Les Chartreux

9

Statistiques

Statistiques descriptives univariées et bivariées

Introduction...

La **statistique** et les **probabilités** sont les deux aspects complémentaires de l'étude des phénomènes aléatoires.

- Les probabilités peuvent être considérées comme une branche des mathématiques pures, basée sur des théories abstraites, déconnectée de la réalité. On cherche parfois à les appliquer en collant des modèles probabilistes sur des phénomènes aléatoires concrets.
- La statistique est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes aléatoires. La statistique se retrouve dans de nombreux domaines : assurance / finance (fixation des primes d'assurance, gestion de portefeuille, évaluation d'actifs financiers...), biologie / médecine (essais thérapeutiques, dynamique des populations...), sciences de la terre (prévisions météorologiques, exploration pétrolière...), sciences humaines (sondages, enquêtes...), sciences de l'information et de la communication (traitement des images, reconnaissance de formes et de parole...)... On en distingue deux classes :
  - la **statistique descriptive** (analyse de données) : elle a pour but de résumer l'information contenue dans les données via des indicateurs numériques (moyenne, médiane, quartiles...), des tableaux, des graphiques...
  - la **statistique inférentielle** : elle a pour but de faire des prévisions et prendre des décisions à partir du traitement des données recueillies. Ces prévisions sont basées sur d'importants résultats de probabilités ; nous en reparlerons lors du chapitre 16.

Pour bien démarrer...

1. Soit X une variable aléatoire telle que $X(\Omega) = \{x_1, x_2, ..., x_n\}$.

   $\mathbb{E}(X) = \sum_{k=1}^{n} x_k \mathbb{P}\big([X = x_k]\big)$

   $\mathbb{V}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \sum_{k=1}^{n} \big(x_k - \mathbb{E}(X)\big)^2 \mathbb{P}\big([X = x_k]\big)$

   $\sigma(X) = \sqrt{\mathbb{V}(X)}$

2. Soient X et Y deux variables aléatoires telles que $X(\Omega) = \{x_1, x_2, ..., x_n\}$ et $Y(\Omega) = \{y_1, y_2, ..., y_m\}$.

   $\mathrm{Cov}(X, Y) = \mathbb{E}\big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big) = \sum_{(i,j)\in[\![1;n]\!]\times[\![1;m]\!]} \big(x_i - \mathbb{E}(X)\big)\big(y_j - \mathbb{E}(Y)\big)\mathbb{P}\big([X = x_i] \cap [Y = y_j]\big)$

   $\rho(X, Y) = \dfrac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$

3. Que dire du coefficient de corrélation linéaire d'un couple (X, Y) de variables aléatoires discrètes ?

   - $-1 \leqslant \rho(X, Y) \leqslant 1$.
   - $\rho(X, Y) = 1$ si, et seulement si, l'une des variables aléatoires est presque-sûrement fonction affine strictement croissante de l'autre ;
   - $\rho(X, Y) = -1$ si, et seulement si, l'une des variables aléatoires est presque-sûrement fonction affine strictement décroissante de l'autre ;

I Vocabulaire statistique

On utilise les statistiques pour décrire un ensemble appelé **population** constituée **d'individus** auxquels est associé un **caractère**. On distingue :

- les statistiques **quantitatives** : le caractère est quantitatif, il peut prendre différentes **valeurs**,
- les statistiques **qualitatives** : le caractère est qualitatif, il peut prendre différentes **modalités**.

Vocabulaire
On parle aussi de **variable quantitative** ou **variable qualitative**.

Lorsque l'on ne peut pas observer toutes les données relatives à la population, il convient alors d'observer seulement un sous-ensemble de cette population, appelé **échantillon**.
Le programme concerne des variables quantitatives... Sur lesquelles nous pourrons calculer des moyennes, médianes, quartiles... Nous ne pouvons pas faire ceci avec des variables qualitatives !
Nous en étudierons deux types :

- les variables quantitatives **discrètes** : les valeurs sont des réels isolés,
- les variables quantitatives **continues** : les valeurs sont des intervalles de nombres.

Exemple 1

On souhaite étudier les étudiants d'ECG.

- Population : l'ensemble des étudiants d'ECG de France.
- Échantillons possibles : les ECG des Chartreux, ou seulement la classe d'ECG4 des Chartreux ; ou les ECG des Maristes...
- Caractères possibles :
  - caractères quantitatifs : notes au dernier DS de mathématiques, âge, nombre de frères et sœurs, taille, revenus des parents, nombre de cafés consommés chaque jour, temps passé au CDI chaque semaine,...
  - caractères qualitatifs : couleur de cheveux / d'yeux, sexe, catégorie socio-professionnelle des parents, admission en fin de 2A,...

