

- Surtout du vocabulaire pour la TIPE.
- Point central: régression linéaire ← toute à l'écrit

CH18 – Statistiques descriptives

Plan du chapitre

1	Vocabulaire	3
	A) Variable (Statistique)	3
	B) Population - caractère	3
	C) Types de caractères	3
2	Statistiques univariées	4
	A) Série statistique	4
	B) Représentations graphiques	4
	C) Caractéristiques de position	5
	D) Caractéristiques de dispersion	6
3	Statistiques bivariées	7
	A) Cadre	7
	B) Données	7
	C) Représentation des données	7
	D) Indicateurs	7
	E) Covariance, coefficient de corrélation	7
	F) Ajustement affine - régression linéaire	8
	G) recherche de lois approchées par changement d'échelle	10

Liste des définitions

Déf.1	Variable - Constante	3
Déf.2	Population - individu - effectif - caractère - modalité - échantillon - taille	3
Déf.3	Variable statistique catégorielle, ordinale, quantitative	3
Déf.4	Série statistique	4
Déf.5	Série chronologique	4
Déf.6	Effectifs, fréquences de la série x	4
Déf.7	Mode d'une série statistique, classe modale	5
Déf.8	Moyenne ou moyenne empirique de la série x	5
Déf.9	Médiane d'une variable non qualitative	6
Déf.10	Étendue d'une série x	6
Déf.11	Variance/écart-type d'une série	6
Déf.12	Point moyen d'un nuage de points	7
Déf.13	Covariance d'une série statistique bivariée	7
Déf.14	Coefficient de corrélation linéaire	8
Déf.15	Données du problème de l'ajustement affine	8
Déf.16	Ajustement affine	9
Déf.17	Vecteurs abstraits associés	9

Liste des techniques de base

Grille d'analyse des exercices

Exercice	Question	T	Référence(s)	Commentaires/remarques

- 1. **T₀** : technique ancestrale. Pas listée dans les techniques de base.
- 2. **Déf** : pas de technique livrée. Revenir à la définition.
- 3. **C** : utilisation d'un résultat de cours (théorème, proposition, etc.)
- 4.

★

 Question discriminante et plus difficile : demande raisonnement et enchaînement de techniques.

- But : définir des indicateurs permettant de résumer, suivant les besoins, des informations sur la distribution des données collectées auprès d'une population étudiée.
- Les variables étudiées seront **quantitatives**.

1 Vocabulaire

Le vocabulaire des statistiques est souvent polysémique, ce qui induit souvent des erreurs de compréhension : soyez donc très attentifs au sens des mots employés

A) Variable (Statistique)

■ **Définition 1** [Variable - Constante]

- **Variable** : Concept en lien à une « réalité » (physique ou non), et se manifestant sous différents aspects, appelés *valeurs* possibles de la variable.
- **Constante** : variable qui se manifeste toujours sous la même valeur.

■ Remarque 1.

Les termes *variable*, et *constante* ont donc ici des acceptions différentes de leur acception mathématique, ou informatique, ou même physique.

B) Population - caractère

dico maths / stats

■ **Définition 2** ... [Population - individu - effectif - caractère - modalité - échantillon - taille]

1. L'ensemble $\mathcal{P}$ (au sens mathématique) des objets sur lesquels les informations seront collectées est appelé **population**. = ensemble
2. Les éléments de la population $\mathcal{P}$ s'appellent les **individus**. = éléments
3. Le cardinal de $\mathcal{P}$ est appelé **effectif** de la population (si il est défini). = cardinal
4. Une variable statistique attachée à $\mathcal{P}$ est aussi appelée **caractère** de la population. = fonction
5. Les valeurs prises par un caractère s'appellent les **modalités** du caractère ou de la variable. = image de la fonction
6. Tout sous-ensemble fini $\mathcal{E}$ de $\mathcal{P}$ est appelé **échantillon** de la population. = sous-ensemble
7. Le cardinal de $\mathcal{E}$ est appelé **taille** de l'échantillon. = cardinal d'un sous-ens.

C) Types de caractères

■ **Définition 3** [Variable statistique catégorielle, ordinale, quantitative]

Le caractère x est :

1. **qualitatif** ou **catégoriel** si les valeurs de x ne peuvent être ordonnées. Exemple : la couleur des yeux, le groupe sanguin.
2. **ordinal** si les valeurs peuvent être seulement ordonnées sans pouvoir donner de sens aux différences entre les valeurs (typiquement : un caractère mesuré sur une échelle de valeurs, telle que 0 – 1 – 2 – 3 – 4 – 5 ou Satisfait – Moyen – Nul).
3. **quantitatif** quand les valeurs sont celles d'une grandeur numérique (typiquement : toute grandeur exprimée dans une certaine unité).

→ mesurage

2 Statistiques univariées

A) Série statistique

Usages la matière.

1. Quitte à numéroté les n individus de l'échantillon $\mathcal{E}$, on peut supposer : $\mathcal{E} = \{1, 2, \dots, n\}$.
2. Dans ce cas : x_j est la valeur du caractère x observée sur l'individu $j \in \mathcal{E}$.

■ **Définition 4** [Série statistique]
La liste $\mathbf{x} = (x_1, \dots, x_n)$ s'appelle *série statistique*.

■ Remarque 2.

1. Ainsi, si le caractère x est quantitatif, $\mathbf{x} \in \mathbb{R}^n$: on a alors à notre disposition tous les concepts de l'algèbre linéaire euclidienne.
2. La série statistique $\mathbf{x}$ est constante (Déf. 4) si pour un certain réel a , on a : $\mathbf{x} = (a, a, \dots, a)$. Ce qui veut dire que $\mathbf{x}$ est colinéaire à la série : $\mathbf{1} = (1, \dots, 1)$.
3. Ainsi, $\text{Vect}(\mathbf{1})$ est le sev des séries statistiques constantes.

