,7,8,10,10,10,11,13,7,2,15,15)$.
   Noter les résultats sur votre feuille sous forme de tableau.

5.1.2 Indicateurs de position

DÉFINITION 5.4 (Moyenne empirique)

On appelle **moyenne empirique** de la série statistique $x = (x_1, ..., x_n)$ le réel :

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

PROPRIÉTÉ 5.5

• Si la série est regroupée par modalités-effectifs alors on a :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} m_i.n_i.$$

• Si la série est regroupée par modalités-fréquences alors on a :

$$\overline{x} = \sum_{i=1}^{p} m_i.f_i.$$

REMARQUE **5.1.**
On notera la nouvelle correspondance entre probabilités et statistiques :

$$\text{Espérance de } X \quad \leftrightarrow \quad \text{Moyenne empirique } \overline{x}$$

DÉFINITION 5.6 (Médiane)

La **médiane** d'une série statistique ordonnée est un réel m partageant la série en deux séries d'effectifs égaux.

REMARQUE **5.2.**
Dans le cas où la série est de taille pair, la médiane de la série sera la moyenne des deux termes du milieu.

EXERCICE 5.2. Soit $x = (2,4,7,11,23,40)$ et $y = (1,5,7,10,12)$. Déterminer la médiane de chacune de ces deux séries statistiques.

DÉFINITION 5.7 (Quartile)

• Le **premier quartile** q_1 de x est la plus petite valeur de x telle que 25% des valeurs lui soient inférieures ou égales.
• Le **troisième quartile** q_3 de x est la plus petite valeur de x telle que 75% des valeurs lui soient inférieures ou égales.

Python Commandes utiles

- `np.mean(x)` renvoie la moyenne empirique du vecteur `x`.
- `np.median(x)` renvoie la médiane du vecteur `x` (non nécessairement ordonné).

EXERCICE 5.3.

1. Reprendre la série statistique x de l'Exercice 1 et déterminer sa moyenne empirique ainsi que sa médiane.
2. En vous inspirant des fonctions créées dans l'Exercice 1, déterminer, sous forme de tableau le tri de x par modalités-fréquences cumulées.
3. En déduire le premier et le troisième quartile de x.

5.1.3 Indicateurs de dispersion

DÉFINITION 5.8

Soit $x = (x_1, ..., x_n)$ une série statistique.

- On appelle **variance empirique** de x le réel positif : $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
- On appelle **écart-type** de x le réel positif : $\sigma = \sqrt{v}$.

REMARQUE **5.3.**

- La variance empirique correspond bien sûr à la variance dans la correspondance entre probabilités et statistiques
- La variance et l'écart-type mesure la dispersion de la série statistique autour de sa moyenne.

PROPRIÉTÉ 5.9 (Formule Koenig-Huygens)

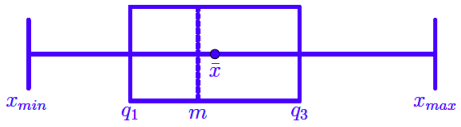
$$v = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\overline{x}\right)^2$$

DÉFINITION 5.10

- On appelle **étendue** d'une série statistique la différence entre la plus grande et la plus petite modalité.
- On appelle **distance inter-quartiles** le réel $q_3 - q_1$.

REMARQUE **5.4.**

La distance inter-quartiles est un indicateur de dispersion : c'est la longueur de l'*intervalle inter-quartiles* $[q_1, q_3]$, lequel contient la moitié des valeurs de la série, réparties autour de la médiane m. On représente parfois une série sous forme d'un diagramme *boîte à moustache* :



Python Commandes utiles

- `np.var(x)` renvoie la variance empirique de la série `x`.
- `np.std(x)` (pour standard deviation) renvoie l'écart-type de la série `x`.
- `np.max(x)-np.min(x)` renvoie l'étendue de la série `x`.

EXERCICE 5.4. Reprendre la série x de l'Exercice 1, déterminer son écart-type et représenter son diagramme à moustache sur votre feuille.

5.1.4 Représentations graphiques

On suppose avoir importé la bibliothèque `matplotlib.pyplot`.