Les éléments d'analyse (indicateurs, représentations graphiques) diffèrent selon le type de variable étudiée : quantitative discrète, quantitative continue, qualitative. Dans la suite, nous nous limiterons aux statistiques quantitatives discrètes.

II Séries statistiques quantitatives discrètes univariées

Dans cette partie, on s'intéresse à un certain échantillon constitué de n individus d'une population. On étudie un caractère quantitatif discret dont le n-uplet des valeurs observées (la série statistique) est noté $x = (x_1, x_2, \dots, x_n)$. On note $z_1, z_2, \dots, z_p$ (avec $p \in \mathbb{N}^*$) les p valeurs possibles *distinctes* de ce caractère.

Remarque
Les valeurs $x_1, x_2, \dots, x_n$ peuvent être égales !

Définitions 1 — Effectif, fréquence

Avec les notations précédentes...

D1 Pour tout $i \in [\![1; p]\!]$, **l'effectif** de z_i, noté e_i, est le nombre de fois où la valeur z_i est prise dans l'échantillon.

D2 **L'effectif total** est le nombre d'individus de l'échantillon, ici n. On a : $n = \sum_{i=1}^{p} e_i$.

D3 **L'effectif cumulé croissant** de z_i, noté ECC_i, est le nombre d'individus dont le caractère a une valeur *inférieure ou égale à celle-ci* : $ECC_i = \sum_{k=1}^{i} e_k$.

D4 **La fréquence** de z_i, notée f_i, est le réel défini par : $f_i = \dfrac{e_i}{n}$.

D5 **La fréquence cumulée croissante** de z_i, notée FCC_i, est le réel défini par : $FCC_i = \dfrac{ECC_i}{n}$.

☞ Pour info...
L'analyse des fréquences (analyse fréquentielle) est utilisée en cryptanalyse pour décrypter des codes !

Remarque
$FCC_i = \sum_{k=1}^{i} f_k$.

Exemple 2

On s'intéresse au nombre de frères et sœurs d'un échantillon de 20 étudiants. La série statistique observée est résumée dans les deux premières lignes du tableau ci-dessous, que nous complétons :

valeurs	0	1	2	3	4	5
effectifs	4	5	6	3	1	1
ECC	4	9	15	18	19	20
fréquence	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{3}{10}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
FCC	$\frac{1}{5}$	$\frac{9}{20}$	$\frac{3}{4}$	$\frac{9}{10}$	$\frac{19}{20}$	1

Remarque
En laissant toutes les fractions sur 20, les calculs sont encore plus rapides !

II.1 Indicateurs

Dans l'étude d'une série statistique quantitative univariée, on peut utiliser deux types d'indicateurs :

- des **indicateurs de position**,
- des **indicateurs de dispersion**.

Commençons par des indicateurs de position :

Définitions 2 — Mode, moyenne, médiane

D1 Le (ou les) **mode(s)** d'une série statistique est la (ou les) valeur(s) de plus grand effectif.

D2 En reprenant les notations des définitions précédentes, la **moyenne (empirique)** de la série x, notée $\overline{x}$, est le réel défini par :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} e_i z_i$$

D3 Une **médiane**, notée $\widetilde{x}$, est un réel tel que la moitié de la série statistique soit inférieure ou égale à $\widetilde{x}$ et la moitié soit supérieure ou égale à $\widetilde{x}$.

Important !

- $\overline{x} = \sum_{i=1}^{p} f_i z_i$
- si $x = (x_1, x_2, \dots, x_n)$ est la série complète (les valeurs peuvent être égales), alors : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

♣ Méthode 1 ♣ **Pour déterminer une médiane d'une série statistique d'effectif total n**, on commence par ranger les valeurs de la série par ordre croissant. Puis :

- si l'effectif total n est impair, la seule médiane alors possible est la valeur de rang central dans la série ordonnée ;
- si l'effectif total n est pair, alors, on choisit comme médiane "habituelle" la moyenne des valeurs des deux rangs centraux dans la série ordonnée.

Exemple 3

Reprenons l'exemple précédent.