■ **Définition 5** [Série chronologique]
Soit $\mathbf{x}$ une série statistique. Si l'ordre dans lequel les mesures ont été prises est important, $\mathbf{x}$ s'appelle *série chronologique*, et la numérotation des éléments de $\mathcal{E}$ est choisie de sorte à respecter l'ordre d'acquisition des données.

■ **Définition 6** [Effectifs, fréquences de la série $\mathbf{x}$]

1. on note n_k le nombre de fois que cette valeur est observée, et on l'appelle effectif des cas où x prend la valeur x_k .
2. Le rapport $f_k = \frac{n_k}{n}$ est appelé fréquence de l'observation x_k .

B) Représentations graphiques

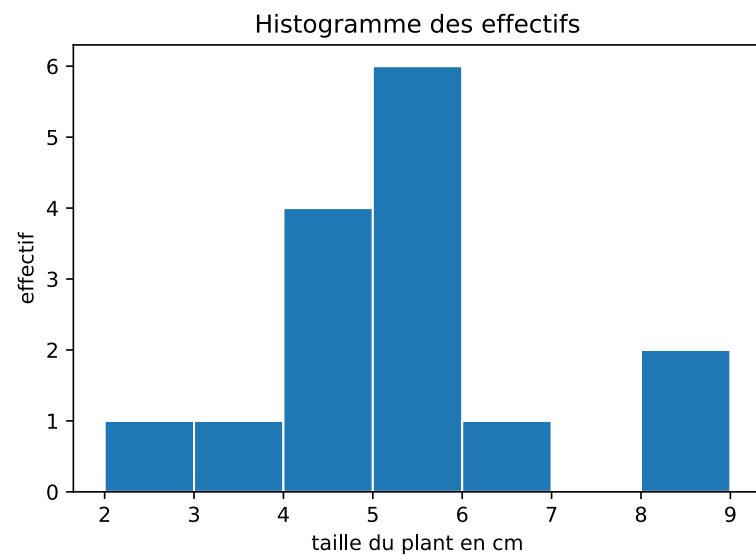
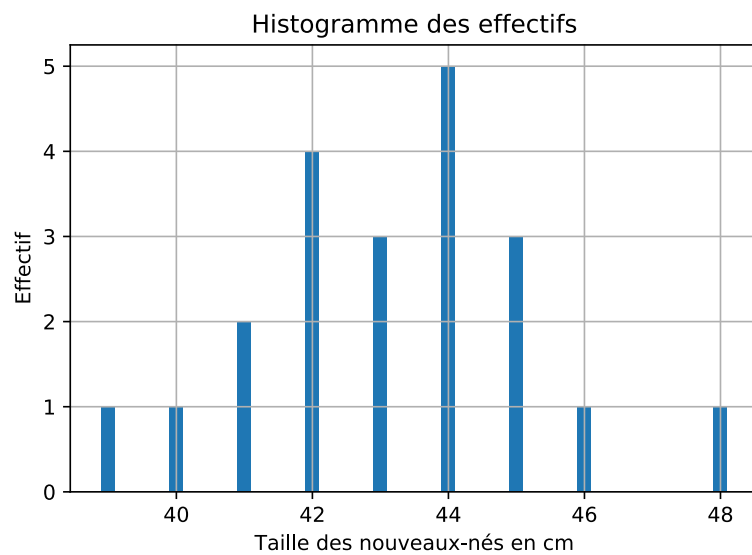
La visualisation des données donne une bonne synthèse des listes de nombres!

Représentations possibles :

1. Boîtes à moustache.
2. Diagramme en bâtons.
3. Histogramme (regroupement par classes dans ce cas).
4. Polygone des fréquences ou effectifs cumulés.

■ Remarque 3.

1. Le choix de l'amplitude des classes ne va pas de soi.
2. En effet, si les classes sont trop amples, la synthèse des données risque de refléter assez peu les tendances de la série.
3. *A contrario*, des classes trop étroites brouillent la lisibilité.
4. On peut toujours essayer de regrouper les classes en classes de largeur de l'ordre d'un écart-type de la série de données : ceci devrait faire ressortir une distribution normale d'après le TCL, lorsque bien entendu ses conditions sont vérifiées. Toutefois, il n'y a pas de règle absolue.



C) Caractéristiques de position

■ **Définition 7** [Mode d'une série statistique, classe modale]

1. Le mode d'une série statistique est la valeur observée le plus grand nombre de fois.
2. En cas de regroupement par classe, la classe modale est la classe d'effectif le plus élevé.

■ **Remarque 4.**

1. S'il y a plusieurs modes ou plusieurs classes modales, n'importe laquelle des valeurs la plus observée constitue un mode.
2. En cas de regroupement par classes, les classes modales sont celles dont les rectangles sont d'aire maximale.