Diagrammes en bâtons

DÉFINITION 5.11

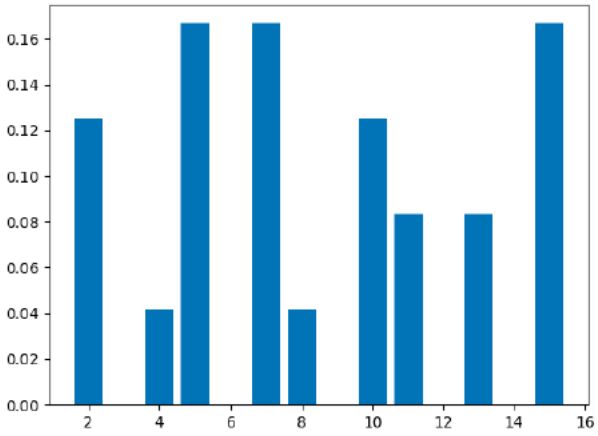
On représente une série statistique *groupée par modalités* à l'aide d'un **diagramme en bâtons** en plaçant sur l'axe des abscisses les modalités et en dressant à la verticale de chacune d'elles, un bâton de hauteur égale à l'effectif ou la fréquence de cet effectif.

Python Commande utile

Si `x` et `y` sont des vecteurs de même dimension, `plt.bar(x,y)` trace le diagramme en bâtons d'abscisses `x` et d'ordonnée `y`.

EXERCICE 5.5.

1. A l'aide de la fonction `mod_freq` précédemment écrite, obtenir le diagramme en bâtons de la série x de l'Exercice 1 et montrer que l'on obtient le diagramme suivant :



2. Ecrire une fonction d'en-tête `bar(x)` qui prend en argument une série statistique `x` et renvoie le diagramme en bâtons de la série.

Histogrammes

DÉFINITION 5.12

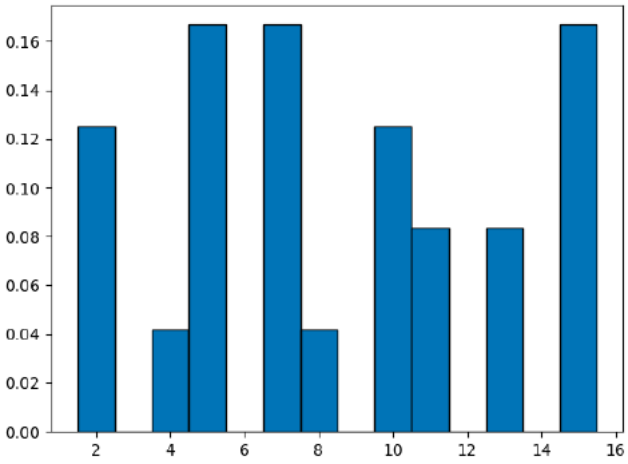
On représente une série statistique *groupée par classes* à l'aide d'un **histogramme**, en plaçant les c_i sur un axe horizontal et en traçant à la verticale un rectangle de base $[c_i, c_{i+1}]$ et d'aire égale à la fréquence de la classe correspondante.

Python Commandes utiles

• `plt.hist(x,n)` renvoie l'histogramme de la série `x` en `n` classes équiréparties entre la plus petite valeur et la plus grande valeur de x.
• `plt.hist(x,c)` renvoie l'histogramme de la série `x` dont les classes sont définies par le vecteur aux composantes strictement croissantes c.
• `plt.hist(x,bins='auto)` renvoie l'histogramme de la série `x` dont les classes sont définies automatiquement.

EXERCICE 5.6.

Montrer que l'histogramme de la série x de l'Exercice 1 est


0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2
4
6
8
10
12
14
16

5.2 Statistiques bi-variées

Compétences attendues

- Représenter un nuage de points associé à une série statistique double.
- Représenter la droite des moindres carrés.
- Calculer le coefficient de corrélation linéaire et interpréter sa valeur.

Objectifs

Les données statistiques ne vont pas toujours toutes seules, et pour un même individu, il est possible de s'intéresser à plusieurs caractères. Dans ce TP, nous nous limiterons à l'étude simultanée de deux caractères. Nous nous poserons alors la question suivante : peut-on exprimer l'un de ces caractères en fonction de l'autre? Plus précisément, l'un est-il une fonction affine de l'autre? De cette recherche de correspondances peuvent découler des analyses fines, explicatives voire prédictives, ou au contraire mettre en évidence une absence de corrélation entre ces caractères.