- Le mode est 2.
- L'effectif de l'échantillon étant pair, toutes les valeurs comprises entre les deux valeurs centrales de la série ordonnée sont des médianes possibles.
  Ici, les deux valeurs centrales sont la 10ème et la 11ème, toutes deux égales à 2. Ainsi :

$$\widetilde{x} = 2$$

Et maintenant, des indicateurs de dispersion :

Définitions 3 — Variance et écart-type

Avec les notations des définitions précédentes...

D1 La **variance (empirique)** de la série statistique x, notée v_x, est le réel défini par :

$$v_x = \frac{1}{n}\sum_{i=1}^{p} e_i (z_i - \overline{x})^2$$

D2 L'**écart-type (empirique)** de la série statistique x, noté s_x, est le réel défini par :

$$s_x = \sqrt{v_x}$$

Important !

- $v_x = \sum_{i=1}^{p} f_i (z_i - \overline{x})^2$
- si $x = (x_1, x_2, \dots, x_n)$ est la série complète (les valeurs peuvent être égales), alors : $v_x = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$
- v_x est une somme de termes positifs, donc $v_x \geqslant 0$, ce qui justifie la définition de l'écart-type.
- On note souvent juste s_x^2 la variance et s_x l'écart-type.

Propriétés 1

P1 La variance d'une série statistique est nulle si, et seulement si, la série statistique est constante.

P2 $s_x^2 = \left(\frac{1}{n}\sum_{i=1}^{p} e_i z_i^2\right) - \overline{x}^2$ *(formule de Koenig-Huygens)*

P3 Soit x une série statistique de moyenne $\overline{x}$, de médiane $\widetilde{x}$ et de variance s_x^2. Soient $a, b \in \mathbb{R}$ et y la série statistique obtenue par transformation affine de x de la sorte : $y = ax + b$. On a :

- la moyenne de y est donnée par : $\overline{y} = a\overline{x} + b$
- la médiane de y est donnée par : $\widetilde{y} = a\widetilde{x} + b$
- la variance de y est donnée par : $s_y^2 = a^2 s_x^2$

Important !

Si $x = (x_1, x_2, \dots, x_n)$ est la série complète (les valeurs peuvent être égales), alors : $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$.

⋆ Démonstration :

P1. Avec les notations précédentes :

$$v_x = 0 \iff \frac{1}{n}\sum_{i=1}^{p} e_i(z_i - \overline{x})^2 = 0$$

↘ une somme de termes positifs est nulle si, et seulement si, tous ses termes sont nuls

$$\iff \forall i \in [\![1;p]\!],\ e_i(z_i - \overline{x})^2 = 0$$

↘ par définition, pour tout $i \in [\![1;p]\!]$, $e_i \neq 0$

$$\iff \forall i \in [\![1;p]\!],\ z_i = \overline{x}$$

D'où le résultat.

P2. Immédiat...

P3. Assez immédiat...

⋆

Définitions 4 — Étendue, quantiles

D1 L'**étendue** d'une série statistique est la différence entre sa plus grande et sa plus petite valeur.

D2 Soit $p \in]0;1[$. Le **quantile d'ordre *p*** est la plus petite valeur de la série statistique pour laquelle la FCC atteint ou dépasse p. En particulier :

- les **quartiles** sont les quantiles d'ordre $0,25$; $0,5$ et $0,75$:
  - le premier quartile, noté Q_1, est la plus petite valeur de la série telle qu'au moins 25% des valeurs prises lui soient inférieures ou égales ;
  - le troisième quartile, noté Q_3, est la plus petite valeur de la série telle qu'au moins 75% des valeurs prises lui soient inférieures ou égales.
- les **déciles** sont les quantiles d'ordre $0,1$; $0,2$; ... ; $0,9$.

D3 L'**intervalle interquartile** est l'intervalle $[Q_1;Q_3]$. L'**écart interquartile** est $Q_3 - Q_1$: c'est l'amplitude de l'intervalle interquartile.

Remarque
Quand on souhaite comparer deux séries statistiques, on peut utiliser :
- le couple (moyenne, écart-type),
- le triplet (médiane, écart-interquartile, étendue).

Important !
Par définition, l'intervalle interquartile contient au moins 50% des valeurs de la série.

Exemple 4

Toujours avec les données de l'Exemple 2 :

- L'amplitude est égale à 5.
- $Q_1 = 1$, car 1 est la plus petite valeur de la série pour laquelle la FCC atteint ou dépasse $0,25$.
- $Q_3 = 2$, car 2 est la plus petite valeur de la série pour laquelle la FCC atteint ou dépasse $0,75$.
- L'intervalle interquartile est $[1;2]$, il contient ici 55% des valeurs de la série.
- L'écart interquartile est égal à 1.
- Le premier décile vaut 0, car 0 est la plus petite valeur de la série pour laquelle la FCC atteint ou dépasse $0,1$.