■ **Définition 8** [Moyenne ou moyenne empirique de la série x]

Soit $x = (x_1, \dots, x_n)$ une série statistique. Alors :

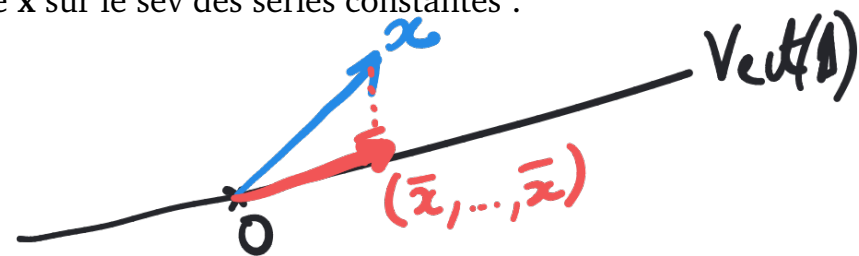
$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j.$$

■ **Remarque 5.**

1. On a donc : $\bar{x} = \frac{x \cdot 1}{\|1\|^2}.$

2. La série $k = (\bar{x}, \dots, \bar{x}) = \frac{x \cdot 1}{\|1\|^2} 1$ est donc le projeté orthogonal de x sur le sev des séries constantes :

$$k = \Pi_{\text{Vect}(1)}(x)$$



■ **Proposition 1** [Stabilité de la moyenne par changement affine]

Si $(a, b) \in \mathbb{R}^2$, $x = (x_j)_{1 \leq j \leq n}$ et $y = (ax_j + b)_{1 \leq j \leq n}$. Alors :

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

■ **Définition 9** [Médiane d'une variable non qualitative]

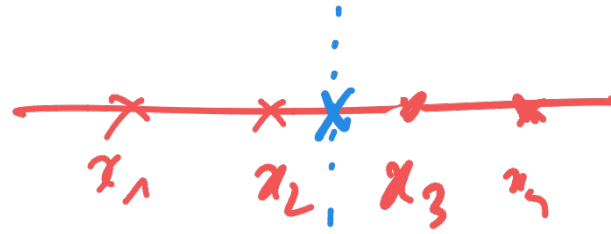
Soit $(x_j)_{1 \leq j \leq n}$ une série statistique. Quitte à renuméroter les modalités, supposons que ces dernières sont rangées dans l'ordre croissant :

$$x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq x_n.$$

La médiane de la série est le nombre noté Q_2 défini par :

1. $Q_2 = x_{\frac{n+1}{2}}$ si n est impair. *← dans la série*

2. $Q_2 = \frac{1}{2} \left(x_{\frac{n}{2}} + x_{1+\frac{n}{2}} \right)$ si n est pair. *← moyenne*



■ **Remarque 6.**

La détermination de Q_2 passe par le classement des données dans l'ordre croissant. Voir la fiche Méthodes de tri.

D) Caractéristiques de dispersion

■ **Définition 10** [Étendue d'une série x]

Soit $x = (x_j)_{1 \leq j \leq n}$ une variable statistique quantitative.

$$\max_{1 \leq j \leq n} (x_j) - \min_{1 \leq j \leq n} (x_j).$$

■ **Remarque 7.**

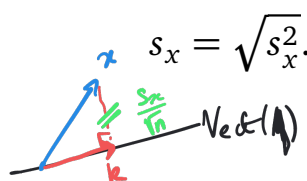
Donne une idée de la «variabilité» des données, mais est extrêmement sensible aux données aberrantes, puisqu'une valeur anormalement petite par exemple (obtenue par erreur de manipulation par exemple) modifie complètement la valeur de l'étendue des données.

■ **Définition 11** [Variance/écart-type d'une série]

◦ Variance des données :

$$s_x^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 = \overline{(x - \bar{x})^2}.$$

◦ Écart-type des données :



■ **Remarque 8.**

1. On peut remarquer, avec les notations de Rem. 5 b., que

$$s_x^2 = \frac{\|x - k\|^2}{n}$$

$$\begin{aligned} x &= (x_1, \dots, x_n) \\ k &= (\bar{x}, \dots, \bar{x}) \\ x - k &= (x_1 - \bar{x}, \dots, x_n - \bar{x}) \\ \|x - k\|^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= n s_x^2 \end{aligned}$$

2. Ainsi, par définition de projeté orthogonal réalisant le minimum de distance, σ_x est la distance de x au sev des variables constantes.

■ **Proposition 2** [Formule de Koenig pour les séries stats.]

$$\overline{x^2} \neq \bar{x}^2$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 \leftarrow \text{théorème de Pythagore}$$

■ **Remarque 9.**

C'est simplement le théorème de Pythagore (cf. fiche Produit scalaire, Prop. 4), et Rem. 5 b., puisque :

$$\underbrace{\|x\|^2}_{n \overline{x^2}} = \underbrace{\|x - k\|^2}_{n s_x^2} + \underbrace{\|k\|^2}_{n \bar{x}^2} \quad \text{car} \quad k = \Pi_{\text{Vect}(1)}(x)$$

■ Proposition 3 [Propriétés de la variance]

1. Si x est en unités u , s_x^2 est en unités u^2 .
2. $\forall a \in \mathbb{R} \ s_{ax}^2 = a^2 s_x^2$ *Homogénéité*
3. $\forall b \in \mathbb{R} \ s_{x+b}^2 = s_x^2$ *Invariance par translation.*

3 Statistiques bivariées

A) Cadre

Sur un échantillon $\mathcal{E}$ donné d'une population $\mathcal{P}$, on est amené à observer plusieurs caractères. Le cas qui nous intéresse ici est celui de deux caractères quantitatifs. Dans ce cas, on étudie des séries statistiques bivariées.

B) Données

On se donne un échantillon $\mathcal{E}$ de taille n d'une certaine population $\mathcal{P}$ sur lequel on a mesuré deux caractères quantitatifs x, y . On note :

1. $x_1 < x_2 < \dots < x_p$ les différentes modalités observées de x sur les individus de $\mathcal{E}$ (quitte à les renuméroter).
2. $y_1 < y_2 < \dots < y_q$ les différentes modalités observées de y sur les individus de $\mathcal{E}$ (idem).
3. On note $n_{i,j}$ l'effectif de l'observation (x_i, y_j) c'est-à-dire le nombre d'individus de $\mathcal{E}$ pour lesquelles l'observation conjointe de x et y donne les valeurs respectives x_i et y_j
4. On a donc $\sum_{i=1}^p \sum_{j=1}^q n_{i,j} = n$.