5.2.1 Présentation des données

DÉFINITION 5.13

Soit $\Omega = \{\omega_1, \ldots, \omega_n\}$ une population d'effectif n, sur laquelle nous étudions deux caractères quantitatifs X, Y : $\Omega \to \mathbb{R}$ avec X supposé non constant. Pour tout $i \in [\![1, n]\!]$, on note :

- $x_i = X(\omega_i)$ la modalité de X prise par l'individu ω_i,
- $y_i = Y(\omega_i)$ la modalité de Y prise par l'individu ω_i.

DÉFINITION 5.14 (Série statistique double ou bi-variée)

On appelle série statistique double (ou bivariée) de la population Ω pour le couple de caractères (X, Y) la donnée du n-uplet $\big((x_i, y_i)\big)_{1 \leqslant i \leqslant n}$ des modalités de (X, Y) sur Ω.

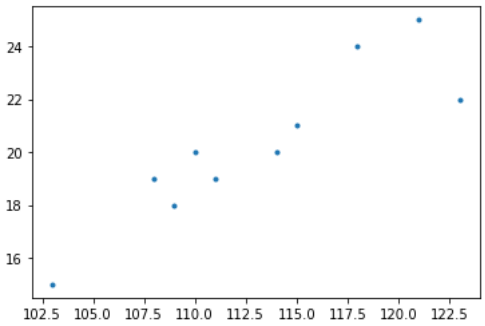
Exemple. 10 enfants de 6 ans d'une même classe sont mesurés et pesés. On note X le caractère désignant la taille de l'enfant (en centimètres) et Y celui désignant le poids de l'enfant (en kilogrammes). On obtient la série statistique double suivante :

Enfant	1	2	3	4	5	6	7	8	9	10
X	121	123	108	118	111	109	114	103	110	115
Y	25	22	19	24	19	18	20	15	20	21

Représentation graphique : nuage de point

DÉFINITION 5.15 (nuage de points)

On représente une série statistique double à l'aide d'**un nuage de points**. C'est l'ensemble des points M_i du plan de coordonnées (x_i, y_i) pour tout $1 \leqslant i \leqslant n$.



Python Commandes utiles

• On représentera une série statistique double sur Python par deux vecteurs `x` et `y` de type `array` ou `list`
⇝ On rappelle que `n=len(x)=len(y)` désigne l'effectif total et commun de chacune des deux séries.

• L'instruction `plt.plot(x,y,'.')` trace le nuage de points d'abscisses données par `x` et d'ordonnées par `y`.
⇝ L'option `'.'` a pour effet de ne pas relier les points. On peut aussi utiliser `'o'` ou `'+'`.

5.2.2 Modèle de régression

Lorsqu'on étudie une série statistique bivariée, on peut penser que l'un des caractères, par exemple X, est une cause de l'autre, par exemple Y. On dit alors que X est le **caractère explicatif** et Y le **caractère expliqué**. On cherche alors un **modèle de régression**, c'est-à -dire une expression de Y en fonction de X :

$$Y = f(X)$$

où la fonction f est appelée **fonction de régression**.

Pour envisager un modèle de régression satisfaisant, on trace le nuage de points de la série double (x, y) afin de proposer une fonction de régression.

EXERCICE 5.7.

1. Pour la série statistique double proposée en exemple, quel est le caractère explicatif et le caractère expliqué ?
2. Représenter le nuage de points associé à cette série statistique double. Quel type de fonction de régression vous semble appropriée?
3. Proposer l'équation d'une droite qui passe "très près" de tous ces points.

Dans la suite de ce TP, nous nous intéresserons plus particulièrement aux modèles de régression linéaire, c'est-à -dire les modèles pour lesquels la fonction de régression f est affine. Nous cherchons ainsi à savoir si le caractère expliqué Y peut s'exprimer comme fonction affine de X, ce qui revient à chercher $(a, b) \in \mathbb{R}^2$ tel que :