II.2 Représentations graphiques

Les trois représentations graphiques les plus fréquentes pour des séries statistiques quantitatives discrètes univariées :

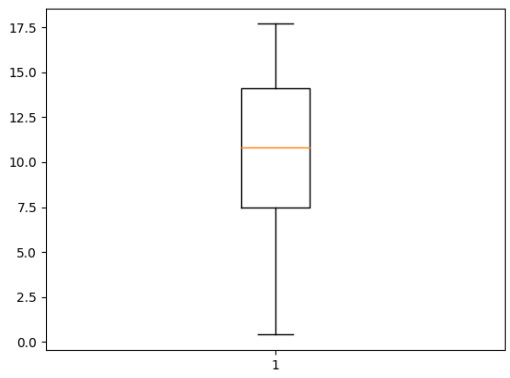
- le diagramme en barres (d'effectifs ou de fréquences) ou en bâtons,
- l'histogramme,
- le diagramme de Tukey, ou boîte à moustaches.

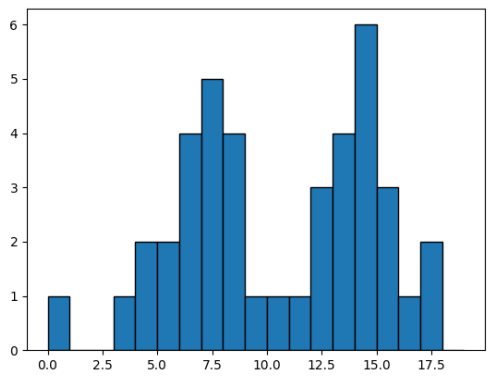
Remarque
Ne pas confondre diagramme en barres et histogramme ! L'histogramme est utilisé lorsque les valeurs obtenues sont regroupées en intervalles...

Un peu d'histoire
John Tukey (1915-2000, américain) est un statisticien à qui on doit des méthodes d'analyse de données ainsi que l'algorithme de transformation de Fourier rapide (FFT en anglais), sur lequel il a travaillé avec James Cooley.

Exemple 5

A gauche, le diagramme de Tukey des notes de l'épreuve 1 du CB, à droite l'histogramme d'effectifs avec des intervalles de la forme $[k;k+1[$, où $k \in [\![0;19]\!]$.





Remarque
Le diagramme en barres n'aurait pas beaucoup d'intérêt ici, car trop peu de notes sont identiques ! Il aurait en revanche un intérêt pour les données de l'exemple 2.

III Séries statistiques quantitatives discrètes bivariées

Comme nous avons pu le constater dans la partie précédente, le principe de l'étude d'une série statistique quantitative discrète univariée est de voir comment les effectifs de chaque valeur peuvent influencer la série et ensuite d'analyser cette série avec des indicateurs tenant compte de ces effectifs.
Dans ce qui va suivre, nous ne nous intéresserons plus aux effectifs ; mais au comportement d'une série de données par rapport à une autre...

On considère à présent une population sur laquelle nous mesurons deux caractères quantitatifs discrets notés X et Y. Sur un certain échantillon de taille n de cette population, nous obtenons ainsi deux séries statistiques quantitatives univariées $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$ **chaque couple (x_i, y_i) étant associé à un même individu de l'échantillon**.
Plutôt que d'étudier chaque caractère (et donc chaque série statistique) séparément, intéressons-nous plutôt à la façon dont X et Y se comportent l'une par rapport à l'autre : existe-t-il une relation entre X et Y ?

Remarque
En fait, on peut voir X et Y comme des variables aléatoires et x et y comme des n-uplets de réalisations de ces variables aléatoires...

III.1 Nuage de points et point moyen

Définitions 5

Soient $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$ deux séries statistiques.

D1 Le **nuage de points** du couple (x, y) est l'ensemble des points du plan de coordonnées (x_i, y_i) pour $i \in [\![1; n]\!]$.