C) Représentation des données

Dans un tableau à double entrée, on représente les modalités de x et de y . À l'intersection de la ligne x_i et de la colonne y_j , on place l'effectif conjoint $n_{i,j}$ de la modalité (x_i, y_j) (Comme pour les lois conjointes de variables aléatoires).

D) Indicateurs

■ Définition 12 [Point moyen d'un nuage de points]

Pour une série statistique (x, y) bivariée, le point moyen est le point de coordonnées $(\bar{x}, \bar{y})$.

E) Covariance, coefficient de corrélation *Adapté aux graphiques A/D.*

■ Définition 13 [Covariance d'une série statistique bivariée]

Soit (x, y) une série statistique double, notons (x_i, y_j) ses modalités distinctes (cf. B) DONNÉES), n la taille de l'échantillon. La covariance de la série est le nombre noté $s_{x,y}$ défini par :

$$s_{x,y} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^q n_{i,j} \overbrace{(x_i - \bar{x})}^{\alpha} \overbrace{(y_j - \bar{y})}^{\beta}.$$

$\alpha \beta$
 $= s_{\bar{x}, \bar{y}}$ d'où $s_{x,y} = s_{\bar{x}, \bar{y}}$

■ Proposition 4 [Formule de König pour la covariance]

$$s_{x,y} = \overline{xy} - \bar{x} \cdot \bar{y}$$

preuve de $S_{xy} = \overline{xy} - \bar{x}\bar{y}$

Posons N : nb de mesures réalisées

et: $x_1 \dots x_N$ les valeurs observées de x
 $y_1 \dots y_N$ les valeurs observées de y

(par rapport aux notations, je ne suppose pas les x_i/y_i deux à deux distincts).

$$S_{xy} \stackrel{\text{def}}{=} \frac{1}{N} \times \sum_{j=1}^N (x_j - \bar{x})(y_j - \bar{y}) \quad (\text{pas besoin de somme double})$$

$$N S_{xy} = \sum_{j=1}^N (x_j y_j - x_j \bar{y} - y_j \bar{x} + \bar{x} \bar{y})$$

$$= \underbrace{\sum_{j=1}^N x_j y_j}_{N \overline{xy}} - \sum_{j=1}^N x_j \bar{y} - \sum_{j=1}^N y_j \bar{x} + \sum_{j=1}^N \bar{x} \bar{y}$$

linéarité de $\sum$ = $N \overline{xy} - \bar{y} \underbrace{\sum_{j=1}^N x_j}_{N \bar{x}} - \bar{x} \underbrace{\sum_{j=1}^N y_j}_{N \bar{y}} + N \bar{x} \bar{y}$

$$\cancel{N} \cdot S_{xy} = \cancel{N} \overline{xy} - \cancel{N} \bar{x} \bar{y} \quad \blacksquare$$

$$x = \boxed{1}, 1, \boxed{3}, 3, 3, 5, 5, 6, 7$$

$$y = \boxed{2}, 2, \boxed{1}, 2, 3, 2, 3, 4, 4$$

$$n_{i,j} = 2$$

Coups observés: (1, 2) 2

(3, 1)

(3, 2)

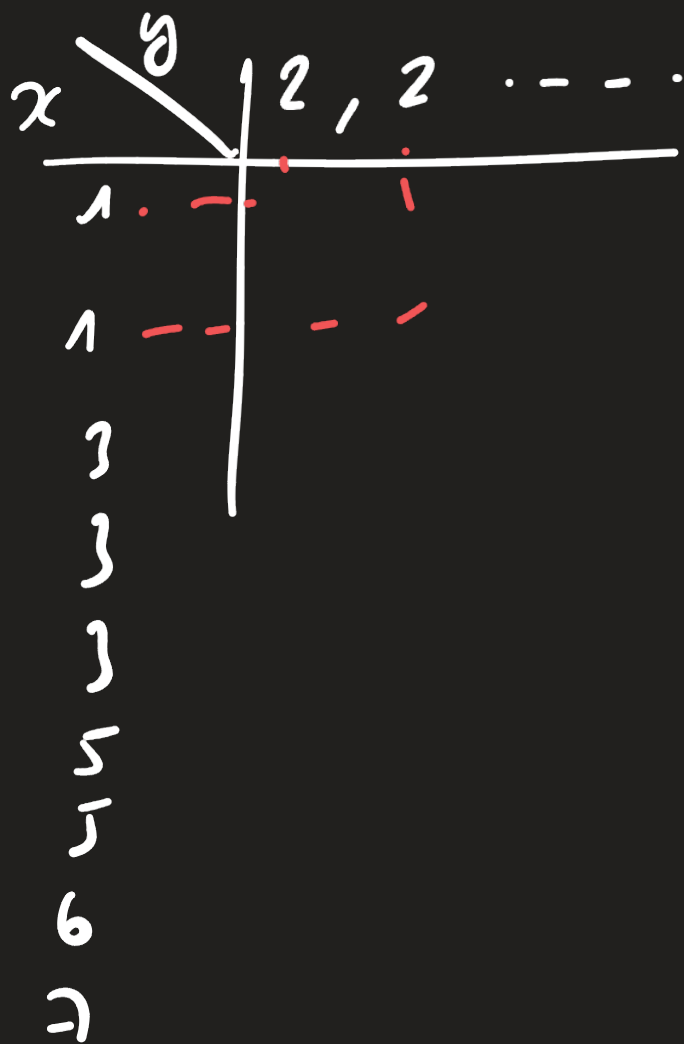
(3, 3)