$$Y = aX + b.$$

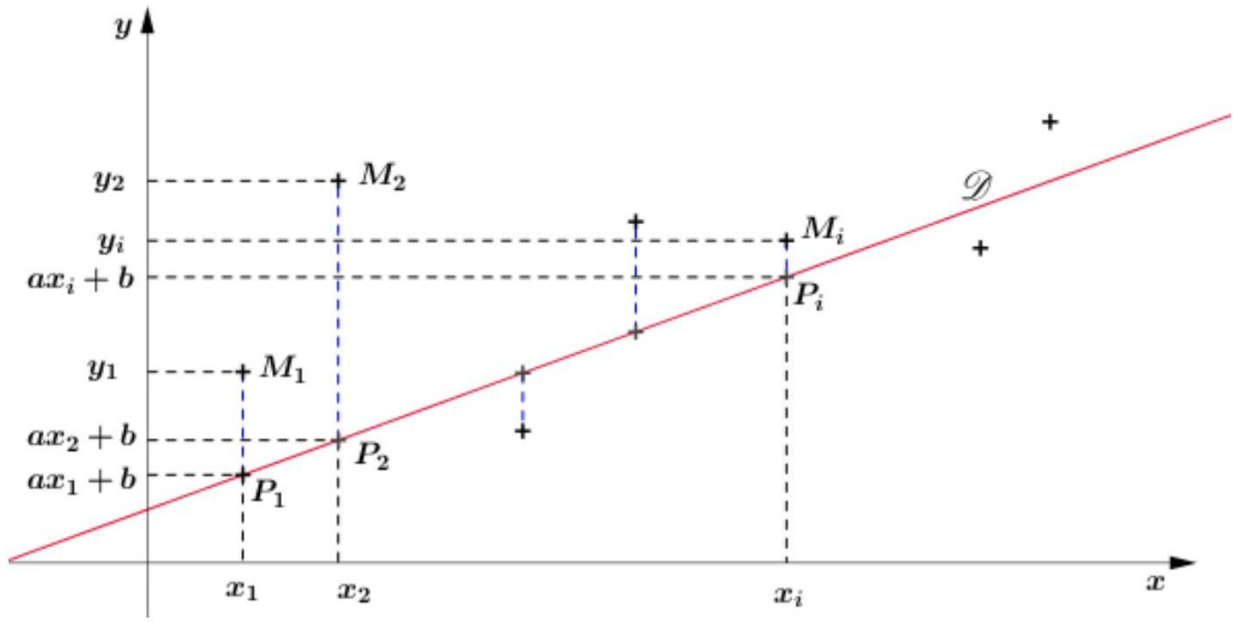
Régression linéaire ou problème des moindres carrés

Étudions l'existence d'une relation linéaire pour la série statistique double de nuage de points $\left(\left(x_i, y_i\right)\right)$. On souhaite ainsi "placer" tous les points M_i de coordonnées $\left(x_i, y_i\right)$ sur une même droite $\mathscr{D}$. On cherche donc $(a, b) \in \mathbb{R}^2$ tel que :

$$\forall i \in [\![1; n]\!], \quad y_i = ax_i + b.$$

Seulement, il y a très peu de chance qu'une telle droite existe, nos points n'étant très probablement pas alignés.

Pour obtenir un modèle de régression linéaire "le plus satisfaisant possible", on va chercher la "meilleure" droite $\mathscr{D}$ approchant l'ensemble des points M_i au sens suivant : pour tout $1 \leqslant i \leqslant n$, on mesure (le carré de) la distance M_iP_i entre M_i et le point $P_i \in \mathscr{D}$ d'abscisse x_i.



On cherche donc $(a,b) \in \mathbb{R}^2$ rendant minimale la quantité :

$$\Delta = \sum_{i=1}^{n} M_i P_i^2 = \sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2.$$

THÉORÈME 5.16 (Admis : Problème des moindres carrés ou régression linéaire))

Considérons une série statistique double $\left((x_i, y_i)\right)_{1\leqslant i\leqslant n}$.
Il existe une et une seule droite minimisant la quantité Δ. On l'appelle **la droite des moindres carrés** associée à la série statistique double $\left((x_i, y_i)\right)_{1\leqslant i\leqslant n}$.

REMARQUE **5.5.**
La droite des moindres carrés est la droite qui passe "la plus près" de tous les points du nuage de points au sens des moindres carrés (c'est-à -dire au sens où elle minimise la quantité Δ). Elle fournit donc "le meilleur" modèle de régression linéaire. Cela ne dit cependant pas si ce modèle est pertinent ou non...