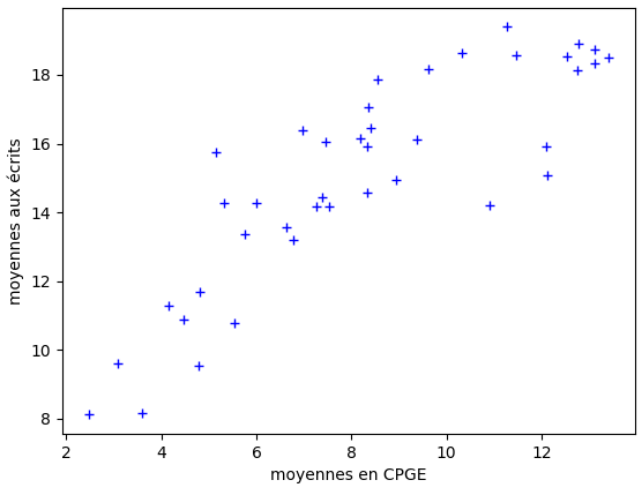
D2 Le **point moyen** du nuage est le point de coordonnées $(\overline{x}, \overline{y})$.

Exemple 6

On considère la promo 2025 de maths appli dont on a récupéré la moyenne en mathématiques en deuxième année de CPGE ECG ainsi que la moyenne aux écrits de mathématiques.
Les commandes **Python** ci-dessous ont permis d'obtenir le nuage de points qui suit.

```
import matplotlib.pyplot as plt
import pandas as pd

notes=pd.read_excel('notes_promo_2025.xlsx')
x=notes["AnneeTotal"]
y=notes["EcritsTotal"]
plt.plot(x,y,"b+")
plt.xlabel("moyennes en CPGE")
plt.ylabel("moyennes aux écrits")
plt.show()
```



Quelles remarques pouvons-nous faire ?

- La corrélation entre la moyenne annuelle et la moyenne aux écrits semble assez forte : plus la moyenne annuelle est élevée, plus la moyenne aux écrits l'est également.
- Le nuage de points est *relativement proche d'une droite*.

III.2 Covariance et coefficient de corrélation linéaire

Définitions 6 — Covariance et coefficient de corrélation linéaire

Soient $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$ deux séries statistiques de moyennes et écart-types respectifs $\overline{x}, \overline{y}$ et s_x, s_y.

D1 La **covariance** du couple (x, y) est le réel, noté $\text{Cov}(x, y)$, défini par :

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

D2 Si σ_x et σ_y sont non nuls, le **coefficient de corrélation linéaire** du couple (x, y) est le réel, noté $\rho(x, y)$, défini par :

$$\rho(x, y) = \frac{\text{Cov}(x, y)}{s_x s_y}$$

Comme dans le cas des couples de variables aléatoires discrètes, on retrouve :

Propriétés 2

Avec les notations précédentes...

P1 $\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$ *(formule de Koenig-Huygens)*

P2 $\text{Cov}(x, y)^2 \leqslant s_x^2 s_y^2$ *(inégalité de Cauchy-Schwarz)*

P3 $-1 \leqslant \rho(x, y) \leqslant 1$

P4 $|\rho(x, y)| = 1 \iff \exists (a, b) \in \mathbb{R}^* \times \mathbb{R} \ / \ y = ax + b$
Et :

- si $\rho(x, y) = 1$, alors $a > 0$
- si $\rho(x, y) = -1$, alors $a < 0$

⋆ Démonstration : Analogues à celles dans le cas des couples de variables aléatoires discrètes... mais plus simples, car toutes les sommes en jeu sont des sommes finies ! ⋆

III.3 Régression linéaire

La propriété précédente (P4), permet de savoir quand y est fonction affine de x ; mais on ne sait en revanche pas encore comment trouver les coefficients a, b de cette fonction affine. Voyons un peu cela...

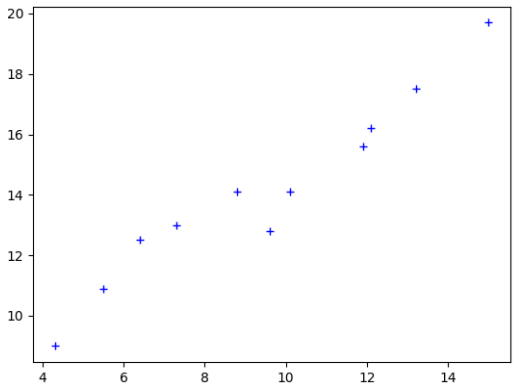
Chercher un modèle de régression consiste à chercher s'il existe une fonction f et une variable aléatoire ε telles que :