(5, 2)

(3, 3)

(6, 4)

(7, 4)



$$\sum_{j=1}^9 x_j y_j$$

$$\sum_{j=1}^9 n_j x_j y_j$$

$$\sum_j n_{i,j} = n$$

$$\sum_i n_{i,j} = n$$

■ Définition 14 [Coefficient de corrélation linéaire]

C'est le nombre noté $r_{x,y}$ et défini par :

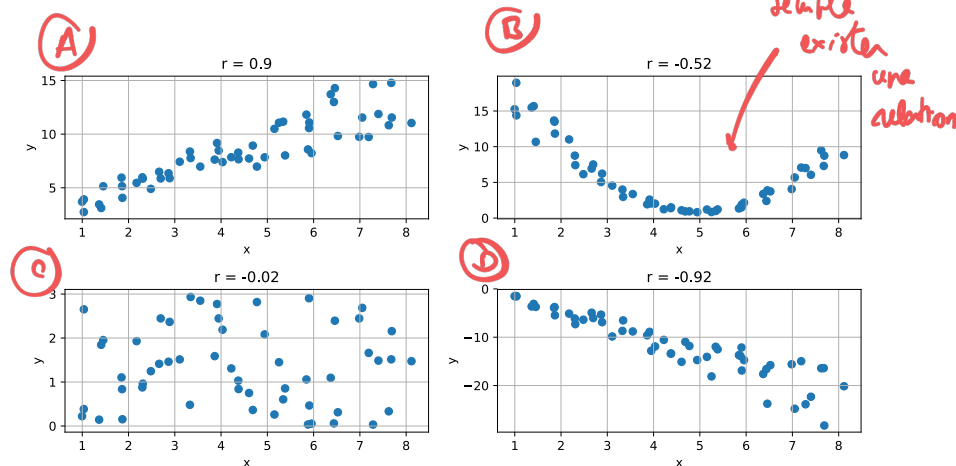
C'est une covariance normalisée :

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = S_{x^*, y^*}$$

$x^ = \frac{x - \bar{x}}{s_x}$ ← série centrée-réduite.*

■ Remarque 10.

Le coefficient de corrélation linéaire renseigne sur la tendance qu'a le nuage de points de la série à se concentrer le long d'une droite : il *ne détecte pas* d'éventuelles corrélations *non linéaires* ! (voir figure ci-dessous, le deuxième cas semble révéler une corrélation du style $y = a(x - b)^2$).



A)D) : semble exister une relation affine

■ Proposition 5 [Propriétés du coefficient de corrélation linéaire]

C) : brouillard/bruit : pas de lien particulier.

1. $r_{x,y} \in [-1, 1]$.
2. $|r_{x,y}| = 1$ si et seulement si il existe deux réels a, b tels que :

$$y = ax + b.$$

(le nuage de points est aligné sur une droite).

■ Remarque 11.

🦋 Ne pas confondre causalité et corrélation !

F) Ajustement affine - régression linéaire *Démos et éléments de compréhension faits en T.D.*

■ Définition 15 [Données du problème de l'ajustement affine]

1. On se donne un nuage de n points du plan (x_i, y_i) ($1 \leq i \leq n$), issu d'une statistique bivariée.
2. Deux réels a, b quelconques, et la droite d'équation $y = ax + b$.
3. On définit l'écart quadratique moyen $EQM(a, b)$ mesurant la «distance» du nuage à la droite.

$$EQM(a, b) := \sum_{i=1}^n \underbrace{\left(y_i - \overbrace{(ax_i + b)}^{\hat{y}_i} \right)^2}_{\delta_i^2}.$$

■ Remarque 12.

Ce choix de EQM est motivé par les considérations suivantes :

1. Les points (x_i, y_i) sont alignés sur la droite d'équation $y = ax + b$ si, et seulement si $EQM(a, b)$ est nul.
2. Comme : $\forall (a, b) \in \mathbb{R}^2, EQM(a, b) \geq 0$, plus $EQM(a, b)$ est petit, mieux la droite d'équation $y = ax + b$ devrait épouser le nuage de points.

preuve de prop 5:

Avec les notations de la preuve de prop 4:

$$S_{x,y} = S_{\bar{x},\bar{y}} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{d'où} \quad NS_{\bar{x},\bar{y}} = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

$$\begin{aligned} \text{Soit je note } \vec{x} &= (x_1, \dots, x_N) \in \mathbb{R}^N \\ \vec{y} &= (y_1, \dots, y_N) \in \mathbb{R}^N \\ \vec{1} &= (1, \dots, 1) \in \mathbb{R}^N, \end{aligned}$$

Je peux interpréter cet indicateur statistique $S_{x,y}$ de façon euclidienne:

$$NS_{\bar{x},\bar{y}} = \langle \vec{x} - \bar{x}\vec{1}, \vec{y} - \bar{y}\vec{1} \rangle$$

D'après Cauchy-Schwarz:

$$|\langle \vec{x} - \bar{x}\vec{1}, \vec{y} - \bar{y}\vec{1} \rangle|^2 \leq \|\vec{x} - \bar{x}\vec{1}\|^2 \|\vec{y} - \bar{y}\vec{1}\|^2$$

$$\text{or } \|\vec{x} - \bar{x}\vec{1}\|^2 = \sum_{j=1}^N (x_j - \bar{x})^2 = N S_x^2$$

variance de x !