Représentation de la droite des moindres carrés avec `Python`

On peut obtenir l'équation de la droite des moindres carrés à l'aide de la commande Python suivante (hors programme).

`Python` Commandes utiles

Soient `x,y` des vecteurs ou listes de même taille, `x` ayant au moins deux coefficients distincts (ce qui en pratique sera hautement le cas).

La commande `a,b=np.polyfit (x,y,1)` renvoie deux réels `a,b` tels que $y = ax + b$ est l'équation de la droite des moindres carrés pour la série statistique double `(x,y)`.

Par conséquent la droite des moindres carrés s'obtient par la commande `plt.plot([x_1,x_2],[a*x_1+b,a*x_2+b])` où x_1 et x_2 sont prises en fonction des valeurs extrémales de la série `x`.

EXERCICE 5.8.

On reprend la série statistique double du premier exemple.

1. Utiliser `Python` pour déterminer les coefficients a et b de la droite des moindres carrés.
2. Représenter la droite des moindres carrés sur le même graphique que le nuage de point. Le résultat est-il conforme à vos attentes ?
3. Quel est le signe du coefficient directeur de cette droite ? Comment l'interpréteriez-vous ?
4. Estimer graphiquement le poids d'un enfant de 6 ans qui mesure $1m20$.

EXERCICE 5.9.

1. Ecrire une fonction Python d'en-tête `def alea():` qui crée deux vecteurs `x` et `y` contenant chacun 1000 nombres tirés au hasard dans l'intervalle $[0;1]$ et renvoie le nuage de point associé. On pourra utiliser la fonction `rd.random()` de la bibliothèque `random`.
2. Modifier votre fonction pour faire apparaître la droite des moindres carrées dans la même fenêtre que le nuage de point.

5.3 Covariance et coefficient de corrélation linéaire

Rappelons que si x désigne une série statistique alors la moyenne empirique est donnée par :

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

DÉFINITION 5.17 (Point moyen)

On appelle **point moyen** de la série statistique double (x, y) le point de coordonnées $(\overline{x}, \overline{y})$.

EXERCICE 5.10.

Représenter le point moyen dans le même graphique que le nuage de points et de la droite des moindres carrés.

1. Dans l'exemple initial.
2. En modifiant la fonction `alea()`.
3. Que remarque-t-on dans chaque cas ?

PROPRIÉTÉ 5.18

Le point moyen d'une série statistique double appartient à la droite des moindres carrés.

DÉFINITION 5.19 (Covariance empirique)

La covariance empirique d'une série statistique double est définie par :

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

PROPRIÉTÉ 5.20 (Formule de Koenig-Huygens)

$$cov(x, y) = \overline{xy} - \overline{x}\,\overline{y}.$$

PROPRIÉTÉ 5.21 (Équation de la droite des moindre carrés (Hors-Programme))

La droite des moindres carrés à pour équation :

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

où σ_x^2 désigne la variance empirique de la série x.
→ en particulier la droite passe par le point moyen $(\overline{x}, \overline{y})$.

DÉFINITION 5.22 (Coefficient de corrélation linéaire)

On appelle coefficient de corrélation linéaire de x et y le réel définit par :

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x, \sigma_y}.$$

PROPRIÉTÉ 5.23 (Propriétés du coefficient $\rho_{x,y}$)

- $|\rho_{x,y}| \leqslant 1$
- $\rho_{x,y} = \pm 1 \iff Y = aX + b$ avec $(a,b) \in \mathbb{R}^2 \iff X$ et Y sont en régression linéaire.

REMARQUE **5.6.**
- Le coefficient $\rho_{x,y}$ est donc une quantité numérique qui permet de mesure "à quel point" les séries sont linéairement corrélées.
- Un coefficient $\rho_{x,y} = 0$ signifie une absente de corrélation linéaire entre x et y.
- En général on estime qu'il y a une forte corrélation linéaire entre x et y lorsque $|\rho_{x,y}| \geqslant 0,9$. Lorsque c'est le cas, un modèle de régression linéaire pourra servir de modèle prédictif.