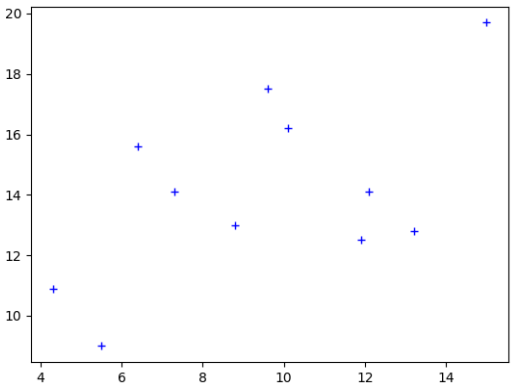
$$Y = f(X) + \varepsilon$$

où f est alors appelée **fonction de régression** et ε **erreur d'ajustement** (ou résidu). Dans ce qui suit, nous nous limiterons au cas de la régression linéaire : cas où f est affine.

✘ Attention !
Régression **linéaire** et pourtant fonction **affine** !

Certains nuages de points ont des formes pouvant penser qu'il existe un ajustement affine entre X et Y...





La forme du nuage de gauche nous laisse penser qu'un ajustement affine est possible entre X et Y ; celle du nuage de droite beaucoup moins.

On cherche à identifier une droite qui ajusterait assez bien le nuage de points, selon des critères à définir. Si l'on suppose qu'il existe des réels a, b tels que pour les séries x et y, on a $y = ax + b$, alors pour tout $i \in [\![1; n]\!]$ l'erreur commise en utilisant la valeur $ax_i + b$ pour estimer y_i est égale à $y_i - (ax_i + b)$.
Pour déterminer les coefficients a et b, on va utiliser ici le **principe des moindres carrés** qui consiste à chercher les valeurs de a et b qui minimisent la somme des carrés de ces erreurs : $\sum_{i=1}^{n}(y_i - ax_i - b)^2$.

Théorème 1

Soient $x = (x_1, x_2, \dots, x_n)$ et $y = (y_1, y_2, \dots, y_n)$ deux séries statistiques de moyennes et écart-types respectifs $\overline{x}, \overline{y}$ et s_x, s_y, avec s_x et s_y non nuls, et de covariance $\mathrm{Cov}(x, y)$.

L'unique couple (a, b) rendant minimale la quantité $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ est défini par :

$$a = \frac{\mathrm{Cov}(x, y)}{s_x^2} \quad ; \quad b = \overline{y} - a\overline{x}$$

Vocabulaire
- Pour ce tel couple, la droite d'équation $y = ax + b$ est appelée **droite des moindres carrés**.
- X est la **variable explicative** et Y la **variable à expliquer**.

À retenir...
Puisque $\rho(x, y)$ et $\mathrm{Cov}(x, y)$ ont même signe, $\rho(x, y)$ et le coefficient directeur de la droite des moindres carrés ont également le même signe.

⋆ Démonstration : Soit $(a, b) \in \mathbb{R}^2$. Considérons la fonction $f : (a, b) \longmapsto \sum_{i=1}^{n}(y_i - ax_i - b)^2$.

1. **Cas particulier.** Supposons que $\overline{x} = \overline{y} = 0$.

1.a. Montrons alors :

$$\forall (a, b) \in \mathbb{R}^2,\ f(a, b) \geqslant a^2 \sum_{i=1}^{n} x_i^2 - 2a \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$$

avec égalité si, et seulement si $b = 0$.
Soit $(a, b) \in \mathbb{R}^2$. On a :

$$\begin{aligned} f(a, b) &= \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\ &= \sum_{i=1}^{n}(y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i) \\ &= \sum_{i=1}^{n} y_i^2 + a^2 \sum_{i=1}^{n} x_i^2 + nb^2 - 2a \sum_{i=1}^{n} x_iy_i - 2b \sum_{i=1}^{n} y_i + 2ab \sum_{i=1}^{n} x_i \qquad \text{↙ } \overline{x} = 0 \text{ et } \overline{y} = 0 \\ &= \sum_{i=1}^{n} y_i^2 + a^2 \sum_{i=1}^{n} x_i^2 + nb^2 - 2a \sum_{i=1}^{n} x_iy_i \end{aligned}$$

Puisque $nb^2 \geqslant 0$, on a bien :

$$f(a, b) \geqslant a^2 \sum_{i=1}^{n} x_i^2 - 2a \sum_{i=1}^{n} x_iy_i + \sum_{i=1}^{n} y_i^2$$

avec égalité si, et seulement si $nb^2 = 0$, autrement dit, si et seulement si, $b = 0$.