$$\text{d'où } |NS_{x,y}|^2 \leq N^2 S_x^2 S_y^2$$

$$\text{donc: } |S_{x,y}|^2 \leq S_x^2 S_y^2$$

en supposant que x et y non constantes:
liées

$$r_{x,y} \rightarrow \left| \frac{S_{x,y}}{S_x S_y} \right|^2 \leq 1$$

Cas d'égalité: les vecteurs $\vec{x} - \bar{x}\vec{1}$ et $\vec{y} - \bar{y}\vec{1}$ sont liés:

$$\exists \alpha, \beta \text{ tq. } \alpha \vec{x} - \alpha \bar{x}\vec{1} = \beta \vec{y} - \beta \bar{y}\vec{1}$$

$$\text{i.e. } \alpha \vec{x} = \beta \vec{y} + (\alpha \bar{x} - \beta \bar{y})\vec{1}$$

En supposant $\beta \neq 0$ (ce qui est le cas si $\vec{y} \neq 0$)

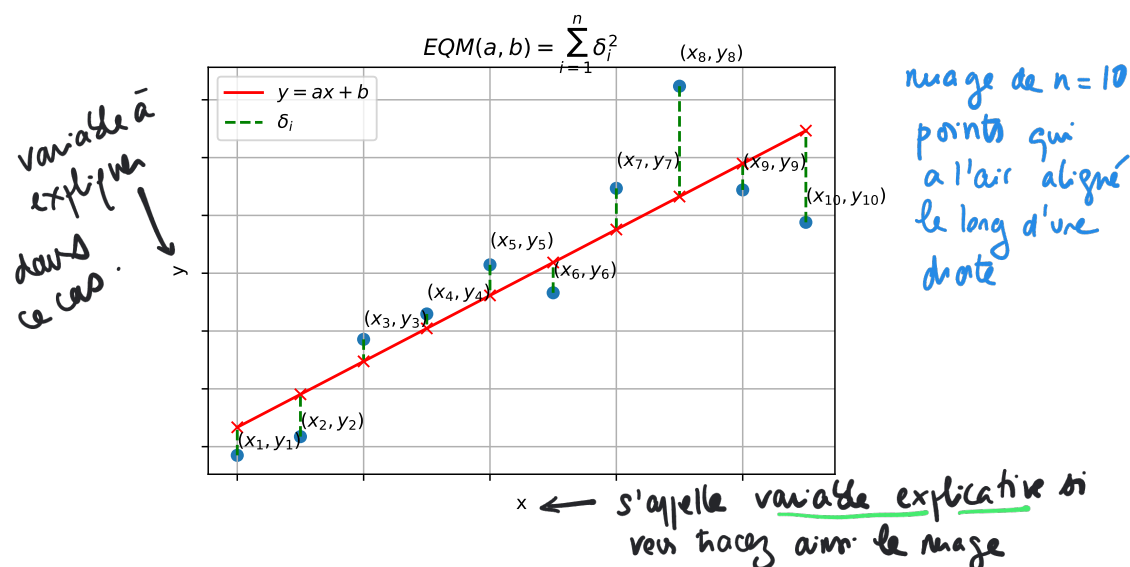
$$\vec{y} = \underbrace{\left[\frac{\alpha}{\beta} \right]}_A \vec{x} - \underbrace{\left(\frac{\alpha \bar{x} - \beta \bar{y}}{\beta} \right)}_B \vec{1}$$

A B res

Ce qui exprime exactement que $\forall i \quad y_i = Ax_i + B$.

→ on voit venir la droite de régression.

3. La «meilleure» droite épousant le nuage de points devrait donc être celle qui rend $EQM(a, b)$ minimal.



■ Définition 16 [Ajustement affine]

On appelle ajustement affine du nuage de points $((x_i, y_i))_{1 \leq i \leq n}$ la recherche des paramètres a, b de sorte à rendre $EQM(a, b)$ minimal.

La résolution du problème de l'ajustement affine se comprend mieux dans un cadre abstrait de projection orthogonale dans $\mathbb{R}^n$.

■ Définition 17 [Vecteurs abstraits associés]

- [Vecteurs des] observations :

$$\mathbf{x} := (x_1, \dots, x_n) \in \mathbb{R}^n$$

$$\mathbf{y} := (y_1, \dots, y_n) \in \mathbb{R}^n$$

- Série certaine :

$$\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$$

- Prévisions du modèle linéaire (a, b) : $\hat{\mathbf{y}} = a\mathbf{x} + b\mathbf{1} \in \mathbb{R}^n$.

- Erreur de prévision du modèle (a, b) :

$$EQM(a, b) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \in \mathbb{R}_+.$$

■ Remarque 13.

1. On appelle les $\hat{y}_i = ax_i + b$ les «prévisions du modèle linéaire (a, b) », puisque si x et y étaient effectivement liés par la relation $y = ax + b$, alors les valeurs de y_i seraient parfaitement prédites à partir de celles de x . En effet, dans ce cas le modèle prédit que : $\underbrace{\hat{y}_i = ax_i + b}_{\text{prévisions}} = \underbrace{y_i}_{\text{observations}}$
2. Résoudre le problème de l'ajustement affine, c'est donc déterminer les couples (a, b) pour lesquels on a : $EQM(a, b) = \min_{(a,b) \in \mathbb{R}^2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, c'est-à-dire modèle linéaire collant le mieux aux observations.
3. L'ensemble des modèles possibles, c'est-à-dire l'ensemble des sets de prévisions possibles $\hat{\mathbf{y}}$ à partir d'observations $\mathbf{x}$, est donc l'ensemble des vecteurs de la forme $\hat{\mathbf{y}} = a\mathbf{x} + b\mathbf{1}$ quand (a, b) décrit $\mathbb{R}^2$. Cet ensemble des modèles possibles est donc le sev de F de $\mathbb{R}^n$ donné par : $F = \text{Vect}(\mathbf{x}, \mathbf{1})$. D'après 2., résoudre le problème de l'ajustement, c'est trouver $\hat{\mathbf{y}}$ tel que $\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \min_{\mathbf{v} \in F} \|\mathbf{y} - \mathbf{v}\|^2$. De thm. 6. du chapitre 15 Géométrie euclidienne, on tire le thm.1 qui suit :

preuve du théorème 1. (TD)

Données : mesures de la variable explicative :

$$x = (x_1, \dots, x_n)$$

• mesures de la variable à expliquer :

$$y = (y_1, \dots, y_n)$$

• En traçant le nuage de points, je constate que celui-ci a l'air de s'aligner le long d'une droite.