Python Commandes utiles

- On rappelle que np.var(x) et `np.std(x)` renvoient respectivement la variance et l'écart-type de la série statistique `x`.
- `np.mean([x[i]*y[i] for i in range(n)])-np.mean(x)*np.mean(y)` renvoie la covariance de la série statistique double `(x,y)`.
- `np.corrcoef(x,y)` renvoie la matrice $\begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix}$.
- `np.corrcoef(x,y)[0,1]` renvoie donc $\rho_{x,y}$.
- en se limitant au programme on obtient $\rho_{x,y}$ ainsi :

```
(np.mean([x[i]*y[i] for i in range(n)])-np.mean(x)*np.mean(y))/(np.std(x)*np.std(y)).
```

EXERCICE 5.11 (Sensibilité aux conditions intiales)**.**

1. Calculer le coefficient $\rho_{x,y}$ pour la série double de l'exemple initial. Cela correspond-il à ce que vous vous attendiez ?
2. Un nouvel arrivant dans la classe prétend mesurer 105cm et peser 27kg.
   (a) Calculer le nouveau coefficient $\rho'_{x,y}$ de cette nouvelle série statistique double. Comparer avec l'ancienne valeur $\rho_{x,y}$.
   (b) Représenter dans un sous graphique (`subplot`) les deux nuages de points correspondant aux deux séries doubles ainsi que leurs droites des moindres carrées associées. Que constatez vous ?

A retenir

La méthode des moindres carrés est très sensible aux valeurs extrêmes : une seule valeur très éloignée de la droite des moindres carrés a une "grosse" influence sur le coefficient de corrélation linéaire et sur la position de la droite. Pour y remédier il peut être avantageux d'exclure au préalable les valeurs aberrantes des séries statistiques avant d'en faire l'étude.

EXERCICE 5.12.

On pose `x=[2*rd.random()-1 for k in range(100)]` et `y=[2*x[i] for i in range(100)]`.

1. Représenter dans un même graphique, le nuage de points, le point moyen et la droite des moindre carrés. Qu'en pensez-vous ?
2. Calculer $\rho_{x,y}$.
3. Les variables x et y semblent-elles indépendantes ?

EXERCICE 5.13. Régression non-linéaire

On pose `x=[np.arrange(1,51)` et `y=[np.log(x[i]+rd.normal(0,1/2) for i in range(50)]`.

1. Représenter le nuage de points, le point moyen et la droite des moindre carrés. Qu'en pensez-vous ?
2. Calculer $\rho_{x,y}$.
3. Représenter la fonction $f : x \mapsto \ln(x)$ dans le même graphique.
4. Etudier la corrélation linéaire de y par rapport à $\log(x)$.

REMARQUE **5.7.**
Le modèle de régression entre x et y n'est pas nécessairement linéaire (penser au cas où $\rho_{x,y}$ est proche de 0), il peut être logarithmique, exponentiel ...
On peut dans ces cas néanmoins étudier la corrélation linéaire entre y et $\log(x)$, e^x, ...

EXERCICE 5.14. Corrélation et causalité

Une bonne corrélation entre deux séries de données ne signifie pas pour autant qu'il existe un lien de cause à effet entre les deux. A titre d'exemple, considérons la série statistique suivante :

Année	1996	1997	1998	1999	2000
Morts	15.85	15.7	15.39	15.32	14.85
Importations de citrons	230	280	360	410	525

Ce tableau donne le nombre de morts (pour un million d'habitants) sur les autoroutes américaines, ainsi que le nombre de tonnes de citrons mexicains importés aux États-Unis de 1996 à 2000.
Calculer le coefficient de corrélation linéaire pour cette série double. En déduisez vous une information pertinente?

Mise en garde :
Attention donc à l'erreur courante, notamment dans les médias, qui est de croire qu'un coefficient de corrélation linéaire élevé (en valeur absolue) induit une relation de causalité entre les deux phénomènes mesurés. Voir à ce sujet cette page des Décodeurs du `https://www.lemonde.fr/les-decodeurs/article/2019/03/01/correlations-ou-causalite-generez-vos-5430063_4355770.html` présentant un outil de corrélation géographique sur la base de données sans rapport, de manière à générer "vos propres cartes pour ne rien démontrer du tout". Vous y apprendrez par exemple que la consommation de fromage est fortement corrélée au nombre de licences de football. Mais ce n'est pas pour autant que les footballeurs mangent plus de fromage.