1.b. Déduisons-en que $f(a, b)$ est minimale lorsque $a = \dfrac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}$ et $b = 0$.

D'après ce qui précède, $f(a, b)$ est minimale si, et seulement si, $b = 0$ et que la quantité $a^2 \sum_{i=1}^{n} x_i^2 - 2a \sum_{i=1}^{n} x_iy_i + \sum_{i=1}^{n} y_i^2$ est minimale.

Or $a^2 \sum_{i=1}^{n} x_i^2 - 2a \sum_{i=1}^{n} x_iy_i + \sum_{i=1}^{n} y_i^2$ est une expression polynomiale de degré 2 (en effet, les x_i ne sont pas tous nuls, sinon, on aurait $s_x = 0$...) en a à coefficient dominant positif ; elle est donc minimale si, et seulement si, $a = \dfrac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}$.

Conclusion : $f(a, b)$ est minimale lorsque $a = \dfrac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}$ et $b = 0$.

☞ Rappels...
- Si $\alpha > 0$, alors la fonction $x \longmapsto \alpha x^2 + \beta x + \gamma$ possède un minimum, atteint en $\dfrac{-\beta}{2\alpha}$.
- Si $\alpha < 0$, alors la fonction $x \longmapsto \alpha x^2 + \beta x + \gamma$ possède un maximum, atteint en $\dfrac{-\beta}{2\alpha}$.

2. **Cas général.** Considérons deux nouvelles séries statistiques x' et y' définies par :

$$x' = x - \overline{x} \quad ; \quad y' = y - \overline{y}$$

2.a. Que dire de $\overline{x'}$ et $\overline{y'}$?
On a alors :

$$\overline{x'} = 0 \ ; \ \overline{y'} = 0$$

2.b. Exprimons $f(a,b)$ en fonction des x'_i et y'_i, en posant $b' = b - \overline{y} + a\overline{x}$.
Pour tout $(a,b) \in \mathbb{R}^2$:

$$\begin{aligned} f(a,b) &= \sum_{i=1}^{n} (y_i - ax_i - b)^2 \\ &= \sum_{i=1}^{n} (y'_i + \overline{y} - a(x'_i + \overline{x}) - b)^2 \\ &= \sum_{i=1}^{n} \left(y'_i - ax'_i - (b - \overline{y} + a\overline{x})\right)^2 \\ &= \sum_{i=1}^{n} (y'_i - ax'_i - b')^2 \end{aligned}$$

↙ $x' = x - \overline{x}$ et $y' = y - \overline{y}$

2.c. Concluons !

D'après le cas particulier (licite car $\overline{x'} = \overline{y'} = 0$), la quantité $\sum_{i=1}^{n} (y'_i - ax'_i - b')^2$ est minimale lorsque $b' = 0$ et $a = \dfrac{\sum_{i=1}^{n} x'_i y'_i}{\sum_{i=1}^{n} x'^2_i}$.

Autrement dit, d'après la question précédente, $f(a,b)$ est minimale lorsque $b = \overline{y} - a\overline{x}$ et $a = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$.

Mais :

$$\begin{aligned} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} &= \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2} \\ &= \frac{\mathrm{Cov}(x,y)}{s_x^2} \end{aligned}$$

Conclusion : la quantité $\sum_{i=1}^{n} (y_i - ax_i - b)^2$ minimale lorsque

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

Remarque

Une autre démonstration pourra être vue dans le chapitre 13 sur les fonctions de deux variables.

★

Propriété 3

La droite de régression linéaire obtenue par la méthode des moindres carrés passe par le point moyen du nuage de points.

★ Démonstration : Notons a le coefficient directeur de la droite des moindres carrés et b sont ordonnée à l'origine. D'après le théorème précédent :

$$b = \overline{y} - a\overline{x}$$

Autrement dit :

$$\overline{y} = a\overline{x} + b$$

Le point moyen, de coordonnées $(\overline{x}, \overline{y})$, appartient donc à la droite des moindres carrés, puisque ses coordonnées vérifient l'équation réduite de cette droite.

★

Propriété 4 — Interprétation de la valeur de $\rho(x,y)$

Plus $|\rho(x,y)|$ est proche de 1, meilleur est l'ajustement affine par la droite des moindres carrés.

Remarque

Cette propriété va dans le sens de Propriétés 3 - P4 !