→ Je suspecte qu'il existe une fonction qui relie les valeurs de y_i à celle de x_i :

$$(y_1, \dots, y_n) = f(x_1, \dots, x_n)$$

$$(y_1, \dots, y_n) = a(x_1, \dots, x_n) + b(1, \dots, 1)$$

En termes des vecteurs :

$$(*) \quad y = ax + b\mathbb{1} \quad \text{CRS}$$

Le couple (a, b) détermine la fonction = modèle
un tel modèle s'appelle modèle linéaire.

En pratique $(*)$ est fautive, puisque les points ne sont pas alignés. Voici l'approche :

• On introduit les valeurs prédites par le modèle :

$$\hat{y}_1 = ax_1 + b$$

...

$$\hat{y}_n = ax_n + b$$

on note $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ = prédiction du modèle

Je cherche donc le modèle (a, b) dont les prédictions collent le mieux aux observations. Pour donner un sens mathématique à "coller le mieux", je définis l'erreur de prédiction du modèle (a, b) :

$$EQM(a, b) = \sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)^2}_{\substack{\text{obs.} \quad \text{pred.} \\ \text{énergie consommée par l'erreur de prédiction}}} \geq 0 \quad \leftarrow \begin{array}{l} \text{le } S_i^2 \\ \text{du graphique} \end{array}$$

Cette grandeur est positive, donc minimale quand elle est nulle.

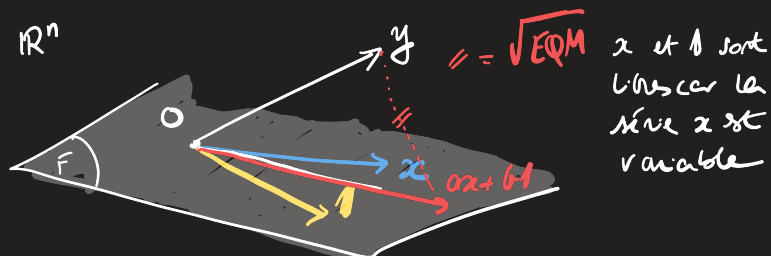
Le meilleur modèle (= meilleure droite) est celui qui rend minimal $EQM(a, b)$.

C'est une grandeur statistique : j'essaie de lui donner un sens euclidien (= en termes de prod. scalaire)

💡 Or Somme de carrés = $\|u\|^2$

$$\begin{aligned} \text{ici } EQM(a, b) &= \|y - \hat{y}\|^2 \\ &= \|y - (ax + b\mathbb{1})\|^2 \end{aligned}$$

requiert un effort d'abstraction important :



Quand (a, b) décrit $\mathbb{R}^2$, le vecteur $ax + b\mathbb{1}$ décrit $\text{Vect}(\mathbb{1}, x) = F$

--- est minimal quand $ax + b\mathbb{1}$ est le projeté orthogonal de y sur F .

Conclusion : calculer (a, b) déterminant la droite de régression, c'est calculer $\Pi_F(y)$.

■ Théorème 1 [Existence et unicité de la droite des moindres carrés]

1. Il existe une unique droite d'équation $y = ax + b$ minimisant la quantité $EQM(a,b)$ (Déf. 4).
2. Les nombres a, b sont les coordonnées sur la base $(\mathbf{x}, \mathbf{1})$ du projeté orthogonal de $\mathbf{y}$ sur le plan $F = \text{Vect}(\mathbf{x}, \mathbf{1})$. En particulier :

$$a = r_{x,y} \frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2}$$

3. Cette droite est appelée **droite de régression**, ou **droite des moindres carrés**, puisque toute autre droite a un EQM qui lui est *strictement* supérieur.
4. Elle passe *toujours* par le point moyen du nuage $G(\bar{x}, \bar{y})$ (Déf. 12), donc $b = \bar{y} - a\bar{x}$.
5. L'équation de la droite de régression est donc :

$$y = r_{x,y} \frac{s_y}{s_x} (x - \bar{x}) + \bar{y}.$$

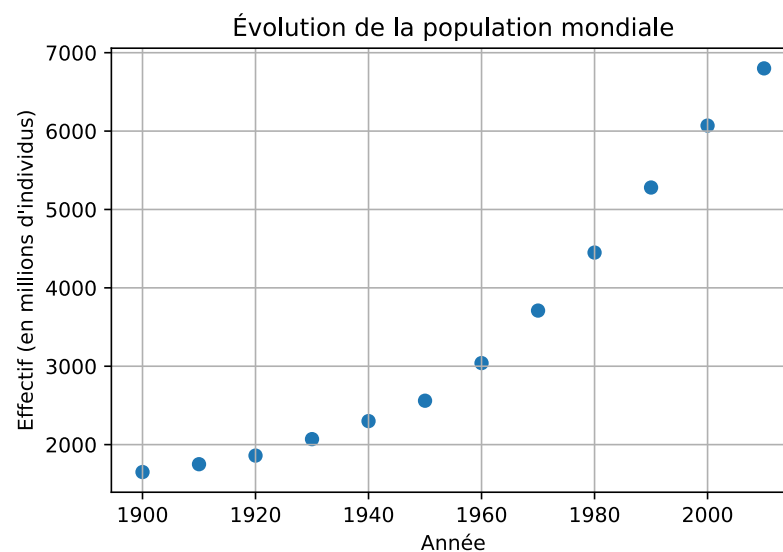
G) Recherche de lois approchées par changement d'échelle

Cas pour lesquels un ajustement affine est envisageable après changement d'échelle :

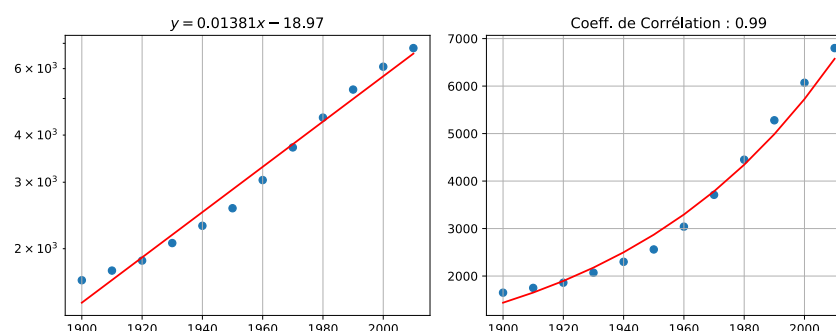
1. **Modèles linéaires du premier ordre.** (qui sont quasiment universels en première approche), la relation est du type : $y = Ce^{\lambda x}$, qui équivaut à : $\begin{cases} Y = AX + B \\ X = x, Y = \ln y, A = \lambda, B = \ln C \end{cases}$

■ Exemple 1.

Évolution de la population mondiale depuis 1900 :



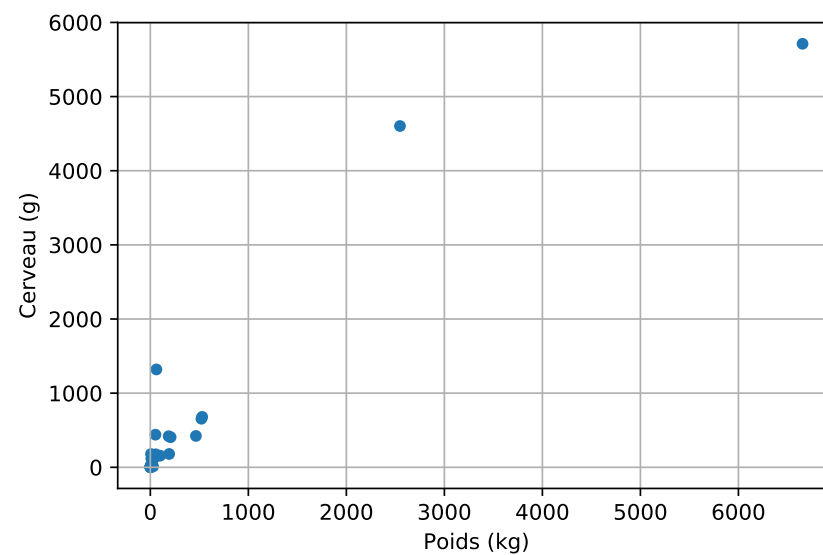
On peut ajuster ces données à un modèle malthusien (cf. fiche Dynamique des populations) :



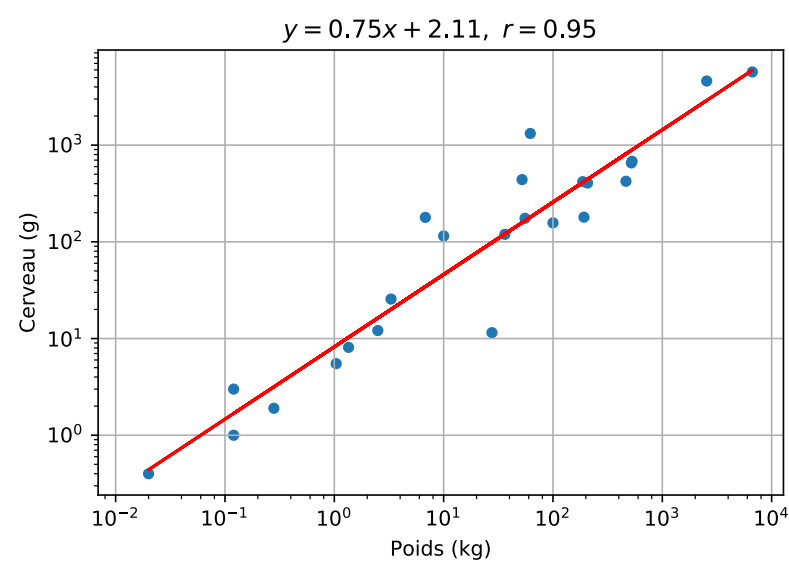
2. **Relations allométriques.** (pour des grandeurs dimensionnées) : On soupçonne un lien du type : $y = Cx^\alpha$, ce qui équivaut à $\begin{cases} Y = AX + B \\ X = \ln x, Y = \ln y, A = \alpha, B = \ln C \end{cases}$

■ Exemple 2.

Corrélation entre la masse corporelle M (en kg) d'espèces de mammifères et celle m_c de leur cerveau (en g).



Les données sont très dispersées (recensées sur 23 espèces allant de la souris (0,02 kg pour un cerveau de 0,4g) à l'éléphant d'Afrique (7 tonnes environ pour un cerveau de 6kg)). En passant aux logarithmes, on voit que l'on peut tenter un ajustement affine :



duquel on tirerait : $m_c \simeq 8.25 \times M^{\frac{3}{4}}$