⋆ Démonstration :

Notons a le coefficient directeur de la droite des moindres carrés et b son ordonnée à l'origine.
Montrons que, plus $|\rho(x,y)|$ est proche de 1, plus l'erreur commise en approchant y par $ax+b$ est faible.
On a, d'après le théorème 1 :

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-ax_i-b)^2 &= \sum_{i=1}^{n}(y_i-ax_i-\overline{y}+a\overline{x})^2 \\
&= \sum_{i=1}^{n}(y_i-\overline{y}-a(x_i-\overline{x}))^2 \\
&= \sum_{i=1}^{n}(y_i-\overline{y})^2 - 2a\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x}) + a^2\sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= ns_y^2 - 2an\mathrm{Cov}(x,y) + a^2ns_x^2 \\
&= n\left(s_y^2 - 2\frac{\mathrm{Cov}(x,y)^2}{s_x^2} + \frac{\mathrm{Cov}(x,y)^2}{s_x^2}\right) \qquad \text{théorème 1} \\
&= n\left(s_y^2 - \frac{\mathrm{Cov}(x,y)^2}{s_x^2}\right) \\
&= n\left(s_y^2 - \frac{\rho(x,y)^2 s_x^2 s_y^2}{s_x^2}\right) \qquad \text{définition de } \rho(x,y) \\
&= ns_y^2\left(1-\rho(x,y)^2\right)
\end{aligned}$$

Important !

La quantité $\sum_{i=1}^{n}(y_i-ax_i-b)^2$ est la somme des carrés des écarts entre la valeur exacte et la valeur approchée obtenue par la méthode des moindres carrés. Plus cette quantité est faible, meilleure est l'approximation.

Par conséquent, plus $\rho(x,y)^2$ est proche de 1, plus $\sum_{i=1}^{n}(y_i-ax_i-b)^2$ est proche de 0 et donc meilleur est l'ajustement affine par la droite des moindres carrés.

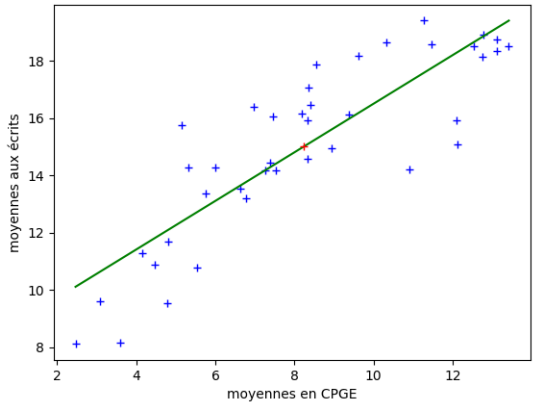
Remarque

De la dernière relation, on retrouve également :
- $-1 \leqslant \rho(x,y) \leqslant 1$
- $|\rho(x,y)| = 1$ ssi $\sum_{i=1}^{n}(y_i - ax_i - b)^2 = 0$ ssi (une somme de termes positifs est nulle ssi tous ses termes sont nuls) : $\forall i \in [\![1;n]\!]$, $y_i = ax_i + b$: tous les points du nuage sont sur la droite des moindres carrés : l'ajustement affine est parfait !

⋆

Exemple 7

Dans l'exemple étudié, on obtient graphiquement :



La droite des moindres carrés est la droite d'équation $y = ax+b$, avec $a \simeq 0,84$ et $b \simeq 8$.
On trouve : $\rho(x,y) \simeq 0,85$. L'ajustement affine est bon.

☞ Pour info...

Parfois, on n'a pas directement $y = ax+b$, mais on peut par exemple avoir :
- $y = ax^2+b$ et l'ajustement affine sera entre x^2 et y
- $y = Ce^{ax}$ et l'ajustement affine sera entre $\ln(y)$ et x...

III.4 Mise en garde !

Il ne faut surtout pas confondre corrélation et causalité ! Deux variables augmentant en fonction d'un même troisième facteur seront assez fortement corrélées, et pourtant il n'y a pas forcément de lien de causalité entre ces deux variables. Il est même possible que la corrélation soit le fruit du hasard !

Pour rire un peu...

Ici : https://www.tylervigen.com/spurious-correlations

Lors d'une analyse statistique révélant un fort coefficient de corrélation linéaire, on peut se poser trois questions :

- une variable a-t-elle un lien de cause à effet avec l'autre ?
- existe-t-il une troisième variable ayant un lien de causalité avec les deux précédentes ?
- la corrélation est-elle fortuite ?